



Universiteit  
Leiden  
The Netherlands

## **Assessing the quality of citizen science in archaeological remote sensing: results from the Heritage Quest project in the Netherlands**

Bourgeois, Q.; Kaptijn, E.; Verschoof-van der Vaart, W.; Lambers, K.

### **Citation**

Bourgeois, Q., Kaptijn, E., Verschoof-van der Vaart, W., & Lambers, K. (2024). Assessing the quality of citizen science in archaeological remote sensing: results from the Heritage Quest project in the Netherlands. *Antiquity*, 98(402), 1662-1678. doi:10.15184/aqy.2024.127





Version: Publisher's Version  
License: [Creative Commons CC BY-NC-ND 4.0 license](https://creativecommons.org/licenses/by-nc-nd/4.0/)  
Downloaded from: <https://hdl.handle.net/1887/4175237>

**Note:** To cite this publication please use the final published version (if applicable).



## Research Article

# Assessing the quality of citizen science in archaeological remote sensing: results from the Heritage Quest project in the Netherlands

Quentin Bourgeois<sup>1,\*</sup> , Eva Kaptijn<sup>2</sup> , Wouter Verschoof-van der Vaart<sup>3</sup>   
& Karsten Lambers<sup>1</sup> 

<sup>1</sup> Faculty of Archaeology, Leiden University, the Netherlands

<sup>2</sup> Het Oversticht, Zwolle, the Netherlands

<sup>3</sup> Netherlands Forensic Institute, The Hague, the Netherlands

\* Author for correspondence ✉ [q.p.j.bourgeois@arch.leidenuniv.nl](mailto:q.p.j.bourgeois@arch.leidenuniv.nl)



Volunteers are a key part of the archaeological labour force and, with the growth of digital datasets, these citizen scientists represent a vast pool of interpretive potential; yet, concerns remain about the quality and reliability of crowd-sourced data. This article evaluates the classification of prehistoric barrows on lidar images of the central Netherlands by thousands of volunteers on the Heritage Quest project. In analysing inter-user agreement and assessing results against fieldwork at 380 locations, the authors show that the probability of an accurate barrow identification is related to volunteer consensus in image classifications. Even messy data can lead to the discovery of many previously undetected prehistoric burial mounds.

Keywords: Western Europe, archaeological prospection, field survey, ground-truthing, data quality

## Introduction

The definition of citizen science is very broad and can encompass many differing degrees of public involvement (Gibb 2019). In recent years, the concept has often been applied to ‘crowd science’, where large groups of volunteers participate in scientific research (Heigl *et al.* 2019; Haklay *et al.* 2021). Although the term ‘citizen scientist’ has faced resistance in archaeology for potentially excluding specific groups, such as indigenous communities, it remains the most widely accepted term (Liebenberg *et al.* 2021). Archaeology has used

Received: 27 October 2023; Revised: 22 February 2024; Accepted: 6 April 2024

© The Author(s), 2024. Published by Cambridge University Press on behalf of Antiquity Publications Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial reuse or in order to create a derivative work.

citizen science for decades, pre-dating its current popularity; volunteer participation occurs at multiple levels in excavation projects and surveys (Smith 2014; Gibb 2019), and the emergence of large-scale, crowd-sourced projects has seen thousands of participants collaborating online. Examples of the latter include the registration of finds by metal-detectorists (Kars & Heeren 2018; Dobat *et al.* 2020) and the creation of peer-to-peer collaboration platforms (Wilkins 2020; Wernke *et al.* 2024). The popularity of high-profile projects in astrophysics such as Galaxy Zoo—where participants identified galaxies from shapes on telescope images (Willett *et al.* 2017)—and similar projects in other scientific fields (Jones *et al.* 2018, 2020) demonstrates the potential for such an approach in archaeology. Specifically, the analysis of high-resolution remote sensing images with the assistance of citizen scientists has gained in popularity (Duckers 2013; Lin *et al.* 2014). Such projects rely on multiple participants classifying a single image and then aggregating the results to obtain expert-quality datasets (Swanson *et al.* 2016).

Although these projects showcase the potential of volunteer scientists, concerns and criticisms regarding the quality of these datasets persist (Dickinson *et al.* 2012; Deckers *et al.* 2018). A significant obstacle is that the datasets generated by untrained and unsupervised volunteers are inherently messy and noisy (Kosmala *et al.* 2016; Swanson *et al.* 2016; Clare *et al.* 2019). This noisiness has led some researchers to dismiss crowd-sourced results as “analytically useless” (Casana 2020: 595). While we fundamentally disagree with this assessment, it must be acknowledged that the quality of crowd-sourced data requires critical assessment before further analysis can proceed (Balázs *et al.* 2021).

In this article, we address this criticism by testing and validating citizen science data drawn from our large-scale citizen science project called Heritage Quest (*Erfgoed Gezocht* in Dutch) where volunteers classified thousands of lidar images via an online platform (Lambers *et al.* 2019). We present our investigations into the overall quality of user classifications, inter-user agreement (consensus) and the quality of the classifications based on a ground-based survey.

## **Background: the Heritage Quest project**

Our citizen science project focuses on two regions in the central Netherlands that share similarities in terms of geology, land-use and history: the Veluwe and the Utrechtse Heuvelrug (Figure 1). These regions are characterised by ice-pushed ridges formed during the Saale glacial period (*c.* 400 000–130 000 years BP), which were subsequently partly covered by cover-sand deposits during the Weichselian glacial period (115 000–11 700 years BP) (Berendsen 2004). The result of these processes is an undulating landscape with significant variation in elevation. Forests and heathland covered the area from the Neolithic period (5500–2000 BC) onwards, surrounded by marshes and river valleys. Gradually increasing deforestation from prehistoric times (Doorenbosch 2013) expanded further in the Middle Ages (*c.* AD 1000–1500), leading to the formation of large drift-sand areas (Koster 2009). In the late nineteenth and early twentieth centuries, large parts of the area were reforested, resulting in today’s extensive forests interspersed with heathlands.

Both study regions contain well-preserved archaeological remains located either in heathland or under forest cover. Prehistoric barrows, Celtic field systems, charcoal kilns, hollow roads and *landweren* (long distance land boundaries dating to the Middle Ages) are among

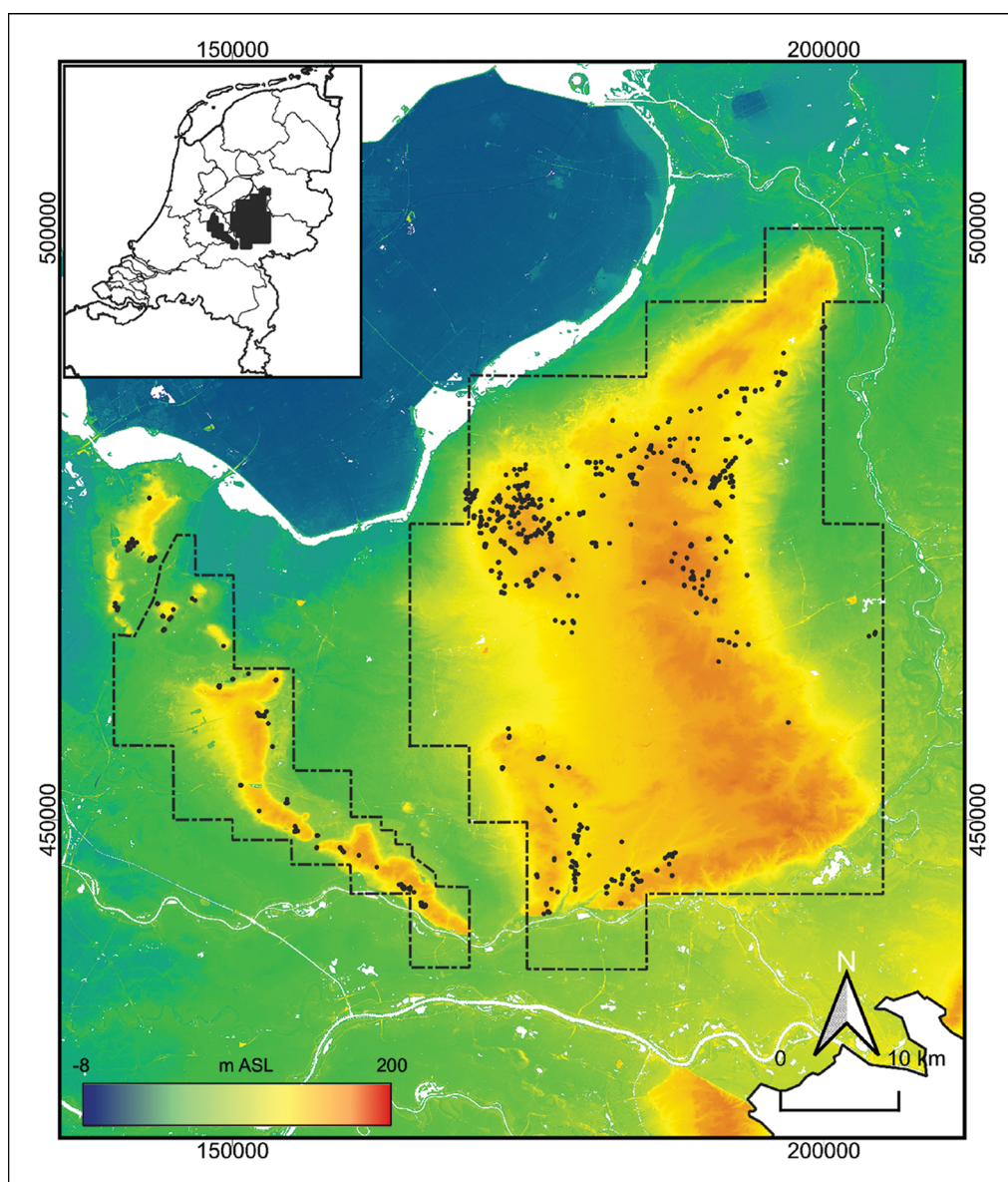


Figure 1. The Heritage Quest research areas (dashed outline, Utrechtse Heuvelrug on the left, the Veluwe on the right) on an elevation map of the Netherlands. Inset) location of research area (black squares) and known barrows (black dots) within the Netherlands (elevation model: Nationaal Georegister 2023; co-ordinates in Amersfoort/RD New, EPSG: 28992) (figure by authors).

the more common objects found in the area. In the field of computer vision, the term ‘feature’ refers to the properties of an image, while an ‘object’ refers to real-world entities (Travaglia *et al.* 2016: 14). Within this article the term ‘objects’ is therefore used for archaeological features, such as barrows. Lidar imagery is essential in detecting these archaeological objects, which are often obscured by dense vegetation cover.



To analyse the lidar imagery (generated from lidar data provided by the *Actueel Hoogtebestand Nederland*; Table 1) from both the Veluwe and the Utrechtse Heuvelrug, we used the Zooniverse, a web-based platform that allows people to participate in citizen science projects or ‘people-powered research’ without any specialised background, training or expertise (Simpson *et al.* 2014). In our Zooniverse project, Heritage Quest, we asked participants to mark any potential barrow, Celtic field, cart track and charcoal kiln within small lidar tiles (Figure 2). These tiles were obtained by dividing the lidar image of the entire area into tiles of 300 × 300m (600 × 600 pixels) with five per cent (30 pixels) overlap to all sides. Participants were presented with two different lidar visualisations (see Figure 2), shaded relief and a simple local relief model (Kokalj & Hesse 2017) to assist them in their classification. The first visualisation was more intuitive to interpret by volunteers, while the second allowed for better visibility of faint traces that would otherwise be difficult to distinguish on a shaded relief map. The latter visualisation improved the detection of Celtic fields significantly.

The user interface was bilingual Dutch/English, ensuring that international citizen scientists, as well as Dutch-speaking volunteers, could participate. Every lidar image was classified by at least 15 different users for the Veluwe and 60 users for the Utrechtse Heuvelrug. A dedicated staff member, assisted by a team of citizen scientists, monitored user-engagement and provided feedback and online support on an accompanying forum throughout the project. When a volunteer joined the project, they were provided with a short tutorial on how to operate the website and how to identify archaeological objects in the images. A comprehensive field guide (see Figure 2) was always available and included many examples of the objects we were interested in, tips on how to identify positive and negative examples (e.g. a roundabout on a road network compared to an actual prehistoric barrow), general background

information on the archaeological objects and the region under investigation, as well as an introduction to archaeological prospection and remote sensing in general.

The project was launched in May 2019 on the Veluwe and succeeded in mapping the entire 1780km<sup>2</sup> of this area in approximately five months. In total, 2063 users participated, between them classifying 396 552 tiles. Volunteers were asked to place points on the locations of potential barrows and charcoal kilns and to draw polygons covering areas where Celtic fields were detected. Each tile was classified by a minimum of 15 volunteers before being retired and removed from the set of available tiles. In April 2020, the Utrechtse

Table 1. Meta-information for the lidar imagery dataset, the so-called *Actueel Hoogtebestand Nederland* (Nationaal Georegister 2023).

Meta-information lidar data	
Purpose	Water management
Time of data acquisition	April 2010
Equipment	RIEGL LMS-Q680i Full-Waveform
Scan angle (whole FOV)	45°
Flying height above ground	600m
Speed of aircraft (TAS)	36m/s
Laser pulse rate	100 000Hz
Scan rate	66Hz
Strip adjustment	Yes
Filtering	Yes
Interpolation method	Moving planes
Point-density (pt per sq m)	6–10
DTM-resolution	0.5m

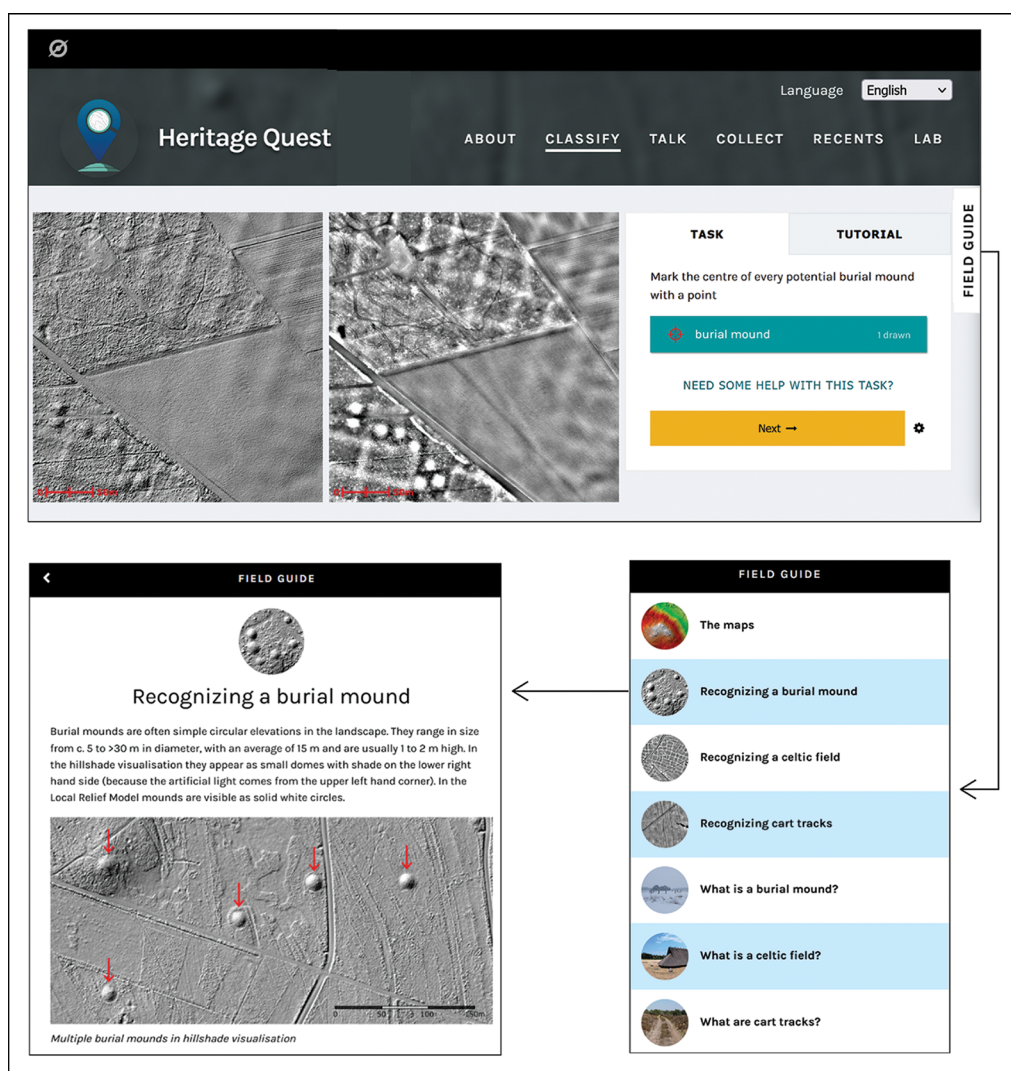


Figure 2. Top image) overview of the interface of the Heritage Quest project on the Zooniverse platform. The image shows both visualisations with a shaded relief image on the left and a simple local relief model on the right. Participants could click on either to mark locations. At any point they could also write comments or questions on this image in the forum. Lower images) segments of the more detailed field guide, providing in-depth information on detecting archaeological objects on lidar imagery (figure by authors).

Heuvelrug project launched with 4572 users who classified 300 971 tiles covering an area of 350km<sup>2</sup>. The workflow for the Utrechtse Heuvelrug focused on identifying barrows, Celtic field-systems and medieval cart tracks, rather than charcoal kilns, as preliminary observations of the images indicated that kilns were very rare in the region. We increased the number of classifications needed before a tile was retired, from 30 to 60, after more than 46 000 tiles were classified on the first day; even so, the entire area was investigated within a month.

Taking the results from both projects together, approximately 6.3 per cent of the entire land surface of the Netherlands was classified.

Aggregating classifications in Heritage Quest

A significant challenge in any citizen science project involving large numbers of participants is creating meaningful data from inherently noisy sources (Swanson *et al.* 2016; Rosenthal *et al.* 2018). Most of the volunteers who participated in our project were not trained as archaeologists; certainly not as remote sensing specialists knowledgeable in the detection of archaeological objects on lidar imagery. Nevertheless, through the repetition of research, the quality of the output data can be increased (Swanson *et al.* 2016; Rosenthal *et al.* 2018). When each image is investigated by multiple participants, a consensus is reached (inter-user agreement). Objects that fit the requirements given to the citizen scientists in the tutorial will be identified by more participants, while more enigmatic objects will be selected by fewer participants. Moreover, accidental errors are easily filtered out as two volunteers will rarely mark the (exact) same location by accident.

In this article, we focus on aggregating the results from barrow classifications. In total, 222 699 individual barrow classifications were registered across both areas (see Table 2). These individual classifications consist of some errors, or misidentifications, but also of locations that were clicked on dozens of times. To aggregate the results of the online project, the data from the Zooniverse (i.e. all individual classifications or ‘clicks’) were converted into geo-spatial entities (i.e. points) with real-world co-ordinates and an additional field (count), containing a single integer (1). Subsequently, a chain of processing tools in the open-access geographic information system software QGIS (v. 3.16, QGIS Development Team 2017) was used to turn these points into polygons incorporating the number of classifications per polygon (Algorithm 1): all points were buffered (with an empirically determined 4m radius) and dissolved to turn them into polygons. These polygons were converted into

Table 2. Factsheet of the results from the Heritage Quest project in both regions.

	The Veluwe region	The Utrechtse Heuvelrug region
Period	May 2019 – September 2019	April 2020 – May 2020
Users	2063	4572
Area (km <sup>2</sup> )	1780	350
Tiles analysed	396 552	300 971
Total number of unique tiles	23 598	5671
Retirement rate	15	60
Total detections	131 874	182 740
Barrows	88 783	133 916
Celtic fields	6728	13 266
Charcoal kilns	26 363	–
Cart tracks	–	35 558

Note that charcoal kilns were researched in the Veluwe region, but since we did not expect these to be common on the Utrechtse Heuvelrug we asked volunteers to mark medieval cart tracks instead. The latter feature was often flagged by volunteers during the Veluwe project in the forum and we therefore decided to mark these as well.

individual, separate features with the Multipart\_to\_singleparts processing tool. Finally, the number of classifications (points) within these polygons was counted through the Join\_attributes\_by\_location (summary) tool (Figure 3).

An analysis of the application of the above workflow demonstrates its usability, even in areas where barrows are close together (Figure 4A–C). Occasionally this processing results in the lumping of disparate objects in very close proximity (within 4m) into a single location (Figure 4D) or the artificial separation of a single location into multiple locations when the classifications are overly spread out (more than 4m away; Figure 4A). These issues occur rarely, however, as the average diameter, shape and form of the barrows in the investigated areas are quite stable (see Bourgeois 2013).

The resulting aggregation created 77 014 individual consensus locations in the Veluwe (37 812) and the Utrechtse Heuvelrug (39 202). These consist of locations identified by between one and 243 individual users, although the higher end of this scale was only reached a few times, if an object was located in the overlap zone between different tiles.

To ensure the accuracy of the aggregated classifications, we corrected the data on the basis of land-use. Two processes have had a negative impact on the preservation and visibility of barrows in the Netherlands. Firstly, (post)medieval agriculture, urbanisation and infrastructure development have systematically erased all above-ground traces of barrows in these regions. Secondly, areas affected by erosion and sedimentation caused by wind (drift-sand) exhibit a negative correlation with the presence of barrows (Bourgeois 2013: 40–47; Verschoof-van der Vaart *et al.* 2020). Almost all preserved prehistoric barrows in the central Netherlands have been found in areas covered with forest or vegetation such as natural grasslands or heather.

As such, we expected to find new potential barrows only in forested areas, heath- or grasslands, which were not affected by drift-sand. The chance of discovering unknown above-ground barrows outside of these areas is close to zero (Verschoof-van der Vaart *et al.* 2020). Therefore, we removed all consensus locations outside of areas covered with forest, heather or grasslands, using the map on present day land-use created by *Centraal Bureau voor de Statistiek* (Statistics Netherlands) (Nationaal Georegister 2023; Figure 5). In most cases the inter-user agreement in these areas was low and in total consisted of 22–30 per cent of the overall consensus locations. We also eliminated all locations in known drift-sand areas based on the geomorphological map of the Netherlands 2019 by the *Basisregistratie*

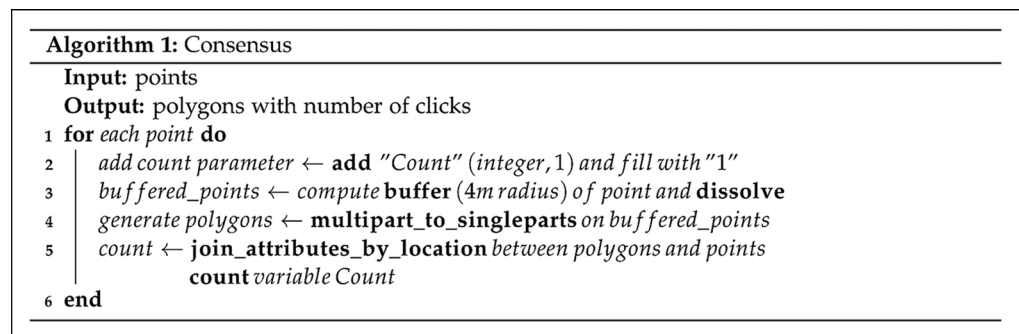


Figure 3. Algorithm showing the aggregation process in QGIS (figure by authors).



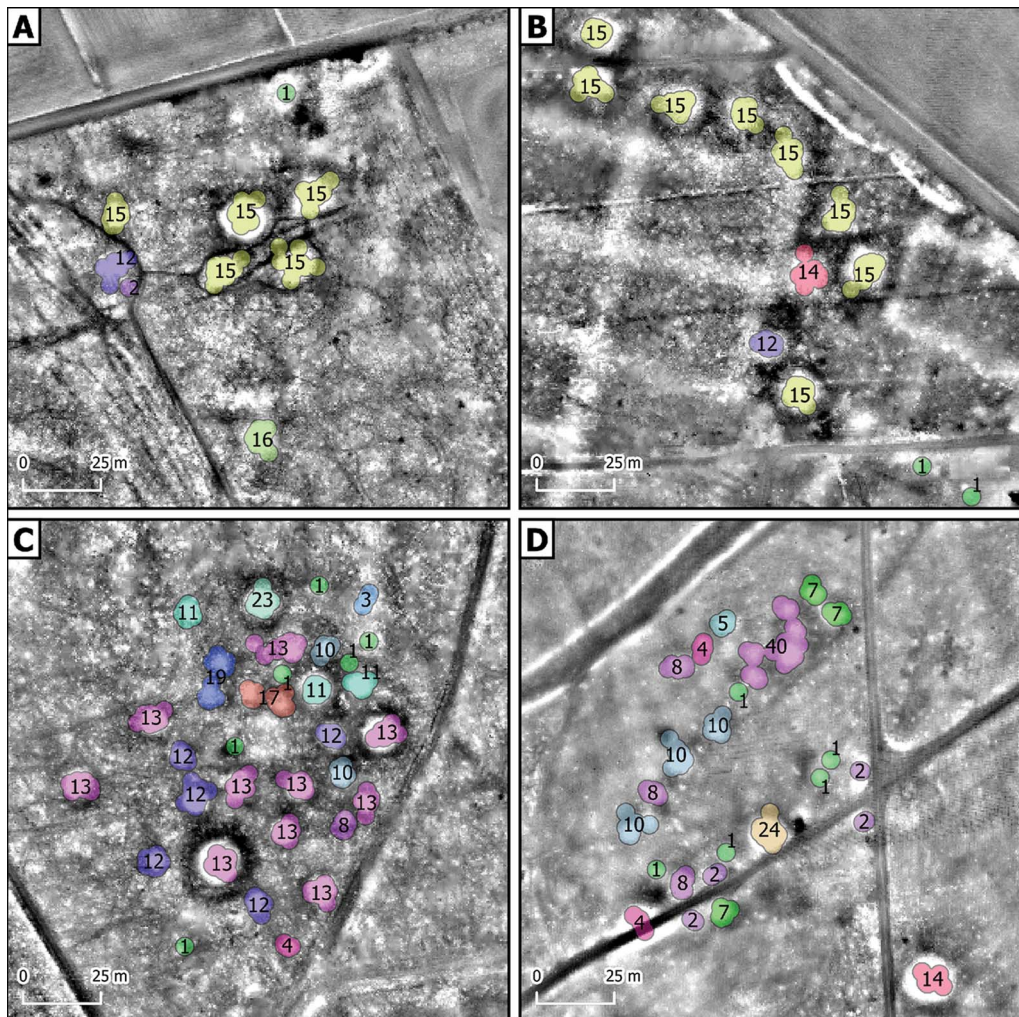


Figure 4. Results from aggregation at dense concentrations of potential barrows: A) aggregation results from a closely spaced group of barrows—all locations with 12 or more classifications are known burial mounds; B) aggregation results of a group of burial mounds within a Celtic field—volunteers were able to correctly identify all known burial mounds (12 or more classifications); C) aggregation results of a previously known urnfield on the Veluwe consisting of at least 28 or more low burial mounds (Verlinde & Hulst 2010; fig. 53). The volunteers have potentially discovered a much larger number of burial mounds than previously known. D) aggregation results of a previously known line of burial mounds. All previously known barrows have been identified, as well as a number of previously unknown mounds. Note that here, the classifications tend to blend into one another if the mounds are close, with closely spaced objects being added together (i.e. providing 24 or even 40 classifications) (figure by authors).

*Ondergrond* (Nationaal Georegister 2023) and the recent drift-sand map by the *Bryologische en Lichenologische Werkgroep* and Wageningen University (Sparrius & Riksen 2019). Our fieldwork identified several additional areas of drift-sand and, as all consensus locations we investigated in drift-sand areas were determined to be sand dunes (Verschoof-van der Vaart *et al.* 2022a), consensus locations were also removed from these newly identified

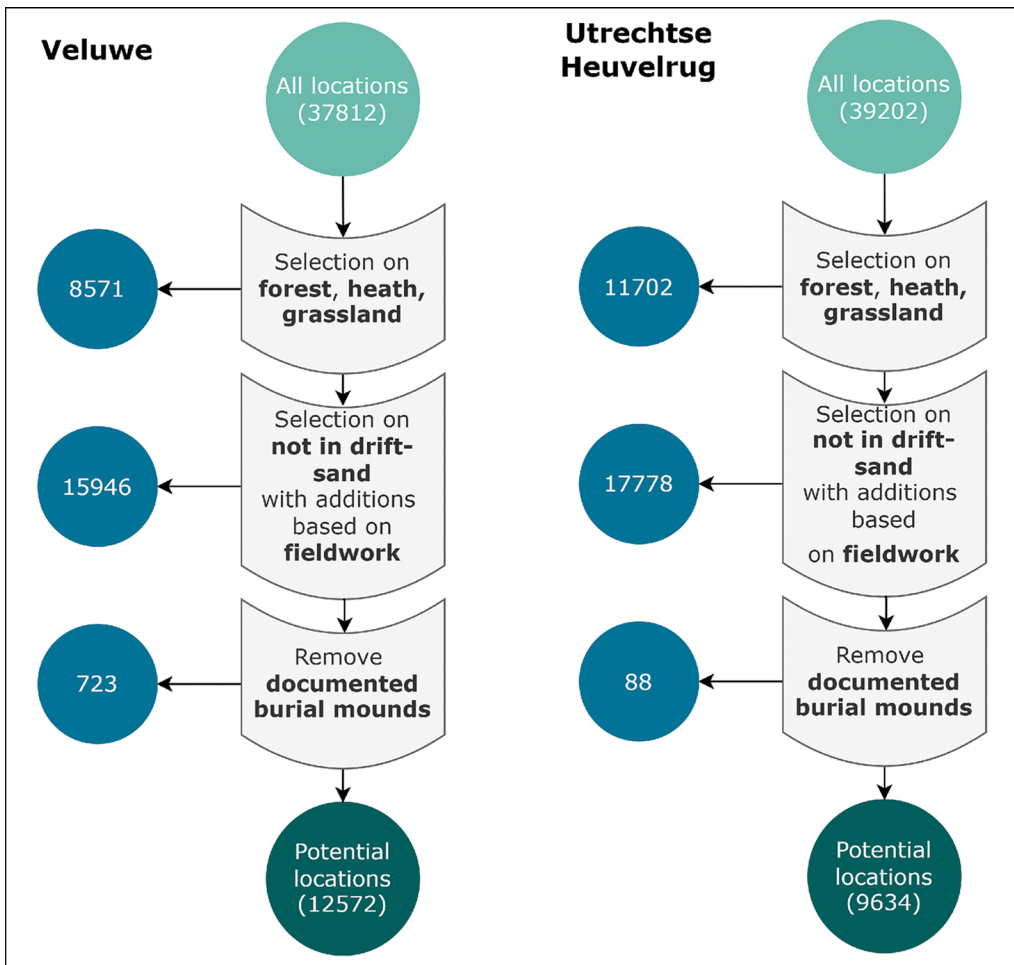


Figure 5. Workflow illustrating the selection processes to identify potential newly discovered locations within the dataset (figure by authors).

areas. Finally, we removed all already known and recorded barrows using the *Archeologisch Monumenten Register* (AMR2) from the State Service for Cultural Heritage (RCE). In total, we retained 22 206 consensus locations (Figure 5).

## Assessing the quality of the consensus locations

The main challenge we faced was assessing the overall quality of the remaining 22 000-plus consensus locations. Evaluating the data based on known barrows in the region proved problematic because these are primarily scheduled monuments that have been restored by adding a layer of material, often sand, on top of the mound. This increases their height significantly, while vegetation is regularly cleared making these restored mounds significantly easier to detect on lidar imagery than the unknown (and therefore unrestored) barrows. Thus, the

detection of these known barrows does not inform us how well the volunteers were able to correctly distinguish natural or modern topographical features from ancient barrows. Using these known barrows to assess the quality would only inform us if the consensus locations conformed to already known and restored prehistoric barrows.

To address this issue, we, together with students and volunteers, investigated 219 individual consensus locations in the field to establish whether they were prehistoric barrows. Using Mergin Maps, an open-source app that uses the internal GPS of a smartphone to present geo-spatial data, we located the consensus locations. We then collected at least three hand coring samples from across each consensus location: one outside, one in the flank and one (slightly off) the centre of the elevation. Coring is required as a simple visual inspection is often not sufficient for the identification of barrows. Many natural topographical features look very similar to burial mounds and present-day dense vegetation obscures their overall shape, making coring and inspection of the internal build-up of the feature necessary for identification. Based on the coring data, each individual consensus location was then interpreted as either an ancient anthropogenic mound or a natural mound, the latter almost always either wind-blown deposits or glacial meltwater relics. Earlier fieldwork by other researchers had previously assessed 161 of our consensus locations either through coring or small test-trenches. These were added to our dataset, resulting in a total of 380 investigated consensus locations, of which 136 were located in the Utrechtse Heuvelrug and 244 in the Veluwe, with the locations distributed throughout landscapes of different geomorphological types. The results of the fieldwork are published elsewhere (Verschoof-van der Vaart *et al.* 2022a, b & c).

Of these 380 consensus locations, 226 were interpreted as barrows, providing an average precision (equal to true positives divided by the sum of true positives and false positives; Verschoof-van der Vaart *et al.* 2020) of 0.59. However, this performance changes significantly when the number of volunteers that agreed on a given consensus location is taken into account (i.e. the inter-user agreement; Figure 6). For consensus locations with an inter-user agreement less than seven, the precision (here calculated as the ratio of barrows versus natural topographic features) fluctuates between 0.21 and 0.41. The precision significantly increases to 0.75 for an inter-user agreement of 10, and further improves to around 0.82 for an inter-user agreement of 12 or greater. An inter-user agreement of 15 or more results in a precision of 0.85 of investigated consensus locations being a prehistoric barrow.

## Discussion

The results of our study demonstrate that citizen scientists can effectively identify prehistoric barrows in lidar imagery, particularly when there is a high level of inter-user agreement. Extrapolation of the precision from our ground-truthed subsample to all consensus locations indicates that the volunteers have likely discovered about a thousand potential barrows—and this is if we only consider the consensus locations identified by more than seven volunteers (Figure 7). These have a greater than 50 per cent chance of indeed being prehistoric barrows. This figure would double the current number of known barrows in the central Netherlands (Bourgeois 2013) and has far reaching implications for research into the communities who built these mounds, as well as for heritage management.

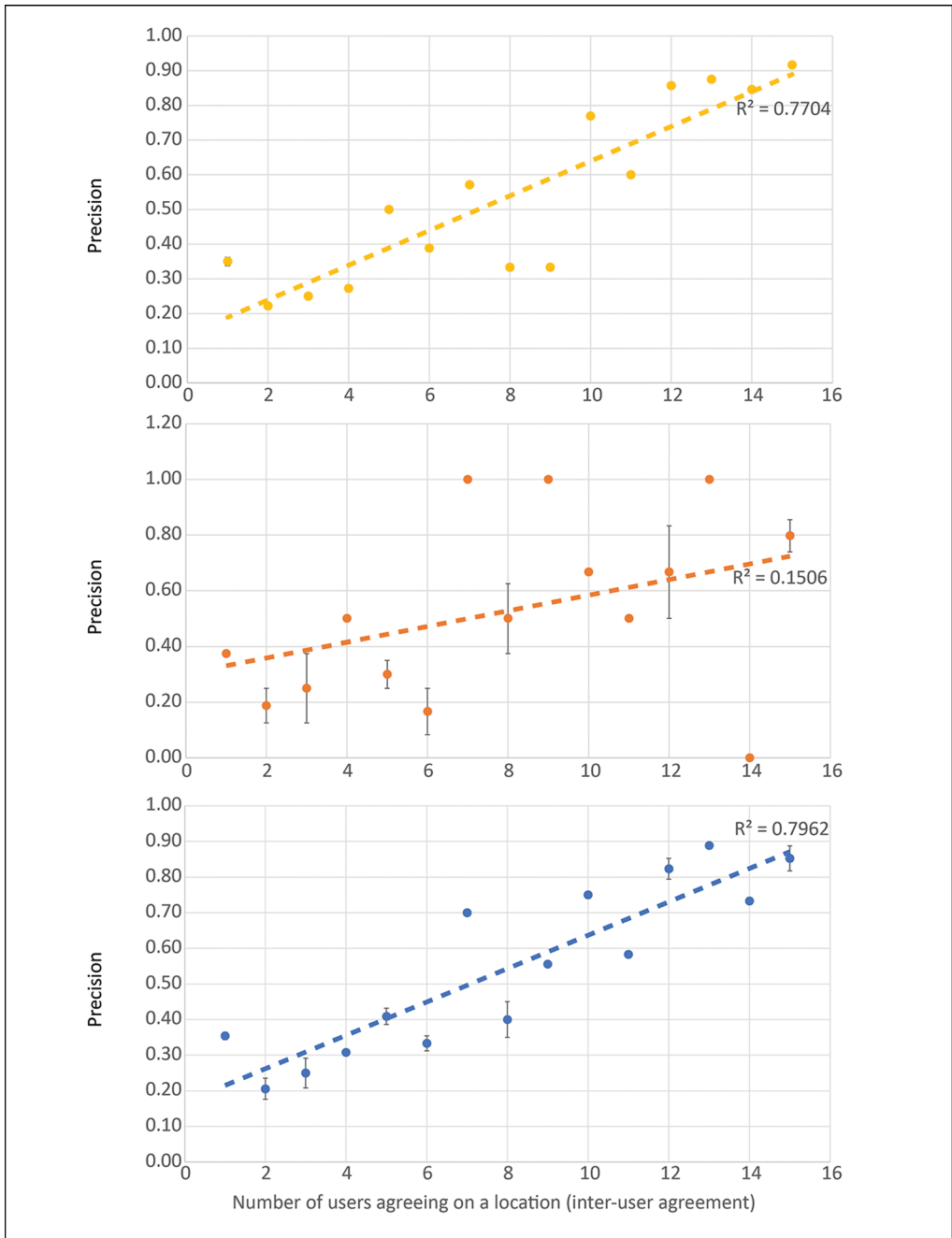


Figure 6. Precision versus inter-user agreement based on fieldwork validated consensus locations. The top panel shows the precision for the Veluwe, the centre panel for the Utrechtse Heuvelrug and the bottom panel the overall precision for the entire project. The error bars indicate examples where the anthropogenic nature of the mound could be established, but not conclusively if it represented a barrow. Note that the 15 classifications also contain objects that have been classified by more than 15 contributors due to overlap between images, aggregation of multiple objects into one, or in the case of the Utrechtse Heuvelrug project, due to a higher retirement rate of the image (figure by authors).



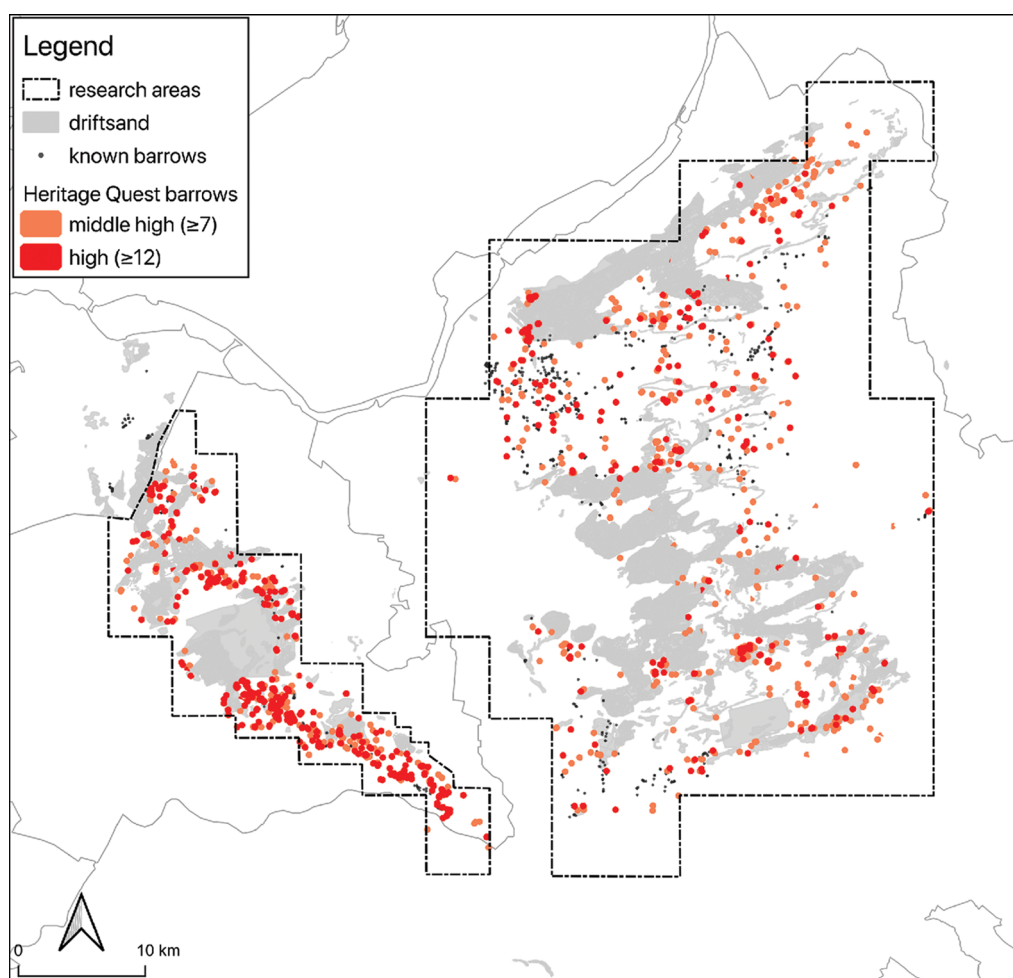


Figure 7. Overview of the locations of all new potential barrows with high (red points) or middle high (orange points) probabilities, known barrows (small black points) and drift-sand (grey shaded areas) (co-ordinates in Amersfoort/RD New, EPSG: 28992) (figure by authors).

However, we also highlight the challenge with using crowd-sourced data for archaeological research; in particular, the noisy nature of the data makes it difficult to assess the quality of the consensus locations. The distinction between natural and anthropogenic elevations on lidar imagery is often difficult to make, even for trained archaeologists (Banaszek *et al.* 2018). Our fieldwork aimed to address this and allowed us to determine the precision of classifications and how this varied depending on the level of inter-user agreement. By validating a sample of the consensus locations, we could determine that the precision increased substantially as the level of inter-user agreement increased. This suggests that our consensus-based approach can be used to complement traditional, expert-led field-survey methods that might otherwise be too time consuming or expensive to cover a similar area. In our project, investigating the dense forests of the Veluwe and Utrechtse Heuvelrug in their entirety with traditional

fieldwork would simply not be possible. Furthermore, expert-led field survey or desk-based assessment may also result in incomplete and error-prone datasets (e.g. Kaptijn 2009: 42–59). Studies using satellite images as base data for archaeological prospection have shown how mapping results may vary considerably between archaeologists (Sadr 2016) depending on their level of training and experience (Snyder & Haas 2024). Thus, independent data, either collected by citizen researchers as in Heritage Quest or by computational approaches (Verschoof-van der Vaart & Lambers 2022), are an important complementary data source for heritage management, as they often include archaeological objects not found by experts. The map we have now generated can be effectively used by local stakeholders, such as municipalities and the state forestry service, to monitor and manage these important features.

Although precision increased with higher levels of inter-user agreement, even low consensus locations still had a non-zero chance of being prehistoric barrows. As the results show, if only one volunteer classified a location as a barrow, it still had a 35 per cent chance of being a barrow. However, it is worth noting that this number is probably inflated owing to the discovery of a Late Prehistoric urnfield during our fieldwork. In this case, the urnfield consisted of a dozen very small mounds of only 0.2–0.3m in elevation that most volunteers did not classify as barrows, yet one of our most prolific users—who classified more than 20 000 images—did identify the small and low mounds as potential barrows. To avoid confusion, the tutorial focused on isolated barrows and omitted the much more difficult to recognise urnfield mounds. Correcting for this urnfield, the actual probability of a location marked only once being a barrow will be lower, and more towards 15–20 per cent. Nevertheless, the discovery of this urnfield suggests that there may be certain types of barrows that are particularly challenging to identify for volunteers, and in fact even for experts. This is an important caveat to keep in mind, and while the consensus-based approach used in this study was effective at identifying a large number of prehistoric barrows, it is possible that some were still missed. Future research could explore ways to improve the detection of these more challenging types of barrows, perhaps through the use of more advanced survey methods or machine learning algorithms.

The latter could be a promising avenue to improve the results obtained through a volunteer-based approach such as the one we have developed. Previously, we have argued for such an approach where human expertise and machine learning algorithms could be used in a collaborative, integrated and iterative process (Lambers *et al.* 2019). We feel this could help improve accuracy and the identification process while still leveraging the knowledge and experience of archaeologists. Such a human-in-the-loop approach could also overcome some of the limitations facing machine learning algorithms, in particular the recognition of patterns that are novel or outside the norm (Verschoof-van der Vaart & Lambers 2022). Within our project we only asked volunteers to classify three different object classes but they were also able to flag specific tiles with their own tags and add them to the forum. One of the categories that they systematically flagged were traces from a new type of prehistoric land parcelling system, comparable to Celtic fields, which was hitherto unknown but suggests landscape organisation on a much larger scale than previously believed and the need for an upward revision of population density estimates (Arnoldussen *et al.* 2022).

Finally, our study has broader implications for citizen science and public engagement within archaeology. By involving thousands of volunteers in the research process, this project not only generated valuable data but also helped to raise awareness and interest in archaeology. Many of our volunteers have become vocal ambassadors of their local archaeological heritage. Local initiatives have been implemented in several of the provinces and municipalities involved, from teaching prehistory and lidar technology in schools (<https://www.reizenindetijd.nl/themas/mijn-huis-staat-in/kaartkijken-met-de-ahn/>) to nature conservation groups incorporating traces of prehistoric landscapes in their excursions. Information on the project also features in several popular books (Neefjes & Bleumink 2021; Tonk 2022). This highlights the potential of citizen science not only as a research tool but also as a means of fostering public engagement and interest in archaeology.

## Conclusion

Results from our study suggest that a volunteer-based approach to identifying prehistoric barrows complements traditional, expert-led field survey methods. With higher inter-user agreement, the accuracy of identifying prehistoric barrows increases significantly. Based on our fieldwork and an assessment of the quality of the consensus locations we can establish that the precision increases above 0.82 with an inter-user agreement of 12 or higher. Even a lower consensus of 7 or higher still has a precision of 0.7 on average. Taking our results into consideration we have arguably doubled the number of known barrows within the research area. Furthermore, our approach proved to be cost-effective in analysing a large research area in a relatively short period of time, producing results that are of fundamental importance for research, heritage planning and wider community awareness of local heritage.

## Acknowledgements

We thank all citizen scientists who devoted so much of their time to the Heritage Quest project. They are the foundation of this research.

## Funding statement

The Heritage Quest project is a collaboration between Leiden University, Erfgoed Gelderland and Landschap Erfgoed Utrecht and has been funded by the provinces of Gelderland and Utrecht, the Fonds voor Cultuurparticipatie, the Cultuur- en Erfgoedpact Noord Veluwe, the National Park Utrechtse Heuvelrug, the Municipality of Arnhem and the KF Heinfonds. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google and a grant from the Alfred P. Sloan Foundation.

## References

- |  |   |
|--|---|
| ARNOLDUSSEN, S., W.B. VERSCHOOF-VAN DER VAART, E. KAPTJIN & Q.P.J. BOURGEOIS. 2022. Field systems and later prehistoric land use: new insights into land use detectability and | palaeodemography in the Netherlands through LiDAR, automatic detection and traditional field data. <i>Archaeological Prospection</i> 30: 283–300. <a href="https://doi.org/10.1002/arp.1891">https://doi.org/10.1002/arp.1891</a> |
|--|---|

- BALÁZS, B., P. MOONEY, E. NOVÁKOVÁ, L. BASTIN, J. JOKAR ARSANJANI. 2021. Data quality in citizen science, in K. Vohland *et al.* (ed.) *The science of citizen science*: 139–57. Cham: Springer.  
[https://doi.org/10.1007/978-3-030-58278-4\\_8](https://doi.org/10.1007/978-3-030-58278-4_8)
- BANASZEK, Ł., D.C. COWLEY & M. MIDDLETON. 2018. Towards national archaeological mapping. Assessing source data and methodology—a case study from Scotland. *Geosciences* 8.  
<https://doi.org/10.3390/geosciences8080272>
- BERENDSEN, H.J.A. 2004. *De Vorming van het Land. Inleiding in de Geologie en de Geomorfologie*. Assen: Van Gorcum.
- BOURGEOIS, Q. 2013. *Monuments on the horizon. The formation of the barrow landscape throughout the 3rd and 2nd millennium BC*. Leiden: Sidestone.
- CASANA, J. 2020. Global-scale archaeological prospection using CORONA satellite imagery: automated, crowd-sourced, and expert-led approaches. *Journal of Field Archaeology* 45: S89–S100.  
<https://doi.org/10.1080/00934690.2020.1713285>
- CLARE, J.D.J. *et al.* 2019. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? *Ecological Applications* 29.  
<https://doi.org/10.1002/eap.1849>
- DECKERS, P., A. DOBAT, N. FERGUSON, S. HEEREN, M. LEWIS & S. THOMAS. 2018. The complexities of metal detecting policy and practice: a response to Samuel Hardy, ‘Quantitative analysis of open-source data on metal detecting for cultural property’ (Cogent Social Sciences 3, 2017). *Open Archaeology* 4: 322–33.  
<https://doi.org/10.1515/opar-2018-0019>
- DICKINSON, J.L. *et al.* 2012. The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment* 10: 291–97.  
<https://doi.org/10.1890/110236>
- DOBAT, A.S., P. DECKER, S. HEEREN, M. LEWIS, S. THOMAS & A. WESSMAN. 2020. Towards a cooperative approach to hobby metal detecting: the European public finds recording network (EPFRN) vision statement. *European Journal of Archaeology* 23: 272–92.  
<https://doi.org/10.1017/ear.2020.1>
- DOORENBOSCH, M. 2013. *Ancestral heaths. Reconstructing the barrow landscape in the central and southern Netherlands*. Leiden: Sidestone.
- DUCKERS, G.L. 2013. Bridging the “geospatial divide” in archaeology: community based interpretation of LIDAR data. *Internet Archaeology* 35.  
<https://doi.org/10.11141/ia.35.10>
- GIBB, J.G. 2019. Citizen science: case studies of public involvement in archaeology at the Smithsonian Environmental Research Center. *Journal of Community Archaeology & Heritage* 6: 3–20.  
<https://doi.org/10.1080/20518196.2018.1549815>
- HAKLAY, M., D. DÖRLER, F. HEIGL, M. MANZONI, S. HECKER & K. VOHLAND. 2021. What is citizen science? The challenges of definition, in K. Vohland *et al.* (ed.) *The science of citizen science*: 13–33. Cham: Springer.  
[https://doi.org/10.1007/978-3-030-58278-4\\_2](https://doi.org/10.1007/978-3-030-58278-4_2)
- HEIGL, F., B. KIESLINGER, K.T. PAUL, J. UHLIK & D. DÖRLER. 2019. Toward an international definition of citizen science. *Proceedings of the National Academy of Sciences USA* 116: 8089–92.  
<https://doi.org/10.1073/pnas.1903393116>
- JONES, F.M. *et al.* 2018. Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project. *Scientific Data* 5.  
<https://doi.org/10.1038/sdata.2018.124>
- JONES, F.M., C. ARTETA, A. ZISSERMAN, V. LEMPITSKY, C.J. LINTOTT & T. HART. 2020. Processing citizen science- and machine-annotated time-lapse imagery for biologically meaningful metrics. *Scientific Data* 7.  
<https://doi.org/10.1038/s41597-020-0442-6>
- KAPTIJN, E. 2009. *Life on the watershed. Reconstructing subsistence in a steppe region using archaeological survey: a diachronic perspective on habitation in the Jordan Valley*. Leiden: Sidestone.
- KARS, M. & S. HEEREN. 2018. Archaeological small finds recording in the Netherlands: the framework and some preliminary results of the Project Portable Antiquities of the Netherlands (PAN). *Medieval Settlement Research* 33: 18–27.  
<https://doi.org/10.5284/1059014>
- KOKALJ, Ž. & R. HESSE. 2017. *Airborne laser scanning raster data visualization: a guide to good practice*. Ljubljana: Research Center of the Slovenian Academy of Sciences and Arts.  
<https://doi.org/10.3986/9789612549848>
- KOSMALA, M., A. WIGGINS, A. SWANSON & B. SIMMONS. 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the*



- Environment* 14: 551–60.  
<https://doi.org/10.1002/fee.1436>
- KOSTER, E.A. 2009. The “European aeolian sand belt”: geoconservation of drift sand landscapes. *Geoheritage* 1: 93–110.  
<https://doi.org/10.1007/s12371-009-0007-8>
- LAMBERS, K., W. VERSCHOOF-VAN DER VAART & Q. BOURGEOIS. 2019. Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection. *Remote Sensing* 11.  
<https://doi.org/10.3390/rs11070794>
- LIEBENBERG, L. *et al.* 2021. Tracking science: an alternative for those excluded by citizen science. *Citizen Science: Theory and Practice* 6: 1–16.  
<https://doi.org/10.5334/cstp.284>
- LIN, A.Y., A. HUYNH, G. LANCKRIET & L. BARRINGTON. 2014. Crowdsourcing the unknown: the search for Genghis Khan. *PLoS ONE* 9.  
<https://doi.org/10.1371/journal.pone.0114046>
- Nationaal Georegister. 2023. Publieke Dienstverlening op de Kaart (PDOK). Available at: <https://www.pdok.nl/> (accessed 7 April 2023).
- NEEFJES, J. & H. BLEUMINK. 2021. *De Veluwe. Biografie van het grootste natuurlandschap van Nederland*. Wageningen: Blauwdruk.
- QGIS Development Team. 2017. QGIS Geographic Information System. Open Source Geospatial Foundation Project. Available at: <http://qgis.org>
- ROSENTHAL, I.S. *et al.* 2018. Floating forests: quantitative validation of citizen science data generated from consensus classifications. *ArXiv*. Available at: <https://doi.org/10.48550/arXiv.1801.08522> (accessed 1 June 2023).
- SADR, K. 2016. The impact of coder reliability on reconstructing archaeological settlement patterns from satellite imagery: a case study from South Africa. *Archaeological Prospection* 23: 45–54.  
<https://doi.org/10.1002/arp.1515>
- SIMPSON, R., K.R. PAGE & D. DE ROURE. 2014. Zooniverse: observing the world’s largest citizen science platform, in C-W. Chung *et al.* (ed.) *Proceedings of the 23rd International Conference on World Wide Web*: 1049–54. New York: Association for Computing Machinery.  
<https://doi.org/10.1145/2567948.2579215>
- SMITH, M.L. 2014. Citizen science in archaeology. *American Antiquity* 79: 749–62.  
<https://doi.org/10.7183/0002-7316.79.4.749749>
- SNYDER, T.J. & R. HAAS. 2024. Unstructured satellite survey detects up to 20% of archaeological sites in coastal valleys of southern Peru. *PLoS ONE* 19.  
<https://doi.org/10.1371/journal.pone.0292272>
- SPARRIUS, L. & M. RIKSEN. 2019. *Evaluatie van elf jaar stuifzandbeheer op de Veluwe 2007–2018* (BLW/G Rapport 23). Wageningen: Bryologische en Lichenologische Werkgroep.
- SWANSON, A., M. KOSMALA, C. LINTOTT & C. PACKER. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology* 30: 520–31.  
<https://doi.org/10.1111/cobi.12695>
- TONK, F. 2022. *Het verhaal van Nederland. Onze geschiedenis van de prehistorie tot nu*. Amsterdam: Nieuw Amsterdam.
- TRAVIGLIA A., D. COWLEY & K. LAMBERS. 2016. Finding common ground: human and computer vision in archaeological prospection. *AARGnews. The newsletter of the Aerial Archaeology Research Group* 53: 11–24.
- VERLINDE, A.D. & R.S. HULST. 2010. *De grafvelden en grafvondsten op en rond de Veluwe van de Late Bronstijd tot in de Midden-IJzertijd*. (Nederlandse Archeologische Rapporten 39). Amersfoort: Rijksdienst voor het Cultureel Erfgoed.
- VERSCHOOF-VAN DER VAART, W.B. & K. LAMBERS. 2022. Applying automated object detection in archaeological practice: a case study from the southern Netherlands. *Archaeological Prospection* 29: 15–31. <https://doi.org/10.1002/arp.1833>
- VERSCHOOF-VAN DER VAART, W.B., K. LAMBERS, W. KOWALCZYK & Q.P.J. BOURGEOIS. 2020. Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from the Netherlands. *ISPRS International Journal of Geo-Information* 9.  
<https://doi.org/10.3390/ijgi9050293>
- VERSCHOOF-VAN DER VAART, W.B., E. KAPTIJN & Q.P.J. BOURGEOIS. 2022a. Archeologisch veldonderzoek Erfgoed Gezocht Veluwe. Onderzoeksgebied Schaarsbergen, gemeente Arnhem. Unpublished report for the Faculty of Archaeology, Leiden University.
- 2022b. Archeologisch veldonderzoek Erfgoed Gezocht Utrechtse Heuvelrug. Onderzoeksgebied Elst, gemeente Utrechtse Heuvelrug & Rhenen. Unpublished report for the Faculty of Archaeology, Leiden University.

- 2022c. Archeologisch veldonderzoek Erfgoed Gezocht Veluwe. Onderzoeksgebieden Noord Veluwe. Unpublished report for the Faculty of Archaeology, Leiden University.
- WERNKE, S. *et al.* 2024. Large-scale, collaborative imagery survey in archaeology: the Geospatial Platform for Andean Culture, History and Archaeology (GeoPACHA). *Antiquity* 98: 155–71.  
<https://doi.org/10.15184/aqy.2023.177>
- WILKINS, B. 2020. Designing a collaborative peer-to-peer system for archaeology: the DigVentures platform. *Journal of Computer Applications in Archaeology* 3: 33–50.  
<https://doi.org/10.5334/jcaa.34>
- WILLETT, K.W. *et al.* 2017. Galaxy Zoo: morphological classifications for 120 000 galaxies in HST legacy imaging. *Monthly Notices of the Royal Astronomical Society* 464: 4176–203.  
<https://doi.org/10.1093/mnras/stw2568>