ORIGINAL ARTICLE

Measuring the Wisdom of the Crowd: How Many is Enough?

Volker Walter[1] · **Michael Kölle**[1] · **David Collmar**[1]

Received: 26 August 2021 / Accepted: 16 March 2022 / Published online: 6 April 2022

Abstract
The idea of the wisdom of the crowd is that integrating multiple estimates of a group of individuals provides an outcome that is often better than most of the underlying estimates or even better than the best individual estimate. In this paper, we examine the wisdom of the crowd principle on the example of spatial data collection by paid crowdworkers. We developed a web-based user interface for the collection of vehicles from rasterized shadings derived from 3D point clouds and executed different data collection campaigns on the crowdsourcing marketplace microWorkers. Our main question is: how large must be the crowd in order that the quality of the outcome fulfils the quality requirements of a specific application? To answer this question, we computed precision, recall, F1 score, and geometric quality measures for different crowd sizes. We found that increasing the crowd size improves the quality of the outcome. This improvement is quite large at the beginning and gradually decreases with larger crowd sizes. These findings confirm the wisdom of the crowd principle and help to find an optimum number of the crowd size that is in the end a compromise between data quality, and cost and time required to perform the data collection.

Keywords Wisdom of the crowd · Paid crowdsourcing · Crowdworker · MicroWorkers · Spatial data collection · Quality evaluation

Zusammenfassung
Messung der Weisheit der Masse—Wie viele sind genug? Die Idee der Weisheit der Masse ist, dass die Integration von mehreren Schätzungen von Einzelpersonen ein Ergebnis liefert, das oft besser ist als die meisten der zugrundeliegenden Schätzungen oder sogar besser als die beste Einzelschätzung. In diesem Beitrag untersuchen wir dieses Prinzip am Beispiel der Erfassung von raumbezogenen Daten durch bezahlte Crowdworker. Hierzu haben wir eine webbasierte Benutzeroberfläche für die Erfassung von Fahrzeugen aus 3D-Punktwolken entwickelt und verschiedene Datenerfassungskampagnen auf dem Crowdsourcing-Marktplatz microWorkers durchgeführt. Unsere Hauptfrage lautet: Wie groß muss die Anzahl von Crowdworkern sein, damit die Qualität der Daten die Qualitätsanforderungen einer bestimmten Anwendung erfüllt? Um diese Frage zu beantworten, berechneten wir Precision, Recall, F1-Score und geometrische Qualitätsmaße für Gruppen mit unterschiedlicher Anzahl von Crowdworkern. Wir haben festgestellt, dass sich die Qualität der Daten durch die Integration der individuellen Erfassungen verbessert. Diese Verbesserung ist zu Beginn groß und nimmt mit zunehmender Gruppengröße ab. Die Ergebnisse bestätigen das Prinzip der Weisheit der Masse und helfen dabei, eine optimale Größe der Crowd zu finden, die letztlich ein Kompromiss zwischen Datenqualität und den für die Datenerfassung erforderlichen Kosten und der Zeit darstellt.

1 Introduction

The term *crowdsourcing* was introduced by Howe (2006) and is a neologism of the terms *crowd* and *outsourcing*. In contrast to outsourcing, where employers outsource tasks to known and well-defined third parties, crowdsourcing involves the outsourcing of tasks to unknown workers (crowdworkers) on the Internet. This gives employers access

✉ Volker Walter
volker.walter@ifp.uni-stuttgart.de

[1] Institute for Photogrammetry (ifp), University of Stuttgart, Geschwister-Scholl-Str. 24D, 70174 Stuttgart, Germany

to a large number of workers who would otherwise not be available.

Many crowdsourcing projects rely on the work of unpaid volunteers, such as Wikipedia (www.wikipedia.org) or Zooniverse (www.zooniverse.org). The voluntary collection of geodata is called Volunteered Geographic Information (VGI) (Goodchild 2007). The most known VGI project is OpenStreetMap (OSM—www.openstreetmap.org), a project to create a map of the world that can be edited by anyone (Haklay and Weber 2008). VGI projects require an active community that is intrinsically motivated to participate. The main motivating factor for users who voluntarily collect OSM data is that their work benefits other users, which in the case of OSM is freely available digital map data, but there are also other motivating factors, such as gaining reputation in the community, fun and inspiration, or gaining new knowledge. Budhathoki and Haythornthwaite (2012) presented a comprehensive discussion of motivational factors for VGI. If such motivations are not available, other (extrinsic) incentives must be provided. A problem of volunteered spatial data collection can be that the number of volunteers in the area to be mapped is often limited, hindering scalability and making it difficult to map large areas within a specific period (Maddalena et al. 2020).

The most common extrinsic motivation for crowdworkers to perform tasks and the one that leads to the fastest results is monetary incentive (Haralabopoulos et al. 2019). In paid crowdsourcing, tasks are published on online marketplaces (Mao et al. 2013) such as microWorkers (www.microworkers.com) (Hirth et al. 2011) or Amazon Mechanical Turk (MTurk—www.mturk.com) (Ipeirotis 2010). MicroWorkers, for example, has access to over 2,700,000 workers worldwide (according to their website—accessed March 2022). The research in this article is based on paid crowdsourcing.

Paid crowdsourcing has been proven a powerful tool for very different applications. It can be used for practically all tasks that can be solved online with a computer. Some examples that show the variety of possible applications: Nguyen et al. (2012) describe an Amazon MTurk based application to detect polyps associated with colorectal cancer in images generated through computed tomographic colonography. Redi and Povoa (2014) compare the reliability in scoring the aesthetic appeal of images by paid and volunteered crowdworkers. Gao et al. (2015) propose a paid crowdsourcing approach that creates translations at much lower cost than hiring professional translators. Lans et al. (2018) developed a technology for a museum app and created content for over 80 museums worldwide with paid crowdsourcing. Koita and Suzuki (2019) propose a method to count traffic without actually being at the survey site by presenting images extracted from videos to paid crowdworkers.

There exists much research on spatial data collection based on the input of volunteers (Goodchild and Li 2012; Antonio and Skopeliti 2015; Albuquerque et al. 2016; See et al. 2016; Fonte et al. 2017; Seratne et al. 2017; Pinhero and Davis 2018) but so far only a limited number of publications on spatial data collection relying on paid crowdsourcing. Estesa et al. (2016) describe a platform for the mapping of crop fields in South Africa realized with Mechanical Turks' Human Intelligence Tasks (HITs). Walter and Soergel (2018) discuss the collection of buildings, forests and streets from aerial images with microWorker campaigns. Koelle et al. (2020) discuss a system for the classification of 3D point clouds where a machine-learning algorithm iteratively improves its performance by learning from paid crowdworkers. Walter et al. (2020) discuss the collection of trees from 3D LiDAR point clouds based on paid crowd campaigns. Maddalena et al. (2020) implemented a system to collect coordinates of points of interest from Street View images by paid crowdworkers. Koelle et al. (2021b) evaluated which 3D data representation (point cloud vs. mesh) is best suited for presenting to paid crowdworkers in context of coupled semantic segmentation of both an ultra-high-resolution point cloud and the derived 3D textured mesh. A well-founded basis for this approach of learning with only a few samples is presented in Koelle et al. (2021a), who also discuss the impact of erroneous answers from the crowd and possible countermeasures. Walter et al. (2020) present a two-level approach for the collection of vehicles from 3D point clouds by paid crowdworkers.

In both paid and free crowdsourcing, quality control is a challenge (Leibovici et al. 2017; Liu et al. 2018) because the quality of crowdsourced work varies widely (Vaughan 2017). The crowd consists of people with unknown and different skills (Daniel et al., 2018). Another problem, especially in paid crowdsourcing, is dishonest workers who try to maximise their income by submitting as many tasks as possible and delivering incomplete or poor results (Hirth et al. 2011). In addition, there may be adversarial workers who compromise the quality of the results (Zhang et al. 2016). Obtaining high-quality ground truth from noisy data collected by non-experts is a major challenge in crowdsourcing (Zhou et al. 2012).

In general, there are two methods for controlling and improving the quality of crowd-sourced data (Zhang et al. 2016): (i) *Quality Control on Task Designing* and (ii) *Quality Improvement after Data Collection*. The first approach stands for methods that guide crowdworkers to provide high-quality data. There are various methods for this purpose, such as skill testing, reputation systems, task assignment, task and workflow optimization, training, real-time quality control or quality checkpoints. An overview of these techniques is presented in (Daniel et al. 2018).

All these methods can improve the quality of the data. In the end, however, they are only partial solutions. Even if they manage to improve the average quality, the quality of individual datasets can still be heterogenous and low quality results cannot be completely eliminated. Many crowdworkers are not familiar with geospatial data collection standards or may not feel the need to follow them (Hashemi and Abbaspous 2015). Even when spatial data are collected by experts, the results can be heterogeneous (Walter and Soergel 2018). When spatial data are collected by non-experts, this effect is even stronger because people with very different expertise are working together (Senaratne et al. 2017). Methods based on *Quality Control on Task Designing* are not included in the research presented in this paper. However, they can be combined with our proposed methods for further quality improvement.

In the second approach (*Quality Improvement after Data Collection*), additional procedures are used to improve the quality of the data after it has been collected. A common idea for this is repeated data collection by different crowdworkers. After data collection, mechanisms are applied to filter out noisy data and to infer the truth. The process of estimating the truth from noisy multiple collected data is called *truth inference* (Zhen et al. 2017).

Surowiecki (2004) has shown on many examples from very different fields that averages of multiple guesses are often better than the best individual guess. Large groups of people are smarter and can solve complex problems even better than experts. To do so, the same data must be collected multiple times (which can be realized without problems with paid crowdsourcing but would be hard to realize with volunteered data collection) and an aggregation rule (such as averaging) to derive a solution (Simons 2004).

Groups can be very smart when their aggregated results are compared with the results of individuals (Lorenz et al. 2010). A similar principle is swarm intelligence where the skills of individuals and the power of the masses are used to solve problems (Krause et al. 2010). This behaviour is also known from groups of insects, fishes, birds, mammals and primates that aggregate their individual decisions into group decisions for various tasks (Zuni and Eckstein 2017). However, swarm intelligence is only partially comparable to crowdsourcing, as the behaviour of participants in a swarm often arises from the behaviour of their physical neighbours (e.g., in the flight of flocks of birds), while participants in a crowdsourcing campaign are usually unaware of the existence of other crowdworkers (again, there are exceptions, such as in collaborative crowdsourcing, where a number of workers form groups and work together (Ikeda 2016)).

The crowd can comprise of any number of individuals (the larger the better) who neither need to know each other nor need to be aware what others are doing (Brown 2015). The wisdom of the crowd is a statistical phenomenon and relies only on mathematical aggregation methods (Hosio et al. 2016).

It is widely acknowledged that Charles Darwin's cousin Sir Francis Galton first observed the wisdom of the crowd in 1907 (Galton 1907). He found out that the average guess of the weight of an ox in a weight-judging contest at a farmer fair outperformed the accuracy of expert opinions (butchers). The average judgement nearly converged to an optimum result. In average, the fair audience estimated the weight of the ox was 1197 pounds. The real weight of the ox was 1198 pounds. Since then, many other researchers verified similar findings (Surowiecki 2004).

Simple majority voting is a widespread mechanism to exploit the wisdom of the crowd (Juni and Eckstein 2017). This mechanism is based on the idea that the majority of the workers are trusted (Kazemi et al. 2013). The employer assigns a task multiple times to many crowd workers. The result with the majority is considered the correct answer (Hirth et al. 2013). Majority voting is a simple but effective method (Zhang et al. 2016) and has been proven useful for different spatial labelling tasks. Salk et al. (2016) used majority vote classification for the identification of croplands in remote sensing images. Hecht et al. (2018) used majority voting to evaluate the results of crowdsourced classification of building footprint data. Herfort et al. (2018) used majority voting for the detection of trees in 3D point clouds by crowdworkers. Liu et al. (2018) applied weighted majority voting as aggregation technique for the crowd-based building of accessibility maps. The weights were derived from the reputation scores of the crowdworker. Koelle et al. (2020) used majority voting for the crowd-based labelling of 3D LiDAR point clouds.

While majority voting can be easily used for labelling tasks, it cannot be applied to the collection of the geometry of spatial objects. Labels are often classified into a finite number of classes (i.e., forming a categorization task) whereas the geometry of spatial objects can have in principle any shape. If different crowdworkers collect the geometry of an object, all resulting representations will be more or less different, which means that it is not possible to determine a majority representation.

If the objects are represented by simple geometric primitives, such as points, circles or rectangles, we can use the average instead of the majority as aggregation rule. This was successfully adapted by Walter et al. (2020) on the example of crowd-based collection of trees from 3D point clouds by means of minimum enclosing cylinders. Each crowdjob was duplicated 10 times. Integrated cylinders were calculated by averaging the centres and the heights of the individual cylinders. The quality of the integrated cylinders was significantly higher compared to the average quality of the individual cylinders.

If the objects are not represented by simple geometric primitives but by free-shaped polygons, the calculation of an average geometry is more difficult because polygons are not defined by a fixed list of parameters that can be averaged but can have any number of intermediate points at any position. The typical approach to calculate an average representation of multiple polygons is that first corresponding parts of the polygons are identified (matching) and then the actual integration is performed (conflation) (Walter and Fritsch 1999). In the past decades, many algorithms for the matching and conflation of geospatial data were introduced. An overview is presented in (Xavier et al. 2016).

Three factors affect the quality of the results of groups of crowdworkers that work together on a problem: The diversity of the group, the expertise of the crowdworkers, and the number of crowdworkers in the group.

If we crowdsource a task via a crowdsourcing marketplace, we have limited influence to the diversity of the workers. For example, on the microWorkers marketplace, it is possible to filter the crowdworkers based on the country of origin, the individual score of a crowdworker based on feedback from employers, the total payment received (top earners) and based on some specific skills some crowdworkers might have (programming, writing, blogging, caption writing, designing, data collecting, image annotating). However, the employer has no detailed control over the diversity of the group since the job assignment is an open process and every registered worker who belongs to the selected group is allowed to perform the tasks.

However, we can exactly control the number of workers. The higher the number of workers in the group, the more likely it is that the diversity of the group is also high. Krause et al. (2011) have shown that high diversity can be more beneficial to the quality than adding expertise. However, also expertise can improve the quality: King et al. (2011) demonstrated that in particular work configurations, experts could improve the quality and prevent significantly inaccurate results. The expertise of the crowdworkers could be checked with qualification tests (qualification-based preselection) (Geiger et al. 2011). However, expertise checks and expertise improvement are not part of our current research. Van Dijk et al. (2020) investigated the improvement of crowdsourced data by aggregating it and examined the relationship between the number of contributions, the parameters of the aggregation method and the resulting quality.

In this paper, our focus is on evaluating the impact of the number of crowdworkers on the quality of the aggregated results. For this, we implemented a graphical user interface for the collection of vehicles from 2D rasterized shadings derived from 3D LiDAR point clouds by means of line segments reaching from the front to the back of the vehicles. Each data collection task was assigned to multiple crowdworkers. For different group sizes, we calculated all possible combinations of the individual results and integrated them with a DBSCAN clustering (Ester et al. 1996). The results are then compared with reference data. The purpose of this work is to answer two questions: (1) How does the number of crowdworkers influence the completeness and correctness? (2) How does the number of crowdworkers influence the geometric quality?

The remainder of this paper is organized as follows. In Sect. 2, we introduce the graphical user interface for the collection of vehicles from rasterized shadings from LiDAR point clouds. The used datasets are presented in Sect. 3. In Sect. 4, we discuss the individual results of the different crowdsourcing campaigns. In Sect. 5, we present our data integration strategy and evaluate the wisdom of the crowd principle. We conclude our findings in Sect. 6 by discussing our results and formulating further research questions.

2 Crowd-Based Collection of Vehicles

Vehicle detection from remote sensing data has gained increasing interest in recent years (Wu et al. 2020) and can be important for many applications, e.g. autonomous driving, traffic management, traffic monitoring, urban planning, parking lot analysis, etc. Numerous approaches have been proposed for vehicle detection. The majority of these approaches can be divided into two classes: (i) detection of vehicles from vehicle-mounted sensors, such as in-vehicle LiDAR (Feng et al. 2019) or in-vehicle cameras (Ponn et al. 2020) and (ii) detection of vehicles from airborne sensors, such as airborne LiDAR (Eum et al. 2017) or airborne cameras (Yang et al. 2018). In recent years, deep learning algorithms have become powerful tools for the automated detection of vehicles from all kind of sensing data and have achieved remarkable results (Wang et al. 2019). Deep learning systems require large amounts of annotated data to implement interpretation concepts and crowdsourcing offers an effective method to provide such data.

Our study is based on an approach that we presented in (Walter et al. 2021). It is a two-level method for the crowd-based acquisition of vehicles from 3D point clouds by means of minimum bounding boxes. We subdivide the approach into two steps to make the data collection task as simple as possible because most crowdworkers have no expert knowledge in the field of geospatial data collection or have worked with 3D point clouds before.

Crowdsourcing tasks are designed in such way, that complex problems are subdivided into smaller sub-problems that can be solved quickly and easily. Paid crowdsourcing tasks take typically only a few minutes to complete and the payment is only several cents (Hirth et al. 2011; Hitlin 2016). For spatial data, subdividing large problems can be understood as geographically dividing the working area into small tiles and assigning these tiles to individual crowdworkers. However, this is only reasonable if the objects to be collected are homogeneously distributed, which is generally not the case for geospatial data. For example, in rural landscapes, the occurrence of vehicles is rather scarce. In this case, we would produce many tiles that contain no vehicles at all.

Therefore, we subdivide the working area into large strips (so that absence of vehicles is very unlikely) and present these strips as 2D rasterized shadings to the crowdworkers in which they must first identify the positions of all vehicles within a strip. The advantage of strips over tiles is that crowdworkers only have to navigate in two directions in the user interface. This makes it less likely, that they will miss certain areas.

In the second step, these positions are used as preliminary information to cut out small parts from the 3D point cloud for each identified vehicle. Each of these small 3D point clouds is then presented to a crowdworker, which is asked to approximate the vehicle with a minimal bounding box. This has the advantage that each crowdworker only needs to download a small portion of the point cloud, which reduces the required bandwidth.

In Walter et al. (2021), we have shown that we can achieve a high degree of completeness and geometric quality in both levels of this approach by multiple data collection and subsequent averaging. With tenfold acquisition, a completeness of 93.3% was achieved. The geometric deviation between the centres of the crowd-based collected minimum bounding boxes and the reference was less than 1 m for more than 95% of the vehicles.

The number of how often the data is collected influences the quality of the results. However, a higher number of redundancies does not necessarily increase the quality. For example, if an object is collected only twice (with different quality characteristics), the quality of the integrated object would be poorer than the quality of the original object with the higher quality. If, however, an object is collected by n crowdworkers, we expect that for a specific number $n > n_0$ the quality of the integrated object will be higher than the quality of the individual objects or at least very close to the quality of the best object.

Higher quality can be achieved through larger number of redundancies, but this also leads to higher cost. If the cost is to be reduced, the number of redundancies can be reduced, but this has a negative effect on the quality. In the end, the optimal number of redundancies is a compromise between data quality versus cost. In this paper, we examine these numbers in more detail to gain a better understanding of the relationship between the number of repeated acquisitions and the results obtained by integrating them.

We investigate the impact of the crowd size on the data quality for the first step of the approach described in (Walter et al. 2021), where crowdworkers are asked to identify vehicles using line segments reaching from the front to the back of the vehicle or vice versa. An example is shown in Fig. 1.

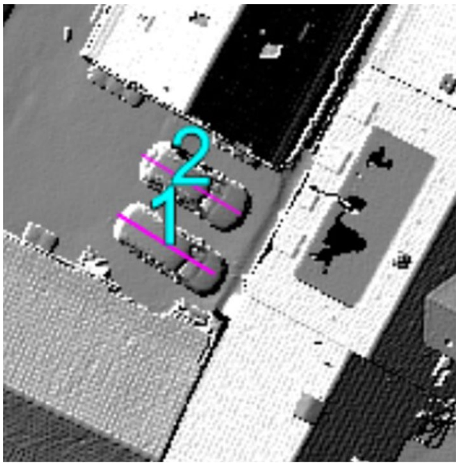
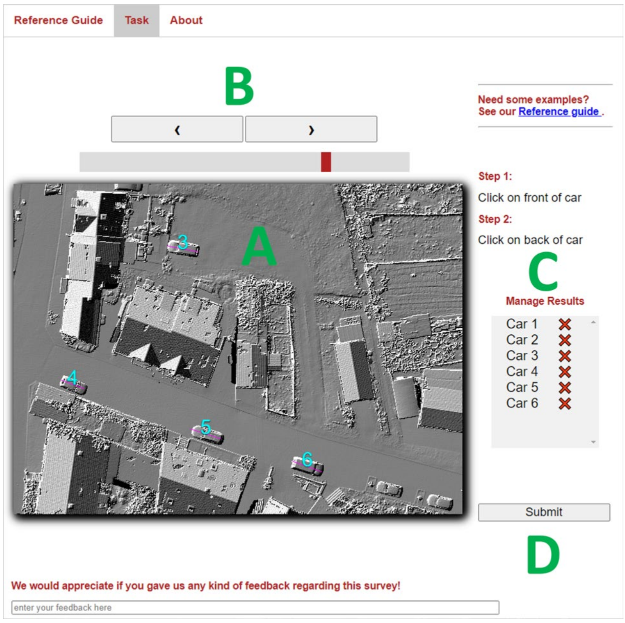


Fig. 1 Example of two vehicles collected by a crowdworker in a 2D rasterized shading



Need some examples?
See our Reference guide .

Step 1:
Click on front of car

Step 2:
Click on back of car

We would appreciate if you gave us any kind of feedback regarding this survey!

enter your feedback here

Fig. 2 GUI for the crowd-based collection of vehicles

The graphical user interface is shown in Fig. 2. It is divided into four areas:

- Data view (A): displays the 2D rasterized shading and the already collected vehicles.
- Data control (B): strips can be moved left and right using two control buttons.
- Management (C): all collected vehicles are displayed in a list. Vehicles that have been collected incorrectly can be deleted. An already collected vehicle can be activated and edited by clicking on the corresponding entry in the list.

- Submit button (D): the crowd job can be ended by clicking on this button.

3 Datasets

As test area, we rely on the newly introduced Hessigheim LiDAR dataset (H3D) (Koelle et al. 2021c). Hessigheim is located in the south of Germany. The point cloud was collected with a RIEGL VUX-1LR LiDAR sensor combined with two Sony Alpha 6000 oblique cameras using the RIEGL RiCopter octocopter. The mean laser pulse density is about 800 points/m^2. The ranging accuracy is 10 mm (Riegl 2018).

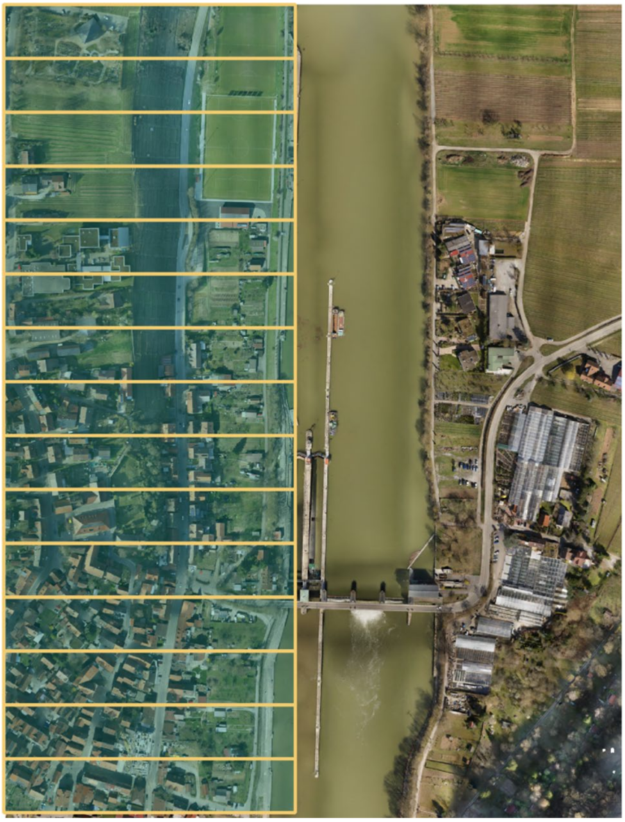


Fig. 3 The Hessigheim dataset subdivided into 15 strips

This area was subdivided in 15 east–west-oriented strips of 250 × 50 m (see Fig. 3), from which we selected the 6 southernmost strips. We did not use the entire data set, but only the part that contains an urban area where vehicles are located. The rest of the data mainly contains agricultural land with very sparse occurrence of vehicles. An example of one strip is shown in Fig. 4. To avoid artefacts of truncated vehicles at the strip boundaries, a second set of overlapping strips (overlap = 50%) is used. Therefore, we have a total of 11 strips.

We use data from two epochs to be able to verify our results on two different dataset and different crowdworkers: Dataset 1 was collected in November 2018 and dataset 2 in March 2019. Figure 5 shows data of both epochs from a part of the study area. The two data sets are very similar due to the use of the same sensor and similar recording conditions, so that similar results are to be expected. If the data sets were more different (e.g., a different sensor, or a different location), it would be expected that the results would also be more different. The main purpose of using a second dataset is to show that running a campaign with a different group of crowdworkers leads to comparable results. The groups of crowdworkers are not completely different but still very distinct. An examination of the worker IDs shows that in the second campaign, 78.2% of the crowdjobs were performed by new crowdworkers who had not already participated in the first campaign.

Whenever mapping 3D data (i.e., point clouds) to 2D raster images, we inevitably lose information due to dimensionality reduction. Especially if an object of interest (i.e., vehicle) is occluded by other structures such as vegetation, simply using the maximum z-value per grid cell is counterproductive, since airborne laser scanning can penetrate such structures. Therefore, we derive the data to be annotated by crowdworkers as follows.

First, we derive a Digital Terrain Model based on filtered ground points, so that for every 3D point an individual height above ground can be calculated. Afterwards we derive two height models: one using the maximum z-value per grid cell (see Fig. 6a) and one only using LiDAR points with a maximum height over ground of 3 m (see Fig. 6b). Due to the elimination of all points that are 3 m above ground, there may be raster cells which contain no 3D point at all (e.g.,

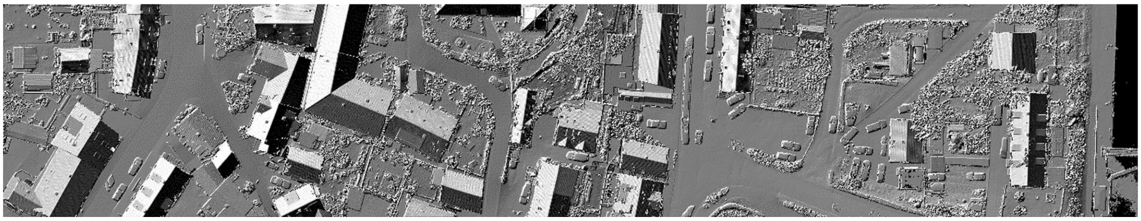


Fig. 4 2D shading of one strip of our test data (size 250 m × 50 m)

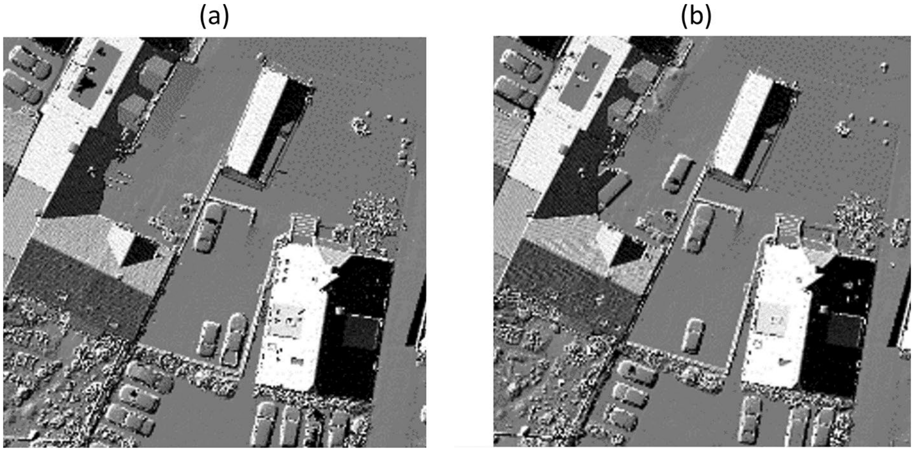


Fig. 5 Comparison of the data from epoch 1 (November 2018) and epoch 2 (March 2019)

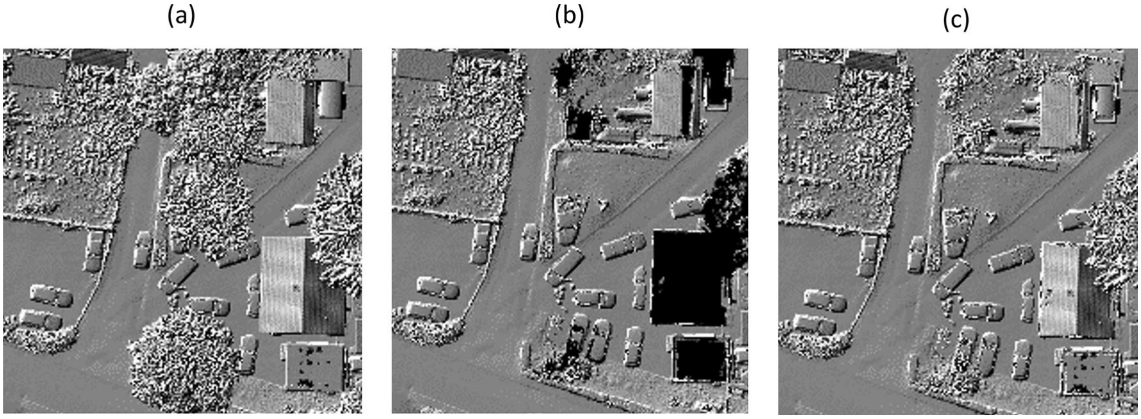


Fig. 6 Calculation of the 2D rasterized shadings (grid size 10 cm) from maximum z-value of **a** all points within each grid cell and **b** all points with a maximum height of 3 m above ground. Data gaps in (**b**) are filled with the maximum z-value inside each cell from **a**) (see **c**))

at buildings). In these cases, we use the maximum z-value of the points inside the corresponding cells to avoid holes in the shading (see Fig. 6c). Following this strategy, we can efficiently filter vegetation points. Please note that vegetation can also be filtered by only considering last echo points, but due to multi-path effects on the vehicle's surface and due to the moving nature of some vehicles during data capturing, the resulting areas in the height model would be extremely noisy and unsuitable for presenting to crowdworkers. All calculations have been carried out using the *Opals* software (Pfeifer et al. 2014).

4 Crowd-Based Data Collection

For each of the two datasets, we created one crowd campaign in which each strip was collected 100 times. Since we have 11 strips (see Sect. 3), this results in 1100 crowd jobs per campaign. We paid \$0.10 per crowdjob, resulting in a total cost of \$110 per campaign. The average time to collect all vehicles in a strip was 335.6 s in the first campaign and 311.3 s in the second campaign. This calculates to an average hourly wage of \$1.07 for campaign 1 and \$1.15 for campaign 2. The time required to execute all 1100 crowdjobs was approximately 10 h for both campaigns.

Table 1 shows the distribution of origin of the crowdworkers sorted by top 10 countries. The top three countries are Bangladesh, India, and Venezuela.

To evaluate the quality of the data, we have very carefully collected reference data ourselves. We developed a MATLAB tool with additional zoom functionalities to collect the vehicles as accurately as possible. Table 2 shows an evaluation of the reference data. The 11 strips of the first campaign contain 309 vehicles and of the second campaign 274 vehicles. Due to the overlap of the strips, this results in 172 unique vehicles in the first campaign and 153 unique vehicles in the second campaign. A total of 26,752 vehicles were collected in all 1100 crowdjobs together in campaign 1 and 24,303 vehicles in campaign 2. Compared to the reference, this results in an average number of vehicles collected per crowdworker relative to the number of vehicles in the reference of (26,752/1100)/(309/11) = 0.868 (Campaign 2:

Table 1 Distribution of origin of the crowdworkers sorted by top 10 countries

Country	Percentage Crowdworker (%)
Bangladesh	50.1
India	11.0
Venezuela	7.6
Philippines	3.5
Pakistan	2.7
Serbia	2.3
Brazil	1.5
Colombia	1.5
Kenya	1.2
Nepal	1.0

Table 2 Comparison of vehicles in reference and number of collected vehicles

Campaign	Vehicles in reference	Unique vehicles in reference	Total number of collected vehicles	Average completeness (%)
1	309	172	26,752	86.8
2	274	153	24,303	88.7

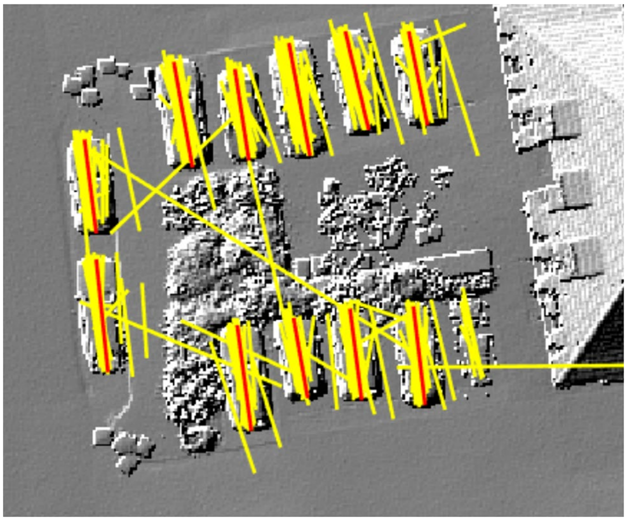


Fig. 7 Example of crowd-based data collection (yellow: crowd-based data collection, red: reference)

(24,303/1100)/(274/11) = 0.887). This means that if only one crowdworker collects the data, we would have to expect an average completeness of less than 90 percent.

Figure 7 shows a typical example of data collection from 100 crowdworkers. It can be seen that although there are individual outliers, the majority of the collected lines can be clearly assigned to a vehicle.

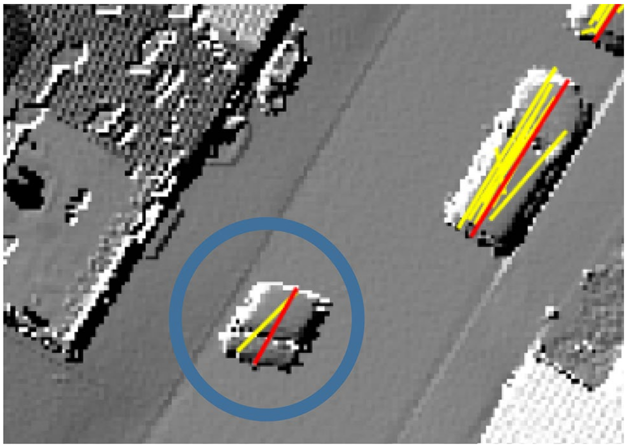


Fig. 8 Example of a distorted vehicle (yellow: crowd-based data collection, red: reference)

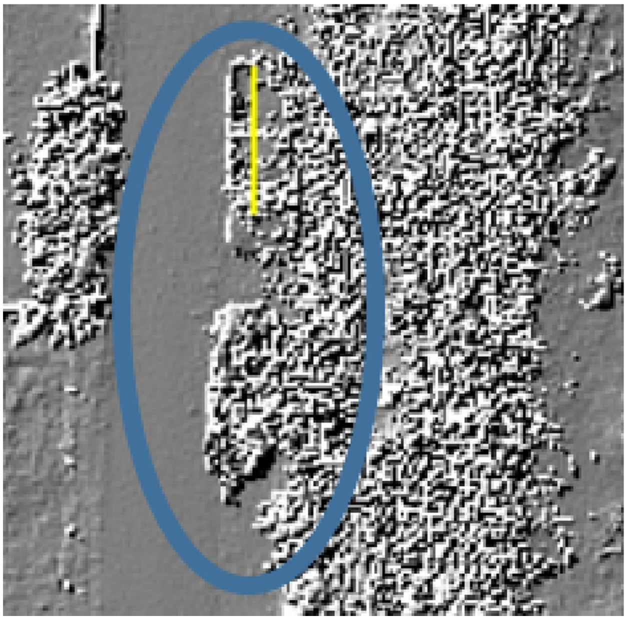


Fig. 9 Example of two vehicles hidden by very low vegetation

Figure 8 shows a problem that occurred sporadically in the data. Vehicles moving while being scanned by the laser scanner may be distorted. In the figure, a vehicle can be seen that is compressed in length (due to opposing movement of the car and the scanner platform). Only one crowdworker collected this vehicle. Dataset 1 contains 11 and dataset 2 15 distorted vehicles. In these cases, it is difficult to decide whether to include the vehicles in the reference, as well as difficult for the crowdworkers to interpret the data correctly.

Figure 9 shows an area where two vehicles are hidden by vegetation. This problem occurs whenever there is dense vegetation close to the ground (within our considered range of 3 m above ground level, see Sect. 3) such as low-hanging branches. The vehicles are very difficult to detect in the data. Only one crowdworker detected one of them. The data in this area are so difficult to interpret that it is very hard to decide whether there are actually vehicles or other objects (at least by only relying on 2D shadings). This problem occurs very rarely, but it shows that it is very difficult even for an expert to achieve a data collection accuracy of 100 percent.

5 Evaluation the Wisdom of the Crowd

5.1 Clustering and Integration

We cluster the centre points of all collected vehicles with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996). DBSCAN requires two parameters: ε specifies the maximum distance between two points in a cluster, and *MinPts* specifies the minimum number of points in each cluster. The value of epsilon results from our application considering the minimum distance of the centres of two neighbouring cars in parking lots. Based on this domain knowledge and visual inspections, we use $\varepsilon = 2$ m. With smaller ε, it can happen that one single vehicle is divided into several clusters (see Fig. 10). With larger ε, it can happen that several vehicles parked next to each other merge into one cluster, so that a cluster no longer represents only one vehicle, but several vehicles (see Fig. 11).

The parameter *MinPts* can be used to control whether vehicles that are difficult to detect appear in the result or not. A vehicle must be detected by at least *MinPts* crowdworkers. Thus, with high *MinPts*, the result would contain only those vehicles that can be detected very reliably. If, on the other hand, vehicles that are difficult to detect should also be included in the result, *MinPts* should be set low. At the same time, *MinPts* must not become too small, since otherwise falsely collected data would be accepted as vehicles. The appropriate choice of *MinPts* depends on n. The larger n, the larger *MinPts* should be chosen. A detailed examination of this relationship is given in Sect. 5.2.

Figure 12a shows the 2D coordinates of the centres of the collected vehicles of a part of our study area on the example $n = 10$. The outliers detected with DBSCAN are marked with red colour. The remaining points are subdivided into clusters (Fig. 12b). The final positions of the clusters are calculated by averaging the x- and y-positions of the collected vehicles within each cluster (Fig. 12c).

In a refinement step, we focus on the individual lines within each cluster and iteratively eliminate all lines that deviate strongly from the averaged value. We perform this step to improve the geometric accuracy of the integrated results. This step has no effect on Precision, Recall and F1 score because it is computed after the cluster are calculated. The elimination of the strongly deviating lines mainly improves the accuracy of the length and has only little effect on position accuracy. A similar approach was used in (Walter et al. 2020) for eliminating outliers from Minimum Bounding Cylinders, which represent trees collected by crowdworkers. We proceed as follows:

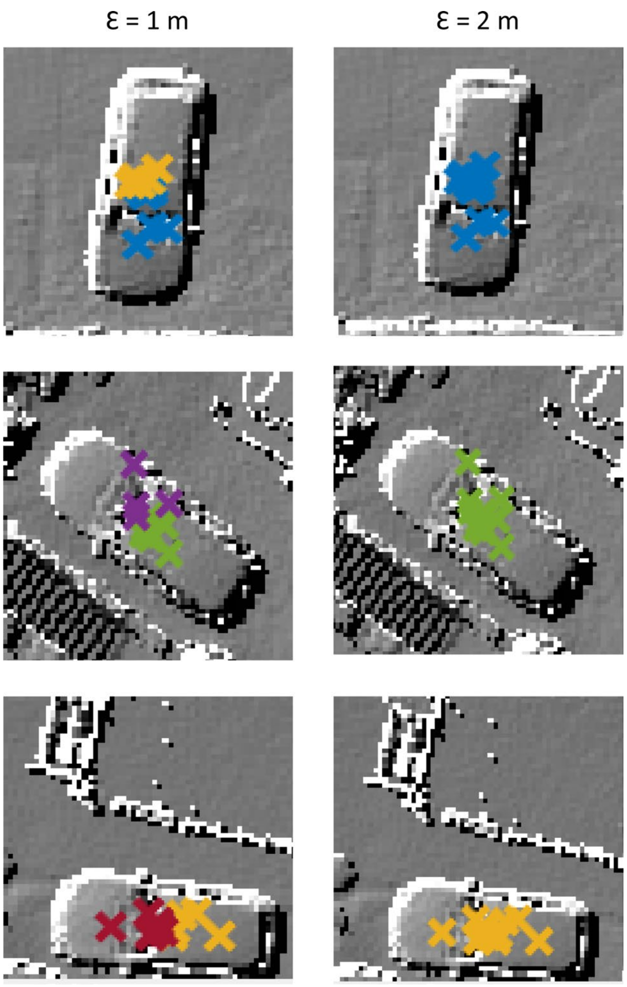


Fig. 10 Examples for cluster calculation with $\varepsilon = 1$ m and $\varepsilon = 2$ m: In the case of $\varepsilon = 1$ m, it can happen that individual vehicles are represented by several clusters; each cross represents the centre point of one crowd-based data collection, different clusters are represented in different colours

FOR all clusters:

- Compute two clusters with k-means ($k = 2$) (i.e., derive one cluster for the start and one cluster for the end point, no matter which one is actually start and end point)
- Implicitly calculate an integrated line by averaging the points within the two clusters
- Compute orthogonal distances of all points (i.e., start and end points) to integrated line
- Omit all lines with a length difference to the integrated line greater 2 m

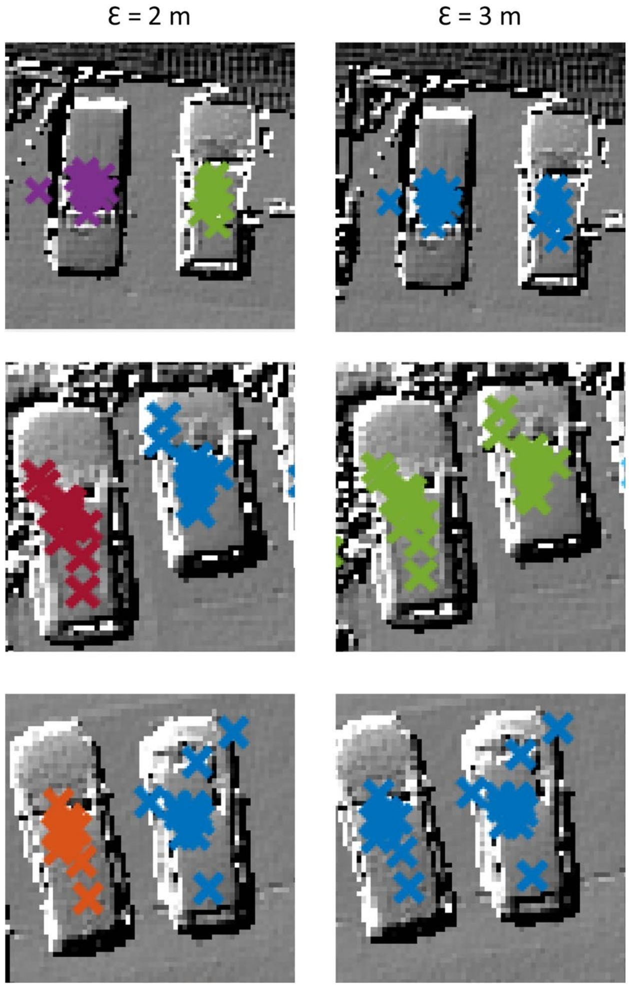


Fig. 11 Examples for cluster calculation with $\varepsilon=2$ m and $\varepsilon=3$ m: In the case of $\varepsilon=3$ m, it can happen that several vehicles parked next to each other merge into one cluster; each cross represents the centre point of one crowd-based data collection, different clusters are represented in different colours

- Repeat until convergence

Figure 13 shows the elimination of strongly deviating lines within a cluster by means of an example. It can be seen that this step eliminates especially short lines that do not go from the front to the back of the vehicle or vice versa, but rather from one side to the other.

5.2 Quality Evaluation

To describe the completeness and correctness of the integrated multiple collected data, we evaluate:

- true positives (TP),
- false negatives (FN),
- false positives (FP),

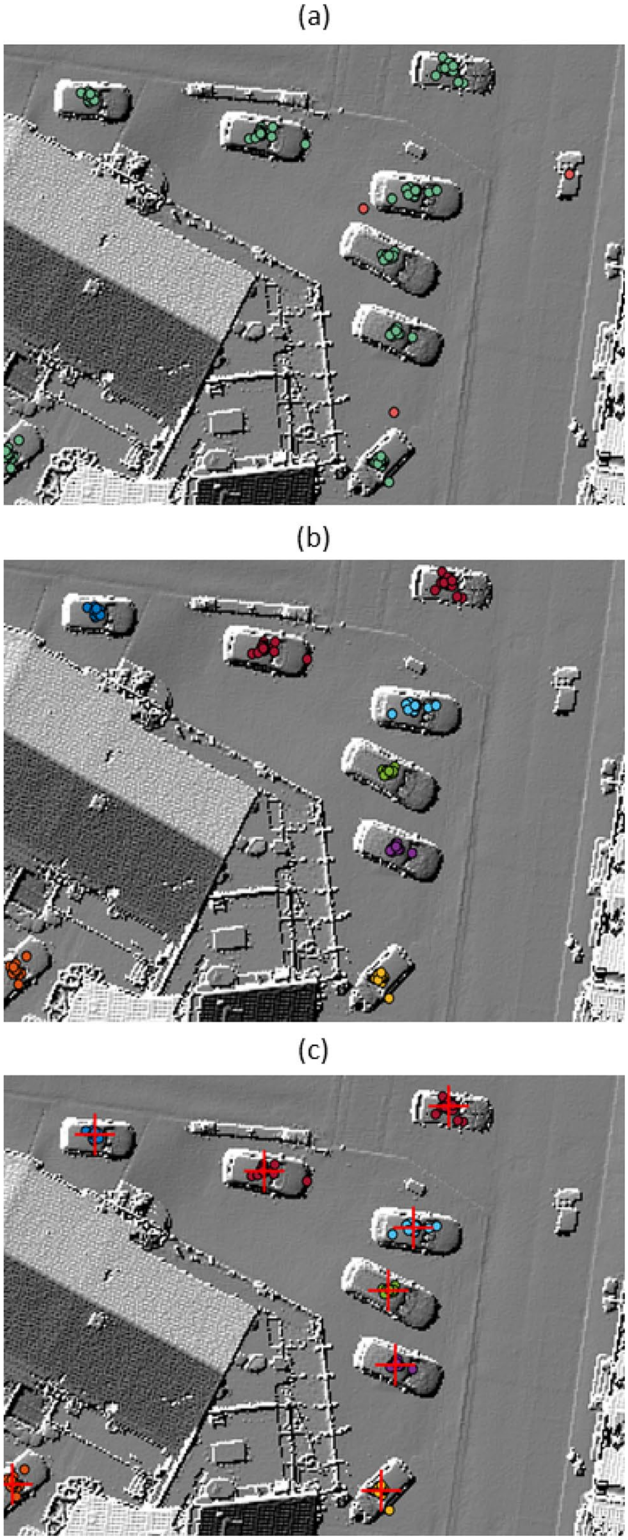


Fig. 12 Clustering of the multiple collected vehicles: **a** Collected centres (*x*- and *y*-positions) of the vehicles of a part of the test area. Outlier detected with DBSCAN are marked with red colour. **b** Remaining centres (*x*- and *y*-positions) of the vehicles are subdivided with DBSCAN into clusters that are marked with different colours. **c** Final result of integration

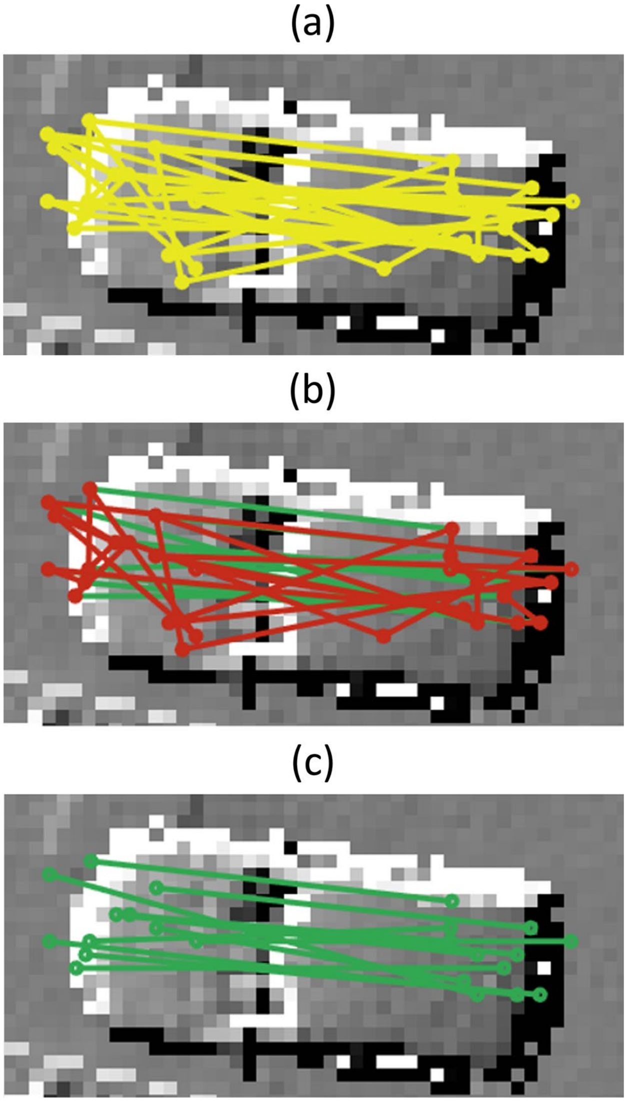


Fig. 13 Example of elimination of strongly deviating lines within one cluster: **a** all lines of the cluster (yellow), **b** lines that deviate strongly (red) and remaining lines (green), **c** result (green)

- precision $= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}$,
- recall $= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$,
- $F1$ score $= 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$, and two geometric features:
- position difference,
- length difference.

A precision value of 1.0 means that every vehicle collected by the crowd is also in the reference (but says nothing about whether all vehicles were collected). A recall value of 1.0 means that all vehicles in the reference were collected by the crowd (but says nothing about how many false vehicles were collected). The $F1$ score is the harmonic mean of precision and recall. The highest value of $F1$ score is 1, representing perfect precision and recall. The lowest value is 0, when either precision or recall is zero.

We divide the quality evaluation into two parts depending on the number n of multiple acquisitions: in the first part we discuss $n \in [1..16]$ and in the second part $n = 20, 40, 60, 80, 100$.

For the first part, we collected the test area 16 times. For each $n \in [1..16]$, we computed all possible $16!/((16-n)!*n!)$ combinations and calculated the mean of each quality measure. In this way, we avoid that the results depend on the selection of individual observations. This combinatorial approach provides more meaningful values than performing the test multiple times with an increasing number of crowdworkers (Test 1: one crowdworker, test 2: two crowdworkers, …). This is especially true for small n. Consider for example $n = 2$: if we repeat the test multiple times with an increasing number of crowdworkers, the result for $n = 2$ would be highly dependent on the two crowdworkers in that test. If one or even both crowdworkers deliver exceptionally good or poor results, this would lead to non-representative results. In the combinatorial approach with 16 crowdworkers, for $n = 2$: $16!/((16–2)!*2!) = 120$, different combinations of crowdworkers are calculated and afterwards their results are averaged. Each individual crowdworker is involved only in 15 of the 120 combination. The combinatorial approach avoids outliers even if some crowdworkers deliver exceptionally good or bad results.

In the second part of the quality evaluation, we did not calculate all combinations, but selected the first n out of 100 observations for $n = 20, 40, 60, 80, 100$, since a combinatorial evaluation is no longer numerically feasible in a reasonable time. Therefore, the results of the second part depend on the selection of observations, but since we now have larger n, this effect is alleviated.

In the following, we discuss the evaluation of dataset 1 (see Sect. 3). The evaluation of dataset 2 led to very similar results. Therefore, we have moved the figures of the evaluation of dataset 2 to the Appendix for better readability of this paper.

5.2.1 Evaluation of TP, FN, FP, Precision, Recall and F1 score

Figure 14 shows the evaluation of TP, FN, FP, Precision, Recall, and F1 score of dataset 1 for $n \in [1..16]$. The quality metrics depend (i) on n and (ii) on the definition of *MinPts*. The larger n, the higher TP and the lower FN. At the same time, it can be seen that FP increases with increasing n, which leads to declining precision and increasing recall. The reason for this is that the more crowdworkers are in the group, the more vehicles that are difficult to recognize will be detected (recall increases). At the same time, vehicles are now also detected in places where there are none at all (precision declines).

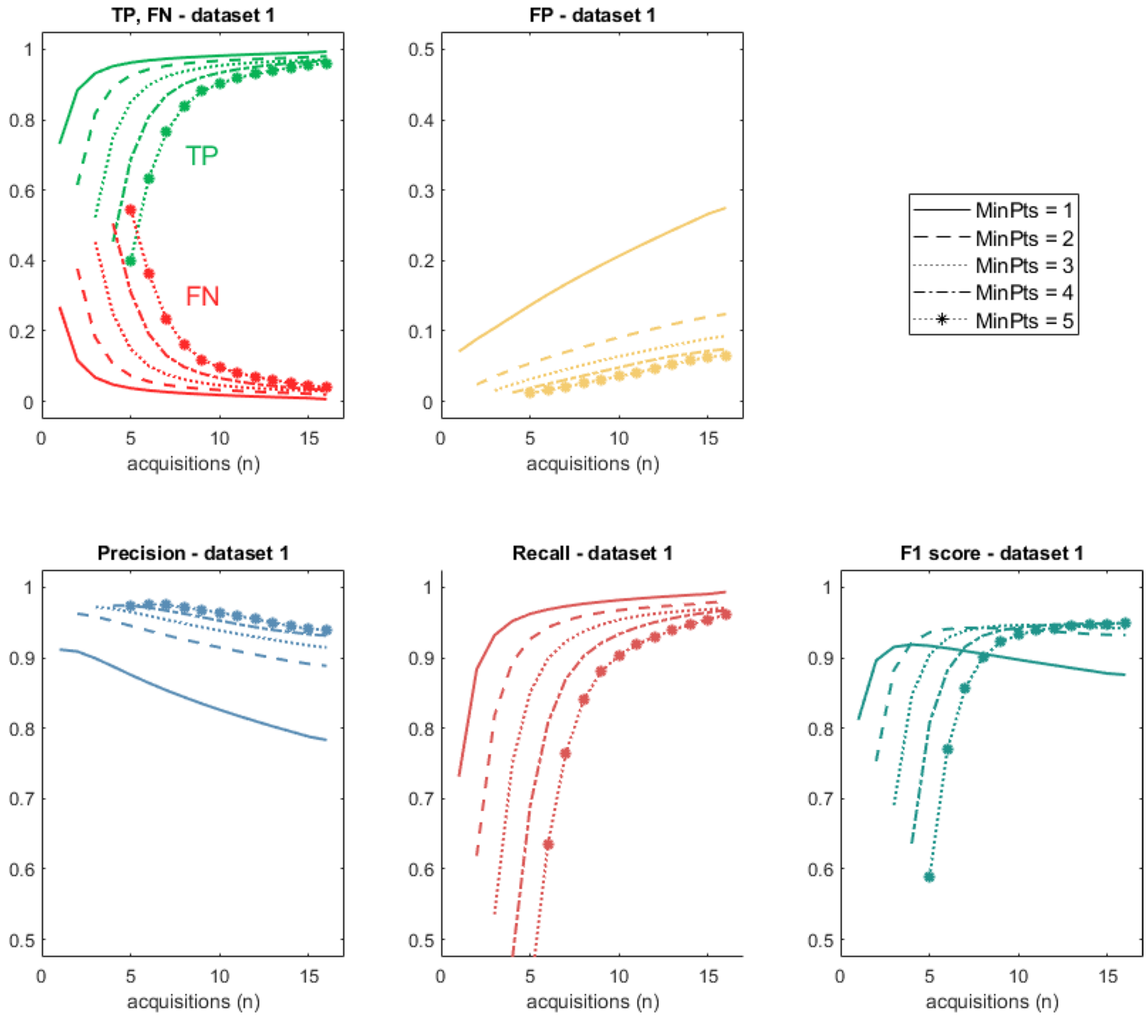


Fig. 14 Evaluation of TP, FN, FP, Precision, Recall and F1 score of dataset 1 for $n \in [1..16]$

This behaviour can be controlled by *MinPts*. If the focus is on high precision, *MinPts* should be chosen large. This leads to the result that only such vehicles are accepted, which were recognized by enough crowdworkers. If the focus is on high recall, *MinPts* should be chosen smaller, so that vehicles that are difficult to recognize are also included in the result.

Figure 15 shows the evaluation of TP, FN, FP, Precision, Recall, and F1 score of dataset 1 for $n = 20, 40, 60, 80, 100$. It can be seen that for large n TP and FP no longer change significantly, but FP increases, resulting in decreasing precision. Compared to $n \in [1..16]$, it can be seen that the F1 score does not improve anymore, but actually decreases. Therefore, with regard to completeness and correctness, it is not reasonable to collect the data more than 20 times.

5.2.2 Evaluation of the Geometric Quality

To describe the geometric quality, we evaluate the position difference and the length difference between integrated lines and reference. The position difference is calculated by the distance between the centre points of the lines. Based on our evaluation in the previous section, we have defined *MinPts* so that F1 score is maximized (compare Fig. 14), resulting in a section-by-section definition:

- $MinPts = 1$ for $n < = 4$
- $MinPts = 2$ for $4 < n < = 8$
- $MinPts = 3$ for $8 < n < = 12$
- $MinPts = 4$ for $n > 12$

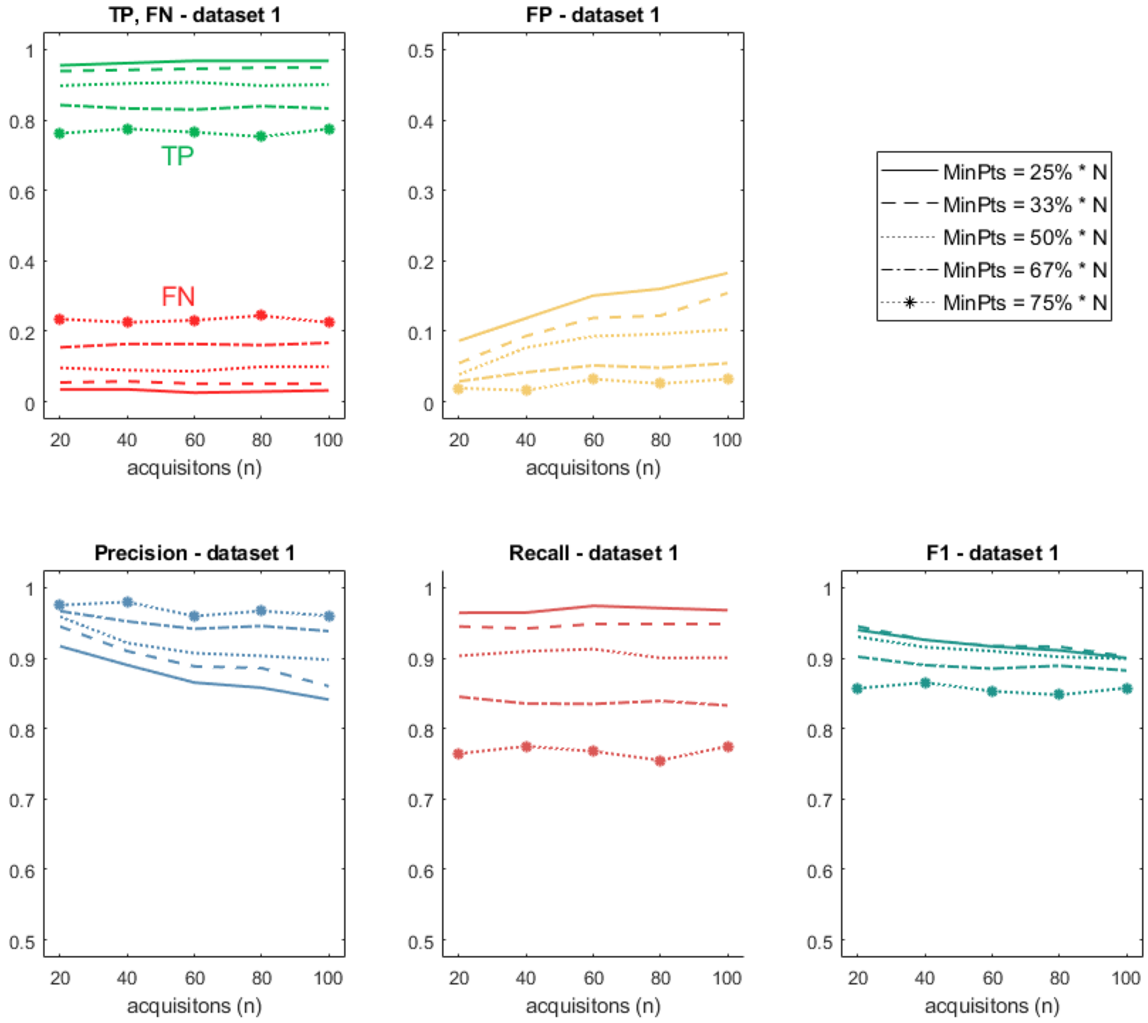


Fig. 15 Evaluation of TP, FN, FP, Precision, Recall and F1 score of dataset 1 for $n = 20, 40, 60, 80, 100$

Figure 16 shows the evaluation of the position difference for $n \in [1..16]$ of dataset 1. When the group consists of only one crowdworker, the mean value of the position difference is on average 0.275 m with a standard deviation of 0.182 m. As n increases, the mean and the standard deviation improve continuously. With 16 crowdworkers in the group, we achieve a position difference of 0.175 m and a standard deviation of 0.106 m.

Figure 17 shows the evaluation of the length difference for $n \in [1..16]$ for dataset 1. When the group consists of only one crowdworker, the mean value of the length difference is on average 0.89 m with a standard deviation of 1.02 m. There is an improvement of the mean value only up to $n = 5$. After that, the mean value deteriorates and then stagnates at around 0.82 m. However, it can be seen that the standard deviation decreases monotonically. This eliminates coarse outliers as n increases.

It is noticeable that the form of the distributions of the position difference and the length difference are different. For small n, there are few position differences that are close to zero, while this is the case for the length differences. The reason for this is that it is difficult to define the exact centre of a line by specifying the beginning and the end of the line. However, entering the exact line length is easier because also lines that are displaced from the reference can still have exactly the same length as the reference. This effect disappears with higher n, since a smoothing effect occurs due to the averaging.

Figure 18 shows the evaluation of the position difference for $n = 20, 40, 60, 80, 100$ of dataset 1. We have defined $MinPts = 0.25 * n$ so that F1 score is maximized (compare Fig. 15). It can be seen that with increasing n, the mean value of the position difference continues to improve, whereas the standard deviation remains about the same.

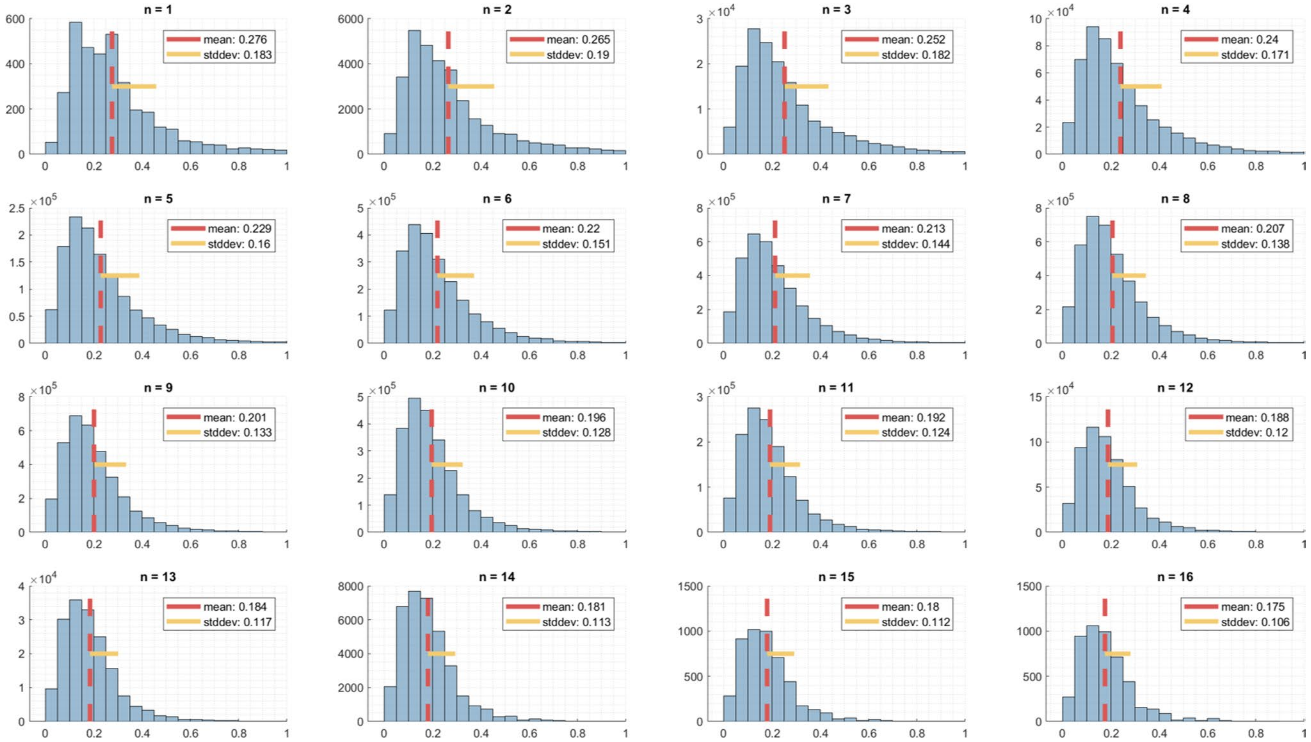


Fig. 16 Evaluation of the position difference of dataset 1 for *n*∈[1..16]. The x-axis shows the average position difference in meter and y-axis the number of combinations (the total number of possible combinations is 16!/((16–*n*)!*n!)

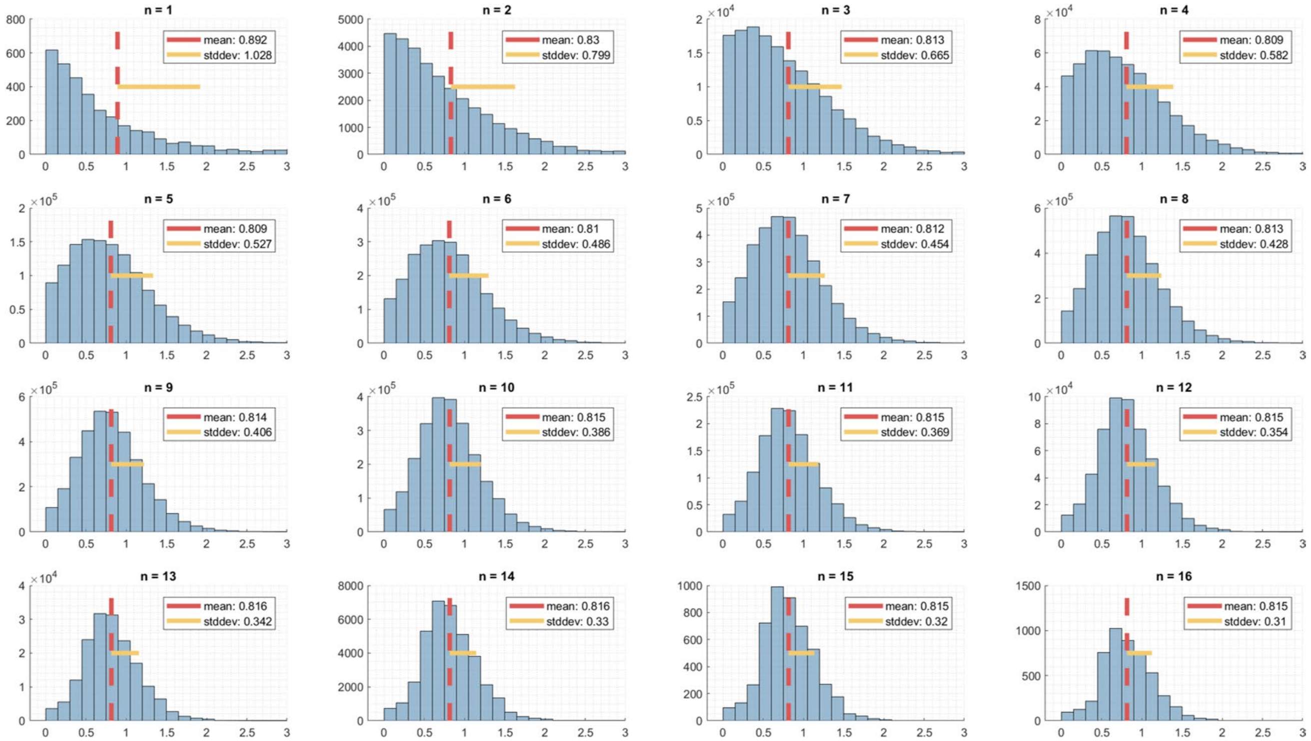


Fig. 17 Evaluation of the length difference of dataset 1 for *n*∈[1..16]. The x-axis shows the average position difference in meter and y-axis the number of combinations (the total number of possible combinations is 16!/((16–*n*)!*n!)

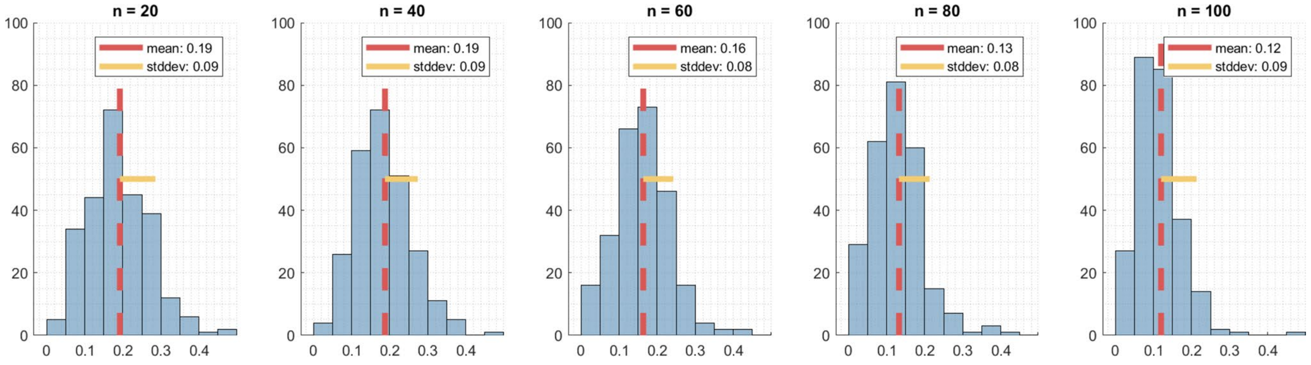


Fig. 18 Evaluation of position difference for $n=20$, 40, 60, 80, 100 of data set 1. The *x*-axis shows the average length difference in meters and the *y*-axis the number of measured values

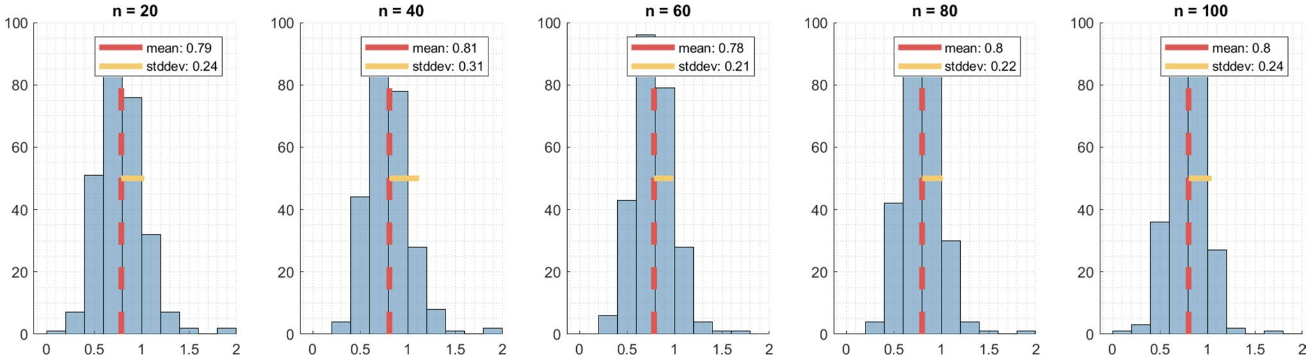


Fig. 19 Evaluation of length difference for $n=20$, 40, 60, 80, 100 of dataset 1. The *x*-axis shows the average length difference in meters and the *y*-axis the number of measured values

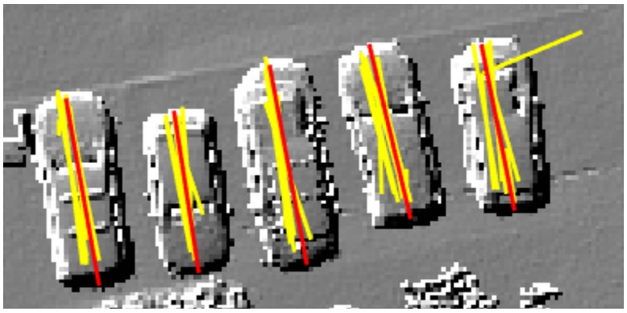


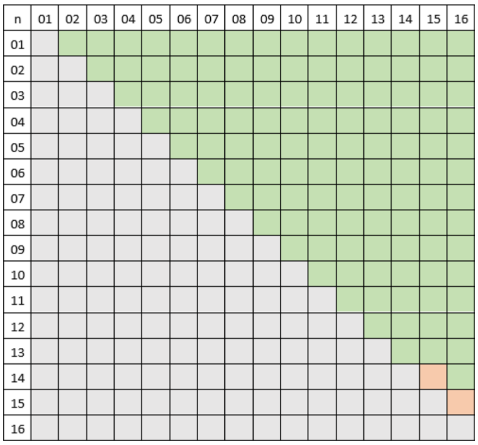
Fig. 20 Typical example of crowd acquisitions: many of the crowd-based acquisitions are shorter than the reference (yellow: 16-fold crowd-based data collection, red: reference)

Figure 19 shows the evaluation of the length difference for $n=20$, 40, 60, 80, 100 of dataset 1. The mean value of the length difference remains approximately constant. Similarly, the standard deviation hardly changes. The reason that the mean length difference does not improve further is that some of the crowdworkers systematically collect the data too short. Figure 20 shows a typical acquisition example. It can be seen that the majority of the lines do not span from edge to edge of the vehicles but are shorter than the reference. A mean value close to zero is therefore not to be expected even with a large n.

5.2.3 Significance Test

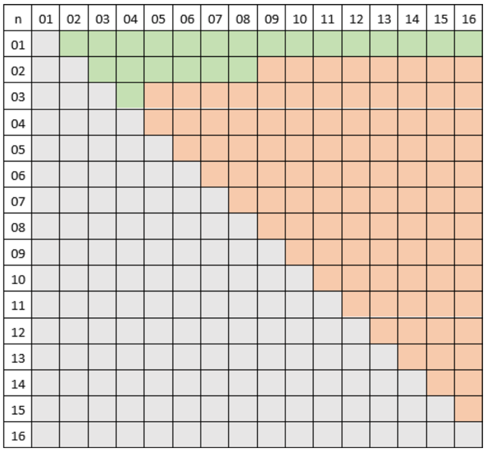
To evaluate whether increasing the number of crowdworkers improves our quality measures in a statistically significant sense, we applied a one-tailed Student's t test, which is formulated to answer the question: "Does the result of more crowdworkers yield to a significant decrease of accuracy errors?" Hence, we compare each distribution against all others and consider the 95% confidence interval as critical threshold to accept or refuse the hypothesis of a significant improvement. The results confirm the observations already made.

Figure 21 shows the evaluation of the position difference of dataset 1 for $n \in [1..16]$. The position error becomes significantly smaller with increasing n. From $n=15$, we run



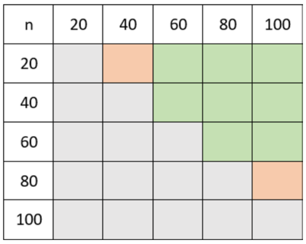
n	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
01	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
02	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
03	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
04	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
05	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
06	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
07	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
08	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
09	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩
10	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩
11	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩	🟩
12	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩	🟩
13	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	🟩	🟩
14	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟩
15	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥
16	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

Fig. 21 One-tailed Student's *t* test for the position difference of dataset 1 for $n \in [1..16]$, green = significant improvement, red = no significant improvement, grey = not evaluated



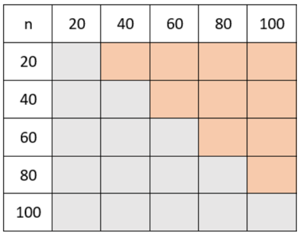
n	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16
01	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
02	⬜	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
03	⬜	⬜	⬜	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
04	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
05	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
06	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
07	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
08	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
09	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥	🟥
10	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥	🟥
11	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥	🟥
12	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥	🟥
13	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥	🟥
14	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	🟥
15	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥
16	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

Fig. 22 One-tailed Student's *t* test for the length difference of dataset 1 for $n \in [1..16]$, green = significant improvement, red = no significant improvement, grey = not evaluated



n	20	40	60	80	100
20	⬜	🟥	🟩	🟩	🟩
40	⬜	⬜	🟩	🟩	🟩
60	⬜	⬜	⬜	🟩	🟩
80	⬜	⬜	⬜	⬜	🟥
100	⬜	⬜	⬜	⬜	⬜

Fig. 23 One-tailed Student's *t* test for the position difference of dataset 1 for $n = 20$, 40, 60, 80, 100], green = significant improvement, red = no significant improvement, grey = not evaluated



n	20	40	60	80	100
20	⬜	🟥	🟥	🟥	🟥
40	⬜	⬜	🟥	🟥	🟥
60	⬜	⬜	⬜	🟥	🟥
80	⬜	⬜	⬜	⬜	🟥
100	⬜	⬜	⬜	⬜	⬜

Fig. 24 One-tailed Student's *t* test for the position difference of dataset 1 for $n = 20$, 40, 60, 80, 100], green = significant improvement, red = no significant improvement, grey = not evaluated

into saturation. The smallest position error is obtained at $n = 16$.

Figure 22 shows the evaluation of the length difference of dataset 1 for $n \in [1..16]$. A significant improvement of the length error occurs only up to $n = 3$. After that, an increase is no longer worthwhile, since the length error does not improve significantly. However, the standard deviation decreases with increasing *n* (see Fig. 17). This eliminates coarse outliers as *n* increases.

Figure 23 shows the evaluation of the position difference of dataset 1 for $n = 20$, 40, 60, 80, 100. Between $n = 20$ and $n = 40$, there is no significant improvement. Thereafter, an increase of *n* up to $n = 80$ leads to significantly better results.

Figure 24 shows the evaluation of the length difference of dataset 1 for $n = 20$, 40, 60, 80, 100. As discussed earlier, increasing *n* in the range between 20 and 100 does not result in smaller length errors.

6 Conclusion

This research proves the effectiveness of the idea of the wisdom of the crowd once more and again demonstrates that it is also applicable for very special tasks such as the collection of spatial objects from remote sensing data. We tested our approach on two datasets depicting the same area but taken at different times. The results of the quality evaluations of both data sets are very similar.

We have shown that increasing the crowd size improves the quality of the data. This effect occurs especially with small *n*. From about $n = 20$, the improvements are only very small and rather not worthwhile in relation to the costs since the costs grow linearly with *n*. For *n* greater than 20, even a deterioration of the F1 score can be observed, which is mainly explained by the increase of FP. To prevent the increase of FP, *MinPts* would have to be chosen larger, which, however, would again have a negative effect on TP and FN. Hence, *MinPts* is an important parameter to control the quality of the results. If data are to be collected where correctness is more important than completeness, *MinPts* should be set rather high. If the focus is on completeness and less on correctness, a low *MinPts* is the better choice.

The position difference is the only measure showing an improvement up to $n = 100$. This is mainly due to the fact that the geometric accuracy can only be calculated for TP and thus an increase of FP has no negative effect to this measure. FP cannot degrade the results because they cannot be matched to a reference and therefore no position difference and no length difference can be measured. It is likely that with even larger n, even more accurate positions can be expected, although improvements would be very slow. For the length difference, no improvement could be observed up from $n = 20$, which is related to the fact that some of the crowdworkers systematically collect the vehicles too short compared to the reference data. This problem could be solved, for example, by training the crowdworkers before the actual task.

Although the collection of the reference data was carried out very carefully, there are some areas that were very difficult to interpret also impacting our quality measures. In these cases, differences between crowd-based collection and reference must always be expected. This problem is of fundamental nature and cannot be solved by increasing n either. Even if we would replace the crowdworkers with experts, this problem would not be completely solved.

We currently see the main application for spatial data collection by paid crowdworkers in providing training data for deep learning methods. The approach presented uses ultra-high-resolution UAV data, which are currently not widely available. Lower-resolution airborne laser data on the other hand are widely available, but vehicle detection in such data sets is limited. However, it is conceivable to apply our approach to image data if we use orthophotos, which are available over large areas in high resolution, instead of the 2D shadings. Another possible data source would be terrestrial laser scanners such as vehicle-mounted laser scanners.

6.1 Limitations

We tested our method on two different campaigns to show that we achieve comparable results with different groups of crowdworkers. Both campaigns were conducted on mircoWorkers. It is conceivable that different results would be obtained if we switched to a different crowdsourcing marketplace due to a deviating pool of workers with different abilities (for instance, the majority of workers on Amazon Mechanical Turk are US citizen, while the ones of microWorkers are mostly situated in Asia).

The two datasets we used are very similar due to the use of the same sensor and similar flight conditions. If the data were more different, it is also to be expected that the results would be more different.

The parameters determined in this project are valid for the task discussed in conjunction with the used data and cannot be directly applied to other tasks or other data. Thus, it is to be expected that for tasks that are more difficult to solve or for data that is more difficult to interpret, more observations are necessary to increase the quality. Similarly, there may be other systematic effects than in our investigation.

We therefore see the approach described here as a guide to how the Wisdom of the Crowd principle can be evaluated qualitatively and quantitatively. The research presented here can be used as a blueprint for other applications.

6.2 Future Work

For further quality improvement, the integration of the individual collected data could be optimized. Currently we use simple averaging as aggregation rule. This could be extended using weights for each individual worker. The weights could be derived from the score that is available for each crowdworker on the crowdsourcing marketplace. Furthermore, the wisdom of the crowd principle is only one way to improve the quality of data from paid crowdworkers. Many other approaches from the fields of psychology and work science are described in the literature. These approaches can be combined with the method described in this paper to achieve further quality improvement.

7 Appendix A

See Figs. A1, A2, A3, A4, A5, A6.

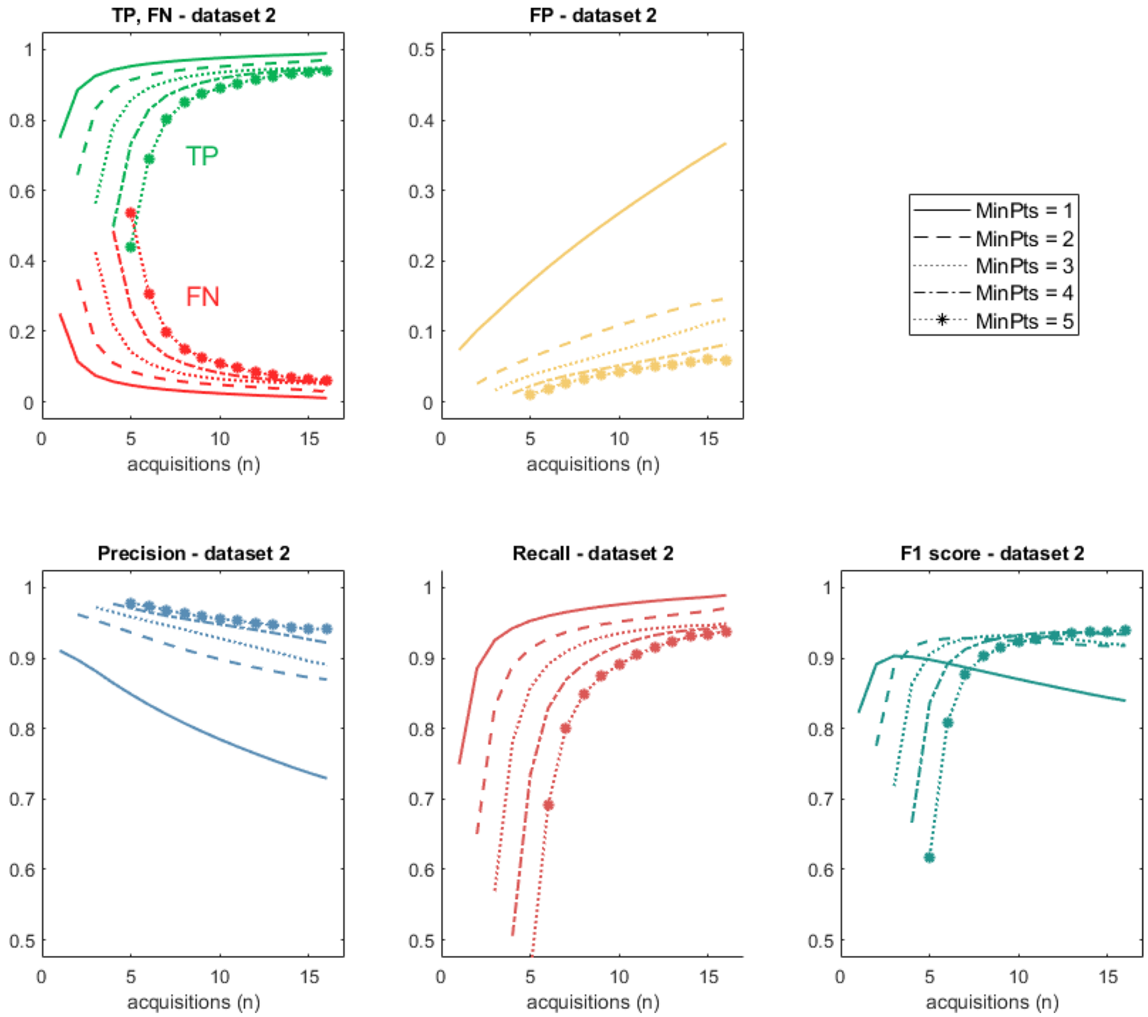


Fig. A1 Evaluation of TP, FN, FP, Precision, Recall and F1 score of dataset 2 for $n \in [1..16]$

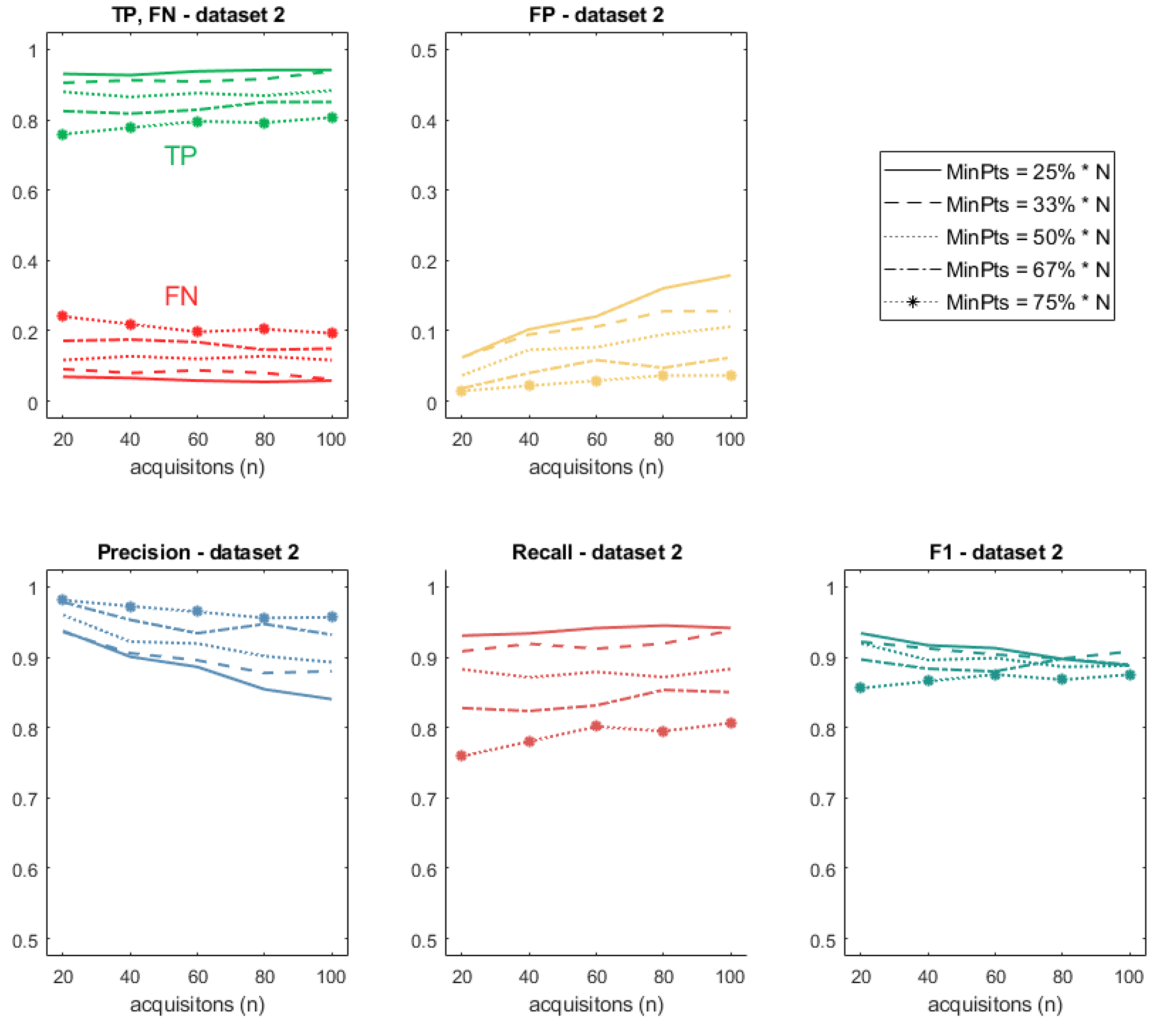


Fig. A2 Evaluation of TP, FN, FP, Precision, Recall and F1 score of dataset 2 for $n = 20, 40, 60, 80, 100$

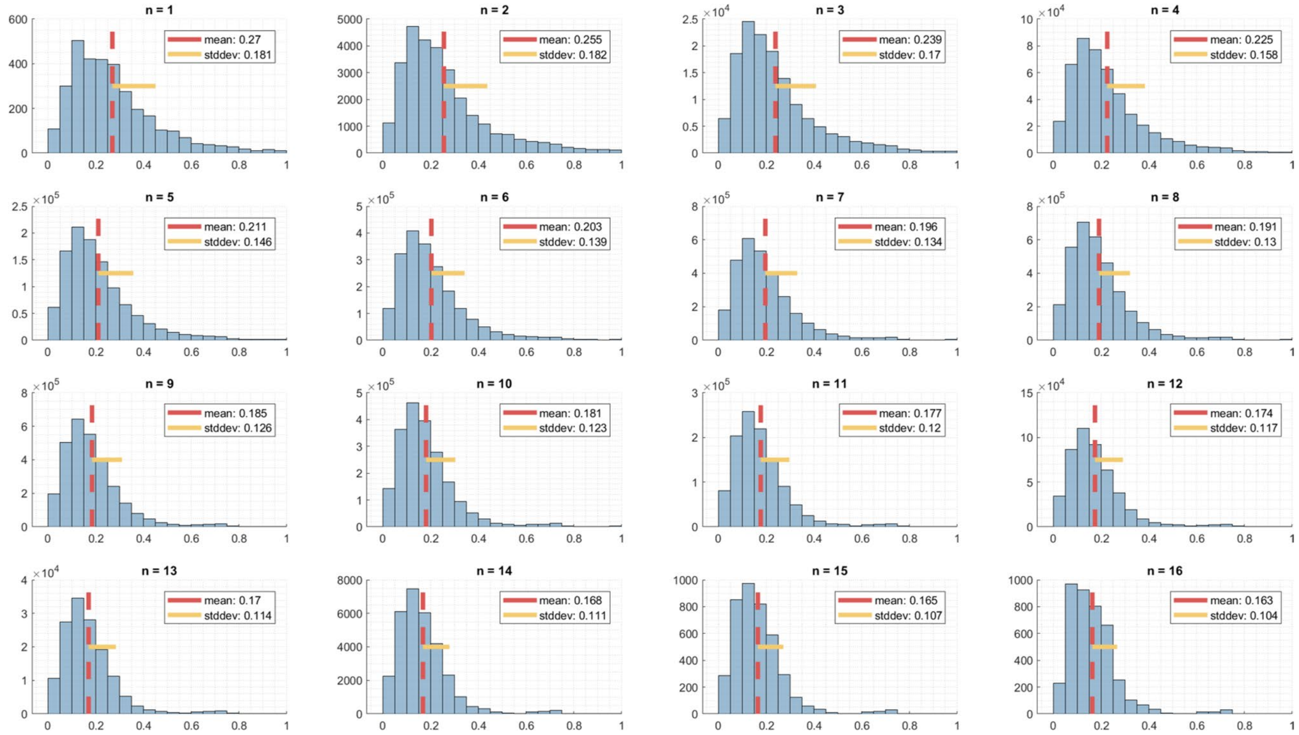


Fig. A3 Evaluation of the position difference of dataset 2 for $n \in [1..16]$. The x-axis shows the average position difference in meter and y-axis the number of combinations (the total number of possible combinations is $16!/((16–n)!*n!)$

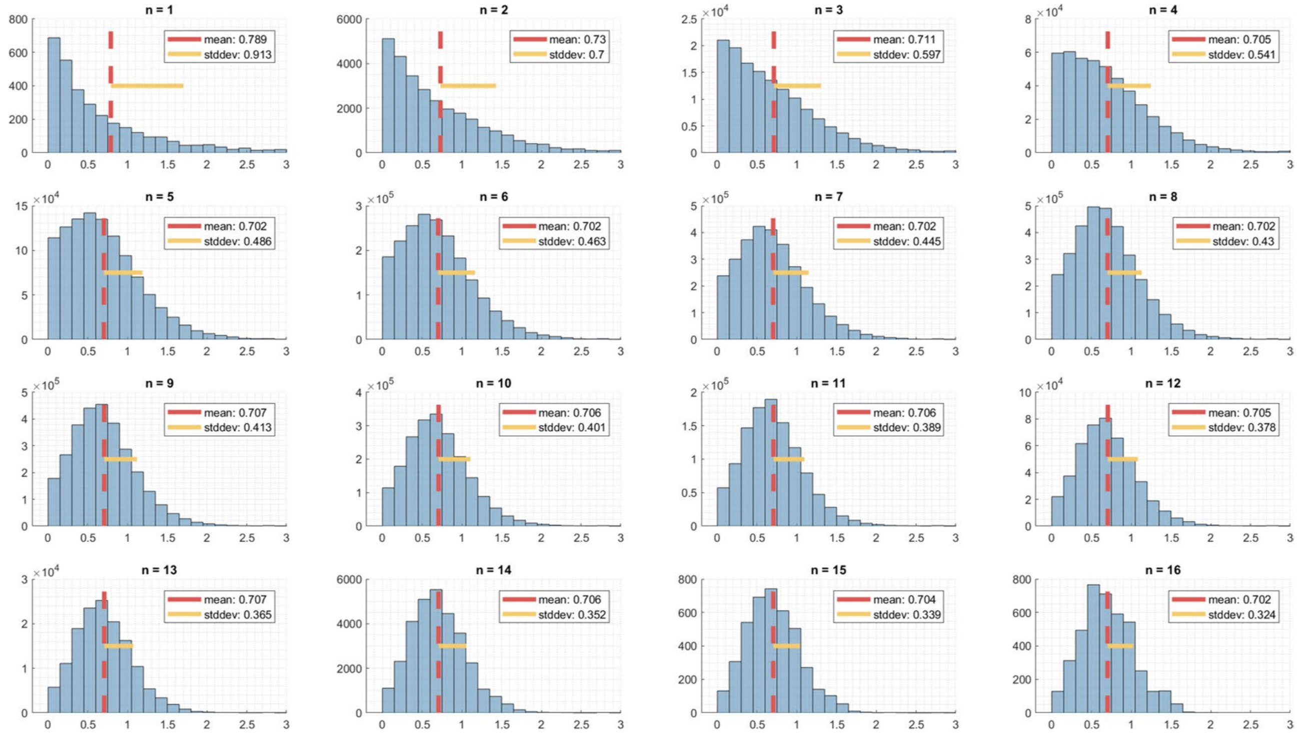


Fig. A4 Evaluation of the length difference of dataset 2 for $n \in [1..16]$. The x-axis shows the average position difference in meter and y-axis the number of combinations (the total number of possible combinations is $16!/((16–n)!*n!)$

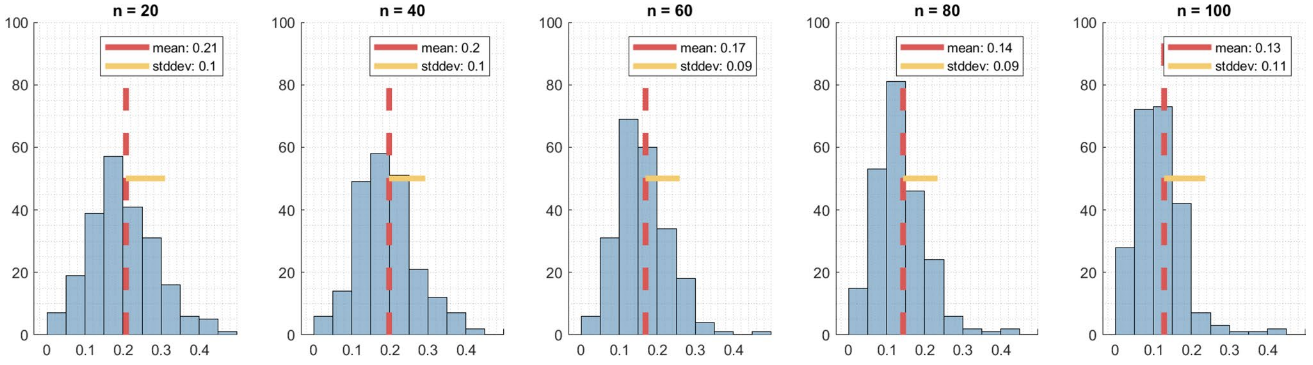


Fig. A5 Evaluation of position difference for $n = 20$, 40, 60, 80, 100 of dataset 2. The x-axis shows the average length difference in meters and the y-axis the number of measured values

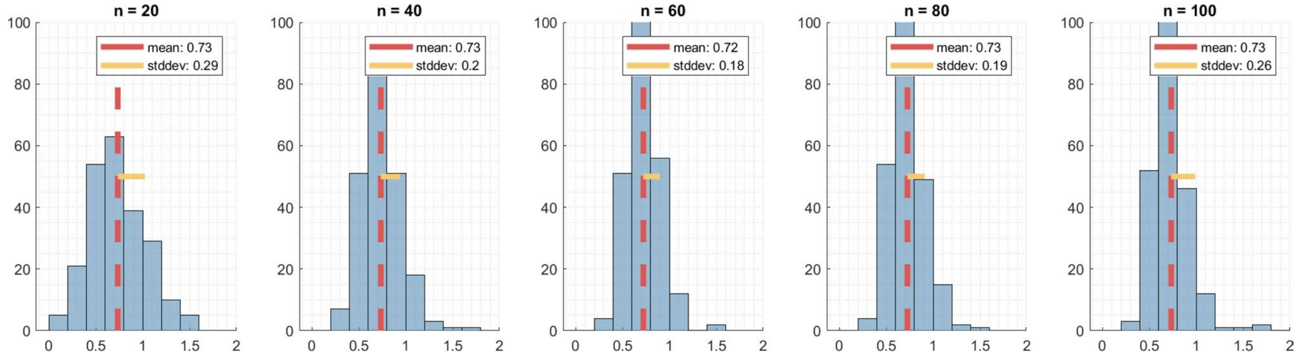


Fig. A6 Evaluation of length difference for $n = 20$, 40, 60, 80, 100 of dataset 2. The x-axis shows the average length difference in meters and the y-axis the number of measured values

Acknowledgements The H3D dataset has been captured in context of an ongoing research project funded by the German Federal Institute of Hydrology (BfG).

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of Data and Material The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

Albuquerque D, 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. In: Capineri C, Haklay M, Huang H, Antoniou V, Kettunen J, Ostermann F, Purves R (eds) European Handbook of Crowdsourced Geographic Information, Chapter 23, Ubiquity Press, pp 309–321. https://doi.org/10.5334/bax.w

Antoniou V, Skopeliti A (2015) Measures and indicators of VGI quality: an overview. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci 1:345–351. https://doi.org/10.5194/isprsannals-II-3-W5-345-2015

Brown G (2015) Engaging the wisdom of crowds and public judgement for land use planning using public participation geographic information systems. Aust Planner 52(3):199–209. https://doi.org/10.1080/07293682.2015.1034147

Budhathoki R, Haythornthwaite C (2012) Motivation for open collaboration: crowd and community models and the case of OpenStreetMap. Am Behav Sci 57(5):548–575. https://doi.org/10.1177/0002764212469364

Daniel F, Kucherbaev P, Cappiello C, Benatallah B, Allahbakhsh M (2018) Quality control in crowdsourcing. A survey of quality attributes, assessment techniques, and assurance actions. ACM Comput Surv. https://doi.org/10.1145/3148148

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering cluster in large spatial databases with noise. In: Simoudis E, Han J, Fayyad UM (eds) Proceedings of the second international conference on knowledge discovery and data mining (KDD-96). AAAI Press, pp 226–231

Estes LD, McRitchie D, Choi J, Debats S, Evans T, Guthe W, Luo D, Ragazzo G, Zempleni R, Caylor KK (2016) A platform for crowdsourcing the creation of representative, accurate landcover maps. Environ Model Softw 80:41–53. https://doi.org/10.1016/j.envsoft.2016.01.011

Eum J, Bae M, Jeon J, Lee H, Oh S, Lee M (2017) Vehicle detection from airborne LiDAR point clouds based on a decision tree algorithm with horizontal and vertical features. Remote Sens Lett 8(5):409–418. https://doi.org/10.1080/2150704X.2016.1278310

Feng D, Rosenbaum L, Dietmayer K (2019) Towards safe autonomous driving: capture uncertainty in the deep neural network for lidar 3D vehicle detection. Int Conf Intell Transp Syst (ITSC). https://doi.org/10.1109/ITSC.2018.8569814

Fonte CC, Antoniou V, Bastin L, Estima J, Jokar Arsanjani J, Laso BJ, See L, Vatseva R (2017) Assessing VGI data quality. In: Foody G, See L, Fritz S, Mooney P, Olteanu-Raimond AM, Fonte CC, Antoniou V (eds) Mapping and the citizen sensor. Ubiquity Press, pp 137–163. https://doi.org/10.5334/bbf.g

Galton F (1907) Vox Populi (the wisdom of the crowds). Nature 75:450–451. https://doi.org/10.1038/075450a0

Gao M, Xu W, Callison-Burch C (2015) Cost optimization for crowdsourcing translation. Human language technologies: the 2015 annual conference of the North American chapter of the ACL. Denver, Colorado, pp 705–713.https://doi.org/10.3115/v1/N15-1072

Geiger D, Seedorf S, Schader M (2011) Managing the crowd: towards a taxonomy of crowdsourcing processes. In: 17th Americas Conference on Information Systems 2011, AMCIS 2011, 5 pages

Goodchild MF (2007) Citizens as sensors. The World of volunteered geography. GeoJournal 69(4):211–221. https://doi.org/10.1007/s10708-007-9111-y

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spatial Stat 1:110–120. https://doi.org/10.1016/j.spasta.2012.03.00

Haklay M, Weber P (2008) OpenStreetMap: user-generated street maps. IEEE Pervasive Comput 7(4):12–18. https://doi.org/10.1109/MPRV.2008.80

Haralabopoulos G, Wagner C, McAuley D, Anagnostopoulos I (2019) Paid crowdsourcing, low income contributors, and subjectivity. In: MacIntyre J, Maglogiannis I, Iliadis L, Pimenidis E (eds) Artificial intelligence applications and innovations. AIAI 2019. IFIP Advances in Information and Communication Technology, Vol 560, Springer. https://doi.org/10.1007/978-3-030-19909-8_20

Hashemi P, Abbaspour RA (2015) Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept. OpenStreetMap in GIScience, Springer, pp 19–36.https://doi.org/10.1007/978-3-319-14280-7_2

Hecht R, Kalla M, Krüger T (2018) Crowd-sourced data collection to support automatic classification of building footprint data. Proc Int Cartogr Assoc. https://doi.org/10.5194/ica-proc-1-54-2018

Herfort B, Hoefle B, Klonner C (2018) 3D micro-mapping: Towards assessing the quality of crowdsourcing to support 3D point cloud analysis. ISPRS J Photogramm Remote Sens 137:73–83. https://doi.org/10.1016/j.isprsjprs.2018.01.009

Hirth M, Hoßfeld T, Tran-Gia P (2013) Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Math Comput Model 57(11–12):2918–2932. https://doi.org/10.1016/j.mcm.2012.01.006

Hirth M, Hoßfeld T, Tran-Gia P (2011) Anatomy of a crowdsourcing platform—using the example of microworkers.com. In: Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Seoul, pp 322–329. https://doi.org/10.1109/IMIS.2011.89

Hitlin P (2016) Crowdsourcing age, a case study. Pew Research Center. July 2016. Available at: https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Hosio SJ, Gonçalves J, Anagnostopoulos T, Kostakos V (2016) Leveraging wisdom of the crowd for decision support. Proceedings of the 30th International BCS Human Computer Interaction Conference. Article No. 38, 1–12. https://doi.org/10.14236/ewic.HCI2016.38

Howe J (2006) The rise of crowdsourcing. Wired Magazine 14(6):1–4

Ikeda K, Morishima A, Rahman H, Roy SB, Thirumuruganathan S, Amer-Yahia S, Das G (2016) Collaborative crowdsourcing with crowd4U. Proc VLDB Endow. 9 13(September 2016):1497–1500. https://doi.org/10.14778/3007263.3007293

Ipeirotis, PG (2010) Analyzing the Amazon Mechanical Turk marketplace. XRDS: Crossroads, The ACM Magazine for Students, December 2010. doi:https://doi.org/10.1145/1869086.1869094

Juni M, Eckstein MP (2017) The wisdom of crowds for visual search. Proc Natl Acad Sci 114:E4306–E4315. https://doi.org/10.1073/pnas.1610732114

Kazemi L, Shahabi C, Chen L (2013) GeoTruCrowd: trustworthy query answering with spatial crowdsourcing. In: SIGSPATIAL'13: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp 314–323. https://doi.org/10.1145/2525314.2525346

King AJ, Cheng L, Starke SD, Myatt JP (2011) Is the true “wisdom of the crowd” to copy successful individuals? Biol Lett 8(2):197–200. https://doi.org/10.1098/rsbl.2011.0795

Koelle M, Walter V, Schmohl S, Soergel U (2021a) Remembering both the machine and the crowd when sampling points: active learning for semantic segmentation of ALS point clouds. ICPR international workshops and challenges. Springer International Publishing, Cham, pp 505–520

Koelle M, Laupheimer D, Walter V, Haala N, Soergel U (2021b) Which 3D Data representation does the crowd like best? Crowd-based active learning for coupled semantic segmentation of point clouds and textured meshes. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2-2021b, pp 93–100. https://doi.org/10.5194/isprs-annals-V-2-2021-93-2021

Koelle M, Walter V, Schmohl S, Soergel U (2020) Hybrid acquisition of high quality training data for semantic segmentation of 3D point clouds using crowd-based active learning. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-2–2020, pp 501–508

Koita T, Suzuki S (2019) Crowdsourcing and its application for traffic survey work. In: IEEE 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, pp 375–378. https://doi.org/10.1109/ICBDA.2019.8712831

Kölle M, Laupheimer D, Schmohl S, Haala N, Rottensteiner F, Wegner JD, Ledoux H (2021) The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. ISPRS Open J Photogramm Remote Sens. https://doi.org/10.1016/j.ophoto.2021.100001

Krause J, Ruxton GD, Krause S (2010) Swarm intelligence in animals and humans. Trends Ecol Evol 25(1):28–34. https://doi.org/10.1016/j.tree.2009.06.016

Krause S, James R, Faria JJ, Ruxton GD, Krause J (2011) Swarm intelligence in humans: diversity can trump ability. Anim Behav 81(5):941–948. https://doi.org/10.1016/j.anbehav.2010.12.018

Lans L, Ansems EL, Khan VJ (2018) Paid crowdsourcing as concept and content generator to enhance museum experiences. In: Vermeeren A, Calvi L, Sabiescu A (eds) Museum experience design. Springer series on cultural computing. Springer, New York. https://doi.org/10.1007/978-3-319-58550-5_7

Leibovici DG, Rosser JF, Hodges C, Evans B, Jackson MJ, Higgins CI (2017) On data quality assurance and conflation entanglement in crowdsourcing for environmental studies. ISPRS Int J Geo-Inf. https://doi.org/10.3390/ijgi6030078

Liu Z, Shabani S, Glassey N, Sokhn M, Cretton F (2018) How to motivate participation and improve quality of crowdsourcing when building accessibility maps. In: 3th International Workshop on Accessible Devices and Services in the 15th Conference of IEEE Consumer Communications & Networking Conference (CCNC), Las Vegas, USA, 6 pages. https://doi.org/10.1109/CCNC.2018.8319237

Lloyd SP (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137. https://doi.org/10.1109/TIT.1982.1056489

Lorenz J, Rauhut H, Schweitzer F, Helbing D (2010) How social influence can undermine the wisdom of crowd effect. Natl Acad Sci USA 108:9020–9025. https://doi.org/10.1073/pnas.1008636108

Maddalena E, Ibanez LD, Simperl E (2020) Mapping Points of Interest through street view imagery and paid crowdsourcing. ACM Trans Intell Syst Technol 1(1). http://arxiv.org/abs/1901.09264

Mao A, Kamar E, Chen Y, Horvitz E, Schwamb ME, Lintott CJ, Smith AM (2013) Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In: Proc. 1st AAAI Conf. Human Computat. Crowdsourcing, pp 1–9

Nguyen T, Wang S, Anugu V, Rose N, McKenna M, Petrick N, Burns J, Summers R (2012) Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. Radiology 262:824–833. https://doi.org/10.1148/radiol.11110938

Pfeifer N, Mandlburger J, Otepka J, Karel W (2014) OPALS—a framework for Airborne Laser Scanning data analysis. Comput Environ Urban Syst 45:125–136. https://doi.org/10.1016/j.compenvurbsys.2013.11.002

Pinheiro MB, Davis CA (2018) ThemeRise: a theme-oriented framework for volunteered geographic information applications. Open Geospatial Data Softw Standards 3(1):2363–7501. https://doi.org/10.1186/s40965-018-0049-4

Ponn T, Kroeger T, Diermeyer F (2020) Performance analysis of camera-based object detection for automated vehicles. Sensors 20(13):3699. https://doi.org/10.3390/s20133699

Redi J, Povoa I (2014) Crowdsourcing for rating image aesthetic appeal: better a paid or a volunteer crowd? In: CrowdMM '14: Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, pp 25–30. https://doi.org/10.1145/2660114.2660118

Riegl (2018) RIEGL VUX-1UAV product specifications. http://www.riegl.com/uploads/tx_pxpriegldownloads/RIEGL_VUX-1UAV_Datasheet_2017–09–01_01.pdf

Salk CF, Sturn T, See L, Fritz S, Perger C (2016) Assessing quality of volunteer crowdsourcing contributions. Lessons from the Cropland Capture game. Int J Digital Earth 9(4):410–426. https://doi.org/10.1080/17538947.2015.1039609

See L, Mooney P, Foody G, Bastin L, Comber A, Estima J, Fritz S, Kerle N, Jiang B, Laakso M, Liu HY, Milčinski G, Nikšič M, Painho M, Pődör A, Olteanu-Raimond AM, Rutzinger M (2016) Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. ISPRS Int J Geo-Inf 5:55. https://doi.org/10.3390/ijgi5050055

Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M (2017) A review of volunteered geographic information quality assessment methods. Int J Geogr Inf Sci 31(1):139–167. https://doi.org/10.1080/13658816.2016.1189556

Simons A (2004) Many wrongs: the advantage of group navigation. Trends Ecol Evol 19:453–458. https://doi.org/10.1016/j.tree.2004.07.001

Surowiecki J (2004) The Wisdom of Crowds—why many are smarter than the few and how collective wisdom shapes business, economics, societies and nations. Doubleday, New York,https://doi.org/10.1111/j.1744-6570.2006.00060_10.x

van Dijk TC, Fischer N, Häussner B (2020) Algorithmic improvement of crowdsourced data: intrinsic quality measures, local optima, and consensus. In: Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20). Association for Computing Machinery, New York, NY, USA, pp 433–436. https://doi.org/10.1145/3397536.3422260

Vaughan JW (2017) Making better use of the crowd. How crowdsourcing can advance machine learning research. J Mach Learn Res 18(193):1–46

Walter V, Fritsch D (1999) Matching spatial data sets: a statistical approach. Int J Geograph Inf Sci 13(5):445–473. https://doi.org/10.1080/136588199241157

Walter V, Soergel U (2018) Implementation, results and problems of paid crowd-based geospatial data collection. PFG J Photogramm Remote Sens Geoinf Sci 86(3–4):87–197. https://doi.org/10.1007/s41064-018-0058-z

Walter V, Koelle M, Yin Y (2020) Evaluation and optimisation of crowd-based collection of trees from 3D point clouds. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-4–2020, pp 49–56

Walter V, Koelle M, Collmar D, Zhang Y (2021) A two-level approach for the crowd-based collection of vehicles from 3D point clouds. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-4-2021, pp 97–104. https://doi.org/10.5194/isprs-annals-V-4-2021-97-2021

Wang H, Yu Y, Cai Y, Chen X, Chen L, Liu Q (2019) A comparative study of state-of-the-art deep learning algorithms for vehicle detection. IEEE Intell Transp Syst Mag 11(2):82–95. https://doi.org/10.1109/MITS.2019.2903518

Wu X, Li W, Hong D, Tian J, Tao R, Du Q (2020) Vehicle detection of multi-source remote sensing data using active fine-tuning network. ISPRS J Photogramm Remote Sens 167:39–53. https://doi.org/10.1016/j.isprsjprs.2020.06.016

Xavier E, Francisco J, Manuel A (2016) A survey of measures and methods for matching geospatial vector datasets. ACM Comput Surv. https://doi.org/10.1145/2963147

Yang MY, Liao W, Li X, Rosenhahn B (2018) Deep learning for vehicle detection in aerial images. IEEE Int Conf Image Process (ICIP). https://doi.org/10.1109/ICIP.2018.8451454

Zhang J, Wu X, Sheng V (2016) Learning from crowdsourced labelled data. A survey. Artif Intell Rev 46(4):543–576. https://doi.org/10.1007/s10462-016-9491-9

Zheng Y, Li G, Yuanbing L, Shan C, Cheng R (2017) Truth inference in crowdsourcing: is the problem solved? Proc VLDB Endowment. https://doi.org/10.14778/3055540.3055547

Zhou D, Platt JC, Basu S, Mao Y (2012) Learning from the wisdom of crowds by minimax entropy. Adv Neural Inf Process Syst (NIPS) 25:2204–2212