Reproducing Geospatial Crowdsourcing: How Consistent Is the Crowd?

David Collmar[1], Volker Walter[1], Uwe Sörgel[1], Roland Ullmann[1]

[1] Institute for Photogrammetry and Geoinformatics (ifp), University of Stuttgart, Germany
(david.collmar, volker.walter, uwe.soergel, roland.ullmann)@ifp.uni-stuttgart.de

Keywords: Geospatial Crowdsourcing, Task Consistency, Reproducibility, Worker Retention, Wisdom of Crowds

Abstract

This paper investigates the long-term consistency and reliability of paid geospatial crowdsourcing on the online platform *Microworkers.com*. Over a five-month period, we conducted three crowdsourcing campaigns, each representing a task typical for remote sensing, i.e., pixel classification, point selection, and geometric outline acquisition, to assess whether repeated worker participation enhances data quality and reproducibility. Beyond individual task performance, we examine the broader question of whether crowdsourcing campaigns can yield reproducible results over extended periods. Despite the large and heterogeneous workforce of *Microworkers.com*, a substantial share of tasks was completed by recurring workers who consistently outperformed one-time participants. Furthermore, across all campaigns, data quality remained largely stable, with only minor variability between epochs. Additionally performed statistical analyses confirm that reproducible outcomes are achievable, highlighting the potential of reliable and reproducible crowdsourcing results for geospatial data acquisition.

1. Introduction

Whereas voluntary crowdsourcing remains an essential technique to leverage the knowledge of a large amount of volunteers (Estellés-Arolas, 2022), paid crowdsourcing has become more and more popular in recent years (Karachiwalla and Pinkow, 2021), partly driven by the rise of machine learning (Zhang, 2022), where paid crowdsourcing is used to collect large amounts of training data. Crowdsourcing is found across different research fields (Lenart-Gansiniec et al., 2023), and, as a result, is used for different applications, with natural language processing and medical applications being prominent examples (Sabou et al., 2012, Wazny, 2018). Possible applications in the field of remote sensing range from simple tasks such as pixel classifications (Saralioglu and Gungor, 2022) to complex implementations including active learning architectures (Sayin et al., 2021), which can be adapted to various different domains and types of geodata.

However, data quality remains a persistent issue in paid crowdsourcing (Kobayashi et al., 2022). One reason is the predominantly externally incentivized nature of tasks in paid crowdsourcing (Hossain, 2012), which diminishes intrinsic motivation, affects quality (Chandler et al., 2013), and makes the implementation of anti-spam measures necessary (Zheng et al., 2017). Malevolent users, such as trolls or spammers, may however circumvent such measures (Zhu and Carterette, 2010), resulting in compromised data. Since spammers can account for up to 45% of contributors (Vuurens et al., 2011), truth inference techniques are often employed as data-quality enhancement strategies (Cui et al., 2021). Potential approaches to truth inference analyze and model submissions either at the task level or at the worker level (Zheng et al., 2017). In task modeling, a difficulty level is assigned to the task itself, enabling the estimation of the probability that it can be solved successfully. Worker modeling, in contrast, aims to characterize the type of individual crowdworker, as different worker characteristics lead to different effects on the quality of their contributions (Chandler et al., 2013). Subsequently, task modeling alone is not sufficient as a truth inference method in most cases, and worker modeling should be included as well for best outcomes (Zheng et al., 2017). In detail, such worker modeling assigns a quality parameter to a worker that represents the ability of the respective worker to solve a specific task. This probability can be estimated through different techniques, including straightforward approaches like a scalar probability value or more intricate methods such as worker bias and variance modeling (Zheng et al., 2017). Whereas many of these approaches partly rely on techniques like qualification tasks and hidden tests, the most reliable assumptions about worker quality can be made if a sufficient number of tasks were performed by the same worker, e.g., over 20 tasks per individual (Zheng et al., 2017), as workers tend to deliver consistent results (Williams et al., 2017). This rather large number, however, might be hard to achieve in a real-world scenario: typical for paid crowdsourcing is the use of crowdsourcing platforms, which facilitate user acquisition, payment, and other administrative tasks (Hirth et al., 2011). However, these crowdsourcing platforms often pride themselves on a large number of registered workers, such as the almost 4,000,000 users on *Microworkers.com* (Microworkers.com, 2024). While such scale suggests crowd heterogeneity due to varying levels of experience, know-how and motivation (Hirth et al., 2011), this large number also indicates a rather slim chance to reach higher levels of redundancy for the same worker over the span of multiple campaigns or a larger timeframe in general, and subsequently questioning the potential of worker modeling. This anticipated low redundancy per worker could not only impede truth inference, but also raises questions in terms of reproducibility and repeatability, if a crowdsourcing campaign is repeated at a later point in time.

Previous research showed that repeatability could be achieved for classification campaigns that were started in a weekly interval (Qarout et al., 2019). However, in that approach, workers were only allowed to participate in a single crowdsourcing campaign, effectively blocking users from participating a second time. As a result, a different crowd was used for every campaign, thereby impeding truth inference through worker modeling. Allowing workers to participate in multiple campaigns over several months therefore does not only enable truth infer-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-69-2026

ence through recurring participation, but also closely emulates the dynamics of real-world crowdsourcing campaigns, where typically no pre-selection or filtering of workers is performed. Not restricting the user base might allow us to observe natural variations in participation, performance, and data quality over time, thereby providing a realistic assessment of reproducibility under authentic conditions. However, it remains uncertain whether a consistency in results can be achieved. Will there be recurring workers for some campaigns, or even every campaign? If so, how many? And will their results be consistent across the different epochs and campaigns? Will the overall task quality remain reproducible? We address these questions in this paper, aiming to evaluate the potential for consistent and reliable outcomes in paid geospatial crowdsourcing.

2. Methodology

To address the aforementioned questions, i.e., whether consistent and reproducible results can be achieved in paid geospatial crowdsourcing, we conducted a long-term study on the platform *Microworkers.com*. The study was designed to examine how worker participation, retention, and task performance evolve over time. To this end, we launched three separate crowdsourcing campaigns with varying levels of difficulty in a monthly interval. Each crowd campaign represents a task typical for applications in remote sensing: a pixel classification task (*Campaign A*), the selection of a target point in an orthophoto (*Campaign B*), and the acquisition of geometric outlines (*Campaign C*). The campaigns were selected to represent complementary remote-sensing task types while using unambiguous imagery to focus the analysis on temporal stability and worker effects rather than scene difficulty. The setup of each campaign is described in detail in the following sections.

2.1 Campaign A - Pixel Classification

The first campaign comprises a straightforward pixel classification task using aerial imagery. Individual pixels belonging to specific object classes were highlighted, and workers were asked to assign each pixel to the correct class via selection from a predefined set of classes. Figure 1 illustrates the graphical user interface (GUI) used for all classification tasks: a red arrow indicates the pixel to be classified, the available classes are displayed as radio-button options on the right-hand side. Each job included five independent classification tasks.

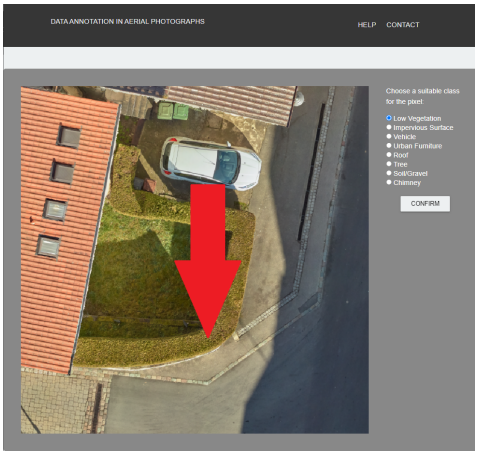

Figure 1. Interface of the pixel classification task with the red arrow indicating the highlighted pixel.

2.2 Campaign B - Point Selection

The second campaign was designed to evaluate the geometric accuracy of worker inputs. In this task, participants were instructed to precisely mark the center of a checkerboard, similar to those used in photogrammetric applications. Figure 2 shows the GUI used for this campaign, with a checkerboard clearly visible. The center point can be placed via simply clicking, and confirmed by using the ”Next” button on the right-hand side. If necessary, the point can be removed using the ”Clear” button. To discourage automated or malicious behavior, the checkerboards were randomly positioned away from the image center. Each crowdsourcing job includes the processing of four checkerboards.

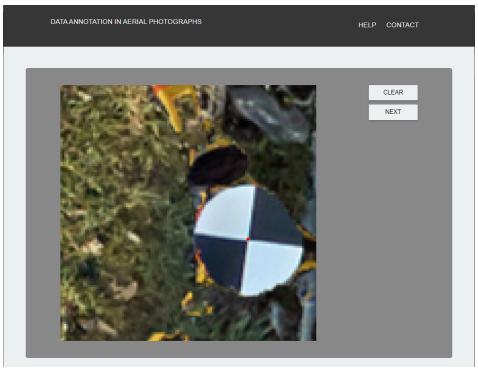

Figure 2. Interface of the geometric accuracy task displaying the checkerboard and point selection controls.

2.3 Campaign C - Polygon Acquisition

The third campaign involved a more complex task compared to the previous ones, combining both semantic and geometric accuracy. In this task, workers are asked to precisely collect tree outlines in orthoimages via polygons. Figure 3 shows the GUI used for this task, including an example of an already acquired tree outline. Polygon vertices can be added by clicking on the image and can be removed using the ”Undo” or ”Clear” buttons on the right-hand side. Clicking the ”Next” button advances to the next tree that needs to be processed. In total, each crowdsourcing job required the acquisition of three such tree outlines.

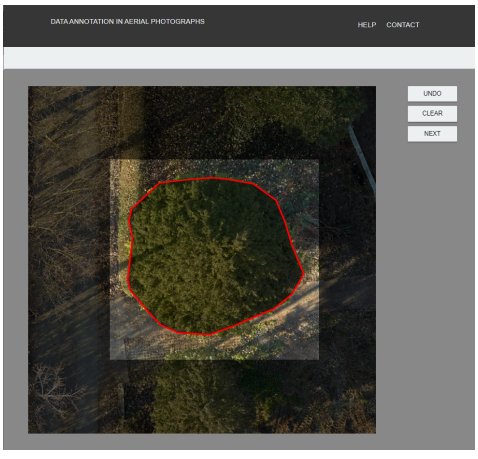

Figure 3. Interface of the geometric acquisition task displaying a tree outline and polygon editing controls.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-69-2026

2.4 Campaign Parameters

Each campaign was conducted during every epoch, for a total of five epochs. To ensure consistent conditions across all epochs, all campaign parameters, including salary and starting time, were kept identical. The campaigns were publicly posted, and participation was open to all workers without invitation or preselection. The time interval between the start of each epoch was set to one month, resulting in a total of five epochs that were conducted between November 2023 and March 2024. This approach not only differs significantly from the five-week study by (Qarout et al., 2019), allowing for verification of their results on a five-month basis, but also allows for an analysis over a longer time period. This timeframe allows for an analysis of short- to medium-term trends that might not be evident in a shorter study.

Each campaign was conducted 50 times per epoch, leading to a total of 150 jobs per epoch over all three campaigns. Given that each job consists of either five classifications, four point selections of checkerboard centers, or the acquisition of three geometries, a total of 250 classifications, 200 checkerboard centers, and 150 tree outline geometries were collected per epoch. Each job was compensated with $0.08, resulting in total costs of $12 plus fees per epoch. All users registered on *Microworkers.com* were eligible to participate in the study, with no prior qualification tests or filtering based on specific characteristics or demographics, reflecting the open participation conditions typical of large-scale crowdsourcing applications.

3. Worker Retention

We are interested in potential fluctuations within the workforce, which can be measured by analyzing worker retention across all five epochs. For this purpose, we define a *recurring worker* as a user who participates in a campaign after having already participated in any campaign during a previous epoch. Within an epoch, we do not distinguish between individual campaigns, but rather analyze retention at the epoch level. Users can choose to participate in one, two, or all three campaigns within an epoch, which may result in varying levels of engagement. This approach enables us to capture overall retention trends without being affected by individual variations in campaign participation, as users may have different task preferences or may choose not to participate in certain campaigns for various reasons. Moreover, focusing on the epoch level emphasizes the assessment of general long-term worker retention.

According to this definition, a worker can be a new worker in only one epoch and a recurring worker in up to four epochs. Table 1 presents both the absolute and relative numbers of new and recurring workers for all five epochs. The total number of NW can be calculated by summing the values across all epochs.

Epoch	NW	NW (%)	RW	RW(%)	Σ	Jobs
1	121	100.0	0	0.0	121	150
2	101	84.9	18	15.1	119	150
3	94	73.4	34	26.6	128	150
4	76	60.8	49	39.2	125	150
5	69	52.3	63	47.7	132	150

Table 1. New workers (NW) and recurring workers (RW) per epoch.

As shown in Table 1, the number of new workers decreases while the number of recurring workers increases with each successive epoch. Although this finding may seem intuitive, it is noteworthy in the context of the vast scale of *Microworkers.com*: despite the large and heterogeneous user base, we observe a steady influx of recurring workers, indicating a consistent core group of participants. This trend is further illustrated in Figure 4, which also includes the total number of unique workers, shown in green.

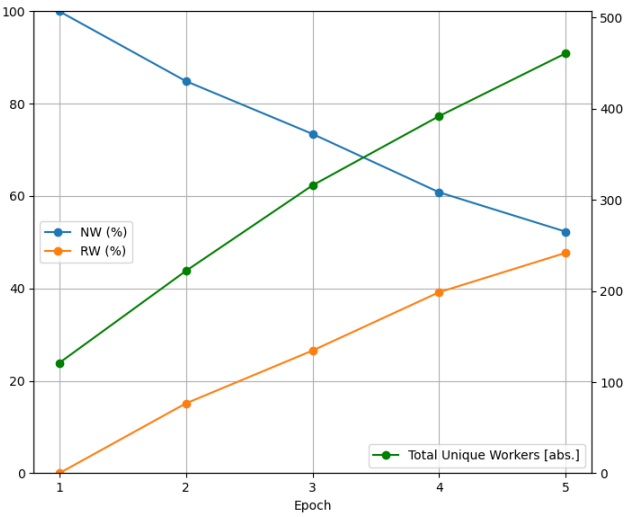

Figure 4. New workers and recurring workers per epoch.

However, Table 1 and Figure 4 only consider whether a worker has participated in a campaign before, without providing further detail on participation patterns across epochs. For instance, both a worker who took part in two epochs and one who participated in all five epochs are classified as recurring, even though their overall level of participation differs substantially. Consequently, this aggregation leads to a certain loss of granularity when analyzing worker retention over time. To address this, we compared user participation on a per-epoch basis, thereby examining how many workers from each epoch continued to contribute in subsequent ones. Figure 5 visualizes these results, offering a more detailed view of worker retention dynamics.

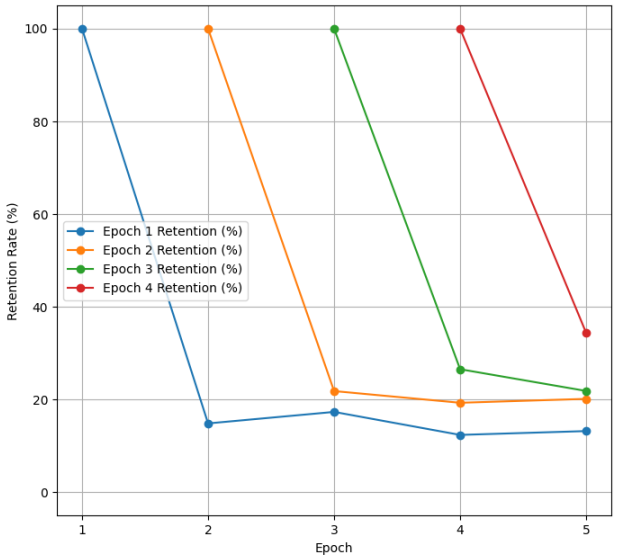

Figure 5. Worker retention across epochs. Epoch 5 is not visualized since only one data point would be included.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-69-2026 | © Author(s) 2026. CC BY 4.0 License.

As can be seen in Figure 5, for Epoch 1, around 15% of the workers also participated in Epoch 2, and approximately 17% of Epoch 1 participants joined Epoch 3. Interestingly, the retention rate increases in the higher epochs. While only 15% of Epoch 1 workers participated in Epoch 2, over 20% of Epoch 2 workers continued into Epoch 3. By Epoch 3, this rate rises to around 27%, reaching a peak of about 34% for Epoch 4. This upward trend indicates the formation of a stable core of recurring workers who regularly contribute across multiple campaigns. Although the exact reasons for this pattern remain uncertain, it highlights a notable level of engagement and continuity within the workforce. However, Figures 4 and 5 do not show how many workers participated in one, two, three, four, or all five epochs overall. These cumulative results are therefore summarized in Table 2.

Number of Epochs	Workers	[%]
1	357	77.4
2	63	13.7
3	27	5.9
4	9	1.9
5	5	1.1

Table 2. Cumulative worker participation for all epochs.

As Table 2 highlights, the vast majority, around 77%, of workers participated in only a single epoch. However, this also means that approximately 1 out of 4 workers participated in more than one campaign, reinforcing the earlier suggestion of a strong core user base. Additionally, about 1 in 11 workers participated in three or more epochs, showing a varied level of engagement among the workforce, with a small but significant proportion demonstrating consistent participation. In summary, despite the enormous number of registered users on *Microworkers.com*, a notable retention rate of 15% to 34% of crowd workers participating in successive campaigns was observed.

Still, beyond participation trends, we also examined whether these recurring workers deliver higher-quality results than one-time participants. For this analysis, we compared workers who contributed in only one epoch to those who contributed across multiple epochs, using task-specific quality metrics: overall classification accuracy (Campaign A), Euclidean distance to the reference point (Campaign B), and intersection over union (IoU) (Campaign C). The results are summarized in the following Table 3.

No. Epochs	Mean Acc.	Mean Dist. [pix]	Mean IoU
1	0.58	9.33	0.80
2	0.60	5.63	0.83
3	0.71	2.35	0.92
4	1.00	1.79	0.84
5	n.a.	n.a.	0.77

Table 3. Mean accuracy, mean distance, and mean IoU for recurring workers across multiple participations.

As shown in Table 3, workers who participated in multiple epochs demonstrated improved performance across both accuracy and distance metrics. Mean accuracy increased from 0.58 for single-epoch participants to 1.00 for those involved in four epochs, while mean distance decreased from 9.33 to 1.79 pixels, indicating enhanced precision. The mean IoU, however, showed a less consistent pattern: it improved from 0.80 for single-epoch participants to 0.92 for those in three epochs, but declined slightly thereafter. This fluctuation likely results from the small sample sizes in the fourth and fifth epochs (1.95% and 1.08% of all workers), and these values should therefore be interpreted with caution. The relatively large mean distances are also influenced by outliers that were potentially caused by malicious workers or spammers, as extreme values skew this metric; thus, later analyses focus on median values instead of means.

Overall, better performance can be observed for workers who participated in multiple campaigns compared to those who participated in only a single campaign. This improvement could suggest a learning or habituation effect, where recurring workers become more familiar with the tasks and interfaces, or simply develop routine over time, as could also be observed in earlier studies (Walter et al., 2022). However, it may also reflect higher motivation among consistently engaged workers. Further research would be needed to distinguish between these effects.

4. Worker and Task Quality

To assess the reproducibility of crowdsourcing results, both worker-level performance and overall task quality were analyzed jointly. While worker quality provides insight into individual consistency, task-level results allow evaluation of the collective reproducibility across epochs. The following subsections summarize these aspects for each of the three campaigns.

4.1 Campaign A - Pixel Classification

The overall accuracy (OA) of all performed classifications is illustrated in Figure 6 for all 5 epochs. Furthermore visible are the proportions of both high- and low-performing workers. Low- and high-performing workers are directly defined from the results over all 5 epochs, where, on average, workers correctly solved around 2.5 of the 5 classifications. Users with four or five correct classifications are counted as high-performing workers, whereas those with only one or zero correct classifications are counted as low-performing workers.

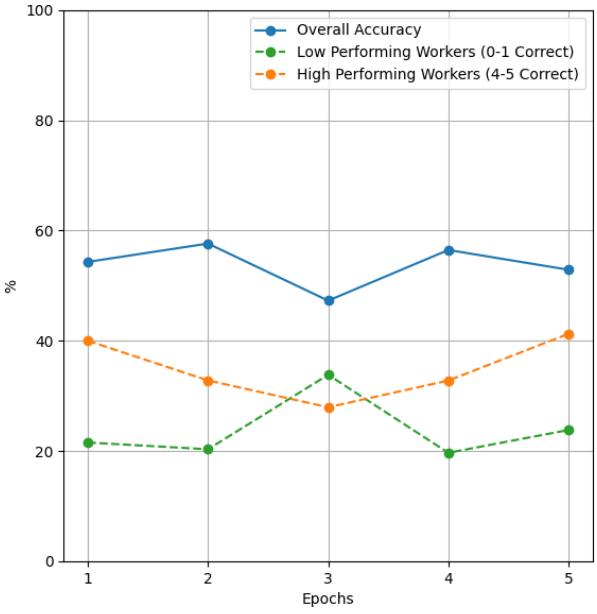

Figure 6. OA in comparison with low-performing & high-performing workers for Campaign A.

As shown in Figure 6, the OA reaches its peak at epoch 2 with approximately 57.64% and its lowest point at epoch 3

with around 47.3%, which appears to be an outlier both in terms of overall accuracy as well as the larger proportion of low-performing workers. No structural changes in task design or campaign parameters occurred in Epoch 3, suggesting that the deviation is most likely attributable to workforce composition effects. The mean OA across all epochs is about 53.7%, which implicitly supports the previous finding that each worker provides roughly 2.5 correct classifications on average. The proportion of low-performing workers remains relatively stable between 20% and 24%, except for epoch 3, where it rises to around 34%. According to an experiment by (Vuurens et al., 2011), approximately 55% of crowdworkers are so-called proper workers, with about 15% being either sloppy workers or semi-random spammers, and 30% being random or uniform spammers. These numbers confirm our findings on low-performing workers as follows: In our scenario, we have 5 classifications with 11 possible classes. Random guessing would result in an expected number of correct classifications of $\frac{5}{11} \approx 0.455$. Hence, a random spammer would be categorized as a low-performing worker.

A semi-random spammer typically tries to evade spam detection by providing some correct answers. If we assume a semi-random spammer gives two correct answers out of five classifications and randomly selects the remaining three, the expected number of correct classifications is $2 + \frac{3}{11} \approx 2.272$. Therefore, semi-random spammers are not considered low-performing workers but rather average-performing ones. The same reasoning applies to sloppy workers. Uniform spammers usually select one primary label to vote for. In our case, this leads to an expected number of correct classifications of $\frac{5}{11} \approx 0.455$, even when they occasionally switch labels. Consequently, uniform spammers are classified as low-performing workers. Given these statistical implications, we can infer that both random spammers and semi-random spammers are part of our low-performing worker group. According to (Vuurens et al., 2011), these two groups constitute 30%. This aligns with our experimental results, where low-performing workers comprise between 20% and 34%.

What cannot be confirmed at first glance through our research, however, is the relatively large number of proper workers. While we achieve around 30% of high-performing workers on average across all 5 epochs, (Vuurens et al., 2011) reported much higher numbers at 55%. However, in that study, an average precision of 0.75 was achieved. In our straightforward case, where only correct classifications are counted, precision and overall accuracy describe the same metric. Our previous definition of high-performing workers includes only those with 4 or 5 correct answers, leading to an accuracy and precision of 0.8 or 1 per worker, respectively. Therefore, to achieve a similar precision, workers with 3 correct answers and thus a precision of 0.6 also need to be included. When including workers with 3 or more correct answers, the proportion ranges between 44.1% and 60.9%, with a mean value of 53.7%, thereby directly confirming the numbers of (Vuurens et al., 2011), i.e., 55%, as well. In summary, previous findings for classifications can be confirmed, and the proportions of high- and low-performing workers remained mostly the same throughout all epochs, with the exception of Epoch 3.

These individual worker trends are also reflected at the task level, where aggregated results show similar patterns of stability and variation across epochs. Each job consisted of five classifications, with 50 jobs performed per epoch, resulting in a total of 250 classifications per epoch. All workers completed the same set of tasks, yielding the OA values shown in Table 4. Following the earlier argument regarding outliers, median values are reported instead of means for the number of correct answers. As the table indicates, the median number of correct classifications remained constant at three across all epochs.

Epoch	OA	Median CA
1	0.54	3
2	0.58	3
3	0.47	3
4	0.57	3
5	0.53	3

Table 4. Overall accuracy (OA) and median of correct answers (CA) per epoch.

A consistent median and overall moderate OA values between 0.47 and 0.58 are observed, both indicating generally stable results. If Epoch 3 is omitted, the remaining OA values, ranging from 0.53 to 0.58, suggest an even more stable overall accuracy. A visualization of the changes in overall accuracy is also provided in Figure 6. Based on the low OA values, it can be concluded that this type of pixel classification is challenging for most crowdworkers and may require further refinement to improve accuracy.

4.2 Campaign B - Point Selection

A direct comparison with the previously discussed study by (Vuurens et al., 2011) is not possible for the geometric case, but similar proportions in terms of high- or low-performing tasks are expected. Spammers are much easier to detect in this type of task: while a random spammer had a $\frac{1}{11}$ chance of guessing correctly in a classification task, the likelihood of clicking the center of a checkerboard (or within a few pixels of it) by blind spamming approaches zero. Therefore, we define workers with a click distance from the reference greater than five pixels as spammers. It should be noted that while this threshold may seem arbitrary, it is justified by the median values around two pixels, as can be seen later. Furthermore, we are more interested in the overall trends and relative changes across epochs in terms instead of absolute numbers, which are independent of the absolute chosen thresholds. Workers were asked to perform four checkerboard center clicks, and we define low-performing workers as those who had at least one of their four acquisitions marked as spam, i.e., with a distance greater than 5 pixels from the reference. High-performing workers are defined as those whose four acquisitions fall within the interquartile range (IQR), with none outside the IQR. The remaining workers are considered average-performing. Table 5 highlights the change in the proportions of high- and low-performing workers over the epochs.

Epoch	Low Performance (%)	High Performance (%)
1	29.4	58.8
2	34.5	55.2
3	34.0	54.7
4	27.5	54.9
5	44.1	40.7

Table 5. Proportions of low-performing and high-performing workers for Campaign B.

As shown in Table 5, both proportions remain relatively stable across the first four epochs. However, Epoch 5 stands out as an outlier, where more low-performing workers than high-performing workers are observed. It should be noted, however,

that this does not necessarily equate to a decline in task quality, as a worker with a single acquisition outside the IQR is no longer considered high-performing, even if the other three acquisitions are perfect. These individual worker trends are also reflected at the task level, where aggregated results reveal similar patterns of consistency across epochs. Each job consisted of four checkerboard acquisitions, with 50 jobs executed per epoch, resulting in a total of 200 points per epoch. The quality of each click was measured by its Euclidean distance to the reference center point, which served as the ground truth. Median values were used instead of means to reduce the influence of outliers. The resulting distributions for all epochs are shown in Figure 7 as boxplots.

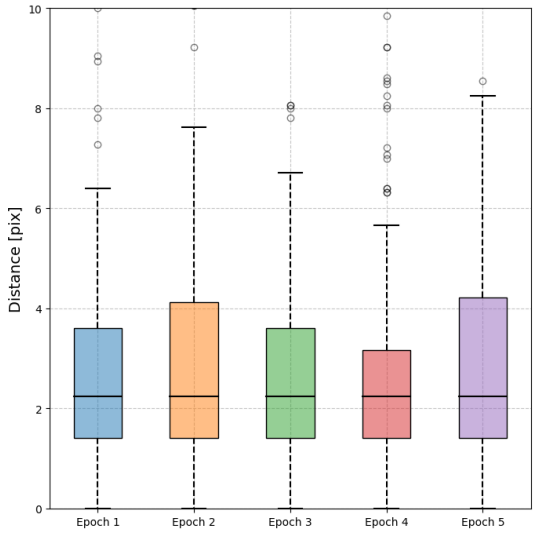

Figure 7. Boxplots of Euclidean distances for every epoch.

As Figure 7 illustrates, worker performance across all epochs is remarkably consistent. The median values are nearly identical, and the IQR share the same lower limits, with only slight variations in the upper limits, up to a single pixel between Epochs 4 and 5. These minimal variations highlight the strong reproducibility of this geometric accuracy task, especially when compared to the pixel classification task, which yielded more variable results.

4.3 Campaign C - Polygon Acquisition

Campaign C involved the acquisition of geometric outlines of three trees in airborne imagery, thereby combining aspects of the previous campaigns. Since reference data are available, the evaluation is performed by calculating the IoU between the acquired geometries and the reference shapes, effectively describing the similarity of the two shapes. The same argumentation as in the previous section applies here. Spammers can be rather easily detected, since IoU values, which serve as the measure of quality, penalize random points with low quality values. Given the nature of the task, which required the acquisition of precise geometric outlines, the identification of low-performing workers was straightforward: workers, whose polygons consistently produced an IoU of less than 0.5 for all three acquisitions, were classified as low-performing, reflecting significant deviations from the reference shapes. Conversely, workers were classified as high-performing if their polygons consistently achieved an IoU of 0.9 or higher across all tasks within an epoch. This threshold was set deliberately high to ensure that only those workers who closely replicated the reference shapes with high precision were actually recognized as high-performing. Still, the previous argumentation also applies here, as we are more interested in relative changes than in absolute numbers, making the choice of threshold values less impactful. Workers who did not meet the criteria for either low or high performance were considered average-performing. Table 6 lists the proportions of high- and low-performing workers across the five epochs of Campaign C.

Epoch	Low Performance (%)	High Performance (%)
1	19.2	35.6
2	12.7	36.7
3	14.0	48.8
4	15.9	37.8
5	10.0	30.0

Table 6. Proportions of low-performing and high-performing workers for Campaign C.

As can be seen from Table 6, the proportions of high- and low-performing workers again varied slightly across the epochs, reflecting minor fluctuations in worker quality over time. Notably, Epoch 3 exhibited the highest proportion of high-performing workers, suggesting a peak in worker accuracy during this period. Interestingly, the proportion of low-performing workers remains relatively constant, even during Epoch 3. However, in Epoch 5, we observe a decline in both low- and high-performing workers. These individual worker trends are also reflected at the task level, where aggregated IoU results demonstrate similar stability across the different epochs. As each job consisted of the acquisition of three tree outlines, and with 50 jobs executed per epoch, this resulted in a total of 150 geometries per epoch. The distributions of the underlying IoU values for all the acquisitions of each epoch are shown in Figure 8, again in the form of boxplots.

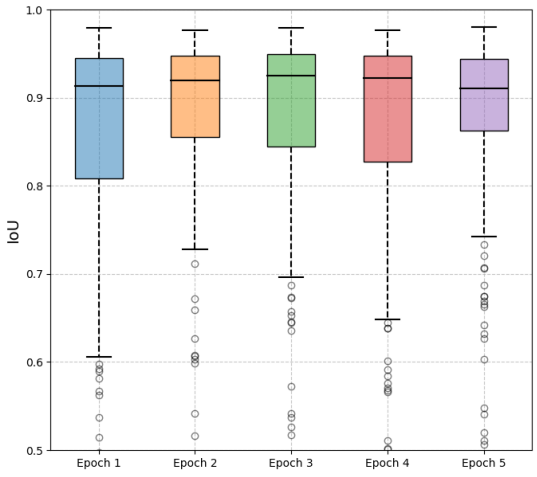

Figure 8. Boxplots of IoU values between reference and worker acquisitions for all epochs.

Figure 8 indicates a consistently high quality of results, with median IoU values exceeding 0.9 in all epochs. Differences in median values are most pronounced in the Epoch 5, accompanied by a generally larger IQR and whiskers extending

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-4-2026-69-2026

up to 1.5 times the IQR, thereby indicating slightly greater variability. Still, despite these variations, the overall results confirm that the geometric acquisition task yields results of high quality across all five epochs.

5. Statistical Analysis

As demonstrated in the previous sections, the results for Campaign C closely align with those of Campaign B, both of which outperformed Campaign A. However, the primary objective of this research is not to compare task types but to assess whether the results obtained across different epochs are statistically reproducible. To this end, a comparative statistical analysis was performed using the Mann–Whitney U test. In essence, the test assesses whether two independent samples originate from the same underlying distribution, as will be described in the following.

5.1 Methodology

To evaluate whether the distributions of overall accuracy (Campaign A), geometric distance (Campaign B), and IoU (Campaign C) differed significantly between epochs, we applied the Mann–Whitney U test (Mann and Whitney, 1947, Nachar et al., 2008, MacFarland and Yates, 2016). This test was selected because it is non-parametric and does not assume an underlying normal distribution. While the data were likely close to normal, this could not be guaranteed; therefore, the Mann–Whitney U test was preferred as it remains valid regardless of distribution shape. This test compares two independent samples to assess whether they originate from the same distribution, thereby making it suitable for evaluating reproducibility rather than mean differences. Pairwise comparisons were performed between all epoch combinations (i, j) for each campaign, with the null hypothesis H_0 stating that both distributions are statistically identical. The significance level was set to $\alpha = 0.05$, representing a relatively strict threshold for significance. A more lenient level could have been chosen, however, the 5% limit ensures a deliberately conservative interpretation of statistical differences. Although no normality was assumed for the data itself, the large sample sizes per epoch (150, 200, and 250 observations) allow the sampling distribution of the U statistic to be approximated by a normal distribution, providing a reliable basis for inference. Because some workers participated in multiple epochs, observations may not be fully independent across comparisons; the Mann-Whitney U test is therefore interpreted at the epoch-distribution level. Consequently, the test results assess distributional stability between epochs rather than individual worker changes.

5.2 Results

The Mann–Whitney U test results for all campaigns are summarized in Table 7. Across all three campaigns, only three out of thirty pairwise comparisons showed statistically significant differences at $\alpha = 0.05$. These findings indicate that the distributions of quality metrics remained largely stable across epochs, supporting the reproducibility of results.

To go a little more into detail, for Campaign A, significant differences were found between Epochs 2 and 3 as well as between Epochs 3 and 4, suggesting temporary instability centered around Epoch 3, making Epoch 3 an outlier. This observation is consistent with the lower OA reported earlier in Section 4.1. Campaign B showed no significant differences

Epochs → Campaign ↓	1,2	1,3	1,4	1,5	2,3
A	0.387	0.055	0.579	0.701	**0.006**
B	0.995	0.358	0.334	0.601	0.362
C	0.113	**0.038**	0.190	0.661	0.607
Epochs → Campaign ↓	2,4	2,5	3,4	3,5	4,5
A	0.771	0.216	**0.017**	0.125	0.356
B	0.335	0.594	0.972	0.180	0.151
C	0.741	0.372	0.426	0.165	0.603

Table 7. Pairwise Mann–Whitney U test p-values across all campaigns and epochs. Bold values indicate statistically significant differences ($\alpha = 0.05$).

across any epoch combinations, with p-values consistently high (up to 0.972). This confirms the temporal consistency and reproducibility of this simpler geometric task. In Campaign C, only Epochs 1 and 3 showed a significant difference, consistent with minor variations observed earlier in the IoU distributions (Figure 8). All other pairs were statistically indistinguishable, indicating strong stability and thereby reproducibility across the complete study period spanning over five months.

6. Limitations

Despite the thorough and systematic evaluations presented, certain limitations of this study must be acknowledged. First, the analysis was conducted exclusively on the *Microworkers.com* platform, which may limit the generalizability of the findings to other crowdsourcing platforms. Different platforms vary in demographics, task presentation, and incentive structures, all of which can influence worker behavior and task outcomes. Consequently, the observed levels of consistency and reproducibility may not completely transfer to other platforms or contexts.

Second, the study incorporated three campaigns representing distinct task types: classification, point selection, and geometric outline acquisition. Although these tasks reflect common remote sensing applications, their differing levels of complexity and, as a result, cognitive demand may have led to variations in worker accuracy and efficiency. This variability may complicate direct cross-task comparisons and limit the generalization of findings beyond the examined task set. Furthermore, the overall sample size could be increased to further strengthen the generalizability of the findings.

Lastly, it remains uncertain whether the performance improvements observed among recurring workers can be attributed to temporal effects such as seasonal variations, a training effect resulting from repeated participation, or to a subset of workers who are inherently more motivated and engaged. Future research could benefit from experimental designs that isolate the effects of learning from those of worker motivation, providing a clearer insight into the mechanisms behind performance enhancements.

7. Summary & Conclusion

This study investigated the consistency and reproducibility of paid geospatial crowdsourcing through a five-month experiment on the platform *Microworkers.com*. Three task types representative of remote sensing were examined: classification (Campaign A), point selection (Campaign B), and polygon acquisi-

tion (Campaign C), thereby examining both quality as well as workforce in terms of reproducibility.

Despite the platform's large and diverse user base, a substantial proportion of recurring workers was observed, with only around 52% new workers by Epoch 5. The overall retention rate was approximately 23%, indicating that nearly a quarter of workers participated in at least two different epochs. Recurring workers generally outperformed those who participated in only one epoch. To evaluate changes in performance over time, workers were classified into low- and high-performing groups. The proportions of these groups remained relatively stable across epochs, with notable exceptions in Epoch 3 for the classification task and Epoch 5 for the geometric acquisition task.

The proportions of high- and low-performing workers were also reflected in the overall task quality, with certain epochs showing slightly lower quality. To evaluate potential reproducibility, statistical testing using the Mann-Whitney U test was conducted, confirming the previous findings and identifying specific outlier epochs. Out of 30 pairwise comparisons, only three yielded statistically significant differences ($\alpha = 0.05$): two in the classification task and one in the polygon acquisition task. The geometric point selection task, in contrast, showed no significant variation across epochs.

In conclusion, the consistency of results across epochs remained high, with only a few exceptions, regardless of the rather high worker retention rate. Overall, these results indicate that, despite worker retention effects and task variability, the crowd remained remarkably consistent; thereby providing an affirmative answer to the leading question posed in this study: *How consistent is the crowd?*

References

Chandler, J., Paolacci, G., Mueller, P., 2013. Risks and rewards of crowdsourcing marketplaces. *Handbook of human computation*, Springer, 377–392.

Cui, L., Chen, J., He, W., Li, H., Guo, W., Su, Z., 2021. Achieving approximate global optimization of truth inference for crowdsourcing microtasks. *Data Science and Engineering*, 6(3), 294–309.

Estellés-Arolas, E., 2022. Using crowdsourcing for a safer society: When the crowd rules. *European Journal of Criminology*, 19(4), 692–711.

Hirth, M., Hoßfeld, T., Tran-Gia, P., 2011. Anatomy of a crowdsourcing platform-using the example of microworkers. com. *2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing*, IEEE, 322–329.

Hossain, M., 2012. Users' motivation to participate in online crowdsourcing platforms. *2012 International Conference on Innovation Management and Technology Research*, IEEE, 310–315.

Karachiwalla, R., Pinkow, F., 2021. Understanding crowdsourcing projects: A review on the key design elements of a crowdsourcing initiative. *Creativity and innovation management*, 30(3), 563–584.

Kobayashi, M., Morita, H., Morishima, A., 2022. Efficient crowdsourcing for semantic segmentation considering human cognitive characteristics. *International Conference on Human-Computer Interaction*, Springer, 300–307.

Lenart-Gansiniec, R., Czakon, W., Sułkowski, Ł., Pocek, J., 2023. Understanding crowdsourcing in science. *Review of Managerial Science*, 17(8), 2797–2830.

MacFarland, T. W., Yates, J. M., 2016. Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R*, Springer, 103–132.

Mann, H. B., Whitney, D. R., 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Microworkers.com, 2024. Microworkers: A crowdsourcing platform. Accessed: 2025-10-28.

Nachar, N. et al., 2008. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1), 13–20.

Qarout, R., Checco, A., Demartini, G., Bontcheva, K., 2019. Platform-related factors in repeatability and reproducibility of crowdsourcing tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7, 135–143.

Sabou, M., Bontcheva, K., Scharl, A., 2012. Crowdsourcing research opportunities: lessons from natural language processing. *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 1–8.

Saralioglu, E., Gungor, O., 2022. Crowdsourcing-based application to solve the problem of insufficient training data in deep learning-based classification of satellite images. *Geocarto International*, 37(18), 5433–5452.

Sayin, B., Krivosheev, E., Yang, J., Passerini, A., Casati, F., 2021. A review and experimental analysis of active learning over crowdsourced data. *Artificial Intelligence Review*, 54, 5283–5305.

Vuurens, J., de Vries, A. P., Eickhoff, C., 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, 21–26.

Walter, V., Kölle, M., Collmar, D., 2022. A gamification approach for the improvement of paid crowd-based labelling of geospatial data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 113–120.

Wazny, K., 2018. Applications of crowdsourcing in health: an overview. *Journal of global health*, 8(1), 010502.

Williams, A., Goh, J., Willis, C., Ellison, A., Brusuelas, J., Davis, C., Law, E., 2017. Deja vu: Characterizing worker reliability using task consistency. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5, 197–205.

Zhang, J., 2022. Knowledge learning with crowdsourcing: A brief review and systematic perspective. *IEEE/CAA Journal of Automatica Sinica*, 9(5), 749–762.

Zheng, Y., Li, G., Li, Y., Shan, C., Cheng, R., 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5), 541–552.

Zhu, D., Carterette, B., 2010. An analysis of assessor behavior in crowdsourced preference judgments. *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 17–20.