Ruprecht-Karls-Universität Heidelberg
Geographisches Institut

Master Thesis

Understanding MapSwipe: Analysing Data Quality of Crowdsourced Classifications on Human Settlements

Benjamin Herfort
Matrikel-Nr. 3040449

Supervisors:
Prof. Dr. Alexander Zipf
Prof. Dr. João Porto de Albuquerque

20.10.2017

Content

List of Figures

List of Tables

List of Abbreviations

AIDR	artificial intelligence for disaster response
API	application programming interface
ARC	American Red Cross
BRC	British Red Cross
DLR	Deutsches Zentrum für Luft- und Raumfahrt
FN	false negatives
FP	false positives
GPS	Global Positioning System
HOT	Humanitarian OpenStreetMap Team
JOSM	Java OpenStreetMap Editor
MSF	Médecins Sans Frontières
OSM	OpenStreetMap
POI	point of interest
SAR	synthetic-aperture radar
TMS	tile map service
TN	true negatives
TP	true positives
UAV	unmanned aerial vehicle
VGI	volunteered geographic information
VIF	variance inflation factor

Abstract

Geodata is missing to populate maps for usage of local communities. Efforts for filling gaps (automatically) by deriving data on human settlements using aerial or satellite imagery is of current concern (Esch et al., 2013; Pesaresi et al., 2013; Voigt et al., 2007). 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 (Goodchild and Li, 2012; Haklay, 2010; Senaratne et al., 2016). Here the quality of data produced by volunteers using the MapSwipe app was investigated. Three different intrinsic parameters of crowdsourced data and their impact on data quality were examined: agreement, user characteristics and spatial characteristics. Additionally, a novel mechanism based on machine learning techniques was presented to aggregate data provided from multiple users. The results have shown that a random forest based aggregation of crowdsourced classifications from MapSwipe can produce high quality data in comparison to state-of-the-art products derived from satellite imagery. High agreement serves as an indicator for correct classifications. Intrinsic user characteristics can be utilized to identify consistently incorrect classifications. Classifications that are spatial outliers show a higher error rate. 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.

1 Introduction

Nowadays, geographical data is everywhere. You're creating millions of data chunks when using your smartphone or browsing on a website. Your consumption habits will be analysed regarding your home location, and free time activities will be suggested based on your work journey. We use Google Maps or OpenStreetMap and in most western countries these maps offer us what we are looking for: street names, building footprints, restaurants and other points of interest. Nevertheless, this is only one part to the story. Indeed, there are many places in the world where people live but the maps are still empty and geodata is missing. This work will focus on these "forgotten" places that don't appear in the official maps yet.

Information on human settlements and spatial distribution of population are important for several use cases, but especially for emergency management and in response to humanitarian disasters (Dobson et al., 2000; Pesaresi et al., 2013). When Doctors Without Borders and other humanitarian organisations responded to the Ebola outbreak in 2014 detailed information on human settlements for the affected areas was hardly available (Lüge, 2014). Also, recent catastrophes like hurricane Irma and the Mexican earthquake, which happened both in September 2017, have emphasized the importance of up to date map data on buildings for disaster response activities (Yin, 2017).

Human settlements can be mapped either directly on the ground or using aerial or satellite imagery. While on the ground mapping can offer high precision, the main disadvantage is that it is expensive and time consuming. Especially in disaster situations the affected area may also be inaccessible. With the rise of new sensors which offer a higher resolution satellite imagery has become a unique source to capture geographic information on human settlements (Voigt et al., 2007). There are several global products based on remote sensing which offer information on human settlements. These products are available at several spatial and temporal resolutions. The Global Human Settlement Layer is produced at regional scale with 50 metre resolution (Pesaresi et al., 2013). The German Aerospace Centre (DLR) provides Global Urban Footprint data derived from global synthetic-aperture radar (SAR) datasets (Esch et al., 2013). Furthermore, companies like Facebook are providing data on human settlements based on optical satellite imagery (Zhang et al., 2016). However, these global products also have disadvantages such as the under-representation of low density built-up areas (Klotz et al., 2016). For land cover maps in general, Schultz et al. (2017) identify another issue regarding the long updates cycles of traditional satellite based products which negatively affects the temporal accuracy and fitness for purpose of the data generated. Today, existing human settlement data products offer a great benefit for many use cases, however data quality issues can hamper their application in disaster

scenarios or humanitarian aid. Thus, further methods and tools to enrich and complement the existing datasets are needed to address the known shortcomings.

Despite the usage of semi-automated methods and pre-processed data products, crowdsourcing is another tool which can help to collect information on human settlements and complement the data produced using satellite imagery analysis. Since the rise of the Web 2.0 in the early 2000's crowdsourced data has already complemented official information in several disaster situations. In 2010, crowdsourced data from OpenStreetMap and other crowdsourcing platforms supported disaster response efforts in Haiti (Zook et al., 2010). Text messages from social networks like Twitter have been used to find and model on the ground information for flood events (de Albuquerque et al., 2015; Vieweg et al., 2010) and earthquakes (Crooks et al., 2013; Earle et al., 2011). Many researchers have highlighted the potential of these new data sources to provide up-to-date data and an in-situ view in disaster situations (de Albuquerque et al., 2016; Meier, 2012; Zook et al., 2010). However, there is also a wide range of research studies investigating and questioning the reliability of the crowdsourced data (Flanagin and Metzger, 2008; Goodchild and Li, 2012; Haklay, 2010). These heterogeneous results and conclusions on data quality of volunteered geographic information demand for further research that focuses on improving specific applications and aspects of data quality.

This recent evolution of crowdsourcing platforms such as OpenStreetMap and the still pressing need for up-to-date and valid map data is also addressed by humanitarian organizations themselves in the scope of the Missing Maps Project, which was founded in November 2014 by American Red Cross (ARC), British Red Cross (BRC), Doctors Without Borders UK (MSF) and the Humanitarian OpenStreetMap Team (HOT). Within the collaboration of humanitarian organizations and together with research institutes a mobile application called MapSwipe was developed. MapSwipe is designed towards crowdsourcing information on the spatial distribution of human settlements in places where other datasets show poor quality or are entirely missing. The app has already supported to map around 400,000 square kilometres and has been used by more than 15,000 individuals. Given this successful evolution since the launch in July 2016, future development of the app, scaling up the crowdsourcing approach and further integration of the data into the existing Missing Maps workflows require a better understanding of the quality of the data produced. This will enable to make better use of each volunteered contribution and directly serves the purpose to fill the blank spots in the global OpenStreetMap.

Therefore, in respect to the existing work on approaches and quality analysis of volunteered geographic information, it is the purpose of this study to investigate the quality of the

geodata produced by MapSwipe volunteers regarding its ability to correctly represent the spatial distribution of human settlements. The approach utilized in this work will examine three different parameters of crowdsourced data and their impact on data quality: agreement, user characteristics and spatial characteristics. Additionally, a novel mechanism based on machine learning techniques will be presented to aggregate data from multiple users.

The remainder of this thesis is structured as follows. In the next section an overview on research towards crowdsourcing geographic information approaches is provided. This section will concentrate on the quality of volunteered geographic information and approaches for combining crowdsourced data and data derived using automated methods. Section 3 covers a description of the datasets used in this study and provides additional information towards the design and fundamental principles of the MapSwipe app. In section 4 the quality analysis methods based on reference datasets and intrinsic data characteristics will be explained in detail. Results and Discussion are presented in section 5. Finally, section 6 concludes the paper and makes recommendations for future research and the design of the MapSwipe app.

2 Related Works

In the past ten years, many researchers analysed what Goodchild (2007) coined "volunteered geographic information" in 2007. The research on these new datasets that are created by volunteers instead of experts is embedded in a wider field of study towards crowdsourcing, a term already introduced in 2006 by Howe (2006). In this part of the thesis approaches to crowdsource geographic information are presented in the first step. Secondly, quality assurance mechanisms will be described. Afterwards, the potential of crowdsourced datasets to support automated information extraction using machine learning techniques is explored. Finally, the section concludes with the identification of research gaps which will be addressed in the next parts of the thesis.

2.1 Crowdsourcing Geographic Information

When Goodchild (2007) introduced the concept of volunteered geographic information he pointed out the dramatic changes to the production of geographic information induced by the widespread engagement of large numbers of private citizens, often with little professional experience or formal qualifications. A long tradition of volunteered geographic data collection can be found in the work of citizen scientists (Haklay, 2013), but in the last ten years OpenStreetMap has become one of the major communities attracting several thousands of active volunteers every day (Neis et al., 2013).

Since 2007 many different crowdsourcing campaigns or tools have been addressing various problems which require different user skills and knowledge. Barrington et al. (2011) highlight three stages of a successful crowdsourcing campaign: (1) Split overall problem into many manageable components (micro task), (2) motivate many contributors to solve the tasks (crowdsource) and (3) combining multiple answers into a single result (aggregation). Micro task crowdsourcing tasks can not only be found in the domain of geography but in many different discipline and cover problems such as information finding or verification and validation (Gadiraju and Demartini, 2015). However, in this work the focus will be on projects incorporating geodata. Albuquerque et al. (2016) differentiate three different types of geographic information crowdsourcing approaches: "Classification", "Digitization" and "Conflation". In the following examples for each type will be presented.

2.1.1 Classification

Classification tasks can be seen as the easiest form of crowdsourcing since most of the time they require only little knowledge by the users (Albuquerque et al., 2016). For the geographic domain Albuquerque et al. (2016) define this category of as follows:

> *"In a classification task, the volunteer recognises features/objects in an existing piece of geographic information and then enriches this by adding an extra attribute that represents a value or category."*

Classification tasks have been deployed for several purposes. The Geo Wiki project aims at classifying satellite imagery to derive land use/land cover information (Fritz et al., 2009). Ali et al. (2016) propose a guided classification system for conceptual overlapping classes in OpenStreetMap. In the past several projects used crowdsourced classification tasks in a humanitarian context. In the aftermath of hurricane Sandy hitting the US coast in 2012 volunteers analysed aerial imagery to classify damaged infrastructures (Chan et al., 2013). The GEO-CAN Initiative coordinated damage mapping efforts after the 2010 Haiti earthquake and the 2011 earthquake in Christchurch, New Zealand (Barrington et al., 2011). The MapSwipe app too incorporates a crowdsourcing approach that builds upon classification tasks (Herfort et al., 2017).

Additionally, there are approaches to classify content derived from social media networks. For instance, the Micro Mappers platform was used by Imran and Elbassuoni (2013) to classify social media messages in regard whether they can provide useful information for disaster response purposes. Damage classification from twitter data for road network assessment is investigated by Schnebele et al. (2014).

2.1.2 Digitization

Digitization is the second type of crowdsourcing tasks. Likewise the automated object-based image analysis in the field of remote sensing, the product of digitization tasks is the delineation of readily usable objects mostly from satellite imagery (Blaschke, 2010). Following Albuquerque et al. (2016) digitization can be defined as:

> *"In a digitisation task, a volunteer also starts with the recognition of a real-world object/feature in an existing piece of information, but then goes further to produce a corresponding digital representation (i.e., usually a vector geographic object)."*

OpenStreetMap is the major project which collects crowdsourced digitisations of topographic map features. Since 2007 a remarkable growth both in data entries and number of documented map features is observed (Mocnik et al., 2017). In the foundation phase of the project mapping the street network gained prime attention (Haklay, 2010; Neis et al., 2011), whereas nowadays also several sophisticated features such as damaged buildings are part of the global OpenStreetMap (Westrope et al., 2014).

However, digitisation tasks are also applied by other crowdsourcing projects. For instance, Wikimapia provides a crowdsourced dataset of georeferenced features annotated with rich descriptions. This kind of dataset can function as crowdsourced place name index and offers an alternative to traditional gazetteers (Goodchild and Glennon, 2010). The Forest Watcher citizen science project crowdsources data on deforestation in Brazil. In this project volunteers digitize areas on satellite imagery which have been cleared due to human activity (Arcanjo et al., 2016). Hillen and Höfle (2015) propose a crowdsourcing approach to digitise building footprints from earth observation data incorporating the reCAPTCHA concept.

In disaster situations and for humanitarian purposes OpenStreetMap provides a unique set of tools combined with an active community and therefore already supported response and preparedness activities in several countries. The Humanitarian OpenStreetMap Team created a crowdsourcing project management tool, the HOT Tasking Manager, which improved the ability to coordinate the mapping activity of hundreds of people. After the Haiti earthquake 2010 more than 490 mappers worked together, and following typhoon Yolanda in November 2013 more than 1,500 volunteers mapped the affected areas in the Philippines (Palen et al., 2015). Since 2014, the Missing Maps Project aims at proactively mapping regions vulnerable to crises and motivated more than 6,000 new volunteers to contribute to OpenStreetMap (Dittus et al., 2017).

2.1.3 Conflation

Conflation is proposed as the third type of crowdsourcing geographic information. The term is related to the fields of spatial data integration and data fusion and describes the process of combining geographic information from overlapping sources to retain accurate data, minimize redundancy, and reconcile data conflicts (Longley, 2005). Albuquerque et al. (2016) provide the following definition:

> *"In a conflation task, volunteers interpret more than one source of geographic information, identify matching objects/features and bring them in relation to produce new geographic information."*

For the OpenStreetMap community, conflation tasks are often performed by users with higher level of experience and for tasks that require specific editors such as JOSM, which are capable of visualising geographic information in several layers. In this manner street level photography derived from Mapillary data (Juhász and Hochmair, 2016), GPS traces and satellite imagery can be combined to map features such as sidewalks or kerbsides (Voigt et al., 2016).

Although the conflation of crowdsourced geographic information with other (authoritative or commercial) data sources has already been proposed and is a promising avenue for future research (See et al., 2016), only few projects have incorporated these tasks in their workflows. In the context of disaster response, Anhorn et al. (2016) present an approach to crowdsource the validation and updating of spontaneous shelter camps in Kathmandu, which evolved after the 2015 Nepal earthquake. The Geo-Wiki platform is another example, in which volunteers use visualisations from multiple land cover datasets which are conflated with geo-tagged pictures to decide which land cover type is found on the ground (Fritz et al., 2009).

2.2 Quality Assurance for Crowdsourced Geographic Information

The emergence of volunteered geographic information and crowdsourcing as a tool to produce new datasets was accompanied by discussions on data quality and credibility right from its beginning (Flanagin and Metzger, 2008; Goodchild, 2007).

The heterogeneity of volunteered geographic information regarding completeness, thematic accuracy or fitness for purpose is stressed by many researchers (Ali et al., 2016; Ballatore and Zipf, 2015; Fonte et al., 2015; Girres and Touya, 2010). In early research quality was often measured in comparison to a reference dataset (Girres and Touya, 2010; Neis et al., 2013). Recent research has extended the quantitative description of data quality of volunteered geographic information by integrating intrinsic characteristics of the datasets and its producers and the spatial context of the data created (Barron et al., 2014). These indicators allow a prediction of the data quality even if no reference dataset exists.

Furthermore, these intrinsic indicators also reveal the impact of different quality assurance strategies in the field of crowdsourcing geographic information. Goodchild and Li (2012) have perceived three different mechanism for assuring data quality. The “crowdsourcing approach” relies on the validation of individual contributions by a larger group of volunteers that mutually reviews their edits. The reputation of individuals to produce reliable data functions as the basis for the “social approach”. Spatial relationships between individual contributions and geographical plausibility are central to the “geographic approach”. In each case, these approaches put a single aspect of the dataset and the underlying crowdsourced contributions at its core. The crowdsourcing approach highlights the importance of agreement among volunteers for data quality. User characteristics are the main driver regarding the social approach and spatial characteristics are condensed by the geographic approach. In the following, research results towards these three aspects of crowdsourced data and the associated quality assurance mechanism will be summarized.

2.2.1 Agreement

Goodchild and Li (2012) describe this approach as the crowdsourcing approach referring to the solution of problem by incorporating several people without respect to their individual qualification. The authors also accentuate the role of Linus’s Law for data quality assurance. In principal the law hypothesizes an increase in data quality if a larger number of volunteers contributes data.

Haklay et al. (2010) have examined the validity of Linus Law for the positional accuracy of OpenStreetMap data in England in comparison to authoritative data from Ordnance Survey. Their results confirm that a higher number of contributors is associated with higher

positional accuracy. However, the authors also show that this relationship is not linear. In their study, data quality remains at a high level for 15 or more contributors per square kilometre, whereas the increase in quality is highest for the first five contributors to an area. The completeness of the street network in Germany can be estimated by the growth in length for individual street categories in combination with the number of active contributors for a specific region (Barron et al., 2014). For the digitisation of deforested areas, Arcanjo et al. (2016) utilize intensity maps together with agreement indices to aggregate multiple contributions. Their results further underline the non-linear relationship between overall accuracy and number of contributors. For their study, the overall accuracy increased with higher number of contributors, but then peaked and started to decrease. The results emphasize the importance of agreement among users for crowdsourcing digitization tasks.

The concept of agreement is even more adopted for the analysis of crowdsourcing classification tasks. Given the nature of this type of tasks, a major challenge lies in the aggregation of multiple answers from different users into a single consistent results (Barrington et al., 2011). Many projects using a classification approach aggregate data based on majority agreement among different contributors. Albuquerque et al. (2016) aggregate classifications if at least 50 % of all contributors agree on the presence of roads or settlements in satellite imagery in a selected task area. Salk et al. (2013) use majority agreement for crowdsourced cropland mapping.

The studies reviewed so far used majority agreement as a mechanism to obtain a single answer from multiple inputs, nevertheless there are also authors who point out the disadvantages of such aggregation methods (Salk et al., 2016). A problem of majority agreement is that it assumes that all experts are equally good. Nevertheless, given a scenario with one expert and several beginners which tend to give incorrect answers, majority agreement leads to a wrong aggregated result (Raykar et al., 2010). The problem is also addressed by Simpson et al. (2012), who use a Bayesian approach to combine multiple ratings from users with varying performance due to radically different expertise and domain knowledge.

2.2.2 User Characteristics

Since volunteered geographic information or crowdsourced data is not produced by experts with known credibility and reputation, but by volunteers with varying levels of experience and skills, user characteristics and their implications for quality have attracted many researchers. This is also addressed by Goodchild and Li (2012) and their description of the social approach.

A central aspect for most crowdsourcing applications lies in a vast contribution inequality amongst different users. In general, for many projects most data is produced by very few contributors. For the OpenStreetMap project several researchers have proved that a minority of less than 5 % of all users accounts for more than 90 % of all contributions (Haklay, 2016; Neis and Zipf, 2012; Yang et al., 2016). High contribution inequality is also observed for crowdsourced classification tasks (Albuquerque et al., 2016). Accordingly, some researchers differentiated volunteers using their number of total contributions. For instance, Neis et al. (2013) group OpenStreetMap users into “Nonrecurring Mapper”, “Junior Mapper” and “Senior Mapper”.

Additionally, the origin of the contributors and its relationship to the edits have been identified as important aspects of user characteristics. Regarding crowdsourced land cover classification Comber et al. (2016b) show that differences between user groups have an impact on the quality of the crowdsourced data and thus may provide a potential source of error and uncertainty. The impact of local and remote mapping on OpenStreetMap data quality is analysed by Eckle and de Albuquerque (2015). The authors discuss the quality of remote mapping. The distinction between local and remote mapping may be important, since local OSM contributors know their area of interest and rely upon local knowledge, whereas the sole basis for remote mapping is often satellite imagery. Furthermore, cultural differences may affect the semantic representation of features within the global OpenStreetMap database and the activity and mapping practices of the contributors (Ballatore and Zipf, 2015; Quattrone et al., 2014).

Expertise and domain knowledge are important user characteristics as well. Exel et al. (2010) identify local knowledge, experience and recognition as important dimension to assess user quality. Its importance is also highlighted by Dittus et al. (2016). The authors show that initial project experience of contributors can be crucial, especially when the crowdsourcing task requires some degree of expertise. In their study, Dittus et al. (2016) quantified each user’s degree of prior OSM experience utilizing the number of days on which they had contributed to OSM before they joined their first HOT project.

2.2.3 Spatial Characteristics

When researchers refer to the diverse quality of crowdsourced data, they often emphasize spatial disparities. Furthermore, geographic laws are applied and investigated to measure and understand these heterogeneities. This research stream has been characterized by Goodchild and Li (2012) as the geographical approach towards assuring the quality of volunteered geographic information.

In general, for projects such as OpenStreetMap data quality may vary a lot regarding which regions will be considered. Hagenauer and Helbich (2012) analyse urban areas in OpenStreetMap and propose a machine-learning approach to assess unmapped or only partially mapped areas. Their results show the spatial heterogeneity of data completeness on the one hand, but on the other they also reveal that the applicability of the proposed solution itself is highly dependent on location. Touya and Antoniou (2016) assess the quality of points of interest from OpenStreetMap intrinsically by utilizing their spatial relations to nearby features such as buildings or streets.

Towards the analysis and filtering of social media data Herfort et al. (2014) propose an approach that combines hydrological data and digital elevation models to prioritize crisis relevant text messages. Yan et al. (2017) use digital elevation models and tourism related POIs from OpenStreetMap to apply a viewshed analysis to differentiate Flickr contribution from locals and tourists.

For a crowdsourcing classification task Albuquerque et al. (2016) show, that geographical features such as the size of a building have significant influence on the likelihood of a correct classification by a volunteer. These results are extended by Herfort et al. (2017) who confirm that disagreement among volunteers is not randomly distributed in space but caused by specific characteristics of the underlying geographical features. For crowdsourced land cover classifications Comber et al. (2016a) use a geographically weighted average approach to infer land cover present at locations on a 50 kilometres grid. Their approach obtains higher correspondence to a reference dataset in comparison to traditional non-spatial aggregation methods.

Citizen science projects dealing with biodiversity monitoring incorporate spatial characteristics of the data produced to assess the plausibility of observations (Jacobs, 2016). Especially for projects with an increasing amount of incoming data, the traditional approach based on expert validation is not feasible anymore. The analysis of the spatial and temporal context of user contributions using existing observation data or information on environmental context can assist to validate new contributions.

2.3 Interplay of Crowdsourcing and Machine Learning

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 et al., 2006). Through the online game users helped to annotate information about what type of object 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. Hara et al. (2014) present another approach that incorporates machine learning, computer vision and crowdsourcing to extract information on curb ramps from google street view images. The results of this study show that the combined approach can achieve results with the same quality as manually labelled datasets, but is able to reduce the time costs by 13 %. Russakovsky et al. (2015) developed a system that collects bounding box annotations through crowdsourcing using imageNet data for learning object detectors, whereas Deng et al. (2013) introduced an online game to crowdsource discriminative features from photographs to train a classifier to distinguish bird types.

Also in the domain of social media research crowdsourcing is applied for the supervision of machine learning tasks. Imran et al. (2014) developed the AIDR (artificial intelligence for disaster response) platform, which collects human annotations of text-based social media messages and then automatically classifies new text messages as either informative or non-informative.

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 combines machine learning for identifying potential villages from very high-resolution satellite imagery and uses a crowdsourcing classification approach to validate the generated polygons. 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 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 OpenStreetMap to

detect crosswalks on satellite imagery. Additionally, their approach utilizes the crowdsourcing application MapRoulette to validate the generated results and to directly add the detected features into the OpenStreetMap database. Hagenauer and Helbich (2012) use a machine learning approach to model unmapped residential areas in OSM. Their approach uses OSM data for training purposes.

In addition to the research projects presented, there are further applications which integrate crowdsourced data. The DeepOSM project (Johnson, n.d.) aims to identify road features in satellite imagery by training neural networks with OSM data. 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 OpenStreetMap.

2.4 Research Gaps

The previous sections have shown, that different types of crowdsourcing can be applied to generate geographic information. Crowdsourced classification tasks enable large numbers of volunteers to contribute, even if they have little or no experience. The MapSwipe app takes advantage of this fact. However, it has also been shown, that the quality of VGI data is still in question and that heterogeneity can be assumed for most datasets. Furthermore, different strategies to assure quality and influencing parameters have been presented. This affirms the conclusion that agreement, user characteristics and spatial characteristics need to be considered. Finally, there is a growing field of applications which use crowdsourcing in combination with machine learning techniques.

The review of the recent research reveals that there are still gaps that need to be addressed to gain further insights about the quality of crowdsourced geographic information. Whereas crowdsourcing classification tasks have been applied to many different research domains, the quality of the MapSwipe dataset and the controlling variables are still underexplored. Furthermore, there are many studies which incorporate crowdsourced datasets to train machine learning models. However, only few work has been done towards utilizing machine learning techniques to improve the crowdsourcing workflow itself. It is still not fully understood how automated classifiers could help to aggregate crowdsourced classifications in respect to agreement, user characteristics and spatial characteristics. This work will therefore focus on the following research questions (RQ), each addressing a specific information need towards better understanding the quality of crowdsourced data. Whereas the questions are explored for the MapSwipe dataset, they are significant for the quality assurance of crowdsourcing geographic information in general.

- **RQ1**: How do agreement and redundancy affect data quality?
- **RQ2**: How do intrinsic user characteristics affect data quality?
- **RQ3:** How do spatial characteristics of the crowdsourced classifications affect data quality?
- **RQ4:** To what degree can automated classifiers considering agreement, user characteristics and spatial characteristics enhance the aggregation mechanism to improve data quality?

The research questions identified will be addressed in a case study on 55 MapSwipe projects. The following sections of this work will further describe the methods applied and datasets used.

3 Case Study and Datasets

This work focuses on crowdsourced data produced by volunteers using the MapSwipe app. For this study, data from 55 MapSwipe projects dealing with the task of building classification are considered. The geographic scope of these projects will be explained in the first part. The app and the data produced will be explained in the second step. Thirdly, reference datasets will be explained. These datasets are used to assess the quality of the MapSwipe data. All datasets used in this study are made available to the public through the research data repository at Heidelberg University (Herfort, 2017).

3.1 Case Study

MapSwipe project from different regions of the world are considered in this research.

Figure 1 provides an overview where projects are located. Whereas most of the projects are designed to map buildings on the African continent, there are also projects in south east Asia and Latin America. The first project of our case study has been launched on July 30th, 2016. The last project has been finished on July 25th, 2017. Accordingly, this research covers about one year of MapSwipe contributions.

The projects analysed in this work have a size ranging from very small ones, e.g. project “Bijagos Islands 6” to very big projects, e.g. “Nigeria for MSF 1”. In total all projects cover an area of more than 200,000 square kilometres. Furthermore, there are several projects which are part of bigger mapping activations coordinated by humanitarian organisations. The French NGO CartONG runs a large-scale mapping project in Madagascar which comprises of 11 projects. MSF UK manages projects in different regions but with a strong focus on sub-Saharan Africa including parts of Nigeria, South Sudan and Sierra Leone. A third group of projects is connected to the Malaria Elimination Campaign organized by the Clinton Health Access Initiative and supported by the Humanitarian OpenStreetMap Team. The projects in Guatemala, Botswana, Laos and Cambodia are part of this program.

A more detailed perspective is chosen for a second study region containing four projects in south west Laos (“6807 - Eliminate Malaria Laos”, “6794 - Eliminate Malaria Laos 2”, “6930 - Eliminate Malaria Laos 3”, “7064 - Eliminate Malaria Laos 4”). These projects are part of the Malaria Elimination Campaign mentioned above. Within the extent of these projects all buildings have already been mapped and validated in OpenStreetMap after finishing the MapSwipe workflow. This provides the unique situation, that building footprint reference data can be extracted from the OpenStreetMap database. Figure 2 shows where these projects are located.

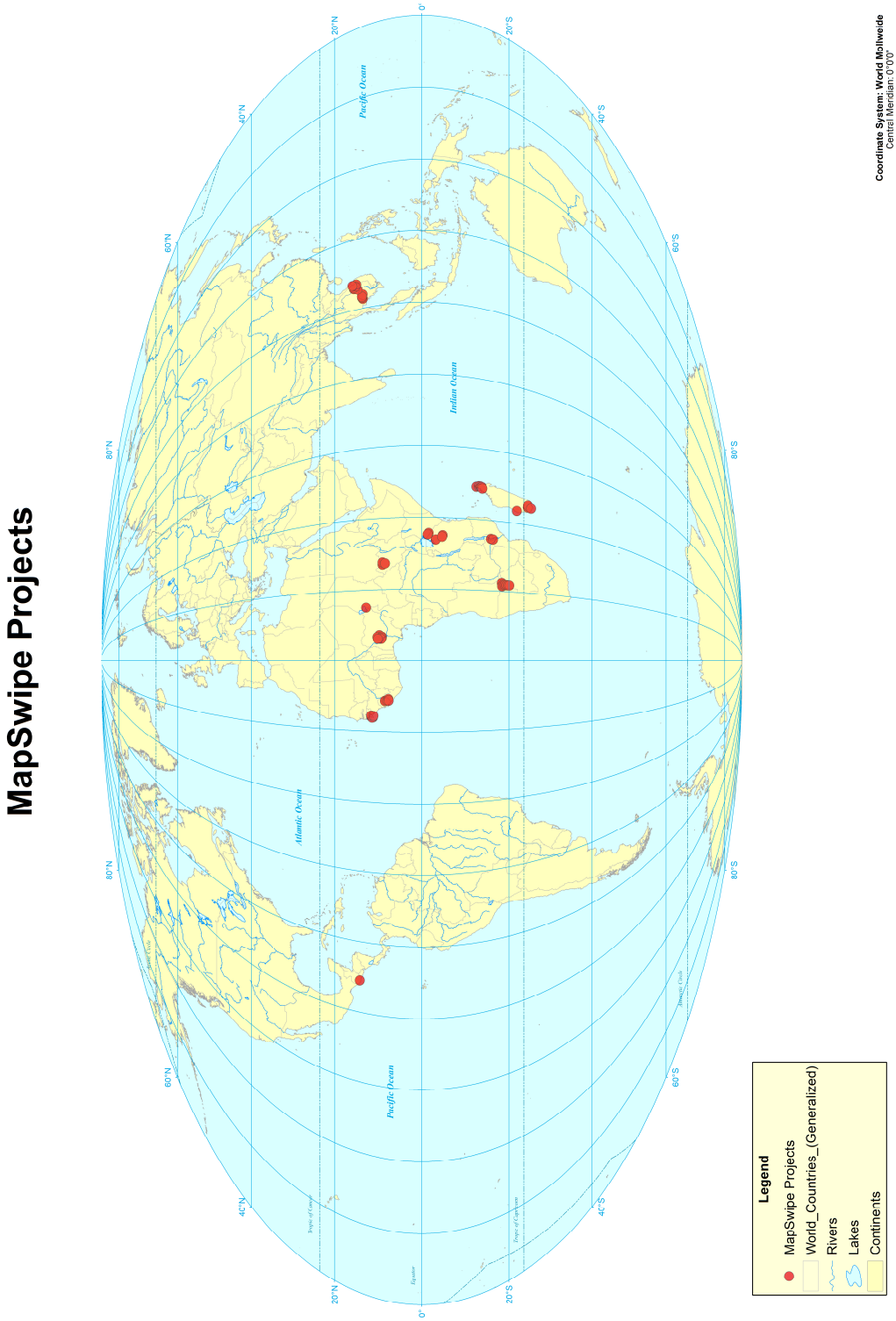

Figure 1 MapSwipe Projects

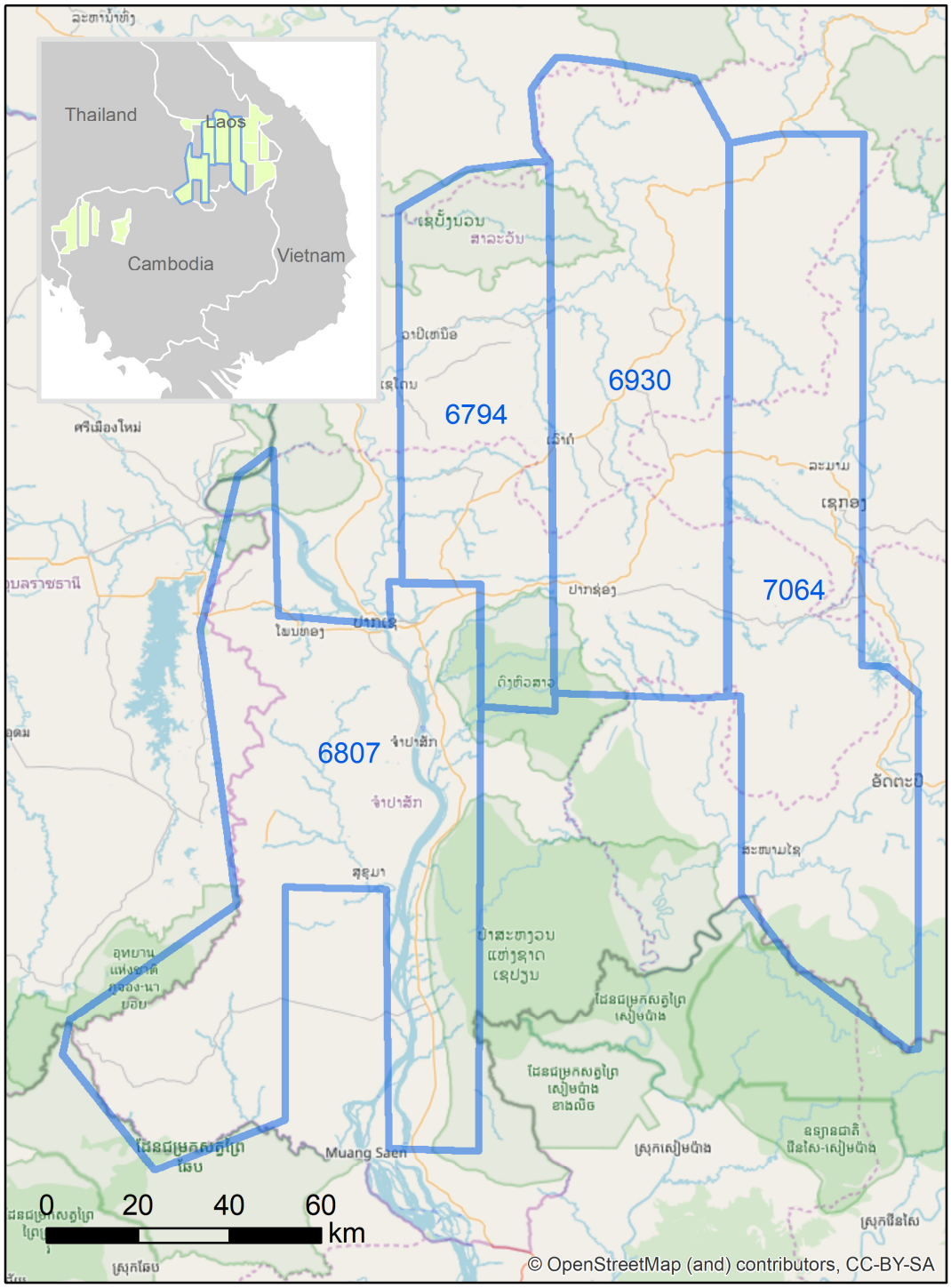

Figure 2 MapSwipe Projects in Laos

3.2 MapSwipe

The MapSwipe app was developed by MSF as part of their support for the Missing Maps project, which was established in November 2014 by the American Red Cross (ARC), the British Red Cross (BRC), HOT and MSF. In this section, the app and the underlying data model will be explained. Data from 55 projects containing more than nine million tasks and more than 30 million individual classification results contributed by about 7,500 users are considered in this study (see Table 1). The selected projects in Laos cover about 10 % of the global dataset.

Table 1 Size of the MapSwipe Dataset

Projects	**Tasks**	**Results**	**Users**
All (55 projects)	9,081,290	31,932,500	7,518
Laos Selection (4 projects)	941,589	3,275,380	1,534

3.2.1 MapSwipe App

MapSwipe is a mobile crowdsourcing application designed for smartphones and tablets to generate geographic information from satellite imagery. It can be downloaded from Google's Playstore or Apple's Appstore. The app incorporates a crowdsourcing classification task. MapSwipe users classify tiles of satellite imagery into four different categories ("No", "Yes", "Maybe", "Bad Imagery"). For the projects analysed in this work users are always asked whether they can spot buildings in the satellite imagery. They can indicate this by tapping and swiping.

The mapping interface consists of six squares, each square representing one tile of satellite imagery. Users can mark each of these squares individually. If no buildings are present in the shown imagery, swiping to the left will load six new squares to be classified. At the top of the interface (see Figure 3) users are instructed what features to look for, in our case this is always buildings. The projects analysed in this study use Bing satellite imagery as the basis for the crowdsourced classifications. Given the different locations of the projects, imagery is provided from different sensors and captured at various timestamps.

At the bottom, a progress bars shows how many tasks the user already processed of the current mapping session. This session is also referred to as "group". In section 3.2.2 the definition of a group will be explained in further detail.

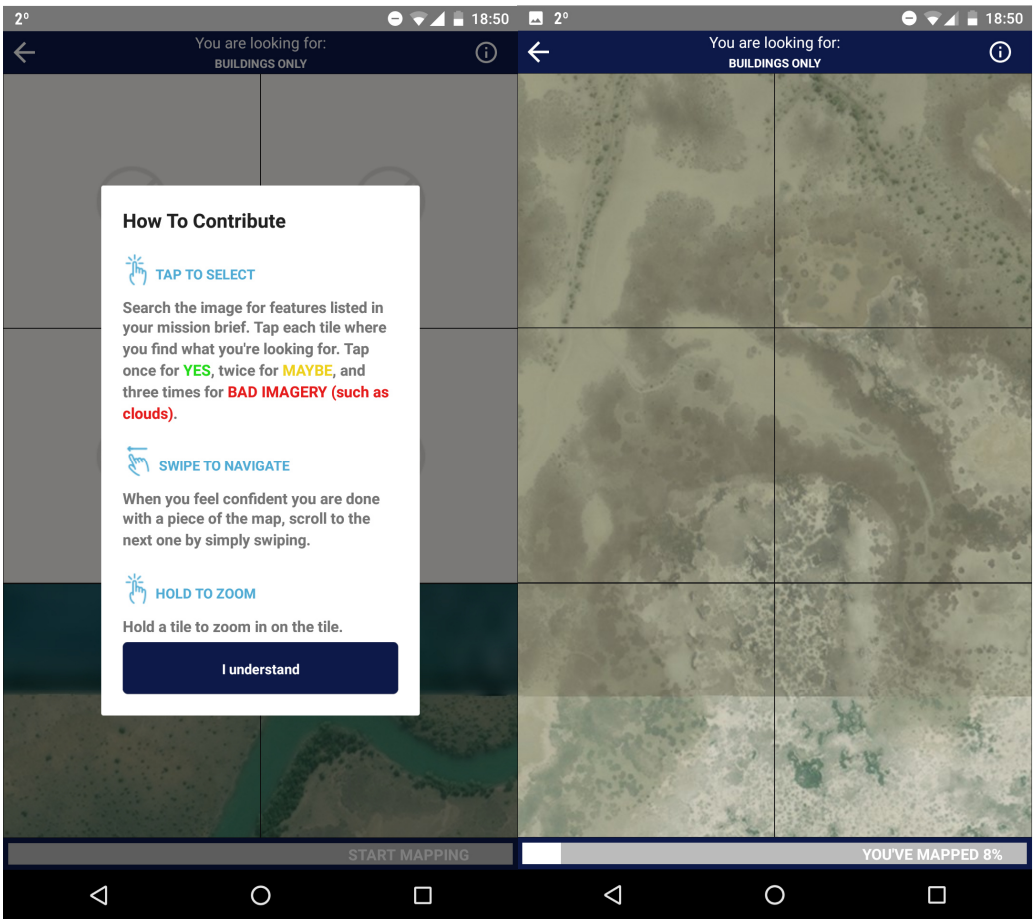

Figure 3 MapSwipe App Interface

3.2.2 MapSwipe Data Model

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 (Figure 4).

A project in MapSwipe is primarily characterized by an area of interest, a set of satellite imagery tiles and a feature type to look for. In addition, each project defines the number of users that are requested to classify each individual satellite imagery tile. Furthermore, background information on the context of the mapping request is provided. The information is uploaded through a web form by project managers. Table 2 provides further details on each project parameter.

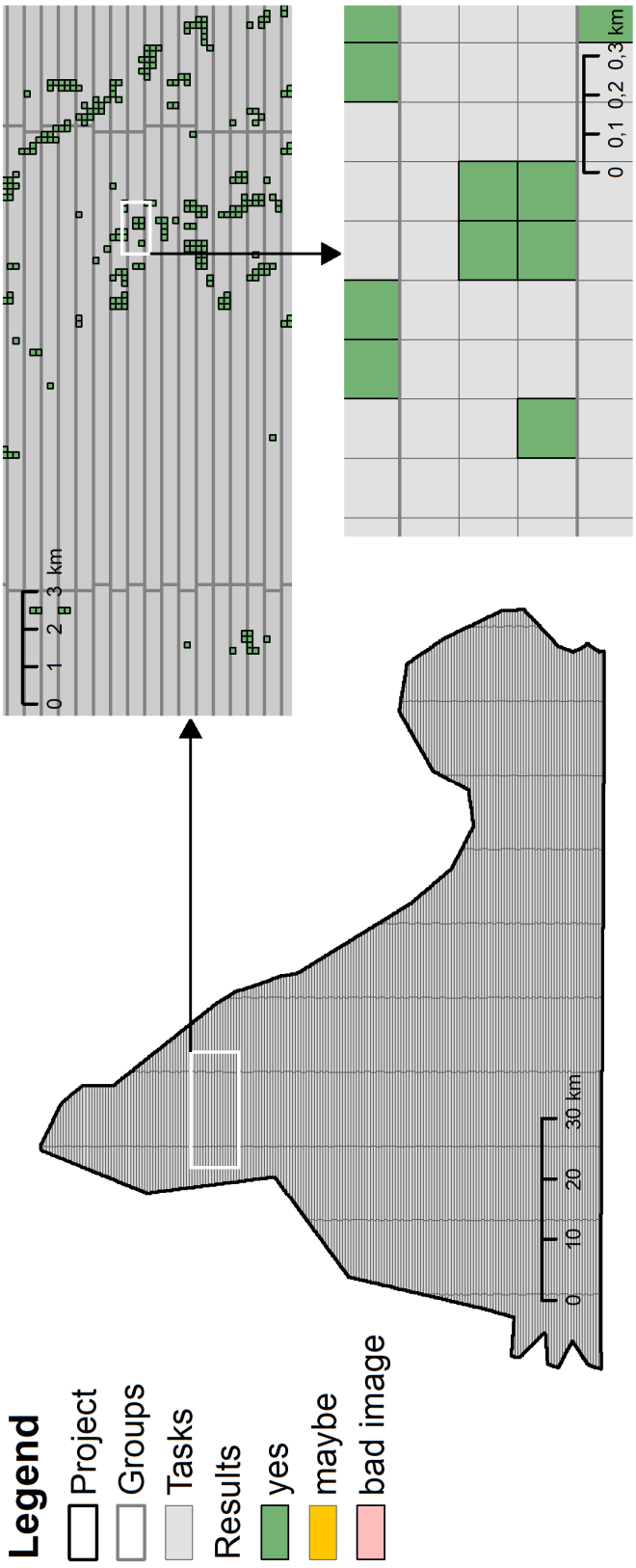

Figure 4 Data Model: Project, Groups, Tasks and Results for Project "Madagascar 9"

Table 2 Project Information

Parameter	Description
Id	Each project has a unique identifier. This Id will be generated automatically and cannot be chosen by the project manager.
Name	Each project has a name, which will be shown in the app. Often project in the same region have similar names, e.g. “Madagascar 1” and “Madagascar 2”.
Geometry	A project is characterized by its multi polygon geometry. Thus, projects can theoretically consist of several distinct regions. Nevertheless, most projects focus on a single area.
Redundancy	Project managers can define how often each individual task will be classified by MapSwipe volunteers at minimum. For most projects this redundancy is set to three.
Imagery Provider	This parameter refers to a provider of a tile map service. For projects of this study imagery is provided by Bing. Tiles of satellite imagery can be obtained from a tile map service endpoint. Each tile can be identified using a quad key representation of its x, y and z coordinates. Tiles are projected in WGS 84 Web Mercator (Auxiliary Sphere). This corresponds to the EPSG code 3857.
Project Details	The project details describe the goal and scope of the project. This is visualized in the app and is important to stir the volunteer’s motivations. In general, the project is described by five to ten sentences.
Look For	The parameter defines which type of features are mapped in the project. This will be visualized in the mapping interface of the app. For this study, volunteers are always asked to look for buildings only.

To create a new mapping task, the overall project extent is split up into many single tasks. Tasks are the smallest unit in the MapSwipe data model. They are derived from the area of interest by gridding it into many small equal-sized rectangular polygons. Each task corresponds to a specific tile coordinate from a tile map service (TMS) using a web Mercator projection as its geographical reference system. Therefore, each task is characterized by a geometry and its tile coordinates, which describe its x, y and z position. For the projects analysed in this project, the tiles for all tasks are generated at zoom level 18. Taking the latitude of each task location into account the satellite imagery has a maximum

spatial resolution of ~ 0.6 meter at the equator. Table 3 describes each task parameter in greater detail.

Table 3 Task Information

Parameter	Description
Id	Each task can be identified by its Id. The Id is a composition of its position in the corresponding tile map system, which can be described by the x, y and z coordinates.
Tile Z	The z coordinate of the tile defines the zoom level. Greater values for z will correspond to higher spatial resolution of the corresponding image. For most regions Bing provides images up to zoom level 18. For aerial imagery or images captured by UAVs even higher z values are valid.
Tile X	The x coordinate characterises the longitudinal position of the tile in the overall tile map system taken the zoom level into account. The x coordinates increase from west to east starting at a longitude of -180 degrees.
Tile Y	The y coordinate characterises the latitudinal position of the tile in the overall tile map system taken the zoom level into account. The latitude is clipped to range from circa -85 to 85 degrees. The y coordinates increase from north to south starting at a latitude of around 85 degrees.
Geometry	Each task has a polygon geometry, which can be generated by its x, y and z coordinates. At the equator the task geometry is a square with an edge length of around 150 metres covering circa 0.0225 square kilometres. Due to the web Mercator projector the task geometry will be clinched with increasing distance to the equator. At the same time the area per task will decrease.
Tile URL	The tile URL points to the specific tile image described by the x, y, and z coordinates. Usually, the image has a resolution of 256 x 256 pixels. However, some providers also generate image tiles with higher resolution (e.g. 512 x 512 pixels).

Single MapSwipe projects can contain up to several hundred thousand tasks. This can pose a challenge to fast and performant communication between clients and server if many volunteers contribute data at the same time. Therefore, groups have been introduced to reduce the amount of client requests on the backend server. Groups consists of several tasks, that will be shown to the user in one mapping session. The grouping algorithm uses the

extent of a project as an input and generates chunks of tasks lying next to each other. Each group has a height of three tasks and a width of approximately 70 tasks. Further details on each group parameter can be obtained from Table 4.

Table 4 Group Information

Parameter	Description
Id	Each group can be identified by its Id.
Tasks	Each group contains several tasks. The information for all tasks (see Table 3) of the group will be stored in an array.
Geometry	The group geometry is defined by the union of all assigned task geometries.
Completed Count	Once a group has been completely mapped by a volunteer the completed count of the corresponding group will be raised by one. The completed count of the group is used to assess the overall progress of each project. For doing so the completed count is compared to the redundancy required (see Table 2). During the mapping process groups will be served in ascending completed count order. Thus, groups with low completed count will be served first.

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. “No Building” classifications can only be modelled retrospectively for groups where a user also submitted at least one “Yes”, “Maybe” or “Bad Imagery” classification. Figure 5 provides an overview on this workflow. Initially, for user A all groups are selected, where this user submitted at least one “Yes”, “Maybe” or “Bad Imagery” result. For these groups all intersecting tasks are selected 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. Therefore, the workflow can only track “No Building” classifications for groups for which a user contributed at least one time differently. For each result task Id, timestamp and user Id are captured additionally. Table 5 describes these parameters in further detail.

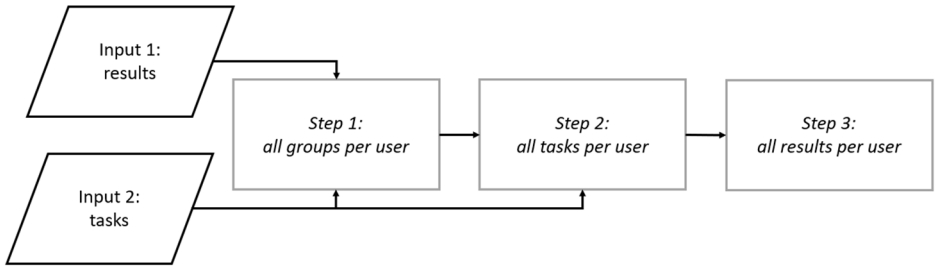

Figure 5 Workflow to derive all results including "No Building Classifications"

Table 5 Result Information

Parameter	**Description**
Id	Each result can be identified by its Id. The Id is a combination of task Id and user Id.
Task Id	Each result corresponds to a specific task, which can be described by its Id.
User Id	Each result is contributed by a specific user. Users can be identified by their Id.
Timestamp	The timestamp provides information on the time the user completed the group and uploaded the result data. Results within the same group are assigned the same timestamp.
Result	This parameter describes the given answer. 1 corresponds to “Yes”, 2 corresponds to “Maybe” and 3 corresponds to “Bad Imagery”. Each user can only submit one result per task.

3.3 Reference Datasets

In this study three reference datasets are deployed to validate the crowdsourced classifications. First a set of manually labelled data will be explained. In the second step, a reference dataset derived from the MapSwipe data itself will be introduced. Thirdly, a reference based on OpenStreetMap building data for four projects in Laos is described.

3.3.1 Manually Labelled Reference Dataset

Given the large size of the MapSwipe dataset, it is impossible to validate every single classification. Therefore, tasks were stratified according to the overall user agreement and afterwards selected randomly. The stratified random sampling approach applied in this study is well described by Card (1982) and will be further explained in section 4.1.2. For the sampling tasks with six or more contributors were selected to enable the analysis of the impact of redundancy on overall data quality. In total, the reference dataset consists of 1,224 tasks (Table 6). The tasks of the reference sample have been labelled manually by two experts. For each task the experts were asked to classify into one of the following three classes: “no building”, “building”, “bad imagery”. Cases of disagreement between the experts were considered separately and discussed between both experts to reach a final decision.

Table 6 Expert Reference Dataset

Projects	Tasks	Results	Users
All (55 projects)	1,224	9,340	2,594

3.3.2 Crowd Based Reference Dataset

Since external reference datasets are not available for most of the regions analysed in this study, the decision was made to use the majority answer for each task as reference to assess user characteristics. Thus, this reference dataset contains 9,081,290 tasks. Nevertheless, this also limits the validity of the results to a certain degree, since the majority answer can lead to incorrect classifications as well. This needs to be considered when interpreting the results based on this reference dataset.

3.3.3 OpenStreetMap Reference Dataset

The OpenStreetMap reference dataset covers the extent of four MapSwipe projects in Laos (“Eliminate Malaria Laos”, “Eliminate Malaria Laos 2”, “Eliminate Malaria Laos 3”, “Eliminate Malaria Laos 4”) depicted in Figure 2. In total, it contains 324,152 individual buildings. The data was obtained from bbbike’s planet.osm extracts in ESRI shapefile

format. The data is directly extracted from the OpenStreetMap planet file which contains all OpenStreetMap edits.

To a great extent the OpenStreetMap data was captured by HOT volunteers. The mapping efforts have been organised using the HOT Tasking Manager tool. The area of interest corresponds to the following Tasking Manager project Ids: 3358, 3359, 3362, 3364, 3383, 3391, 3392, 3393, 3399 and 3400. The mapping has been part of Clinton Health Access Initiative's malaria program. All projects have been completely mapped and validated in the Tasking Manager. Thus, a first quality assurance was already applied. The HOT Tasking Manager projects rely on aggregated and processed MapSwipe data. Therefore, only built up areas that have been identified by the MapSwipe volunteers are considered for the detailed mapping in OpenStreetMap. 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 OpenStreetMap 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. For task which contain a building, the number of buildings and the sum of the area of all buildings is generated additionally (Table 7).

Table 7 Reference Dataset based on OpenStreetMap Building Information

Parameter	**Description**
Task Id	The reference dataset is computed for MapSwipe task geometries. The corresponding MapSwipe task Id is captured.
Building	Boolean value (0,1). All tasks which intersect with at least one building geometry from OpenStreetMap will be assigned 1.
Building Count	This parameter refers to the number of individual buildings that intersect with the task geometry.
Building Area	This parameter refers to the sum of the area of all buildings that intersect with the task geometry.

4 Methods

In this section the overall methodology will be presented. The analysis is divided into four parts. In the first step the quality of the MapSwipe dataset will be evaluated using agreement amongst contributors. The second part investigates intrinsic user characteristics. In the third step spatial characteristics of the dataset are analysed. Finally, the performance of a machine learning approach to combine crowdsourced classifications from several users is assessed.

4.1 Agreement Analysis

Agreement is a central variable to aggregate multiple answers into a consistent single result. This section investigates the quality of the MapSwipe dataset in respect to different agreement-based aggregation methods and number of classifications per task in comparison to an expert reference dataset. The total number of classifications and the distinct count for each individual class are captured for each task to calculate agreement. In this study, three different variables are computed.

4.1.1 Definition of Agreement

Scott's Pi is calculated using Equation 1, as the proportion of agreeing pairs of classifications out of all the possible pairs of assignment. In the following, this will be referred to as agreement. This is computed following Fleiss (1971). Accordingly, *n* is the number of ratings per task, *k* is the number of classes in which assignments are made, and n_{ij} is the number or users which assigned the *i*-th task to the *j*-th class. The building classification index is defined as described by Equation 2. The index indicates the proportion of building classifications on all classifications per task. Likewise, the no building classification index is calculated as the proportion of no building classifications on all classifications per task (Equation 3).

Equation 1: Scott's Pi	$$P_i = \frac{1}{n * (n-1)} * \sum_{J=1}^{k} n_{ij}^2 - n_{ij}$$
Equation 2 Building Classification Index	$$\text{BI} = \frac{n_{building}}{n}$$
Equation 3 No Building Classification Index	$$NBI = \frac{n_{no\ building}}{n}$$

4.1.2 Analysis of Agreement for Correct and Incorrect Classifications

The analysis starts with a description of the distribution of agreement, building index and no building index for the MapSwipe dataset. For doing so, violin plots are computed, which combine a density distribution visualization and boxplot.

Furthermore, correct and incorrect user contributions are defined using the expert reference dataset. A user contribution is considered as correct, if it matches the expert answer. “Maybe” contributions by volunteers are considered as correct only if they contain a building. “Maybe” contributions for which the experts classified as “no building” or “bad imagery” are considered incorrect. For each agreement variable a conditional density plot is generated. These plots depict how the conditional distribution of the categorical variable *y* (in our case “correct”, “incorrect”) changes over a numerical variable *x* (in our case agreement, building index, no building index).

4.1.3 Aggregation and Redundancy Analysis

Majority agreement is a common method to aggregate multiple results into a single answer. Besides majority agreement several further aggregation rules can be defined considering different number of users per task and varying aggregation thresholds. The aggregation threshold defines the minimum number or proportion of classifications of type *x* (e.g. building), to aggregate the overall result into *x* (e.g. building).

For example, given three contributions from different users, three possible aggregation methods regarding building classifications exist. A strong aggregation threshold would require that all users agree (consensus agreement) on a building classification to aggregate the final answer as building. Majority agreement would only require minimum 50 % of the users to choose building (two out of three users for our example). Additionally, aggregation can also incorporate building classifications from a minority. For our example, one building classification per task would be enough, to aggregate as building in the result. According to this logic the number of users per task determines how many different aggregation methods can be utilized. The higher the number of contributors, the more potential aggregation thresholds exist. In this research, tasks with six users are considered. Therefore, the analysis is conducted for up to six different aggregation methods.

To evaluate the quality of the MapSwipe dataset depending on the aggregation threshold and number of users the expert reference dataset (see section 3.3.1) is utilized. A confusion matrix is generated using the aggregated answers and the answers from the expert reference dataset. The confusion matrix illustrates true positives (TP), false positives (FP), true

negatives (TN) and false negatives (FN). Since three different classes are investigated these FP, TN and FN split up into two groups for each class.

True positives are features for which the aggregated classification result and the reference dataset coincide. For each class, features indicated in the reference dataset but not in the aggregated classification result are regarded as false negatives. For building classifications these false negatives split up into $FN_{\text{no building}}$ and $FN_{\text{bad image}}$. Consequently, false positives are features that are indicated as such in the aggregated classification result, but not in the reference dataset. An incorrect building classification can therefore either be $FP_{\text{no building}}$ or $FP_{\text{bad image}}$. For true negatives both datasets agree that the specific class is not present.

Each field in the confusion matrix is further processed to account for the bias produced by the stratified sampling approach (Card, 1982). This approach is chosen due to the heterogeneous proportion of "No Building", "Building" and "Bad Image" classifications and between "Consensus" and "Disagreement" tasks in the MapSwipe dataset. Figure 6 provides an overview on these proportions for tasks with six classifications and considering the majority answer. A simple random sample would oversample classes with high frequency (e.g. "No Building Consensus"), but undersample categories of low frequency (e.g. "Building Consensus"). The proportions are also listed in Table 8. The table further depicts the number of samples selected for each stratum.

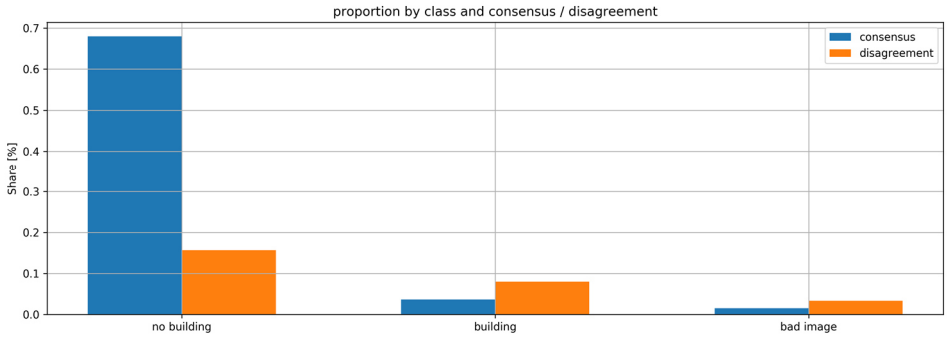

Figure 6 Proportion of Disagreement and Consensus Tasks per Class

Table 8 Distribution of Samples

	Consensus			**Disagreement**		
Class	No Building	Building	Bad image	No Building	Building	Bad image
Proportion	67.9 %	3.6 %	1.5 %	15.7 %	7.9 %	3.3 %
Samples	530	84	36	321	176	77

Table 9 Contingency Table Format (following (Card, 1982)*)*

		M				
T		*1*	*2*	*...*	*r*	
	1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
	2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
	...	...	...	...	...	...
	r	n_{r1}	n_{r2}	...	n_{rr}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	...	$n_{.r}$	n

To evaluate the quality of the aggregated contributions we utilize the following metrics, commonly applied to measure the performance of machine learning and information retrieval approaches: accuracy, sensitivity, precision, f1 score. For doing so, the following variables are defined. The total number of tasks is referred to as n. Furthermore, $n_{i.}$ is defined as the total number of tasks in the i-th row (e.g. total number of building classifications in the reference dataset) and $n_{.j}$ is defined as the total number of tasks in j-th column (e.g. total number of building classifications in the MapSwipe dataset, see Table 9). The marginal distributions of T (true, "reference") and M (map, "crowdsourced") are defined as p_i and π_i. The joint probability distribution of T and M is described by p_{ij}.

Accuracy is calculated for the whole dataset, whereas sensitivity, precision and f1 score are computed separately for each class ("no building", "building", "bad image"). The statistics are computed as depicted by Equation 4, Equation 6, Equation 8 and Equation 10 (Card, 1982; Sokolova and Lapalme, 2009). The uncertainty of each parameter will be derived following Card (1982) as described in Equation 5, Equation 7 and Equation 9. The functions are implemented in the python programming language.

The derived statistics will be interpreted and compared among the different aggregation methods. In total, 21 methods will be explored and visualized in the style of an annotated heatmap. Finally, rules for aggregation will be defined based on the results obtained. These rules consider different number of users per task and will be applied to all tasks of the dataset. In the following this aggregated answer will be referred to as "crowd answer".

Equation 4 Accuracy	$\hat{P}_c = \sum_{j=1}^{r} \pi_j \, n_{jj} \, / n_{.j}$
Equation 5 Variance of accuracy	$\mathrm{V}(\hat{P}_c) = \sum_{i=1}^{r} p_{ii} \, (\pi_i - \, p_{ii}) \, / \, n_{.i}$
Equation 6 Precsion	$\hat{\lambda}_{ii} = \, n_{jj} \, / \, n_{.j}$
Equation 7 Variance of Precision	$\mathrm{V}(\hat{\lambda}_{ii}) = p_{ii} \, (\pi_i - \, p_{ii}) / (\pi_i^2 \, n_{.i} \,)$
Equation 8 Sensitivity	$\hat{\theta}_{\mathrm{ii}} = \frac{\pi_i}{\hat{p_i}} \frac{n_{ii}}{n_{.1}}$
Equation 9 Variance of Sensitivity	$\mathrm{V}(\hat{\theta}_{\mathrm{ii}}) = \mathrm{p_{ii}} \, \mathrm{p_i^{-4}} \left[\mathrm{p_{ii}} \sum_{\mathrm{j \neq i}}^{\mathrm{r}} \frac{\mathrm{p_{ij}}(\pi_\mathrm{j} - \mathrm{p_{ij}})}{\mathrm{n_{.j}}} + (\pi_\mathrm{i} - \mathrm{p_{ii}})(\mathrm{p_i} - \mathrm{p_{ii}})^2 / \mathrm{n_{.ji}} \right]$
Equation 10 F1 Score	$F1_{ii} = \frac{2tp_i}{2tp_i + \, fp_i + \, fn_i}$
Equation 11 Standard deviation	$\mathrm{S(x)} = \sqrt{\mathrm{V(x)}}$

4.2 User Characteristics Analysis

Previous research has shown, that crowdsourcing results are influenced by the characteristics of individual users. The analysis of intrinsic user characteristics may therefore open new opportunities to understand and improve the quality of crowdsourced geographic information. This part investigates to what degree the probability of correct and incorrect classifications can be explained by characteristics of the users at the point in time they contributed the data. First, this section provides several definitions and variables of user activity and user performance. In the second part, the contribution quality of MapSwipe users is analysed towards their performance. Finally, it is tested whether users improve their performance with ongoing activity.

4.2.1 Definition of User Activity and User Performance

A characterization of users according to their contribution history is essential to estimate the quality for each user. This understanding enables to generalize the results so that they can be used to improve mapping instructions or the aggregation procedure. Therefore, variables of user characteristics are defined towards intrinsically quantifying user activity and user performance. User activity describes the amount of data or time each individual volunteer contributed to MapSwipe (Table 10). A mapping session is defined as a continuous usage of the MapSwipe app without breaks longer than 12 hours. The concept of an edit session has been introduced by Geiger and Halfaker (2013) to measure participation in Wikipedia. Dittus et al. (2017) applied it to evaluate crowdsourcing in humanitarian mapping. User performance variables are defined by constructing a confusion matrix for each user's classifications using the crowd answers as reference. They are computed cumulatively for each user and each completed group. Changes in the user performance over time can be captured following this approach. For example, for a user who contributed 15 groups, user performance variables are computed 15 times. First, a descriptive analysis of contribution inequality is performed. Furthermore, the distribution of user activity and user performance is examined using violin plots. Based on this analysis a group of variables will be selected for further use in the upcoming sections of this study.

Table 10 User Activity and Performance Variables

User Activity	**User Performance**
number of contributions	accuracy
number of completed groups	(no building, building, bad image) sensitivity
number of different projects	(no building, building, bad image) precision
number of mapping sessions	(no building, building, bad image) f1 score.

4.2.2 Analysis of User Performance and Data Quality

The impact of user performance on data quality is analysed using the expert reference dataset. All user classifications are grouped into either “correct” or “incorrect” as indicated by the reference. For no building classifications the impact of overall accuracy, no building precision and no building sensitivity are investigated. Likewise, accuracy, building precision and building sensitivity are explored for building classifications. Correct and incorrect bad image classifications are analysed towards potential influence of accuracy, bad image precision and bad image sensitivity.

For each group, violin plots of the distribution of user performance variables are generated to investigate differences between correct and incorrect user classifications. In the next step mean values are compared by conducting a Mann-Whitney-U test. This nonparametric test is used to determine whether correct and incorrect classification are significantly different in respect to the central tendencies of user performance. Thirdly, conditional density plots are derived to visualize how the conditional distribution of correct and incorrect classifications changes with increased user performance.

The same methods are applied to the Laos dataset. However, due to the limitations of the OSM reference dataset only building classifications are considered. The results for both reference datasets are compared and interpreted to identify user characteristics that are associated with high or low data quality.

4.2.3 Analysis of Performance Improvement with Higher Activity

It is assumed that by contributing more data users gain experience in solving mapping tasks. This section investigates whether this hypothesis can be validated and how the learning curve can be quantified. This analysis is conducted using the cumulative user performance variables defined in section 4.2.1. In total, this dataset contains information on 7,518 users and up to several hundred timestamps for each contributor.

For every number of completed groups the mean performance off all users is computed. Additionally, the standard deviation is derived to estimate variability for each step. A Mann-Kendall test is applied to test whether the mean performance increases with higher activity. The Mann-Kendall test investigates whether there is a statistically significant upward or downward trend of a variable over time.

4.3 Spatial Characteristics Analysis

The geographic approach described by Goodchild and Li (2012) emphasizes that spatial relations can be used to estimate the quality of crowdsourced datasets. It is the goal of this part to quantify to what degree the probability of correct and incorrect classifications can be explained by the spatial distribution of all crowdsourced classifications. First, an analysis of the spatial distribution of the MapSwipe dataset is conducted. Spatial characteristics for each task are defined in the second part. Then, an analysis is performed to test for statistically significant differences between correct and incorrect classifications.

4.3.1 Analysis of Spatial Distribution

The spatial distribution of agreement and building index is analysed by performing Moran's I statistics for spatial autocorrelation (Moran, 1950). The metrics are applied for all 55 projects considered in this study and provide a measure for the degree of clustering of tasks with high agreement. Differences among projects are investigated using boxplots. Additionally, an exemplary visual interpretation of the data is conducted for the project "Laos 6".

4.3.2 Definition of Spatial Characteristics per Task

For the definition of the spatial context a kernel density estimation is calculated. The kernel density is computed using the algorithm proposed by Silverman (1986) and the implementation in ArcMap 10.4.1. For each task polygon the kernel density of the centre point is considered. The search radius of the function is set to 440 metres, which corresponds to a neighbourhood of 24 features for each task. The input value for the density estimation is a binary layer. Thus, the kernel density of building classifications in computed using all tasks as input where at least one user assigned the "building" class (hatched tasks in Figure 7). Accordingly, the kernel densities for "no building" and "bad image" classifications are calculated. In the following the kernel density for each class will be referred to as "no building class density", "building class density" and "bad image class density". Figure 7 provides an example for the distribution of building class density. Blue colours highlight a high kernel density. For example, task [11] is surrounded by 23 other tasks which have been classified as building and is characterized by a high kernel density of around 43 per square kilometre. On the contrary, for task [1] no building classification can be observed in the spatial surrounding. The kernel density for this task is zero. Spatial outlier tasks, e.g. task [5], are described by a kernel density greater than 0 and smaller than 10. Tasks at the edge of larger clusters of building classifications or with some further building classifications in their neighbourhood are characterized by density values ranging from 10 to 25.

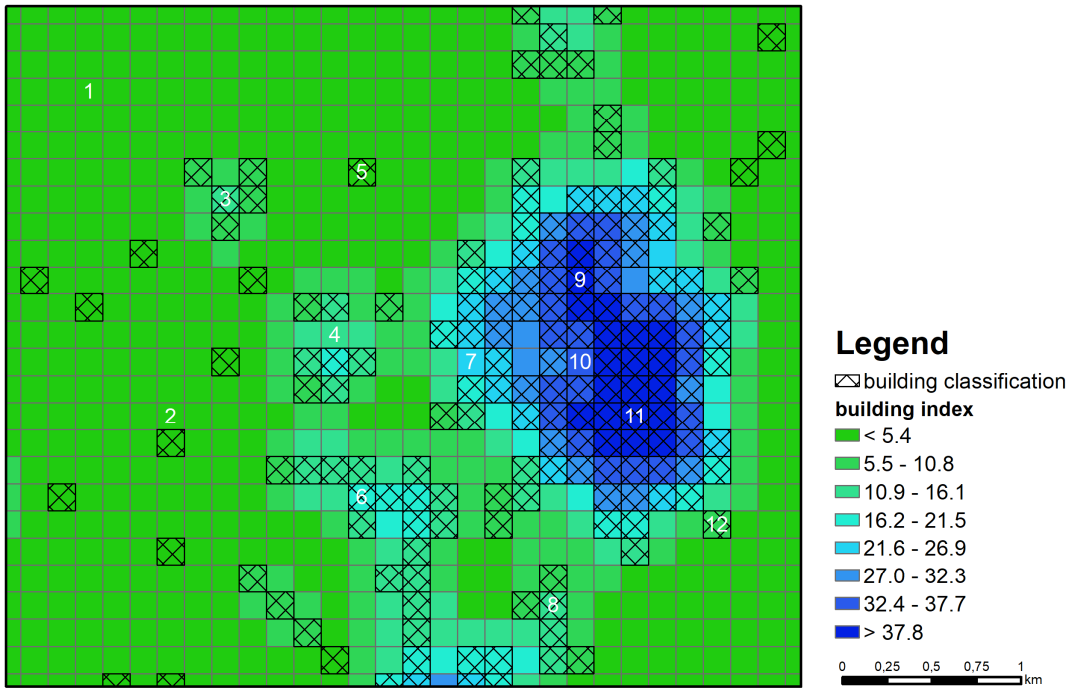

Figure 7 Kernel Density for Building Classifications

4.3.3 Analysis of Spatial Characteristics for Incorrect and Correct Classifications

The impact of spatial characteristics on data quality is analysed using a random sample of 1,000,000 classifications from the crowd reference dataset. All classifications are grouped into either "correct" or "incorrect" as indicated by the reference. For the analysis all contributions are further differentiated into "no building classifications", "building classifications" and "bad image classifications". For these groups the impact of no building class density, building class density and bad image class density on data quality is investigated.

Violin plots are computed to investigate differences in kernel density for correct and incorrect classifications. The distributions are tested for statistically significant differences applying a Mann-Whitney-U test. The impact of kernel density on data quality is further analysed by computing conditional density plots. These plots show how the conditional densities of correct and incorrect classifications change when increasing kernel density variables. The same methods are applied to the Laos dataset. However, due to the limitations of the OSM reference dataset only building classifications are considered. The results for both reference datasets are compared and interpreted to identify spatial characteristics that are associated with high or low data quality.

4.4 Machine Learning Models

The last part of the methodology section explores to what degree automated classifiers considering agreement, user characteristics and spatial characteristics can improve the performance of aggregating multiple classifications into a single result. The results are compared to the results of a simple agreement-based aggregation method. The goal of this section is to bring together “crowdsourcing”, “social” and “geographic” quality assurance approaches and characteristics of the dataset.

4.4.1 Logistic Regression Analysis

After the exploratory analysis of the factors that drive data quality a logistic regression model is utilized to assess how good task containing buildings can be predicted. The logistic regression model is performed on a task basis. The model incorporates agreement, building index and no building index. Furthermore, no building class density, building class density and bad image class density are considered. However, for each task several different users with varying user characteristics contributed data. Therefore, the individual user characteristics are aggregated into single variables per task. In the study, user characteristics of each task are 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 are computed for each task using all no building classifications. Likewise, user characteristics are generated from building and bad image classifications. Table 11 depicts nine user characteristics considered for each task. Missing values are imputed using the overall mean of each variable. In the pre-processing variables are tested for independence and multicollinearity using a correlation matrix and by inspecting variance inflation factors (VIFs). The logistic regression analysis is performed for the Laos dataset.

Table 11 User characteristics per task

No building	**Building**	**Bad Image**
Average accuracy	Average accuracy	Average accuracy
Average no building precision	Average building precision	Average bad image precision
Average no building sensitivity	Average building sensitivity	Average bad image sensitivity

4.4.2 Performance of Machine Learning based Aggregation

In the next step three different machine learning based aggregation methods are reviewed. In this study, random forest, logistic regression and neural network classifiers are applied to aggregate crowdsourced classifications from MapSwipe. The machine learning models classify each task into either “no building” or “building”. The analysis is performed for the

Laos dataset and implemented in the python programming language. Random forest and logistic regression functions are applied as provided by the open-source scikit-learn library. Neural network functionality is provided by the python deep learning library Keras. Keras is a high-level neural network API.

In the initial phase the tasks of the Laos dataset are split up into training and testing samples. The fraction of the training sample is set to 0.3 which corresponds to circa 280,000 training samples. The samples are chosen randomly. Accordingly, about 660,000 tasks (70 %) of the Laos dataset are used for testing. The derived models are validated using the testing dataset.

The performance of the machine learning based aggregation is investigated by constructing a confusion matrix. The data obtained from OSM functions as a reference. The derived information of true positives, true negatives, false positives and false negatives are the basis for the assessment of overall accuracy, building precision, building sensitivity and building f1 score. A confusion matrix is also constructed to assess the performance of the aggregation method proposed in section 4.1.3. This simple crowd answer aggregation is only based on agreement and provides a baseline to evaluate the performance of the machine learning based aggregation.

Finally, the random forest based aggregation is analysed towards its performance in respect to decreasing training sample size. For doing so, accuracy, building precision, building sensitivity and building f1 score are examined for training sample sizes from 280,000 samples (30 %) to about 500 samples (0.05 %).

5 Results and Discussion

This section presents the results of the analysis towards investigating the research questions identified in section 2.4. The first part wraps up the agreement analysis (RQ1). In the next two sections the findings for user characteristics (RQ2) and spatial characteristics (RQ3) are presented. Finally, the performance of the machine learning classifiers is shown (RQ4).

5.1 Agreement Analysis

In the first step the distribution of agreement, building index and no building index are analysed. The violin plots (Figure 8) illustrate the distribution for all tasks of the MapSwipe dataset. The majority of all tasks shows a high agreement. Most of these tasks even reach consensus agreement. However, this distribution is dominated the high number of tasks, for which no building classifications have been assigned.

Figure 8 Distribution of Agreement, No Building Index and Building Index

5.1.1 Analysis of Agreement for Correct and Incorrect Classifications

The impact on agreement, building index and no building index on classification accuracy is depicted by the conditional density plots (Figure 9). The figures confirm the hypothesis that tasks with a higher agreement contain fewer incorrect classifications. Vice versa, tasks with a considerable disagreement (e.g. building index of 0.5, or no building index of 0.5) are most likely to contain many incorrect classifications. Remarkably, for tasks with a building index greater than 0.8 almost perfect accuracy can be observed. A high agreement is strongly associated with good data quality.

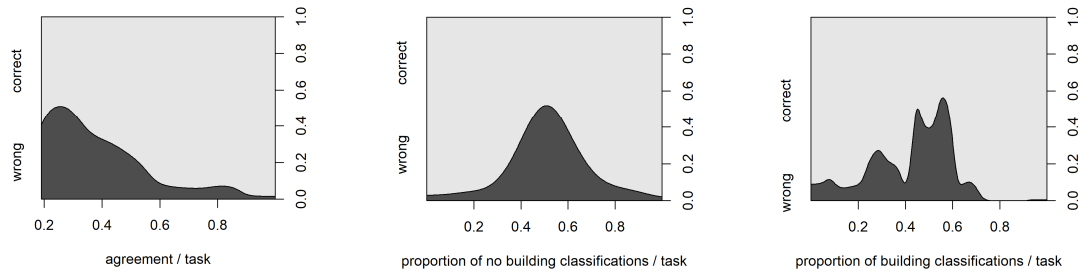

Figure 9 Conditional Density of Agreement, Building Index and No Building Index

5.1.2 Aggregation Threshold and Redundancy Analysis

In the next step, the performance of different aggregation methods regarding aggregation threshold and number of contributors per task is investigated. The overall accuracy varies between 82 % and 95 % and shows a slight increase with higher number of users (left to right in Figure 10). At the same time the accuracy varies for different aggregation thresholds (from bottom to top in Figure 10). Whereas the lowest accuracy is always achieved for a soft aggregation method (e.g. a building count threshold of one), the differences in overall accuracy for higher thresholds are only marginal. Statistically significant differences cannot be obtained for the higher thresholds as indicated by the 95 % confidence interval. Due to the biased distribution of tasks regarding map classes in the MapSwipe dataset, the analysis of accuracy alone is not sufficient to judge the fitness for purpose of each aggregation method.

Since the MapSwipe projects considered in this study are designed towards the identification of inhabited areas, the ability of each aggregation method to achieve a high quality for the building class is of prime importance. The analysis of the f1 scores for building classifications reveals greater distinctions among the different aggregation methods (Figure 13). However, likewise to the previous findings a low building count threshold always produces worst results. But, the results also indicate that with higher number of users data quality increases. This increase is not linear but rather stepped. There is considerable growth up to three users and when increasing the number of users from four to five. Hence, no differences in quality are obtained if three and four or five and six users classify each task.

The insights from the analysis of the building classification f1 score can be reinforced by looking at the statistics for building precision and building sensitivity. Choosing a low building count threshold will always produce a low building precision regardless the number of users per task. Contrasting building precision and building sensitivity puts emphasize on another critical issue when analysing data quality. The correct choice of an aggregation method depends on the purpose the data should be utilized for. Almost perfect building precision (with high confidence) is obtained choosing three or more users and a very high building count threshold (Figure 11). For example, when using building consensus aggregation for five users an aggregated building classification will be correct with a probability of 99 % ± 1.2 %. At the same time, building precision will always decrease when choosing a lower building count threshold (e.g. 35 % ± 4.2 % for five users and building count threshold of one).

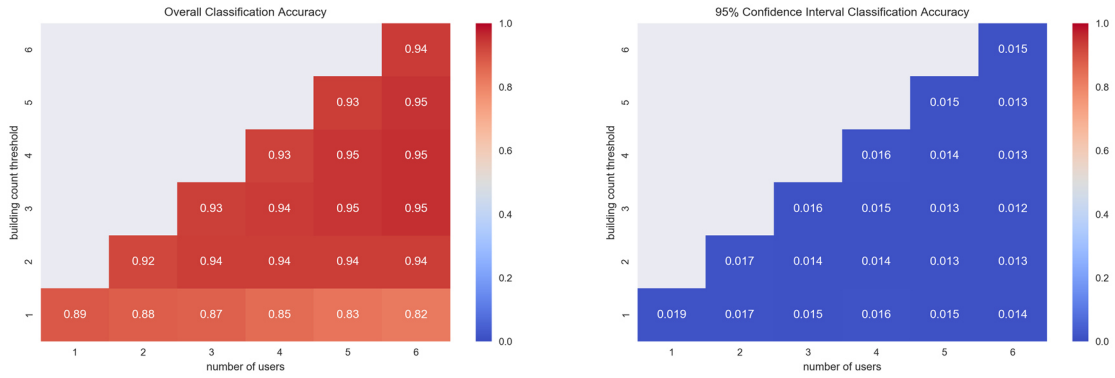

Figure 10 Overall Accuracy

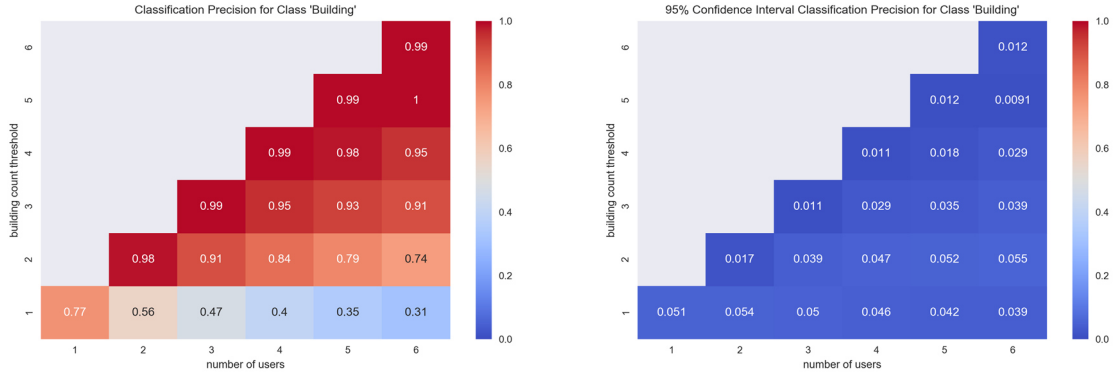

Figure 11 Building Classification Precision

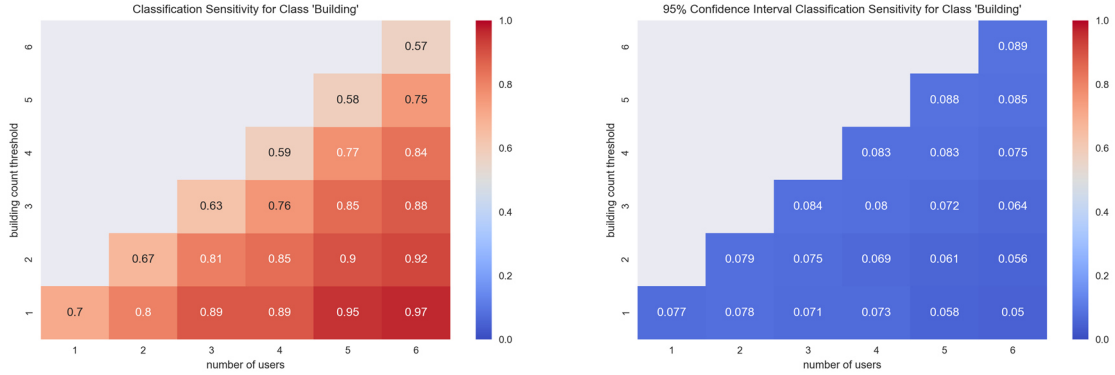

Figure 12 Building Classification Sensitivity

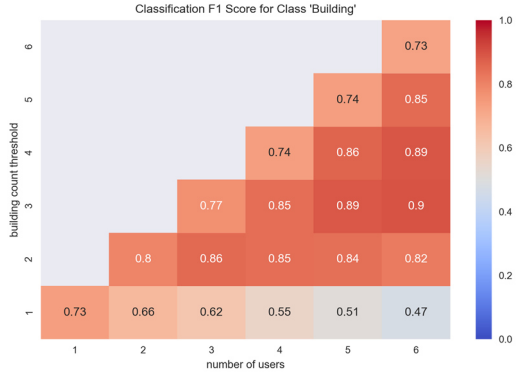

Figure 13 Building Classification F1 Score

When analysing building sensitivity, the opposite characteristics can be observed (Figure 12). Highest sensitivity is obtained for a low building count threshold and a high number of users. For instance, a sensitivity of 97 % ± 5 % is reached for six users and a building count threshold of one. Crowdsourcing projects which require a very high completeness of a specific class may therefore prefer a higher number of users, although this also induce a higher workload.

Nevertheless, the main challenge lies in finding a good compromise. The analysis reveals that for the MapSwipe dataset quality increases given a higher number of users. Soft majority agreement produces the best results regardless the number of users. The best results in respect to the trade-off between building precision and building sensitivity can be reached when choosing a minimum number of five contributors and a building count threshold of three. In the following of this study, the individual classifications per task will be aggregated using these findings. Thus, soft majority agreement will be applied. This aggregated result will be referred to as crowd answer from here on.

5.1.3 Discussion

The conditional densities of agreement and building index for correct and incorrect classifications show that majority agreement may lead to incorrect results for task with high disagreement. For these uncertain tasks using agreement as the only parameter to aggregate may not lead to the best solution.

These initial findings confirm what has been shown by Albuquerque et al. (2016). The analysis of agreement and redundancy shows that the MapSwipe dataset provides information on human settlements with an accuracy of more than 90 % for most of the tested aggregation methods. These results show higher performance in comparison to the one of automated approaches to detect buildings from very high resolution remote sensing data. For instance, Vakalopoulou et al. (2015) achieved an average sensitivity of 80% and an average precision of 90%. Klotz et al. (2016) provide information on the accuracy of the global urban footprint data and global human settlement layer for central Europe. Their results show, that accuracy ranges between 0.8 and 0.9, and built-up areas can be identified with a sensitivity of 70 – 85 %. Gueguen et al. (2017) present an approach to detect village boundaries, which has an average sensitivity of 84 % and an average precision of 70 %. In comparison to these products the MapSwipe approach can provide a valuable source of information on human settlements. However, a limitation of the approach presented in this study is determined by the small size and geographic scope of the expert reference dataset. This is expressed in the large confidence intervals for building precision and building sensitivity. Furthermore, it is not fully understood how the agreement varies for different

regions. Future research should therefore dig deeper into potential spatial variations incorporating a larger expert reference dataset.

Furthermore, it is important to notice that these results need to be set in context to the necessities of each individual project and managing organization. This includes the intended usage for the data and the related data quality requirements. Increasing the number of users per task manifolds the total number of classifications needed. This is a clear limitation of the crowdsourced approach in comparison to readily available datasets such as the global human settlement layer. In scenarios where timeliness is crucial, but it is not possible to motivate enough volunteers it may be wise to set the number of users required to three. Choosing this redundancy will still generate data with reasonable quality (e.g. building completeness of 80 %). On the other hand, in disaster response scenarios or other situations when many volunteers are keen to help increasing the number of users may be a good strategy. This is especially valid for cases where completeness (e.g. for mapping of damaged buildings) is important.

The results of the agreement based aggregation show that a potential building sensitivity of 97 % can be reached when asking six different users per task. Hence, most buildings are detected. Preserving this high sensitivity, but increasing the precision is therefore a main requirement for smarter aggregation methods. The results of the next two section will provide first insights whether intrinsic user characteristics and spatial characteristics can help to reach this milestone.

5.2 User Characteristics Analysis

The analysis of user characteristics is performed to understand how data quality is influenced by individual users or groups of users showing similar mapping behaviour. Users have different levels of expertise and domain knowledge which may influence the quality of their contributions. Nevertheless, this kind of user related characteristics are difficult to measure and collect. Therefore, in this study only user characteristics which can be derived intrinsically from the MapSwipe dataset itself are analysed.

5.2.1 Descriptive Statistics

As for many crowdsourced datasets the contributions per volunteer in the MapSwipe data are highly diverse. Figure 14 depicts user contribution inequality from two different perspectives. On the one hand, the major amount of data (around 80 %) is produced by only a small fraction of all users (around 20 %). Hence, few very active users have a high influence on the overall data quality. However, on the other hand about 50 % of the data is also contributed by users which have not completed more than 25 groups in total at the point in time when they submitted their results. This may point out that a great portion of the users are still less experienced with the app itself or the crowdsourcing task. This two-sided view on user contribution inequality already points out that both very active and novice users are central to understand data quality and asks for further measures of user characteristics.

In the first step, user activity is investigated. User activity is defined using four different parameters: number of contributions, number of completed groups, number of projects worked on and number of mapping sessions. All parameters show the same relationship. Most users contribute only very little. Vice versa there are only few users with a high activity. Figure 15 further portrays that user activity may be an inappropriate parameter to model differences among users due to its narrow distribution. In respect to activity most users have very similar characteristics.

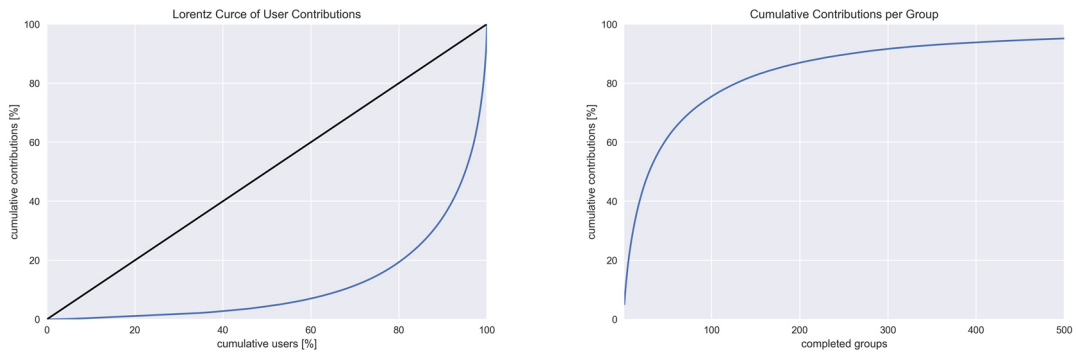

Figure 14 Contribution inequality by users and user activity

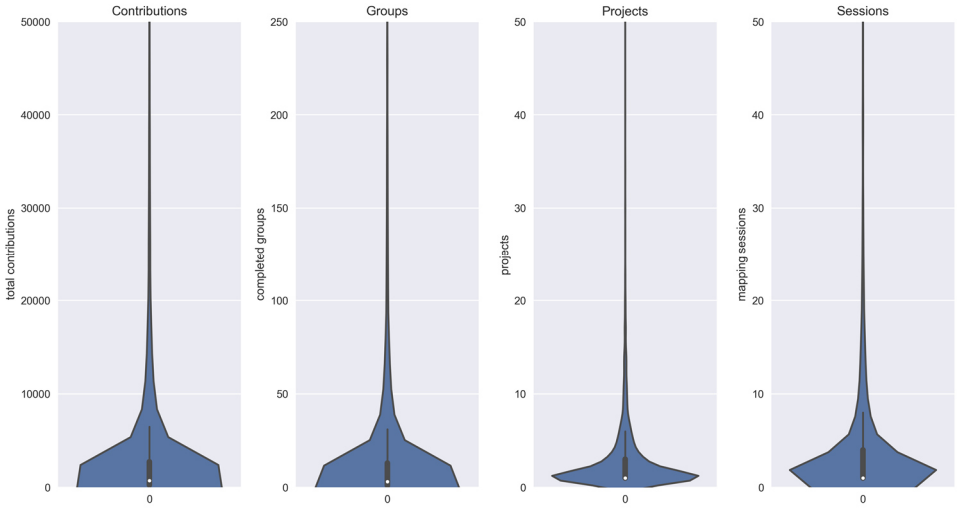

Figure 15 Distribution of User Activity

In the second part, user performance is assessed using the crowd answer as a reference. Figure 16 illustrates the distribution of overall accuracy and precision, sensitivity and f1 score for each map class. As for user activity, there is only small variety regarding overall accuracy among users. The average MapSwipe user obtains an accuracy of about 90 %. Likewise, the values of f1 score, precision and sensitivity for no building classifications are distributed. This approves the findings of the agreement based analysis: there is a high agreement among users on no building classifications. Since most classifications are no building contributions, this class dominates the overall accuracy statistics.

The distributions for building and bad image classes show considerable differences compared to the ones for the no building classifications. First, the distributions of f1 score, precision and sensitivity show a more dispersed shape. Thus, there is a greater diversity between users considering these parameters. Secondly, the averages are lower in comparison to the no building counterparts. These findings are backed by the agreement analysis conducted in section 5.1 as well. For tasks containing building or bad image classifications the agreement is considerably lower. Interestingly, for bad image classifications and to a certain degree for building classifications as well, there is a fraction of users with performance values close to zero. These outliers are an indicator for systematic misinterpretation of the crowdsourcing tasks. For example, users who mix up building and bad image classifications by tapping three times to (incorrectly) indicate that there is no building will obtain a very low bad image sensitivity score. Users who intendedly mark everything as building without checking the image are likely to get a building precision score close to zero.

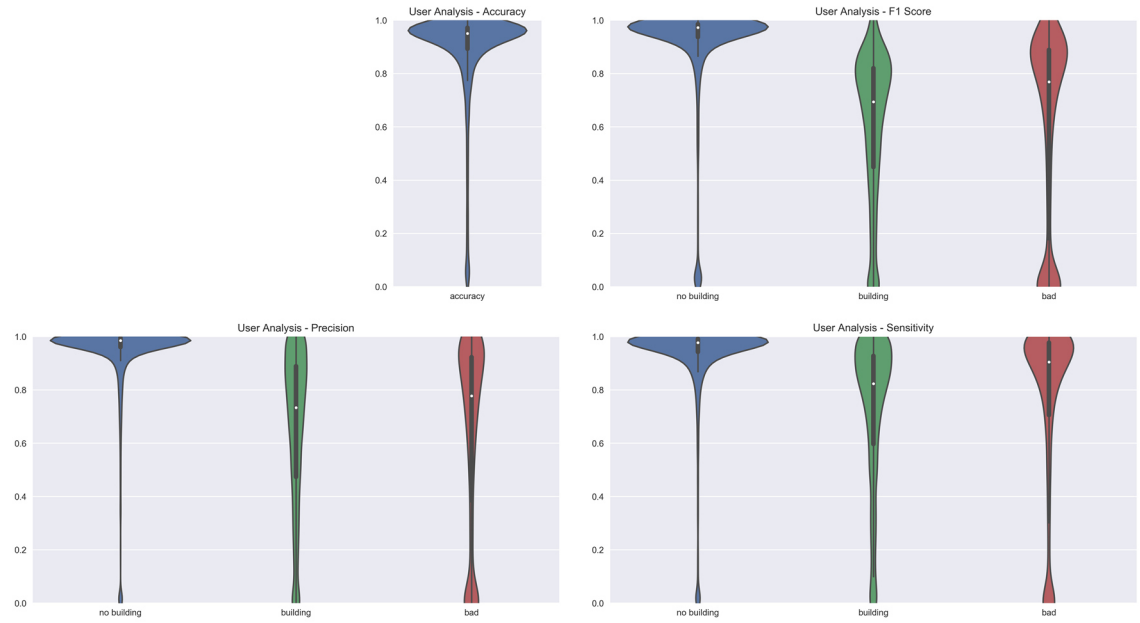

Figure 16 Distribution of User Characteristics

5.2.2 Global Analysis of User Quality using Expert Reference

This part analyses to what degree correct and incorrect user classifications can be separated by the individual performance of each user. For doing so, the dataset is split into three groups. The first group consists of all contributions that have been assigned “no building”, the second group where users chose “building” and the third group contains “bad image” classifications.

The violin plots for no building classifications show that correct and incorrect classifications have a very similar distribution of accuracy, no building precision and no building sensitivity (Figure 17). The visual impression suggests that mean accuracy for correct no building classifications is not higher compared to the mean accuracy for incorrect no building classifications. Furthermore, there are only very minor differences for the mean no building precision and no building sensitivity values as well. The violin plots indicate that incorrect classifications are hardly influenced by user characteristics. These initial findings are backed by the results of the Mann-Whitney-U test which prove statistical significance (Table 12). The test shows that there are no significant differences for accuracy and no building sensitivity. The difference for building precision is very small, albeit significant. The results are complemented by the conditional density plots. The interpretation of the figures reveals, that it is not possible to obtain any clear trend. No building classifications with lower user accuracy and lower building sensitivity seem to be more likely to be incorrect. However, regarding building precision the effect is very reduced or not observed.

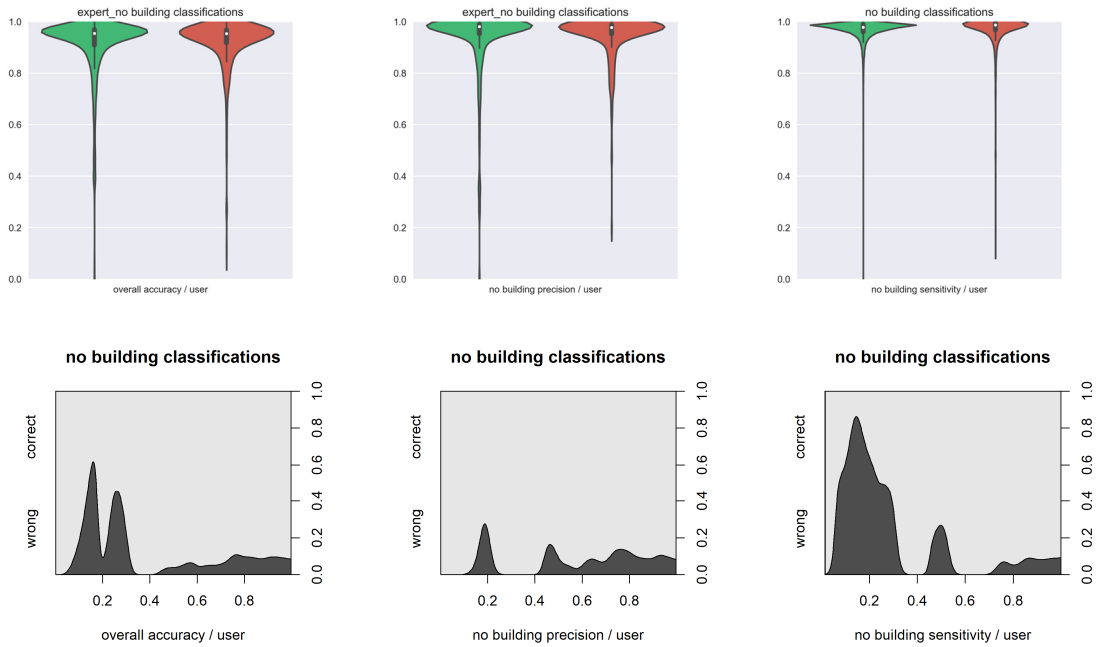

Figure 17 User Characteristics for Correct and Incorrect No Building Classifications

Table 12 Mann-Whitney Statistics of User Characteristics for Correct and Incorrect No Building Classifications

	Correct	Wrong	Mann-Whitney-U	p
count	5,524	537	-	-
mean accuracy	0.91	0.92	1479790.0	0.465
mean no building precision	0.93	0.94	1357873.5	0.001
mean no building sensitivity	0.97	0.96	1446765.0	0.173

The analysis of building classifications allows more definite conclusions. The violin plots demonstrate that building precision is a good parameter to describe differences between correct and incorrect classifications. Correct building classifications are associated with a high building precision score, whereas incorrect classifications show the opposite characteristics. Furthermore, incorrect building classifications show lower overall accuracy. These correlations are also statistically significant as indicated by the results of the Mann-Whitney-U test (Table 13). The distributions of building sensitivity reveal no great difference for correct and incorrect classifications (Figure 18). The conditional density plots for overall accuracy and building precision show a similar trend. The probability of incorrect classifications is continuously increasing with lower user accuracy and user building precision. Especially for very low values of building precision of less than 0.4 there is a high probability, that a building classification will be incorrect. The probability of incorrect classifications does not change much with increased building sensitivity.

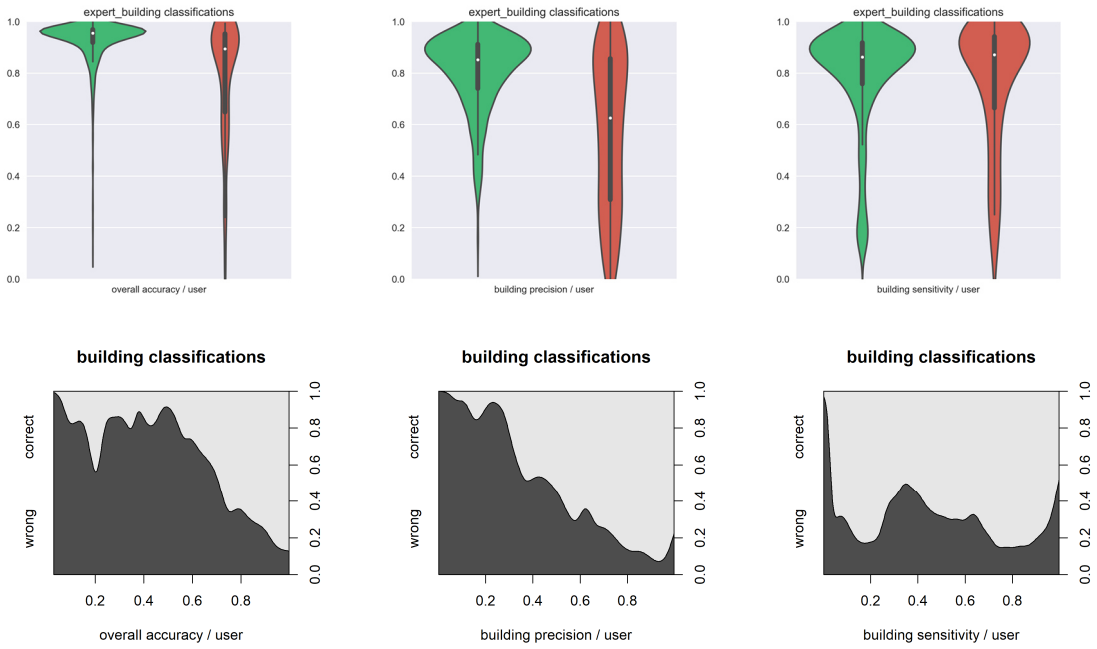

Figure 18 User Characteristics for Correct and Incorrect Building Classifications

Table 13 Mann-Whitney Statistics of User Characteristics for Correct and Incorrect Building Classifications

	correct	wrong	Mann-Whitney-U	p
count	1,502	227	-	-
mean accuracy	0.92	0.78	99864.0	0.0
mean building precision	0.80	0.58	94964.0	0.0
mean building sensitivity	0.79	0.76	164255.5	0.187

The visual interpretation of the violin plots for bad image classifications indicates that differences between correct and incorrect classifications can be observed regarding the characteristics of overall accuracy and bad image precision (Figure 19). The results of the Mann-Whitney-U test reveal, that statistically significance at 95 % confidence interval level is observed for overall user accuracy and bad image precision (Table 14). Correct classifications are characterized by higher mean overall accuracy. The association for bad image precision shows into the same direction. For bad image sensitivity a clear interpretation cannot be drawn. The results depicted by the violin plots are complemented by the conditional density plots. In contrast to the impression of the violin plots, the figures of conditional density show consistent characteristics for all three variables. The proportion of incorrect tasks gradually decreases with higher scores. This relationship is best pronounced for bad image precision.

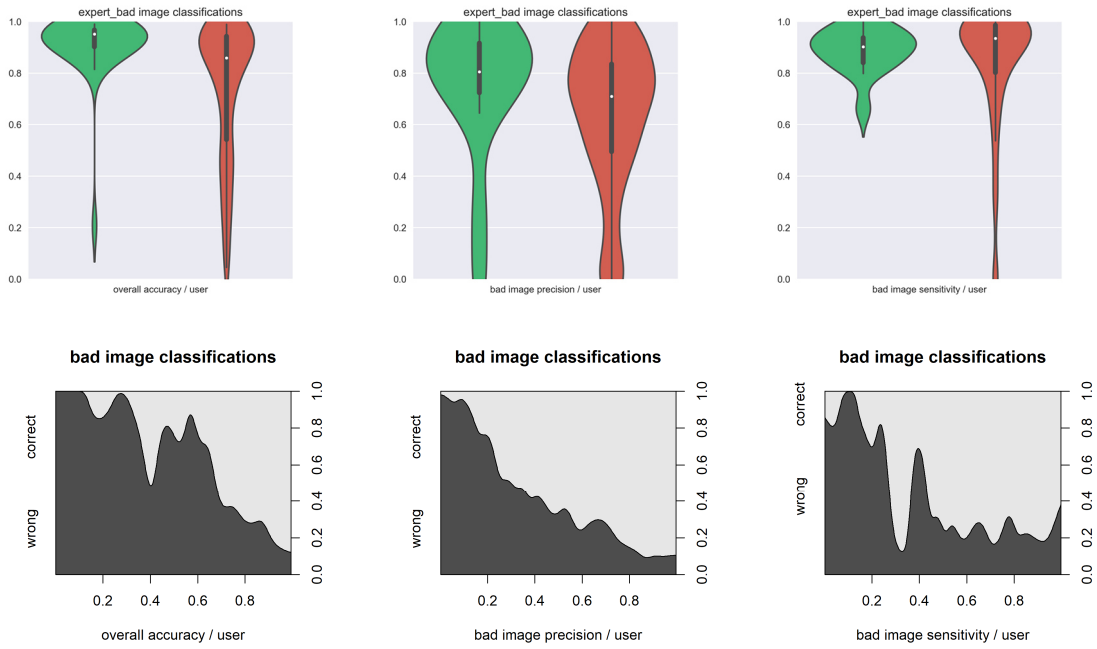

Figure 19 User Characteristics for Correct and Incorrect Bad Image Classifications

Table 14 Mann-Whitney Statistics of User Characteristics for Correct and Incorrect Bad Image Classifications

	correct	**wrong**	**Mann-Whitney-U**	**p**
count	30	275	-	-
mean accuracy	0.91	0.74	2221.5	0.0
mean bad image precision	0.74	0.63	3233.0	0.026
mean bad image sensitivity	0.89	0.83	3723.5	0.191

The analysis of user characteristics for correct and incorrect classification using the expert reference dataset implies that unambiguous trends or correlations can be identified for building and bad image classifications. Therefore, the study results suggest that user characteristics can be a complementing factor to describe and understand the data quality of the MapSwipe dataset. However, the study of user characteristics for no building classifications reveals that they are less suited to differentiate correct and incorrect contributions of this class. Especially the high number of no building classifications with very high agreement make it difficult to evaluate how user characteristics influence data quality.

5.2.3 Local Analysis of User Quality using OSM reference

In this section the results of the previous part are reviewed for the selected projects in Laos. Given the limitations of the OSM reference dataset only building classifications are

considered in this part. Albeit, this offers the possibility to use a large sample of more than 200,000 building classifications.

The violin plots show only small differences between the distributions of overall accuracy and building sensitivity (Figure 20). For building precision, the visual interpretation indicates that incorrect classifications are associated with lower scores. These impressions are consistent with the findings of the global analysis using the expert reference dataset. Especially the stretched distribution of the study results suggest that user characteristics can be a complementing factor to describe and understand the quality of the MapSwipe dataset. The results of Mann-Whitney-U tests show statistical significance for all three variables (Table 15). Nevertheless, the difference between the mean values is highest for building precision. Also, the correlations depicted in the conditional density plots reinforce the findings of the global analysis. The probability of incorrect classifications decreases almost linear with higher building precision scores. To a certain degree this can also be observed for user accuracy and even for building sensitivity.

The local analysis of the influence of user characteristics in Laos underlines the general findings of the previous sections. User characteristics are suited to model incorrect building classifications. Due to the limitations of the reference dataset the findings for no building classifications and bad image classifications could not be verified.

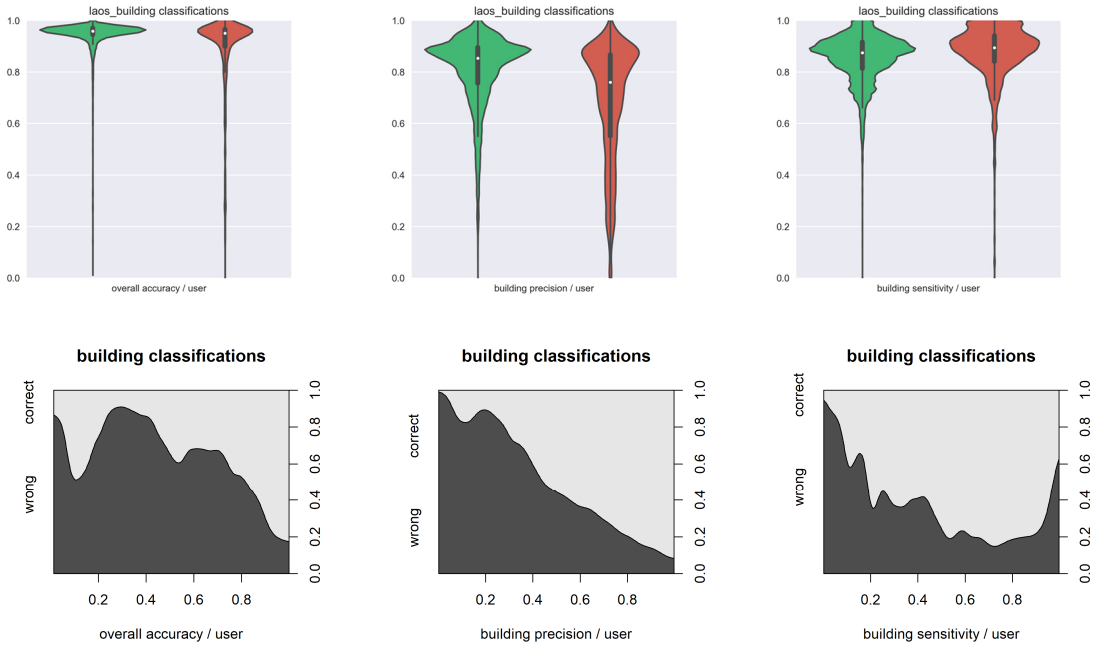

Figure 20 User Characteristics for Correct and Incorrect Building Classifications in Laos

Table 15 Mann-Whitney Statistics of User Characteristics for Correct and Incorrect Building Classifications in Laos

	correct	wrong	Mann-Whitney-U	p
count	193,104	34,270	-	-
mean accuracy	0.94	0.89	2564827741.0	0.0
mean building precision	0.81	0.68	2204083488.0	0.0
mean building sensitivity	0.85	0.87	2697620231.0	0.0

5.2.4 Analysis of Performance Improvement with Higher Activity

Finally, in the last part of this section on user characteristics the relationship between user activity and user performance is investigated. The results of this analysis are important to understand whether users can learn directly in the process of using the MapSwipe app. For doing so, learning is understood as an increase in user performance. For example, if a user increases his building precision score with more and more completed groups, this is interpreted as positive learning process. Furthermore, the analysis provides information on differences between beginners and more active volunteers.

Figure 21 visualizes the evolution of the average of the user performance variables per completed groups. The number of completed groups functions as an indicator of user activity. Users will increase their number of completed groups over time, however most users will drop out quickly. Therefore, the number of user contributing to each average value decreases with higher completed groups count. To limit the influence of individual users on the average this analysis is only performed up to 250 completed groups. Fewer than 1 % of all users contributed more.

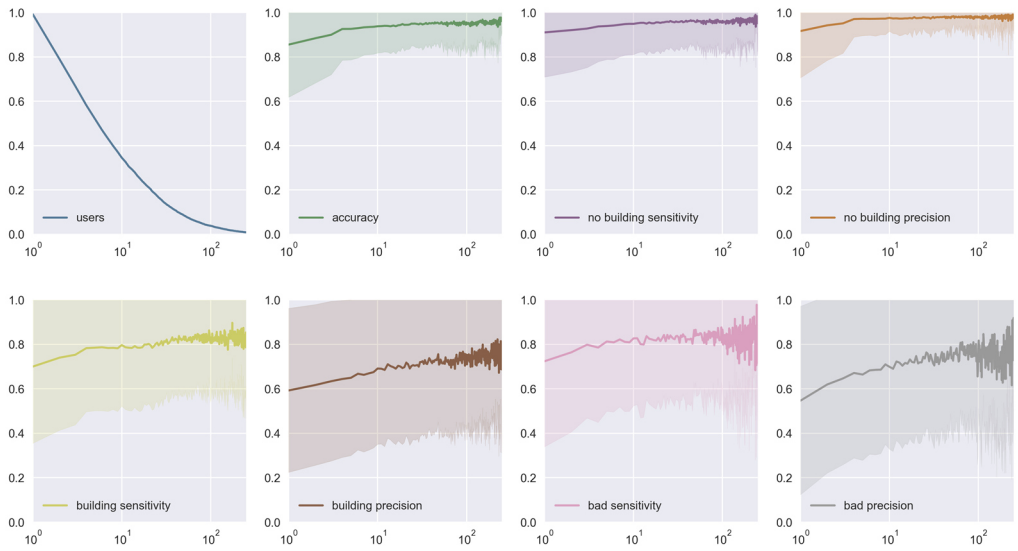

Figure 21 User Performance Evolution with higher Activity

The analysis demonstrates that user performance increases with higher activity. In general users with higher activity (in terms of completed groups) have a higher agreement with the crowd answer. Especially regarding building sensitivity, building precision, bad image sensitivity and bad imagery precision the users increase their performance within the first ten completed groups. This might be an indicator that users quickly learn to classify the images accordingly. On the other hand, this also shows, that beginners tend to do more incorrect classifications compared to user who have already contributed more. The overall trend, that user performance increases with activity, is confirmed by the Mann-Kendall test. All user performance variables show a statistically significant increase over the first 250 completed groups (Table 16).

Table 16 Results of Mann-Kendall-Test for Monotonic Trend

User Dimension	Trend	h	p	z
Accuracy	increasing	True	< 0.005	10.078
No Building Sensitivity	increasing	True	< 0.005	9.006
No Building Precision	increasing	True	< 0.005	5.436
Building Sensitivity	increasing	True	< 0.005	4.71
Building Precision	increasing	True	< 0.005	10.199
Bad Image Sensitivity	increasing	True	< 0.005	3.295
Bad Image Precision	increasing	True	< 0.005	3.405

5.2.5 Discussion

The results of this section have shown that there is an association of probability of incorrect classifications and user performance variables. However, the effect of user characteristics is more pronounced for building and bad image classifications than for no building classifications. Consequently, for quality estimation of no building classifications the impact of user characteristics needs to be further evaluated.

The MapSwipe projects examined in this study are designed towards the identification of buildings. A perfect crowdsourcing approach should be able generate high quality data in respect to sensitivity and precision. Sensitivity is affected by the ability of users to identify no building tasks. The uncertain impact of user characteristics on no building classifications constitutes a serious limitation to the analysis of building completeness. But, precision is shaped by the accuracy of building classifications. For this group, the analysis proved that quality can be estimated utilizing intrinsic user characteristics. This is backed by the findings of Salk et al. (2016), who assess user performance for the cropland capture game. The authors conclude that overall user performance can be used to model the quality of

their individual ratings. User performance characteristics are also incorporated by Gueguen et al. (2017) for village boundary detection. The authors model the reliability of each user in terms of a user score. In their study, users who consistently agree on a number of features form a consensus group and are considered reliable.

Another limitation arises when looking at absolute numbers of incorrect classifications rather than considering probabilities. Although there is a strong association of incorrect classifications and low user performance characteristics (e.g. low building precision for incorrect building classifications), looking at the absolute numbers of incorrect classifications may lead to divergent results. The overall distribution of user performance scores shows that only very few users are characterized by low performance values (Figure 16). At the same time, the large number of users with high performance scores will submit many incorrect classifications as well, although they have a lower probability. Consequently, the intrinsic analysis of user characteristics only allows to detect users with very low data quality. The results of Arcanjo et al. (2016) confirm these findings for the ForestWatcher citizen science project. In their study, the authors use the accuracy of the volunteer's contribution history to evaluate their future contributions. They show that this can be effective approach to filter out low-performance or malicious volunteers. Designing approaches that help to identify expert users remains a major challenge for future research.

Taking this into account the results still help to understand the MapSwipe data quality and might be relevant for the design of future mapping projects. The analysis allows to differentiate types of classification errors. On the one hand, incorrect classifications by users with low user performance scores are very likely to have systematic causes. One reason for this might be, that users didn’t understand mapping instructions correctly. Therefore, user characteristics could be utilized to directly address users who do it consistently wrong. Assisting unexperienced users is an essential way to limit incorrect contributions and ensure high data quality right from the beginning. The importance of instant feedback is also highlighted by Salk et al. (2016) for crowdsourced cropland mapping. On the other hand, the analysis also reveals that some tasks are classified incorrectly although users show good performance in general. These cases, which are likely to be missed even by experienced users, could improve the guidance material. By highlighting difficult cases users could be trained specifically. This seems to offer a great potential for further data quality improvement.

5.3 Spatial Characteristics Analysis

Agreement and user characteristics can explain the quality of the crowdsourced results to a certain degree. However, these characteristics are non-spatial and they can be generated for any kind of crowdsourced classification results. MapSwipe is designed towards geospatial applications on purpose. Projects, groups and tasks in MapSwipe are not only a set of features, they also refer to specific places on the earth and are interconnected by their spatial arrangement. The spatial characteristics are therefore in the focus of this section.

5.3.1 Descriptive Statistics

Spatial auto correlation was investigated for all 55 MapSwipe projects considered in this study. In Figure 23 boxplots show the distribution of Moran's I index of spatial auto correlation of agreement and building index. For both variables an average Moran's I of ~ 0.35 can be found in the data. For all projects the p-value was smaller than 0.005, thus there is statistically significant positive spatial auto correlation of agreement and building index. This implies, that both variables are not distributed randomly in space, moreover they tend to cluster.

Figure 22 visualizes the spatial distribution for agreement for the MapSwipe project "Laos 6". This example shows how diverse agreement is distributed for a single project. There are large areas with relatively high agreement values, sometimes interspersed by single tasks with low agreement. In the central western part of the project a cluster of tasks with low agreement can be estimated. Furthermore, there are stripes of tasks from west to east, overlapping with the geometry of individual groups, with consistently lower agreement values. At the south-western border of the project a stripe ranging from north to south with lower agreement values is indicated as well. This visual interpretation confirms that MapSwipe data is spatially autocorrelated. Due to the fact, that building index is not distributed randomly in space, there are regions where no building classifications tend to cluster and other regions where building and bad image classifications are concentrated. But, there are also spatial outliers, e.g. tasks with a high building index surrounded by many tasks with very low building index and vice versa. These different spatial patterns are described by the kernel density of no building, building and bad image classifications and further investigated in the next section.

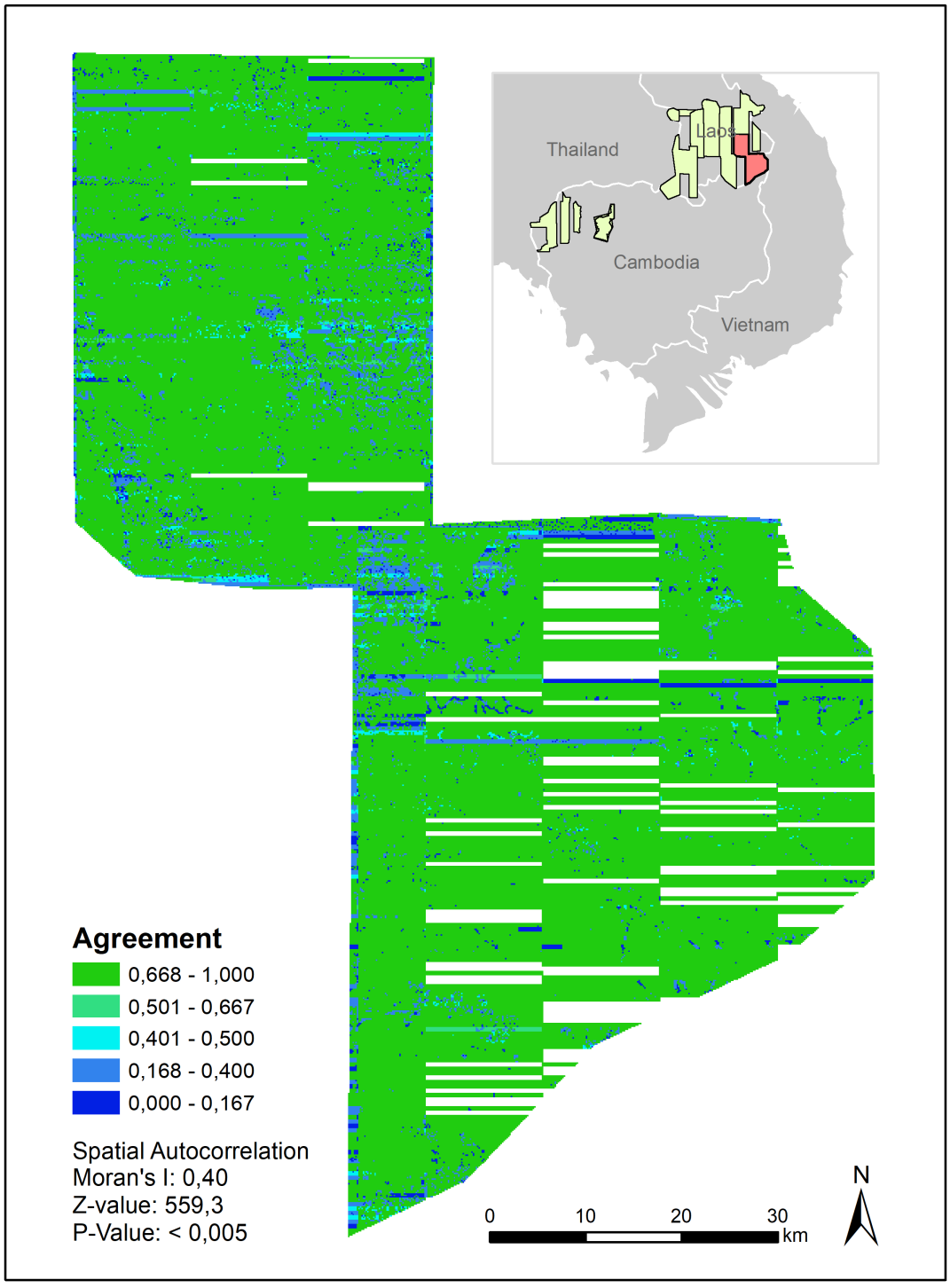

Figure 22 Spatial Autocorrelation of Agreement for Project Laos 6

Figure 23 Moran's I Index of spatial autocorrelation for agreement and building index

5.3.2 Global Analysis of Spatial Characteristics using Crowd Reference

The influence of kernel density on data quality is analysed in the next step. The violin plots show the different characteristics of kernel density for no building classifications (Figure 24). Whereas, the mean values of no building density and bad image density are similar for correct and incorrect no building classifications, it is obvious that the mean values of building density aren't. Correct no building classifications are characterised by low building density values. Incorrect no building classifications show significantly higher values. This indicates, that no building classifications which have many building classifications in their neighbourhood may have a higher probability of being incorrect. The Mann-Whitney-U test confirms these findings and shows that differences are statistically significant (Table 17).

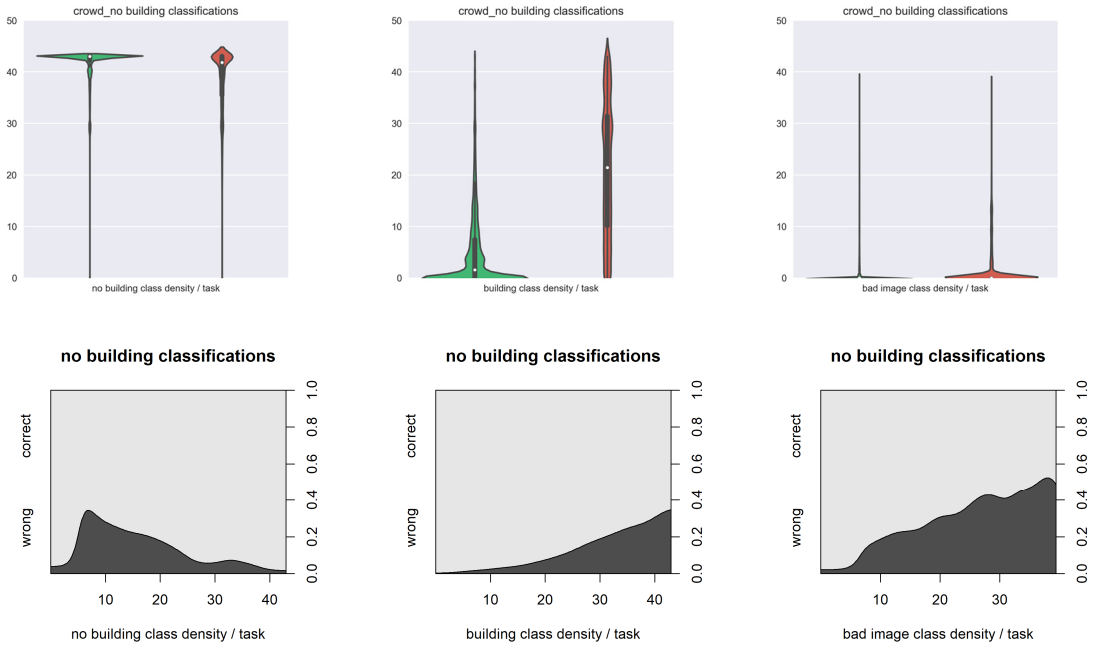

Figure 24 Spatial Characteristics for Correct and Incorrect No Building Classifications

Table 17 Mann-Whitney Statistics of Spatial Characteristics for Correct and Incorrect No Building Classifications

	correct	wrong	Mann-Whitney-U	p
count	845,576	21,188	-	-
mean no building class density	41.45	38.39	5961594562.0	0.0
mean building class density	5.21	21.22	2508651511.0	0.0
mean bad image class density	0.15	1.34	7636270285.5	0.0

The conditional density plots extend these findings. For all three kernel density variables a trend can be estimated. Low no building class density is associated with higher probability of being incorrect. At the same time, higher building class density and higher bad image class density contribute to the probability of incorrect classifications.

For building classifications differences among correct and incorrect classifications are easier to observe. As depicted by the violin plots in Figure 25 the mean values for no building density and building density vary considerably for correct and incorrect classifications. Correct classifications show lower no building class density and higher building class density on average compared to incorrect classifications. The differences for bad image class density cannot be quantified by the violin plots. For both groups the mean bad image density has a value close to zero. The results of the Mann-Whitney-U test proof the statistical significance of the results (Table 18).

Especially very high no building class density is associated with a higher probability of incorrect classifications as depicted by the conditional density plots. Spatial outliers, building classifications that are surrounded by no building classifications and characterised by very high no building density values, are very probable of being incorrect. Likewise, the higher the building density, the higher is the probability of a classification to be correct. Finally, the conditional density of bad image class density reveals, that for density values greater than five there is a sharp increase in the probability of being incorrect. This indicates that building classification which are surrounded by bad image classifications are likely to be wrong. This may be caused by the spatial characteristics of the bad image class. Bad image classifications are either assigned to areas covered by clouds or areas where satellite imagery is missing completely. Both tend to be solid blocks and circular structures are hardly probable. Therefore, a building classification in the centre of a region covered totally by bad image classifications has a very little chance of being correct.

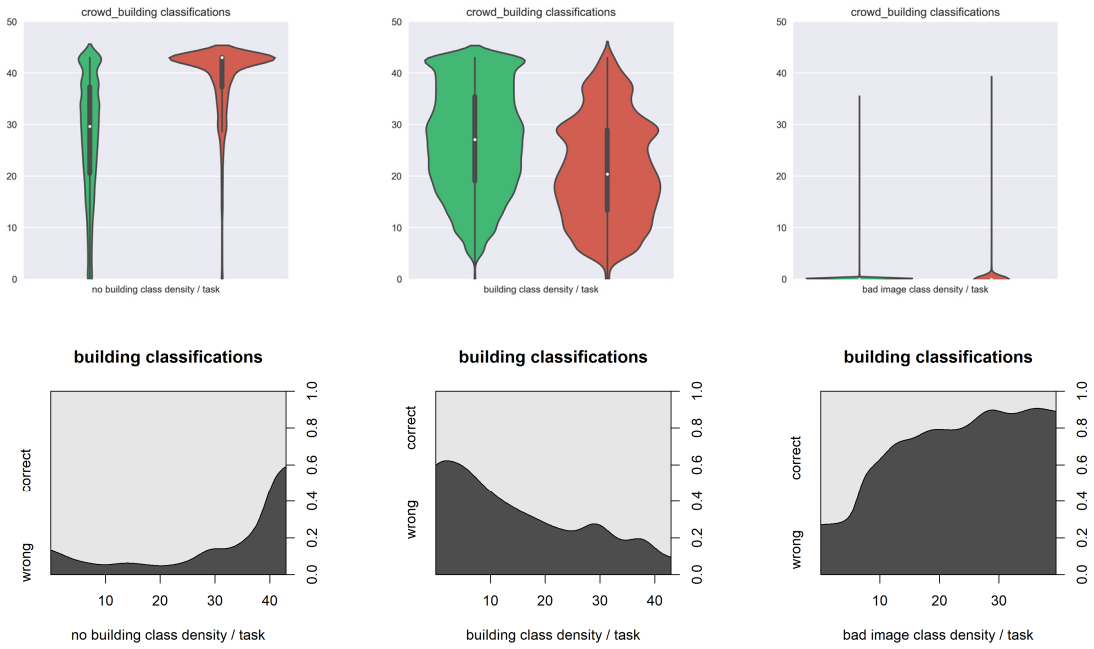

Figure 25 Spatial Characteristics for Correct and Incorrect Building Classifications

Table 18 Mann-Whitney Statistics of Spatial Characteristics for Correct and Incorrect Building Classifications

	correct	wrong	Mann-Whitney-U	p
count	52,412	10,252	-	-
mean no building class density	27.9	38.75	100964296.5	0.0
mean building class density	26.92	21.2	185591991.0	0.0
mean bad image class density	0.1	0.8	255625999.0	0.0

The interpretation of the violins plots for bad image classifications discloses that differences for correct and incorrect classifications can be explained by no building density, whereas the results for building density and bad image density are not as conclusive (Figure 26). Correct bad image classifications have a lower mean no building density than incorrect ones. The means for building density and bad image density are very similar. The Mann-Whitney-U test confirms that differences are statistically significant (Table 19). The conditional density plots underline these conclusions. There is a distinct increase in the probability of being incorrect for classifications with a no building density greater than 20. Since clouds and areas with missing imagery tend to cover multiple tasks in MapSwipe, task which are surrounded by no building classifications entirely are most likely not to show a cloud or missing satellite imagery. Furthermore, the plots show that building class density has negative influence on accuracy of bad image classifications. For higher bad image densities, the probability of being correct increases.

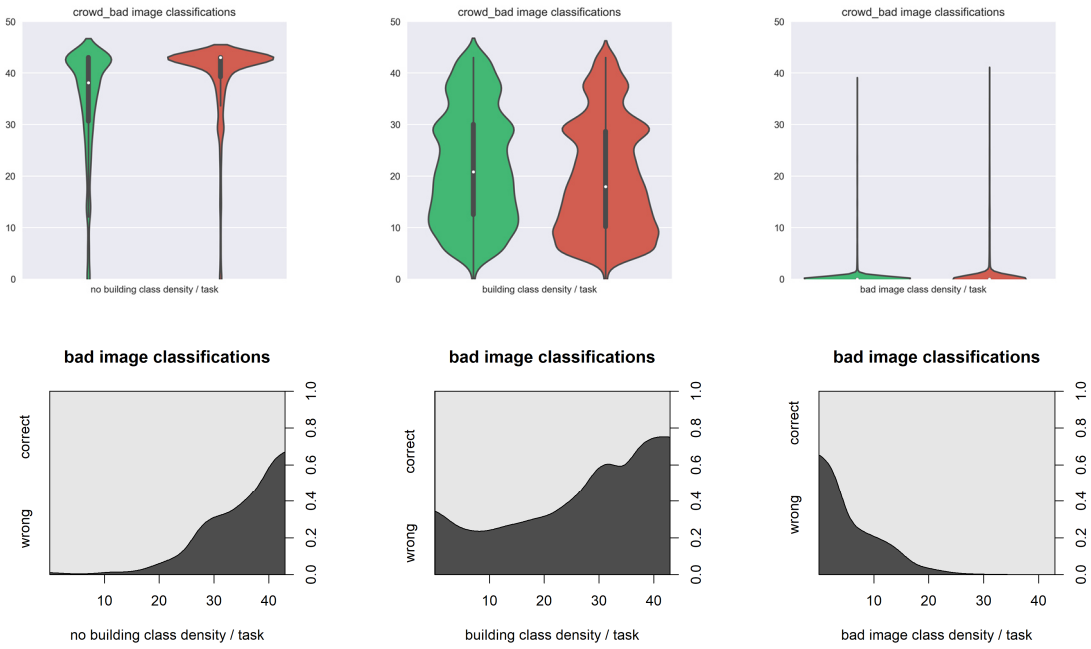

Figure 26 Spatial Characteristics for Correct and Incorrect Bad Image Classifications

Table 19 Mann-Whitney Statistics of Spatial Characteristics for Correct and Incorrect Bad Image Classifications

	correct	wrong	Mann-Whitney-U	p
count	6,706	13,252	-	-
mean no building class density	34.68	39.06	5961594562.0	0.0
mean building class density	21.86	19.69	2508651511.0	0.0
mean bad image class density	0.5	1.25	7636270285.5	0.0

5.3.3 Local Analysis of Spatial Characteristics using OSM reference

In this section the impact of spatial characteristics on data quality is analysed for the selected MapSwipe projects in Laos to review the findings of the global analysis. Due to the constraints of the OSM reference dataset only building classifications are investigated.

The violin plots (Figure 27) confirm, what has been shown for the spatial characteristics of building classifications in the previous section. Correct building classifications tend to have lower no building class density that incorrect classifications. Vice versa, a high building class density can be obtained for correct classifications and lower values for incorrect classifications. The mean values for the distributions of bad image class density are located close to zero. Visually no clear effect can be observed. Table 20 shows the results of the Whitney-Mann-U test. The conditional density plots verify, that the probability of incorrect classifications decreases with higher building class density. Also, the influence of bad image class density is described as for the global expert reference sample.

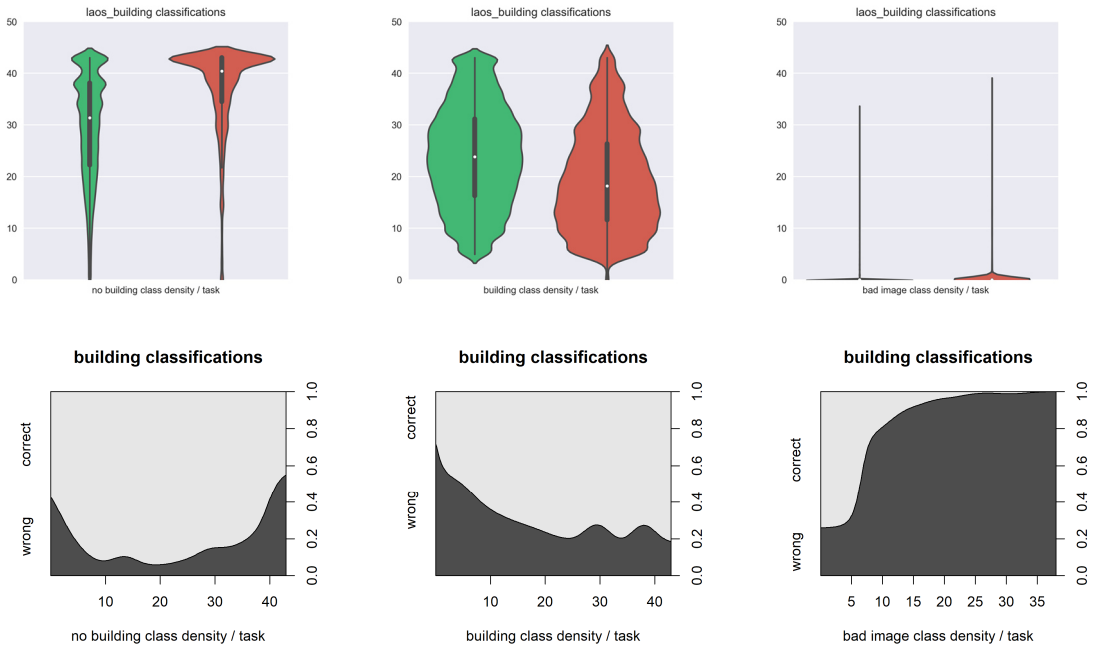

Figure 27 Spatial Characteristics for Correct and Incorrect Building Classifications in Laos

Table 20 Mann-Whitney Statistics of Spatial Characteristics for Correct and Incorrect Building Classifications in Laos

	correct	wrong	Mann-Whitney-U	p
count	193,104	34,270	-	-
mean no building class density	29.55	37.1	1707282071.5	0.0
mean building class density	23.9	19.4	2444575888.0	0.0
mean bad image class density	0.05	0.79	3142532917.5	0.0

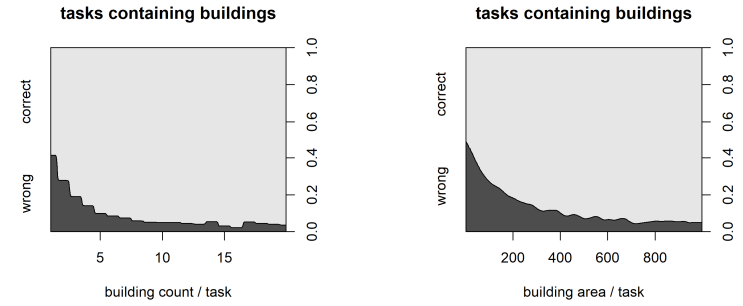

Figure 28 Impact of Building Count and Building Area per Task on Accuracy

In a second step it is analysed to what degree the number of buildings and the total building area per task relate to the probability of correct and incorrect classifications. Figure 28 indicates a clear trend for both interrelated variables. A low number of buildings per task and thus a small building area are described with a higher probability of incorrect classifications. Tasks containing only one building are most likely to be classified

incorrectly. Additionally, the graphs show that the probability of incorrect classifications decreases rapidly for tasks with up to five buildings and remains constant for tasks which contain more than five buildings. This demonstrates, as one would expect, that difficult tasks are those which contain few and small buildings.

5.3.4 Discussion

The analysis of the spatial characteristics of the MapSwipe dataset has pointed out that spatial relationships are important to explain variations in the quality of the data. Consequently, spatial characteristics such as kernel density of classifications should be further investigated to model the uncertainty of crowdsourced classifications. This is also highly recommended due to the small size of the expert reference dataset and the limited validity of the OSM reference dataset.

The results reveal that spatial outliers, e.g. building classifications surrounded by a bunch of no building classifications, are more likely to be incorrect. This underlines that users have problems to interpret the size, shape and appearance of features present in the satellite imagery if no other man-made objects are present in the surrounding. The effect is even stronger if buildings are small. By applying a logistic regression model Albuquerque et al. (2016) confirm that the size of geographical features is significantly associated with the likelihood of correct classifications. However, also the accuracy of global settlement products such as the global human settlement layer is affected by differences in building material, construction type, settlement structure and physical surrounding (Klotz et al., 2016).

Other authors have investigated the influence of spatial resolution of the satellite imagery on crowdsourced data quality (e.g. Battersby et al. (2012)). It can be assumed that differences in satellite imagery data quality have an impact on the spatial distribution of incorrect classifications. However, in this study this was not further elaborated. Future research should therefore investigate properties of the satellite imagery in more detail to account for potential bias in the classification results.

Furthermore, spatial characteristics might be influenced by the individual user characteristics as well. Comber et al. (2016b) show that the spatial distribution of land cover derived from crowdsourcing classifications may change significantly between different groups of users. For the MapSwipe dataset the high proportion of incorrect bad image classifications could be explained by the mapping behaviour of individual users. Some users tend to misunderstand the mapping instructions. These contributors tap three times to indicate no building classifications, although the tasks are annotated as bad image.

Therefore, incorrect bad image classifications are observed visualized as long stripes from east to west (Figure 22). For these tasks the calculated bad image density will be high, although the results are very likely to be incorrect. This limits the validity of using spatial characteristics for data quality assurance. Nevertheless, at the same time, this shows the potential of an integrated quality assessment considering both user characteristics and spatial characteristics.

Therefore, further mechanisms should be considered to analyse the impact of spatial characteristics on data quality. Kernel Density is a simple method to characterize the spatial context of individual classifications with limited explanatory power. For instance, Lesiv et al. (2016) show that geographically weighted logistic regression can be applied successfully to develop a hybrid forest cover map incorporating crowdsourced classifications. Therefore, the applicability of further geography aware statistical methods should be evaluated to understand the spatial patterns of quality variations in the MapSwipe dataset.

5.4 Machine Learning Models

The results of the first three parts of the analysis have shown how agreement, user characteristics and spatial characteristics affect data quality. However, most crowdsourcing workflows still rely on a simple majority aggregation to combine results from multiple users. Also, MapSwipe data is currently processed without considering user characteristics and spatial characteristics. To address this gap, machine learning models are applied in this section to enable the combinations of multiple dimensions of data characteristics which affect quality. The analysis is carried out for the selected MapSwipe projects in Laos and using the OSM reference dataset. In total this dataset contains 941,589 tasks. In the following the dataset is split up into a training dataset with 282,476 tasks and a validation dataset with 659,113 tasks (Table 21).

Table 21 Sample Sizes for Training and Validation Data

total	**training**	**validation**
941,589 (100%)	282,476 (30%)	659,113 (70%)

5.4.1 Logistic Regression Analysis of MapSwipe Data Quality

In the first step, a logistic regression model is 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 checking variables for multi collinearity and investigating variance inflation factors (VIFs) seven variables have been chosen as an input for the analysis. The correlation matrix plot (Figure 29) shows that no critical correlation between variables is observed. This is confirmed by the small VIFs close to 1.0.

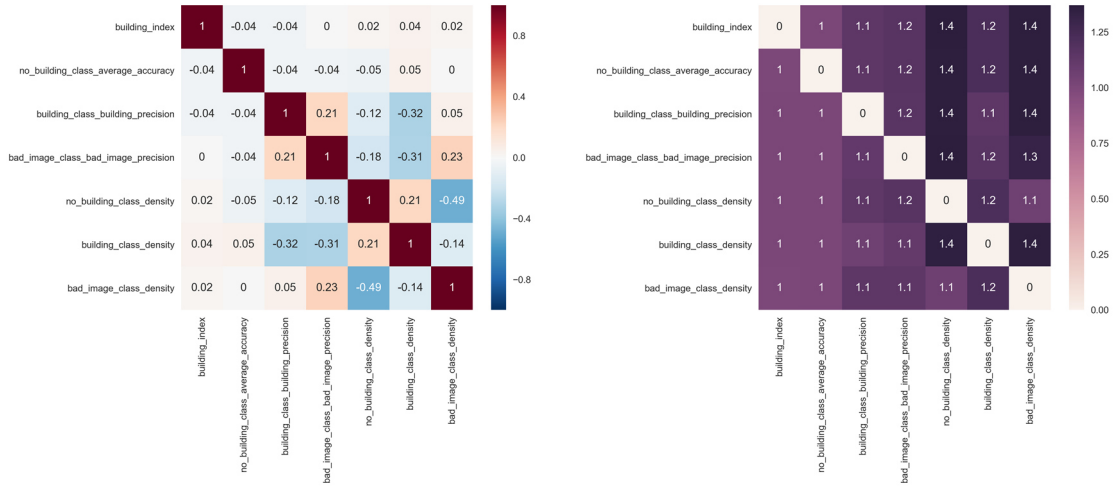

Figure 29 Correlation Matrix (left) and Variance Inflation Factors (right) for Logistic Regression Input Variables

The results for the whole reference dataset containing 941,589 observations are presented in Table 22. The logistic regression model performed was statistically significant with $\chi^2(6) = 940{,}640$ and $p < 0.005$. The model explains 71.5 % of the variance in the crowdsourcing performance (Nagelkerke pseudo R^2). As already indicated by the previous analysis, increases in the building index and average building precision for building classifications are associated with a strong and significant increased likelihood 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 are associated with a significant decreased likelihood of presence of buildings within the MapSwipe task.

Table 22 Results of the Logistic Regression Analysis

	Coefficients	**Std. Error**	**Significance**	**Odds Ratio**
Building Index	78.735	0.027	0.000	2626.8337
Average Accuracy (No Building Results)	-84.139	0.076	0.000	0.0002
Average Building Precision (Building Results)	62.469	0.051	0.000	516.406
Average Bad Image Precision (Bad Image Results)	-14.723	0.059	0.000	0.2294
No Building Class Density	-0.0148	0.001	0.000	0.9853
Building Class Density	0.0919	0.001	0.000	1.0962
Bad Image Class Density	-0.2360	0.005	0.000	0.7897

5.4.2 Performance of Machine Learning based Aggregation

The results of the logistic regression analysis show that the various intrinsic parameters on agreement, user characteristics and spatial characteristics can be used to estimate which tasks will contain buildings. The next step is conducted to evaluate the performance of three different machine learning classifiers and the naïve agreement based method (proposed in section 5.1.2) towards their ability to correctly depict which tasks contain buildings.

Table 23 Confusion Matrix for Crowd Answer

		Crowd Answer	
		No Bui	Bui
Ref	No Bui	584,735	5,262
	Bui	14,943	54,173

Table 24 Confusion Matrix for Random Forest Classifier

		Random Forest Classifier	
		No Bui	Bui
Ref	No Bui	589,114	883
	Bui	4,467	64,649

Table 25 Confusion Matrix for Logit Classifier

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

Table 26 Confusion Matrix for Keras Classifier

		Keras Classifier	
		No Bui	Bui
Ref	No Bui	588,653	1,344
	Bui	4,457	64,659

The performance of each classifier can be interpreted by looking at the corresponding confusion matrixes (Table 23, Table 24, Table 25, Table 26). Given the unbalanced distribution of building tasks in the dataset, it is no surprise that all aggregation methods classify most tasks as “no building”. About 90 % of all tasks are assigned to this category. The high proportion of correct no building classifications on the overall number of tasks is the main reason for the very high accuracy values obtained by all classifiers (Table 27). Using the crowd answer an accuracy of 96.7 % is reached, for the machine learning based aggregation methods even higher values greater than 99% are observed.

When looking at the building classifications greater heterogeneity for the different approaches is revealed. The crowd answer shows a considerable higher number of false positive classifications in comparison to the other classifiers. 5,262 tasks are incorrectly classified as building using the crowd answer, whereas only 883 to 1,514 false positives are recorded for the machine learning based aggregation. This fact is also captured by differences in the building precision score for each method. The best building precision could be achieved for the random forest classifier (98.6 %). The building precision for the crowd answer aggregation only reaches 91.1 %. Differences between the aggregation methods are also reflected in the number of tasks incorrectly classified as “no building”. False negatives are again most dominant for the crowd answer aggregation. 14,943 tasks were aggregated as no building, although a building was mapped in the OSM reference. For the machine learning based aggregation less than a third of this number is achieved. For instance, only 4,457 tasks with buildings are missed applying the deep learning classifier. These characteristics are also described by the building sensitivity score. The score reaches only 78.4 % for the crowd answer aggregation, but up to 93.5 % for the machine learning based aggregations.

Table 27 Performance of Crowd Answer and Machine Learning Classifiers

	Crowd Answer	**Random Forest**	**Logit**	**Keras**
Overall Accuracy	0.9693	0.9919	0.9906	0.9912
Building Sensitivity	0.7838	0.9354	0.9325	0.9355
Building Precision	0.9115	0.9865	0.9770	0.9796
Building F1 Score	0.8428	0.9603	0.9543	0.9571

The results for building precision and building sensitivity show that machine learning based aggregation outperforms crowd answer aggregation in both directions. Aggregation based on a random forest, logit or deep learning classification method generate data with a better quality. This is expressed by the high values for f1 score, which is the harmonic mean of building sensitivity and building precision. For the crowd answer aggregation 84.3 % are obtained, whereas the random forest method derives a value of 96 %. Based on this metric the random forest classifier generates results with the highest quality.

The dissimilarity between crowd answer aggregation and random forest based aggregation are visualized in Figure 30 and Figure 31 which show the spatial distribution of false negatives, false positives, true negatives and true positives. The maps depict that larger settlement structures are well captured by both approaches. The white vertical stripes within the project extent are characterized by complete consensus of no building classifications. However, due to the limitations of the MapSwipe data model (see section 3.2.2) for these tasks no user information was obtained and thus these areas are not subject of this study. As described already by the confusion matrixes 90 % of the study area are uninhabited. For crowd answer aggregation bands from east to west marked as false positive are present, e.g. in the northern part of the study area. These incorrect classifications could be successfully removed using the random forest based approach which also incorporates intrinsic user characteristics. Additionally, the high number of false negatives for the crowd answer aggregation is depicted in the map, for example in the mid-western part of the study area. Also in these regions the random forest based approach showed a better performance.

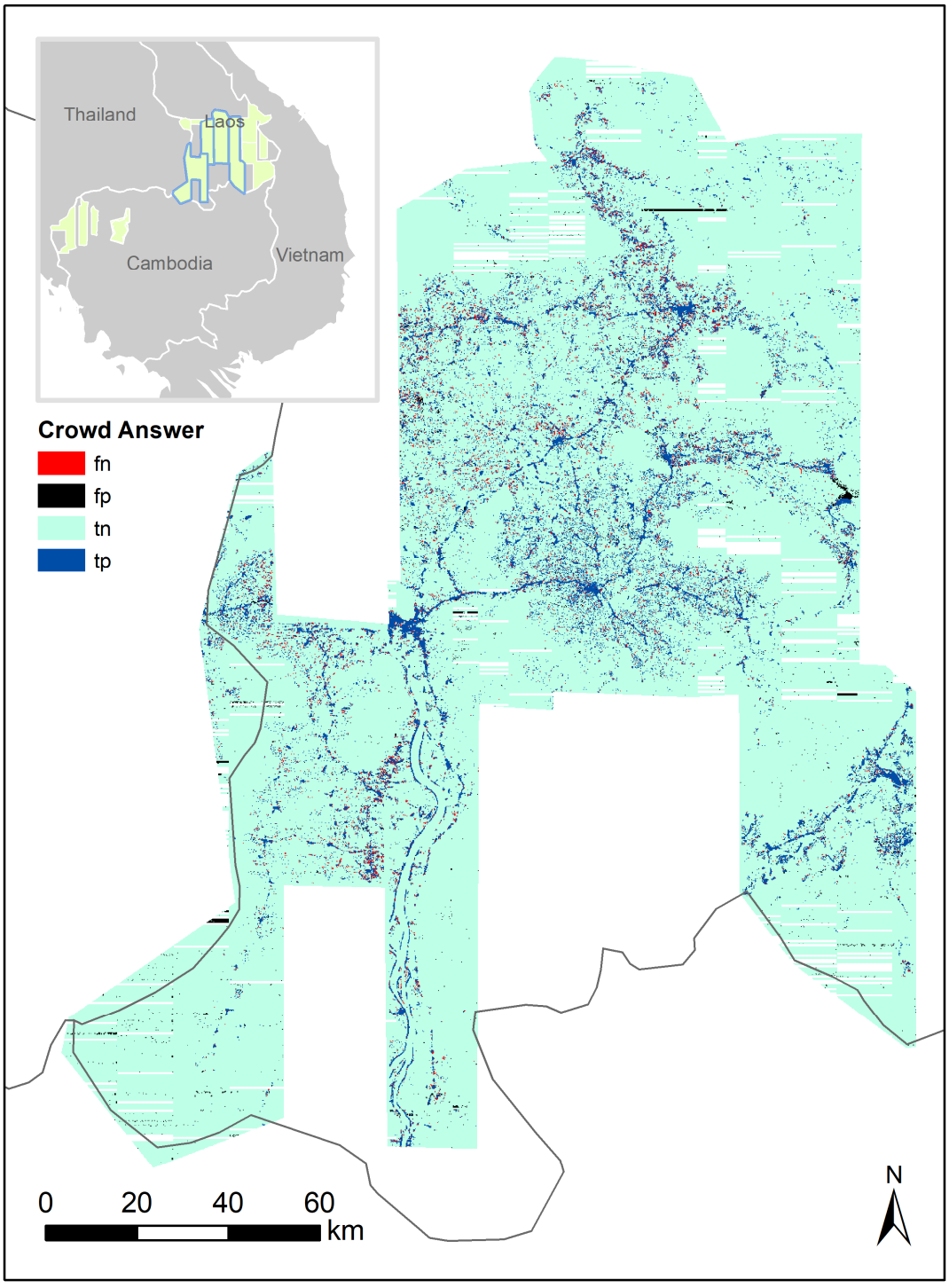

Figure 30 Spatial Distribution of Classification Results using Crowd Answer

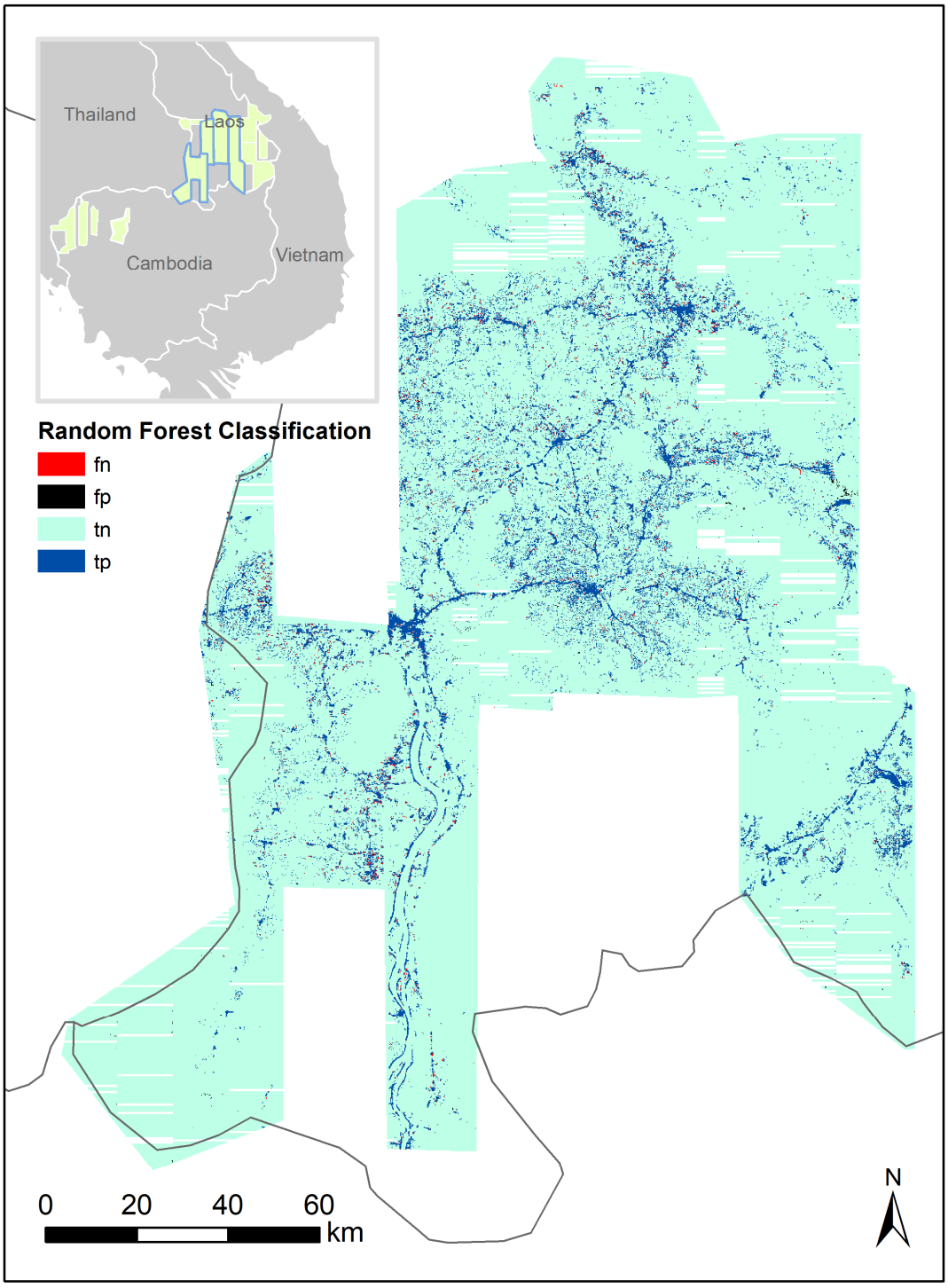

Figure 31 Spatial Distribution of Random Forest Classification Results

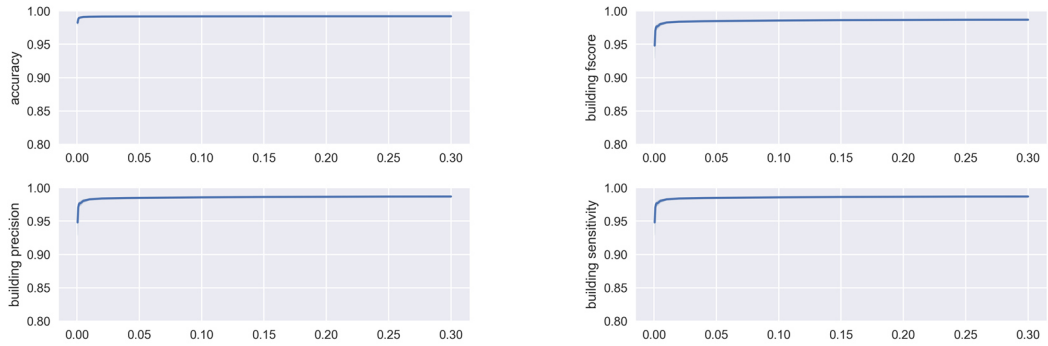

Figure 32 Performance of Random Forest Aggregation regarding Training Sample Size

Finally, the analysis of the performance of the random forest based aggregation regarding training sample size shows that even for small training samples high data quality can be obtained (Figure 32). The performance of the random forest classifiers starts to decrease rapidly for training sample proportions smaller than 0.1%. This corresponds to a total number of about 950 training samples.

5.4.3 Discussion

Machine learning based aggregation methods have potential to generate a high-quality settlement layer from crowdsourced MapSwipe data. The logistic regression model proved that the intrinsic characteristics of the dataset can be used to 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 or maximum likelihood.

Characteristics of the satellite imagery are 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 opens further potential to improve the performance of the machine learning models and resulting data quality.

The lack of reference data of sufficient quality limits the findings of this study. Since the reference dataset used in this study was derived from OpenStreetMap, the data quality might vary given the large size of the examined area. The quality of OpenStreetMap data

has been investigated by many authors and spatial variations in data quality are well described (Ali et al., 2016; Ballatore and Zipf, 2015; Fonte et al., 2015; Girres and Touya, 2010). 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.

The random forest based aggregation outperformed the naïve aggregation method applied so far significantly regarding building precision and building sensitivity. The results describe the performance for four selected MapSwipe projects. Given the global distribution of MapSwipe projects further validation of the automated classification and its transferability is needed.

6 Conclusion

In this study, crowdsourced classifications contributed by volunteers using the MapSwipe app were analysed regarding data quality. The work focused on the impact of agreement and redundancy on data quality (RQ1). Results have shown that high agreement on building classifications serves as an indicator for correct contributions. User characteristics were analysed towards their effect on data quality (RQ2) where it was revealed that he analysis of intrinsic user characteristics can be utilized to identify consistently incorrect classifications. Inaccurate users can be identified by low user performance values. The impact of spatial characteristics on data quality was examined (RQ3) and the results proof that the kernel density of classifications can explain incorrect contributions. Spatial outliers are probable of being incorrect. This work aimed at analysing the performance of machine learning based aggregation of crowdsourcing classifications (RQ4). A proposed random forest based aggregation of MapSwipe data to generate settlement information produces high quality data in comparison to state-of-the-art products derived from satellite imagery.

Low temporal resolution and old capture dates of available very high-resolution satellite imagery (e.g. provided by Bing) for some regions limit the applicability of MapSwipe data on the ground. In situations where populations are highly dynamic such as in parts of sub-Saharan Africa as recent remote sensing data as possible should be utilised. Whereas in disaster situations high-resolution satellite imagery is made available for free, this is rarely the case for ongoing and lingering crises. Accessibility to remote sensing data still poses serious limitations to crowdsourcing approaches. By solving this issue, time-aware crowdsourcing approaches could also support continuous land use monitoring (Schultz et al., 2017; Zhu and Woodcock, 2014). Therefore, the potential of crowdsourcing applications like MapSwipe for updating existing geographic information should be evaluated.

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. Especially the implementation of dynamic user characteristics should be addressed to provide instant feedback (Salk et al., 2016). Initial testing of the user's performance and characterisation of each user's personal difficulties may provide a novel approach towards personalised mapping guidance material.

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. 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. For example, slum mapping and slum type classification from satellite imagery still face many challenges (Kuffer et al., 2016).

MapSwipe data is used to derive geographical information on human settlements. Currently, it has been used by humanitarian organisations such as MSF, CartONG or HOT to delineate inhabited areas. The Current crowdsourcing approach provided only a binary settlement layer. More detailed information is added through the detailed mapping of buildings in OpenStreetMap. This setting offers further potential to generate detailed information on population distribution and population density. In many natural disasters or humanitarian catastrophes information on affected populations has been crucial for disaster response activities (Pesaresi et al., 2013; Voigt et al., 2007). Preventive risk assessment or the implementation of disaster risk reduction strategies require detailed information on the spatial-temporal population distribution. The potential of MapSwipe data as an input for population mapping should therefore be a focus for future research.

It is important to notice, that MapSwipe is more than just geographical data. MapSwipe also builds a community of volunteers that want to support humanitarian organizations by using their free time to map. The motivation of MapSwipe volunteers is still underexplored. Finding out more about the MapSwipe community and fostering discussions within the community could provide great benefits for the future development of the application. Building sustainable communities is a major topic for crowdsourcing projects (Dittus et al., 2016). Tools that visualize the data produced by volunteers are crucial in this manner. Projects like “MapSwipe Analytics” already provide first views on MapSwipe data. However, in the future more efforts should be put into user centric visualizations.

Using MapSwipe each volunteer helps to fill the gaps in the missing maps, that are urgently needed to tackle global inequalities and support the United Nation’s sustainable development goals to transform our world. MapSwipe was launched in July 2016 and one year of swiping and tapping have shown how an easy to use applications can directly affect the work of humanitarian organisations. Let’s build upon the achieved and bring MapSwipe and its community to the next level.

Acknowledgement

I would like to thank all those who contributed to the success of this thesis. First, I thank Prof. Dr. João Porto de Albuquerque and Prof. Dr. Alexander Zipf who supported this work with their comments and ideas and gave me the opportunity to become a part of the GIScience Research Group at Heidelberg University. I would like to thank Marcel Reinmuth for his aid on MapSwipe data validation and great discussions on MapSwipe related topics in general. Finally, I would like to thank the MapSwipe community. Without the thousands of people using the app this research would not have been possible. Furthermore, all project managers and the members of the "MapSwipe Working Group" contributed to this thesis with their ideas and views we regularly exchange.

References

Albuquerque, J., Herfort, B., Eckle, M., 2016. The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. Remote Sens. 8, 859. doi:10.3390/rs8100859

Ali, A., Sirilertworakul, N., Zipf, A., Mobasheri, A., 2016. Guided Classification System for Conceptual Overlapping Classes in OpenStreetMap. ISPRS Int. J. Geo-Information 5, 87. doi:10.3390/ijgi5060087

Anhorn, J., Herfort, B., Porto de Albuquerque, J., 2016. Crowdsourced Validation and Updating of Dynamic Features in OpenStreetMap An analysis of Shelter Mapping after the 2015 Nepal Earthquake. Proc. ISCRAM 2016 Conf. 0, 22–25.

Arcanjo, J.S., Luz, E.F.P., Fazenda, Á.L., Ramos, F.M., 2016. Methods for evaluating volunteers' contributions in a deforestation detection citizen science project. Futur. Gener. Comput. Syst. 56, 550–557. doi:10.1016/j.future.2015.07.005

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

Barrington, L., Ghosh, S., Greene, M., Har-Noy, S., Berger, J., Gill, S., Lin, A.Y.M., Huyck, C., 2011. Crowdsourcing earthquake damage assessment using remote sensing imagery. Ann. Geophys. 54, 680–687. doi:10.4401/ag-5324

Barron, C., Neis, P., Zipf, A., 2014. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Trans. GIS 18, 877–895. doi:10.1111/tgis.12073

Battersby, S.E., Hodgson, M.E., Wang, J., 2012. Spatial resolution imagery requirements for identifying structure damage in a Hurricane disaster: A cognitive approach. Photogramm. Eng. Remote Sensing 78, 625–635.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. doi:10.1016/j.isprsjprs.2009.06.004

Card, D., 1982. Using know map category marginal frequencies to improve estimates of thematic map accuracy. Photogramm. Eng. Remote Sens. 48, 432–439.

Chan, J., Crowley, J., Elhami, S., Erle, S., Munro, R., Schnoebelen, T., 2013. Aerial Damage Assessment Following Hurricane Sandy.

Chen, J., Zipf, A., 2017. DeepVGI: Deep Learning with Volunteered Geographic Information. WWW '17 Companion Proc. 26th Int. Conf. Companion World Wide Web 771–772. doi:10.1145/3041021.3054250

Comber, A., Fonte, C., Foody, G., Fritz, S., Harris, P., Olteanu-Raimond, A.M., See, L., 2016a. Geographically weighted evidence combination approaches for combining discordant and inconsistent volunteered geographical information. Geoinformatica 20, 503–527. doi:10.1007/s10707-016-0248-z

Comber, A., Mooney, P., Purves, R.S., Rocchini, D., Walz, A., 2016b. Crowdsourcing: It matters who the crowd are. The impacts of between group variations in recording land

cover. PLoS One 11, 1–19. doi:10.1371/journal.pone.0158329

Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J., 2013. #Earthquake: Twitter as a Distributed Sensor System. Trans. GIS 17, 124–147. doi:10.1111/j.1467-9671.2012.01359.x

de Albuquerque, J.P., Eckle, M., Herfort, B., Zipf, A., 2016. Crowdsourcing geographic information for disaster management and improving urban resilience: an overview of recent developments and lessons learned. Eur. Handb. Crowdsourced Geogr. Inf. 309.

de Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A., 2015. A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int. J. Geogr. Inf. Sci. 1–23. doi:10.1080/13658816.2014.996567

Deng, J., Krause, J., Fei-fei, L., 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587. doi:10.1109/CVPR.2013.81

Dittus, M., Quattrone, G., Capra, L., 2017. Mass Participation During Emergency Response: Event-centric Crowdsourcing in Humanitarian Mapping. Proc. 2017 ACM Conf. Comput. Support. Coop. Work Soc. Comput. 1290–1303. doi:10.1145/2998181.2998216

Dittus, M., Quattrone, G., Capra, L., 2016. Analysing Volunteer Engagement in Humanitarian Mapping: Building Contributor Communities at Large Scale. Cscw '16 108–118. doi:10.1145/2818048.2819939

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: A global population database for estimating populations at risk. Photogramm. Eng. Remote Sensing 66, 849–857. doi:10.1016/j.scitotenv.2008.02.010

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006. Review: A gentle introduction to imputation of missing values. J. Clin. Epidemiol. 59, 1087–1091. doi:10.1016/j.jclinepi.2006.01.014

Earle, P.S., Bowden, D.C., Guy, M., 2011. Twitter earthquake detection : earthquake monitoring in a social world. doi:10.4401/ag-5364

Eckle, M., de Albuquerque, J.P., 2015. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. Proc. ISCRAM 2015 Conf. - Kristiansand, May 24-27 1–8.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenbock, H., Muller, A., Dech, S., 2013. Urban footprint processor-Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. IEEE Geosci. Remote Sens. Lett. 10, 1617–1621. doi:10.1109/LGRS.2013.2272953

Exel, M. Van, Dias, E., Fruijtier, S., 2010. The impact of crowdsourcing on spatial data quality indicators. Proc. GIScience 2011 1–4.

Flanagin, A., Metzger, M., 2008. The credibility of volunteered geographic information. GeoJournal 72, 137–148. doi:10.1007/s10708-008-9188-y

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. doi:10.1037/h0031619

Fonte, C.C., Bastin, L., See, L., Foody, G., Lupia, F., 2015. Usability of VGI for validation of land cover maps. Int. J. Geogr. Inf. Sci. 29, 1269–1291. doi:10.1080/13658816.2015.1018266

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., Obersteiner, M., 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. Remote Sens. 1, 345–354. doi:10.3390/rs1030345

Gadiraju, U., Demartini, G., 2015. Human beyond the machine: challenges and opportunities of microtask crowdsourcing. Intell. Syst.

Geiger, R.S., Halfaker, A., 2013. Using edit sessions to measure participation in wikipedia. Proc. 2013 Conf. Comput. Support. Coop. Work - CSCW '13 861. doi:10.1145/2441776.2441873

Girres, J.-F., Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. Trans. GIS 14, 435–459. doi:10.1111/j.1467-9671.2010.01203.x

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. doi:10.1007/s10708-007-9111-y

Goodchild, M.F., Glennon, J.A., 2010. Crowdsourcing geographic information for disaster response: a research frontier. Int. J. Digit. Earth 3, 231–241. doi:10.1080/17538941003759255

Goodchild, M.F., Li, L., 2012. Assuring the quality of volunteered geographic information. Spat. Stat. 1, 110–120. doi:10.1016/j.spasta.2012.03.002

Greenland, S., Finkle, W.D., 1995. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. Am. J. Epidemiol. 142, 1255–1264. doi:10.1093/oxfordjournals.aje.a117592

Gueguen, L., Koenig, J., Reeder, C., Barksdale, T., Saints, J., Stamatiou, K., Collins, J., Johnston, C., 2017. Mapping Human Settlements and Population at Country Scale from VHR Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10, 524–538. doi:10.1109/JSTARS.2016.2616120

Hagenauer, J., Helbich, M., 2012. Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. Int. J. Geogr. Inf. Sci. 26, 963–982. doi:10.1080/13658816.2011.619501

Haklay, M., 2013. Citizen science and volunteered geographic information: Overview and typology of participation. Crowdsourcing Geogr. Knowl.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environ. Plan. B Plan. Des. 37, 682–703. doi:10.1068/b35097

Haklay, M. (Muki), Basiouka, S., Antoniou, V., Ather, A., 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. Cartogr. Journal, 47, 315–322. doi:10.1179/000870410X12911304958827

Haklay, M.M., 2016. Why is participation inequality important? Eur. Handb. Crowdsourced Geogr. Inf. 35.

Hara, K., Sun, J., Moore, R., Jacobs, D., Froehlich, J., 2014. Tohme, in: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology - UIST '14. ACM Press, New York, New York, USA, pp. 189–204. doi:10.1145/2642918.2647403

Herfort, B., 2017. MapSwipe: global dataset of crowdsourced classifications on human settlements and reference datasets. doi:10.11588/data/3GYJUG

Herfort, B., de Albuquerque, J.P., Schelhorn, S., Zipf, A., 2014. Exploring the Geographical Relations Between Social Media and Flood Phenomena to Improve Situational Awareness, in: Huerta, J., Schade, S., Granell, C. (Eds.), Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Cham, pp. 55–71. doi:10.1007/978-3-319-03611-3_4

Herfort, B., Reinmuth, M., de Albuquerque, J.P., Zipf, A., 2017. Towards evaluating crowdsourced image classification on mobile devices to generate geographic information about human settlements. Proc. 20th Agil.

Hillen, F., Höfle, B., 2015. Geo-reCAPTCHA: Crowdsourcing large amounts of geographic information from earth observation data. Int. J. Appl. Earth Obs. Geoinf. 40, 29–38. doi:10.1016/j.jag.2015.03.012

Howe, J., 2006. The Rise of Crowdsourcing. Wired Mag. 14, 1–5. doi:10.1086/599595

Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S., 2014. AIDR: Artificial intelligence for disaster response. Proc. companion Publ. 23rd Int. Conf. World wide web companion 159–162. doi:10.1145/2567948.2577034

Imran, M., Elbassuoni, S., 2013. Extracting information nuggets from disaster-related messages in social media, in: Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany, May 2013. pp. 1–10.

Jacobs, C., 2016. Data quality in crowdsourcing for biodiversity research: issues and examples. Eur. Handb. Crowdsourced Geogr. Inf. 75–86.

Johnson, A., n.d. DeepOSM [WWW Document]. URL https://github.com/trailbehind/DeepOSM

Juhász, L., Hochmair, H.H., 2016. User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service. Trans. GIS 20, 925–947. doi:10.1111/tgis.12190

Keller, S., Bühler, S., Kurath, S., 2016. Erkennung von Fußgängerstreifen aus Orthophotos 162–166. doi:10.14627/537622023.Dieser

Klotz, M., Kemper, T., Geiß, C., Esch, T., Taubenböck, H., 2016. How good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from Central Europe. Remote Sens. Environ. 178, 191–212. doi:10.1016/j.rse.2016.03.001

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016. Slums from Space — 15 Years of Slum Mapping Using Remote Sensing. Remote Sens. 8, 1–29. doi:10.3390/rs8060455

Lesiv, M., Moltchanova, E., Schepaschenko, D., See, L., Shvidenko, A., Comber, A., Fritz, S., 2016. Comparison of data fusion methods using crowdsourced data in creating a hybrid forest cover map. Remote Sens. 8. doi:10.3390/rs8030261

Levin, G., Newbury, D., McDonald, K., Alvarado, I., Tiwari, A., Zaheer, M., 2016. Terrapattern: Open-Ended, Visual Query-By-Example for Satellite Imagery using Deep Learning [WWW Document]. URL http://terrapattern.com

Longley, P., 2005. Geographic information systems and science. John Wiley & Sons.

Lüge, T., 2014. GIS Support for the MSF Ebola response in Guinea in 2014.

Meier, P., 2012. New information technologies and their impact on the humanitarian sector. Int. Rev. Red Cross 93, 1239–1263. doi:10.1017/S1816383112000318

Mocnik, F.-B., Zipf, A., Raifer, M., 2017. The OpenStreetMap folksonomy and its evolution. Geo-spatial Inf. Sci. 20, 219–230. doi:10.1080/10095020.2017.1368193

Moran, P.A.P., 1950. Notes on Continuous Stochastic Phenomena. Biometrika 37, 17–23. doi:10.1093/biomet/37.1-2.17

Neis, P., Zielstra, D., Zipf, A., 2013. Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. Futur. Internet 5, 282–300. doi:10.3390/fi5020282

Neis, P., Zielstra, D., Zipf, A., 2011. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Futur. Internet 4, 1–21. doi:10.3390/fi4010001

Neis, P., Zipf, A., 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. ISPRS Int. J. Geo-Information 1, 146–165. doi:10.3390/ijgi1020146

Palen, L., Soden, R., Anderson, T.J., Barrenechea, M., 2015. Success & Scale in a Data - Producing Organization : The Socio - Technical Evolution of OpenStreetMap in Response to Humanitarian Events. Chi 2015 4113–4122.

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M.A., Ouzounis, G.K., Scavazzon, M., Soille, P., Syrris, V., Zanchetta, L., 2013. A global human settlement layer from optical HR/VHR RS data: Concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2102–2131. doi:10.1109/JSTARS.2013.2271445

Quattrone, G., Mashhadi, A., Capra, L., 2014. Mind the Map : The Impact of Culture and Economic Affluence on Crowd-Mapping Behaviours. Proc. 17th ACM Conf. Comput. Support. Coop. Work Soc. Comput. - CSCW '14 934–944. doi:10.1145/2531602.2531713

Raykar, V.C., Yu, S., Zhao, L.H., Hermosillo Valadez, G., Florin, C., Bogoni, L., Moy, L., Org, L.M., 2010. Learning From Crowds. J. Mach. Learn. Res. 11, 1297–1322. doi:10.2139/ssrn.936771

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252. doi:10.1007/s11263-015-0816-y

Salk, C.F., Sturn, T., See, L., Fritz, S., 2016. Limitations of Majority Agreement in Crowdsourced Image Interpretation. Trans. GIS 0, n/a-n/a. doi:10.1111/tgis.12194

Salk, C.F., Sturn, T., See, L., Fritz, S., Perger, C., 2013. Assessing quality of volunteer

crowdsourcing contributions: Lessons from the Cropland Capture game. J. Chem. Inf. Model. 53, 1689–1699. doi:10.1017/CBO9781107415324.004

Schnebele, E., Cervone, G., Waters, N., 2014. Road assessment after flood events using non-authoritative data. Nat. Hazards Earth Syst. Sci. 14, 1007–1015. doi:10.5194/nhess-14-1007-2014

Schultz, M., Voss, J., Auer, M., Carter, S., Zipf, A., 2017. Open land cover from OpenStreetMap and remote sensing. Int. J. Appl. Earth Obs. Geoinf. 63, 206–213. doi:10.1016/j.jag.2017.07.014

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., Liu, H.-Y., Milčinski, G., Nikšič, M., Painho, M., Pődör, A., Olteanu-Raimond, A.-M., Rutzinger, M., 2016. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS Int. J. Geo-Information 5, 55. doi:10.3390/ijgi5050055

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., Haklay, M. (Muki), 2016. A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 8816, 1–29. doi:10.1080/13658816.2016.1189556

Silverman, B.W., 1986. Density estimation for statistics and data analysis. CRC press.

Simpson, E., Roberts, S.J., Psorakis, I., Smith, A., 2012. Dynamic Bayesian Combination of Multiple Imperfect Classifiers.pdf. Decis. Mak. with Imperfect Decis. Makers Springer 1–27.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. doi:10.1016/j.ipm.2009.03.002

Touya, G., Antoniou, V., 2016. Assessing Crowdsourced POI Quality : Combining Methods based on Reference Data , History , and Spatial Relations 1–29.

Vakalopoulou, M., Karantzalos, K., Komodakis, N., Vakalopoulou, M., Karantzalos, K., Komodakis, N., Building, N.P., 2015. Building detection in very high resolution multispectral data with deep learning features. Igarss 1873–1876. doi:10.1109/IGARSS.2015.7326158

Vieweg, S., Hughes, A., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Voigt, C., Dobner, S., Ferri, M., Hahmann, S., Gareis, K., 2016. Computers Helping People with Special Needs 9759, 257–264. doi:10.1007/978-3-319-41267-2

Voigt, S., Kemper, T., Riedlinger, T., Kiefl, R., Scholte, K., Mehl, H., 2007. Satellite image analysis for disaster and crisis-management support. IEEE Trans. Geosci. Remote Sens. 45, 1520–1528. doi:10.1109/TGRS.2007.895830

von Ahn, L., Liu, R., Blum, M., 2006. Peekaboom: A Game for Locating Objects in Images. Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '06 55. doi:10.1145/1124772.1124782

Westrope, C., Banick, R., Levine, M., 2014. Groundtruthing OpenStreetMap Building Damage Assessment. Procedia Eng. 78, 29–39. doi:10.1016/j.proeng.2014.07.035

Yan, Y., Eckle, M., Kuo, C., Herfort, B., Fan, H., 2017. Monitoring and Assessing Post-Disaster Tourism Recovery Using Geotagged Social Media Data. ISPRS Int. J. Geo-Information 6, 144. doi:10.3390/ijgi6050144

Yang, A., Fan, H., Jing, N., Sun, Y., Zipf, A., 2016. Temporal Analysis on Contribution Inequality in OpenStreetMap: A Comparative Study for Four Countries. ISPRS Int. J. Geo-Information 5, 5. doi:10.3390/ijgi5010005

Yin, A., 2017. A Mapathon to Pinpoint Areas Hardest Hit in Puerto Rico. New York Times 1.

Zhang, A., Liu, X., Tiecke, T., Gros, A., 2016. Population Density Estimation with Deconvolutional Neural Networks, in: Workshop on Large Scale Computer Vision at NIPS 2016.

Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171. doi:10.1016/j.rse.2014.01.011

Zook, M., Graham, M., Shelton, T., Gorman, S., 2010. Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. World Med. Heal. Policy 2, 6–32. doi:10.2202/1948-4682.1069

Affidavit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht wurden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Heidelberg, 20. Oktober 2017

……

Benjamin Herfort