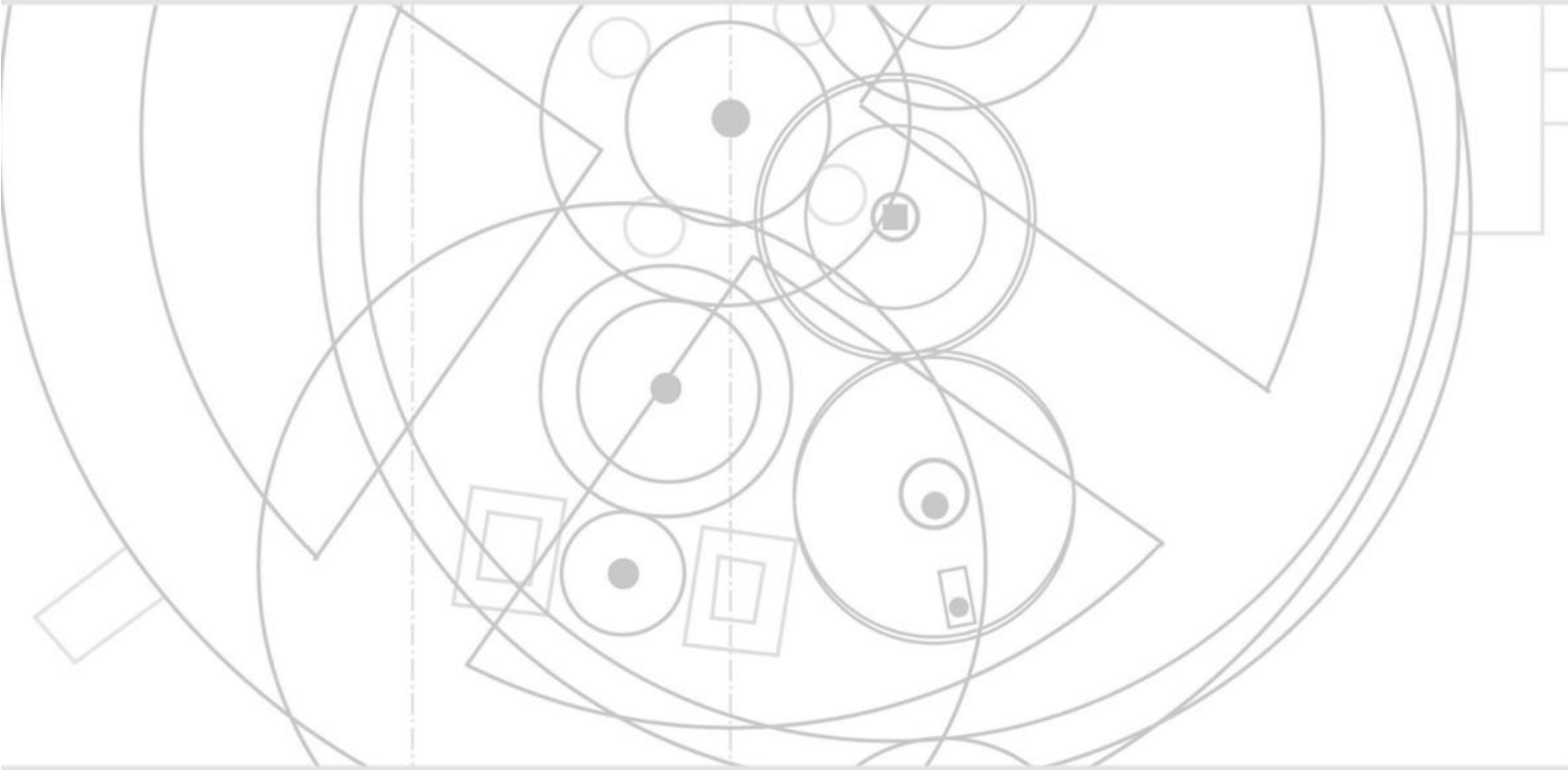
NEW COMPUTATIONAL METHODS FOR AUTOMATED LARGE-SCALE ARCHAEOLOGICAL SITE DETECTION

Iban Berganzo Besga

UNIVERSITAT
ROVIRA i VIRGILI

New Computational Methods for Automated Large-Scale Archaeological Site Detection

IBAN BERGANZO-BESGA



DOCTORAL THESIS
2023

UNIVERSITAT ROVIRA I VIRGILI
NEW COMPUTATIONAL METHODS FOR AUTOMATED LARGE-SCALE ARCHAEOLOGICAL SITE DETECTION
Iban Berganzo Besga

New Computational Methods for Automated Large-Scale Archaeological Site Detection

PhD in Classical Archaeology

Author:
Iban Berganzo-Besga

Directed by:
Hèctor A. Orengo
Felipe Lumbreras

Landscape Archaeology Research Group (GIAP)
Catalan Institute of Classical Archaeology (ICAC)
Universitat Rovira i Virgili (URV)

Computer Vision Center (CVC)
Computer Science Department
Universitat Autònoma de Barcelona (UAB)

Para María, Ticky y Heckels

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke, Profiles of the Future (1973)

Co-authorship Statement

As an article-based thesis, the following is composed of four papers preceded by a unifying introduction, source, and methods and followed by a unifying discussion and conclusions. My academic supervisors, Drs. Hèctor A. Orengo and Felipe Lumbreras are listed as co-authors, Dr. Hèctor A. Orengo for all articles and Dr. Felipe Lumbreras specifically for Papers 2 and 3 in reflection of their direction and guidance. Dr. Maria Carme Belarte and Dr. Joan Canela are listed as co-author for Paper 1 in recognition of their prior analysis and guidance, with specific mention to the first for directing the project. Dr. Miguel Carrero-Pazos, Dr. João Fonte and Dr. Benito Vilas-Estévez are listed as co-author for Paper 2 for providing training data and input during the evaluation of the results. Dr. Cameron Petrie, Dr. Rebecca Roberts and their team are listed as co-authors for Paper 3 for providing training data and guidance throughout the process. Dr. Arnau Garcia-Molsosa, Juergen Landauer and Sergi Tres-Martínez are listed as co-authors for Paper 4 for each setting up their part of the common workflow, and Dr. Paloma Aliende for her work and support in its development. Likewise, I would like to mention Dr. Monica N. Ramsey for her direction in the second paper of the annexes. Finally I would like to declare that the generation of data and models, and the writing of the articles was mainly my responsibility.

Acknowledgements

Funding for this research was gratefully received from the Fundación BBVA; Spanish Ministry of Science, Innovation and Universities; Spanish Ministry of Economy and Competitiveness; the European Union’s Horizon 2020 research and innovation programme; and the Arcadia Fund.

Thanks are given to my supervisors Drs. Hèctor A. Orengo and Felipe Lumbreras for all the support provided during the years in which I have been working on this PhD. Likewise, I would like to thank Dr. Patricia Meehan, Dr. Adriana Velazquez, Dr. Darwin Carabeo, Dr. Marta Sancho i Planas, Dr. Walter Alegría, Dr. Maria Carme Belarte, Dr. Jaume Noguera, Dr. Joan Canela, and Dr. Josep Maria Palet for the trust placed in me, thanks to which I have been able to advance as an archaeologist. Thanks are also extended to the URV, and especially to the ICAC, the CVC and the GIAP team for their support and help during these years. Finally, special mention to my friends and family, who have always been there.

Acronyms

AI	Artificial Intelligence
CL	Curriculum Learning
CNN	Convolutional Neural Network
Colab	Google Colaboratory
CV	Computer Vision
DA	Data Augmentation
DEM	Digital Elevation Model
DL	Deep Learning
DSM	Digital Surface Model
DTM	Digital Terrain Model
FN	False Negative
FP	False Positive
GCP	Ground Control Point
GEE	Google Earth Engine
Lidar	Light Detection and Ranging
ML	Machine Learning
MSRM	Multi-Scale Relief Model
R-CNN	Region-based CNN
RF	Random Forest
RS	Remote Sensing
RVT	Relief Visualization Toolbox
SLRM	Simple Local Relief Model
TP	True Positive

INDEX

New Computational Methods for Automated Large-Scale Archaeological Site Detection

Abstract

This doctoral thesis presents a series of innovative approaches, workflows and models in the field of computational archaeology for the automated large-scale detection of archaeological sites. New concepts, approaches and strategies are introduced such as multitemporal lidar, hybrid machine learning, refinement, curriculum learning and blob analysis; as well as different data augmentation methods applied for the first time in the field of archaeology. Multiple sources are used, such as lidar, multispectral satellite imagery, RGB photographs from UAV platform, historical maps, and several combinations of sensors, data, and sources. The methods created during the development of this PhD have been evaluated in ongoing projects: Urbanization in Iberia and Mediterranean Gaul in the First Millennium BC, Detection of burial mounds using machine learning algorithms in the Northwest of the Iberian Peninsula, Drone-based Intelligent Archaeological Survey (DIASur), and Mapping Archaeological Heritage in South Asia (MAHSA), for which workflows adapted to the project's specific challenges have been designed. These new methods have managed to provide solutions to common archaeological survey problems, presented in similar large-scale site detection studies, such as the low precision achieved in previous detection studies and how to handle problems with little amount of training data. The validated approaches for site detection presented as part of the PhD have been published as open access papers with freely available code so can be implemented in other archaeological studies. Likewise, this doctoral thesis has contributed to the development of a Computational Archaeology Laboratory, which has allowed us to build these models, with which large amounts of data, never before achieved in archaeological site detection studies or survey work, have been obtained. To sum up, this PhD presents a comprehensive guide on the design, application and validation of methods for automated large-scale archaeological site detection within the field of computational archaeology.

INTRODUCTION

1. Introduction

1.1. What is this PhD about

Large-scale archaeological site detection has traditionally been a labor-intensive and time-consuming process that involves sifting through vast amounts of data, such as aerial and satellite imagery, to identify potential sites of interest. However, advances in computational methods have the potential to revolutionize this process, making it faster, more accurate, and more cost-effective. One example of a computational method that has been applied to large-scale archaeological site detection is machine learning, which uses algorithms to analyze data and make predictions or decisions based on that analysis. Machine learning has been used to identify patterns in satellite imagery that may indicate the presence of an archaeological site, such as changes in vegetation or soil composition. Other computational methods that have been explored for large-scale archaeological site detection include geospatial analysis, which involves the use of geographic information systems (GIS) to visualize and analyze spatial data, and remote sensing, which uses sensors on aircraft or satellites to collect data about the earth's surface. Overall, the use of computational methods for automated large-scale archaeological site detection has the potential to greatly enhance the efficiency and accuracy of archaeological research and discovery, and to enable researchers to identify and study previously unknown sites that may hold important clues about our past.

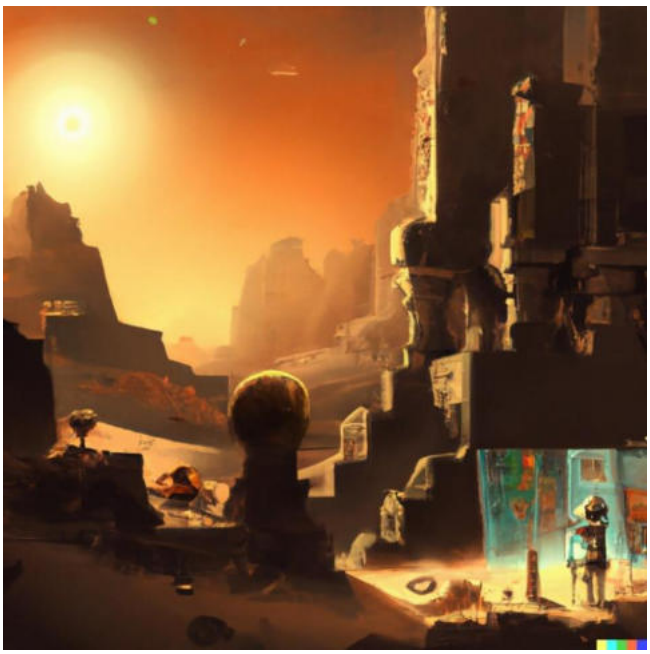


Figure 1. “Archaeology & Artificial Intelligence” image created using DALL·E (DALL·E 2022).

This previous text, which has been created by an artificial intelligence (AI) algorithm (ChatGPT, 2022), as well as the image that accompanies it (DALL·E, 2022) (**Figure 1**), allow us to introduce the topic of this thesis: the development of new computational methods for automated large-scale archaeological site detection and information retrieval. Early concepts of robots and AI emerged from Greek mythology thanks to divine crafts and represented the first impulses to create artificial life and amplify natural faculties, among which the golden maidens (*biotechne*) of the blacksmith god Hephaestus, characterized by their minds (*noos*), stand out. Likewise, the origin of computing dates back to the 2nd century BCE with the Antikythera Mechanism created on the island of Rhodes, a mechanical calculation model for astronomical purposes (Mayor, 2019).

However, it was not until the twentieth century, when the impulse of technology allowed the development of scientific research, with one of its milestones being the development of the radiocarbon technique, an isotopic dating method, by Willard Libby who was awarded the Nobel Prize in Chemistry in 1960. At the same time, great advances were made in surveying techniques using aerial photography thanks to the studies of the archaeologist O.G.S. Crawford in Britain (Crawford, 1923). In turn, landscape archaeology, has progressed enormously in recent years thanks to the introduction of new technologies (Argyrou and Agapiou, 2022).

1.2. Importance of Site Detection Methods

According to processual approaches, the development of site detection methods has evolved from art to scientific method thanks to the introduction of computing (Raczkowski, 2001), and is now an essential part of archaeological processes. These mathematically validated workflows help researchers locate and identify potential sites, just as is done with pedestrian survey. The location of archaeological sites brings with it: 1) advances in research on human past settlements and their patterns, human demographics, cultural group distribution, etc. (Carrero-Pazos, 2019), 2) increased management and protection of heritage (Bewley et al., 2005), and 3) survey and excavation cost savings (Peréx et al., 2012). If sites are not known, they cannot be protected and are subject to human and natural hazards. Heritage catalogues are a basic tool for the management of sites. Archaeological survey is the only way to manage this cultural asset which can be culturally exploited with routes and reconstructions that can be derived from the information recovered thanks to the survey. Also, there is a strong economic aspect, the location of potential sites will allow a better planning and therefore save costs in

surveying and excavation when intervention is unavoidable. This is critical because excavation can be costly and time-consuming, so it is important to be as efficient as possible in selecting sites and areas to investigate. The retrieved information is then used, along with site analysis and excavation if necessary, to help interpret the history and culture of the people who lived at the site, and to understand how they interacted with their environment, a valuable insight into the past.

The current archaeological survey methods are: 1) pedestrian survey, which can confirm the presence of artifacts or other evidence of human activity in situ, 2) photointerpretation using remote sensing (RS) techniques such as aerial photography, satellite imagery and lidar data, which can be used to identify potential sites based on surface features and topography like cross-banks (Bewley et al., 2005), 3) geophysical research, which can be used to detect subsurface features of archaeological interest (Conyers, 2006), 4) predictive modelling, which involves spatial analysis and geostatistical modeling (Willey, 1953), and 5) machine learning (ML), a field within AI characterized by learning algorithms, models which can be trained to learn and recognize patterns and features indicative of archaeological sites such as long-term settlement mounds (Menze and Ur, 2012); where the algorithms of the latter have been the most recent addition to the archaeological methodology.

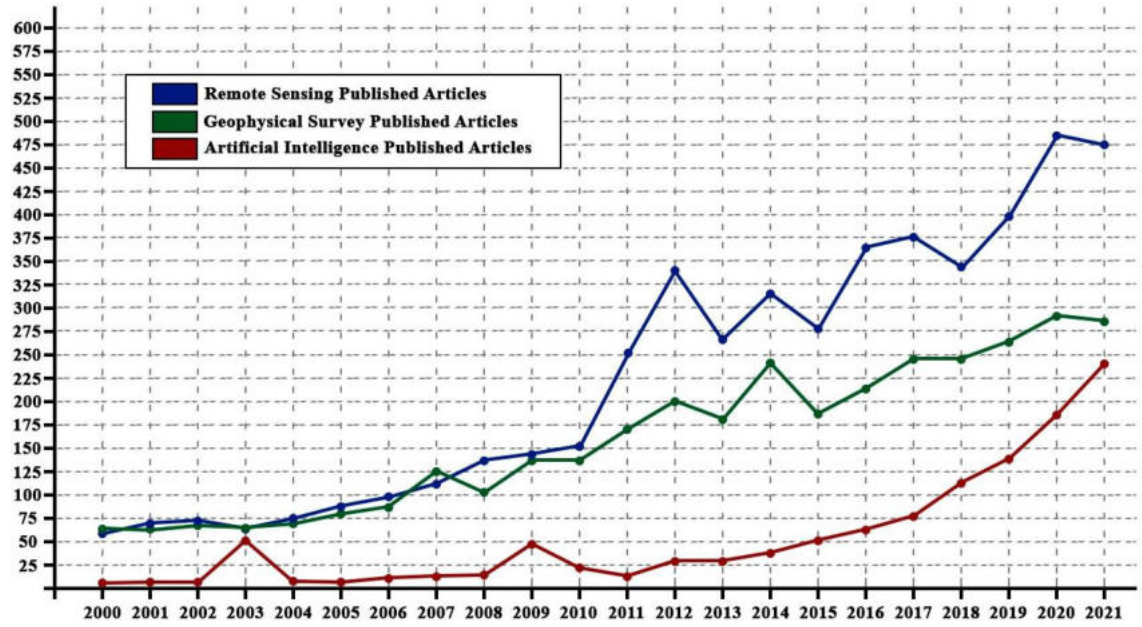


Figure 2. Evolution of the number of articles published per year since 2000 related to geophysical survey (green), RS (blue) and AI (red) in archaeology, according to the Dimensions website.

According to Dimensions website (Dimensions, 2022), the number of articles published on RS and on AI in archaeology has had a quadratic growth, and on geophysical survey in archaeology a linear growth, which has a slope growth around 10% (research category 4301 on the website) (**Figure 2**). This fact confirms the aforementioned growth of technologies in landscape archaeology with a substantial growth of RS since 2010 and even greater of AI in recent years. The late growth experienced by AI compared to the other two archaeological survey methods is due to the fact that the other technologies were developed years before ML algorithms and that AI requires more training but also there is very little expertise available within the archaeological community. This highlights the need to collaborate with other institutions in the field of engineering, as has been done in this thesis with the collaboration with the Computer Vision Center (CVC), a non-profit research center dedicated to carrying out cutting-edge research in the field of computer vision (CV) and immersed in the transfer of knowledge to industry and society.

1.3. Potential of Remote Sensing

During the last 20 years RS-based site location methods have progressively become dominant in the field (**Figure 2**). This is due to: 1) more sources have been publicly available (global digital terrain models (DTMs), Landsat, Sentinel, etc.), 2) these have progressively increased their spatial resolution sources and currently allow the visualisation of individual archaeological sites, 3) these include new sensor types (such as lidar, synthetic-aperture radar (SAR) and multispectral cameras with more spectral, radiometric, and temporal resolution), 4) new platforms, mainly drones, 5) increased availability of computational resources, such as parallel cloud computing platforms, 6) new methods and algorithms, such as ML, and 7) a new generation of archaeologists with a specific training in RS, who have the capacity to design their own approaches specifically tailored to archaeological problems.

The widely used aerial photography is of great interest for archaeological surveys because it allows the identification of buried features via photointerpretation of crop, shadow, moisture and other marks of archaeological elements. However, it has its limitations such as the archaeological landscapes under dense vegetation cover (Devereux et al., 2015). Therefore, it corresponds to current researchers to use new technological methods, such as those exposed in this thesis, which will complement the information obtained through aerial photography. Bewley et al. demonstrated that the use of lidar survey showed new details such as cross-banks

even in well-studied landscapes, adding new sites and extensions to sites that were previously known in Stonehenge (UK). They visualized these features in digital surface model (DSM) using relief shading, one of the most widely used techniques for the visualization of relief, and also they introduced the DTM, created by filtering tree cover from the DSM. Likewise, in the identification of sites they used multi-spectral CASI imagery, which can also indicate the state of the vegetation (Bewley et al., 2005). Evans et al. used a technology similar to the lidar one, the SAR, for the creation of an archaeological map of the medieval settlement of Angkor (Cambodia). They revealed thanks to this survey technology an extensive settlement landscape inextricably linked to water resources (Evans et al., 2007). The research by Opitz et al. is another example of the use of airborne lidar for the identification of archaeological landscape and architectural feature. In this case, the visualization tools used were sky view factor (SVF) and the previously mention hillshade technique. It allowed the detection of 33 archaeological features, ranging from prominent stone-built windmill towers to nearly invisible terracing, trackways and other earthworks (Opitz et al., 2015). Parcero-Oubiña and Nión identified 25 new sites in Galicia (Spain) using lidar and different visualization algorithms. These settlements corresponded to hillsforts. Also, satellite imagery and aerial photography were used to verify these sites (Parcero-Oubiña and Nión, 2021).

This is why RS makes it possible to recover information from non-accessible territories, from areas under administrations that require archaeological evidence to grant access permits, and for projects for which there are not enough funds available to carry out a pedestrian survey (Peréx et al., 2012).

1.4. Potential of Machine Learning

The Homeric narratives with their automatons (*automatai*) show us that the human impulse to automate is as old as its myths (Mayor, 2019). Humans have always sought the development of instrumentation to save time from daily tasks. Today, AI constitutes one more step in that direction, automating repetitive tasks, thus reducing the times of said processes to dedicate the mind to the development of ideas. The progress of technology characterized by AI was discussed in 1956 at the Dartmouth Conference organized by the computer scientist John McCarthy where AI was differenciated between general and narrow AI. The general one defines the theory on the consciousness of machines closer to that imagined by science fiction writers and the second, where we are really making progress, is process automation. “Really

science and art come to men through experience", Aristotle described in *Metaphysics* (350 BCE) the philosophical bases of the operation of what would be AI, ideas which were developed by empiricists such as Locke and Hume to finish with the question proposed by Alan M. Turing in *Computing Machinery and Intelligence* (1950): "Can machines think?". A challenge which is simplified every day with more consistent and faster paths allowing us to get closer every day to the knowledge (*episteme*) defined by Aristotle.

There are several reasons why site detection methods could benefit from a more automated approach using multiple data types and ML: 1) increased efficiency, automated approaches using ML can process large amounts of data quickly, allowing for faster site detection and analysis, 2) greater scalability, automated approaches using ML can be easily scaled up to handle larger datasets (which usually translate in large areas) or to analyse data from multiple sources, even Big Data, which are impossible to analyse manually, 3) continuous monitoring, the automation of processes allows to control the state of the features of interest as often as necessary or possible, 4) improved reproducibility, ML results are consistent because they offer repetitive and reproducible results, unlike manual approaches, making it easier to reproduce and verify the results, 5) it is relatively cheap in comparison with manual approaches, 6) can be applied to most data sources, even those that are not susceptible to manual photointerpretation (such as radar, non-RGB bands and combination of sources), 7) they have the capacity to use measures that are non-distinguishable by humans and therefore they can provide not just detection results but also more accurate classifications, and 8) increased accuracy, they can be combined with traditional approaches to provide complementary information.

The many approaches in the field of ML can be broadly grouped into three categories: supervised learning, unsupervised learning, and reinforcement learning. The supervised algorithm is trained with known outputs for certain inputs and generates predictions that iteratively refine the model. In this way, it will be able to determine the output for new inputs. The unsupervised approach will try, without this knowledge, to classify the information by looking for patterns in it. In the third place, reinforcement learning is a model trained through rewards and penalties. The most common algorithms used for supervised learning are regression and classification, whose main difference is that the first generates continuous outputs and the latter discrete ones. For unsupervised learning we find algorithms such as clustering. Deep learning (DL) is a step further within ML with models which follow the same aforementioned methodologies but based on artificial neural networks. The main difference

between the previous classical ML techniques and DL is that the latter performs feature extraction by itself (Torres, 2020). ML, and even more DL, have shown their potential for automated archaeological site detection. Menze and Ur created a random forest (RF) algorithm, a supervised ML model based on decision trees used in this case for classification, for the detection of long-term anthroposoils in northern Mesopotamia trained with ASTER multispectral satellite imagery (Menze and Ur, 2012). Orengo et al. created a RF algorithm for automated detection of archaeological mounds in the Cholistan Desert in Pakistan trained with Sentinel satellite imagery (Orengo et al., 2020). Davis et al. showed that the use of DL allows to significantly increase the number of known archaeological features, shell rings in their case. For this, they used Mask R-CNN, a DL segmentation algorithm, which results where verified using a RF algorithm trained with Sentinel imagery (Davis et al., 2021). On the other hand, ML has demonstrated its application to other archaeological fields beyond site detection, for example, the identification of recovered data, such as phytolith detection and classification (Berganzo-Besga et al., 2022a).

The physicist Stephen Hawking warned that “the genie is out of the bottle (...) I fear that AI may replace humans altogether”. Like any tool, AI must be used ethically. For this reason, a series of principles were established at the Asilomar Conference on Beneficial AI in 2017, as an extension of Asimov's laws, where the last one indicated that "superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization". In any case, currently, the main problem we find with ML is the incorporation of our own biases as training data, which, if not taken into account, can produce real misinterpretations. Also, ML might present the wrong idea that all sites or a high percentage of those are located when good results are obtained. Only the types of sites used for the training will be located and we do not know the total percentage of those. Therefore, there is a need to be careful with interpretations drawn from these approaches and not to take them as the only way to produce archaeological or heritage catalogues. Likewise, it also presents other practical drawbacks focused mainly on the necessary computational resources, despite some freely available online platforms, most of the computing is expensive, and on the expertise needed for its conscious application, which is not common in archaeology and is more difficult to acquire than traditional approaches such as photointerpretation. This PhD will explore both advantages and drawbacks of the application of ML to specific case studies and explore how to balance these to reach practical applications.

1.5. PhD Organization and Objectives

Some of the latest advances in computational methods for site detection, at the time of starting this thesis, were focused on lidar (Opitz and Herrmann, 2018), multisource multitemporal satellite data (Orengo and Petrie, 2017), automated detection of barrows (Verschoof-van der Vaart et al., 2020) and drone imagery (Orengo and Garcia-Molsosa, 2019). This doctoral thesis, builds up from all these advances, and aims to go one step further by introducing new computational methods for site detection on a large-scale. It does not apply previously developed methods and, although composed by different case studies, cannot be considered as a case study since, as a whole, focuse. As such, this thesis is experimental and explorative in its nature. Nevertheless, all research conducted as part of this PhD has produced new methods that have significantly improved previous approaches, are fully reproductible, and can be applied anywhere with minor adaptations.

The main objectives of this thesis are:

1. To investigate new emerging computational technologies for the detection of archaeological sites,

2. To develop new methods that can substantially improve current site detection techniques,

3. To apply these methods in selected case studies linked to specific projects under development,

4. To test and validate these methods using standard statistical measures,

5. To analyse their possible application and standardisation within mainstream archaeological research.

Moreover, given the fact that the methodologies developed have been applied to specific cases of ongoing projects, a series of more specific objectives have been established per article:

1. Detection of the walled structures of Iberian sites in North-Eastern Iberia,

2. Detection of pre- and proto historic burial mounds in North-Western Iberia

3. Detection of mound features which would correspond to settlements of the Indus civilization and later cultures,

4. Detection of potsherds for the identification for new sites.

Likewise, this doctoral thesis has led me to develop specific algorithms based on the already known ones, to create my own code to automate processes such as the creation of data augmentation (DA), and to contribute to the development of a computational archaeology laboratory based at the ICAC.

This is a paper-based PhD. All papers deal with a particular independent topic but they all together form an integrated methodological and theoretical discourse on the use of RS, geospatial analysis, ML, and associated methods for the detection of archaeological sites over large areas. The general scheme of the thesis is divided in:

1. Introduction, which aims to contextualise the research,

2. The methods section explains in detail the logic behind the methodological developments selected for each paper and how they fit together as a single approach to site detection,

3. The results section compiles all published papers related to the topic of the thesis as published (or submitted at the moment of the delivery of the thesis),

4. The discussion section will discuss the results obtained by the different papers to what extreme they were successful, what were the problems drawbacks and successes and how to improve them and implement them in future works,

5. The conclusions will summarise the aspects developed during the PhD showing the relevance of its contributions to the field.

As a paper-based PhD, they constitute the core of the thesis and the other sections will endeavour to contextualise and provide them with a more global meaning:

Potential of Multitemporal Lidar for the Detection of Subtle Archaeological Features under Perennial Dense Forest *(Published in Land journal in November 2022, DOI: 10.3390/land11111964)*

The first article seeks to show how the combined use of multitemporal lidar datasets, rarely applied for archaeological research, and multiscale approaches to micro-relief visualisation can improve the visualisation of subtle archaeological remains present as micro-reliefs in the

ground surface even in complex environments such as forested mountain Mediterranean areas. The workflow is tested in six Iberian Iron Age sites, all of them located in mountain areas with dense Mediterranean perennial forests and shrub vegetation. New trends such as the development of data fusion methods for multitemporal and multisource data in virtual satellite constellations and the creation of virtual data through data augmentation processes hold much potential for the development of more holistic analysis with deeper analytical capabilities.

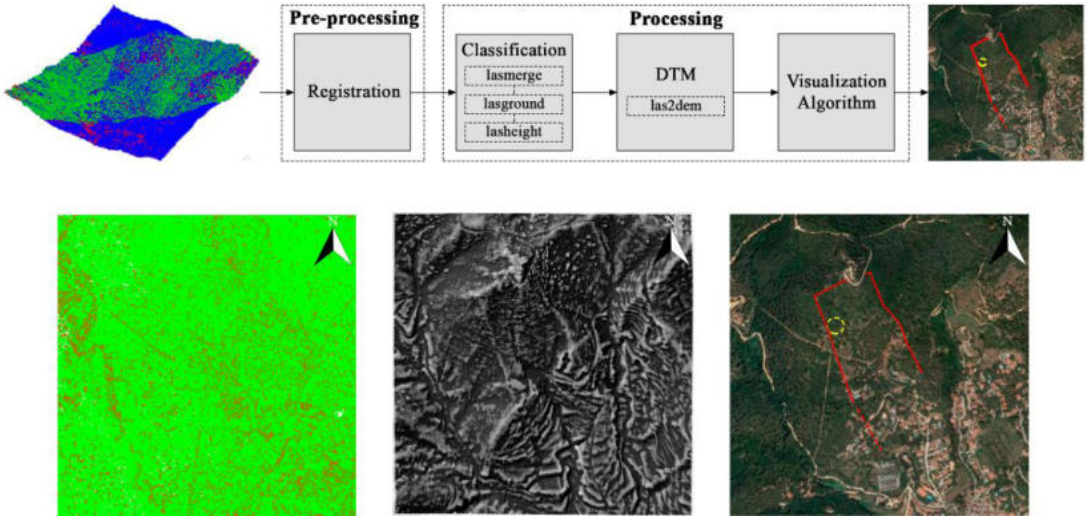


Figure 3. Graphical abstract of the first paper.

Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel-2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia *(Published in Remote Sensing journal in October 2021, DOI: 10.3390/rs13204181)*

The second one aims to show how the combined use of classical ML methods such as RF and DL approaches can improve the large-scale automatic detection of burial mounds, one of the most common types of archaeological sites globally, using lidar and multispectral satellite data. The workflow is tested in North-Western Iberia, an area of near 30,000 km^2, the largest in which such an approach has ever been applied. This article introduces new concepts practically unknown in the archaeological field but widely used in the CV one, such as transfer learning and refinement (with its positive and negative training concepts). Given the performance of the training algorithm presented here for the detection of burial mounds, our method already

constitutes a practical tool that can be applied to any other areas where tumuli are present with few modifications, thus making it a general tool for archaeological research and cultural heritage management in many areas of the world.

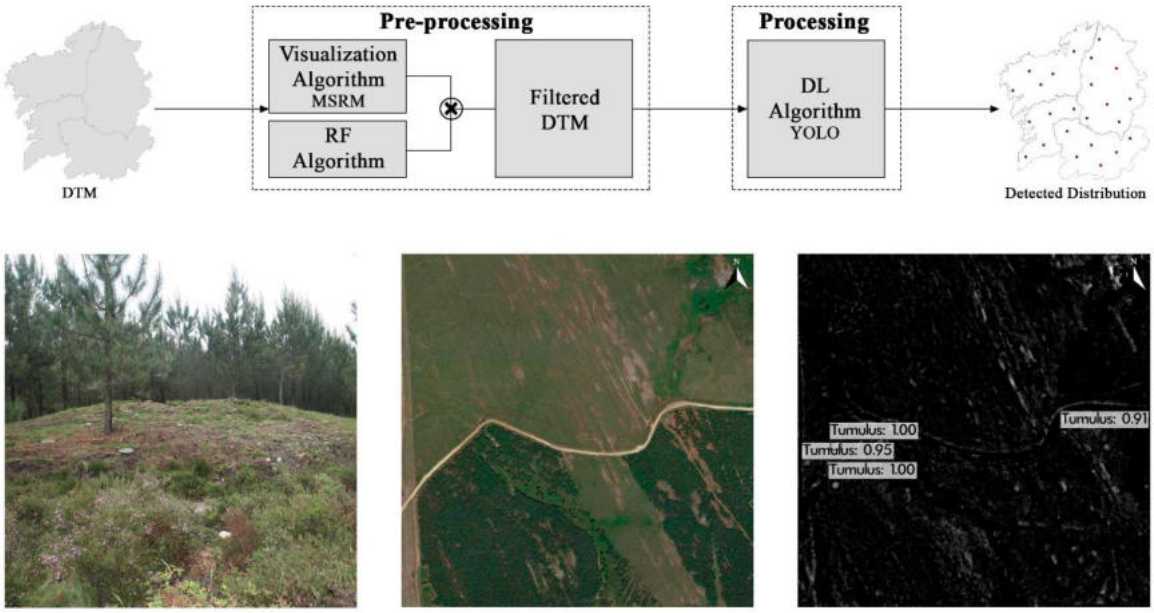


Figure 4. Graphical abstract of the second paper.

Curriculum learning-based strategy for Archaeological Mound Features detection from Historical Maps in low-density areas in India and Pakistan *(Manuscript submitted to Scientific Reports journal in February 2023, DOI: Not available yet)*

The third paper shows the importance of historical maps as a source of information about lost archaeological sites and of the instance segmentation algorithms for large-scale automated detection of mounds in these digitalised and georeferenced maps. The workflow is tested in India and Pakistan, an area of 470,500 km^2, the largest in which such an approach has ever been applied, including, for the first time, different types of maps with a wide variety of symbols. This article introduces new concepts practically unknown in the archaeological field but widely used in the CV one, such as curriculum learning (CL) and blob analysis. This workflow has the potential to be applied to other maps from other countries.

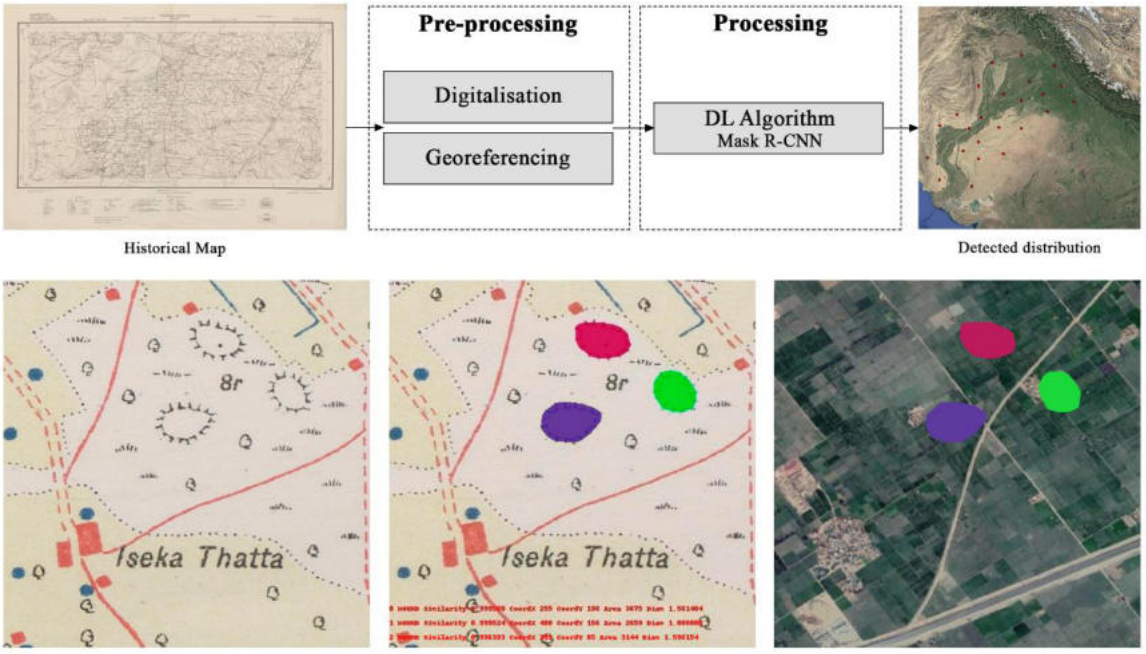


Figure 5. Graphical abstract of the third paper.

New Developments in Drone-Based Auatomated Surface Survey: Towards a Functional and Effective Survey System *(Published in Archaeological Prospection journal in May 2021, DOI: 10.1002/arp.1822)*

The last one aims to probe how the combined use of drone survey and DL approaches can improve the automated detection of individual items or fragments of material culture visible on the ground surface. The combination of approaches allows flexibility to carry out surveys with different objectives, from rapidly sampling a large area with a minimum of computational resources to very intensive survey, which includes the extraction of the shape of each fragment of material culture detected. This article introduces new concepts practically unknown in the archaeological field but widely used within CV, such as instance segmentation instead of semantic one (more common). The method can offer a way forward for the quantifiable large-scale use of survey data in the analysis of past human occupation.

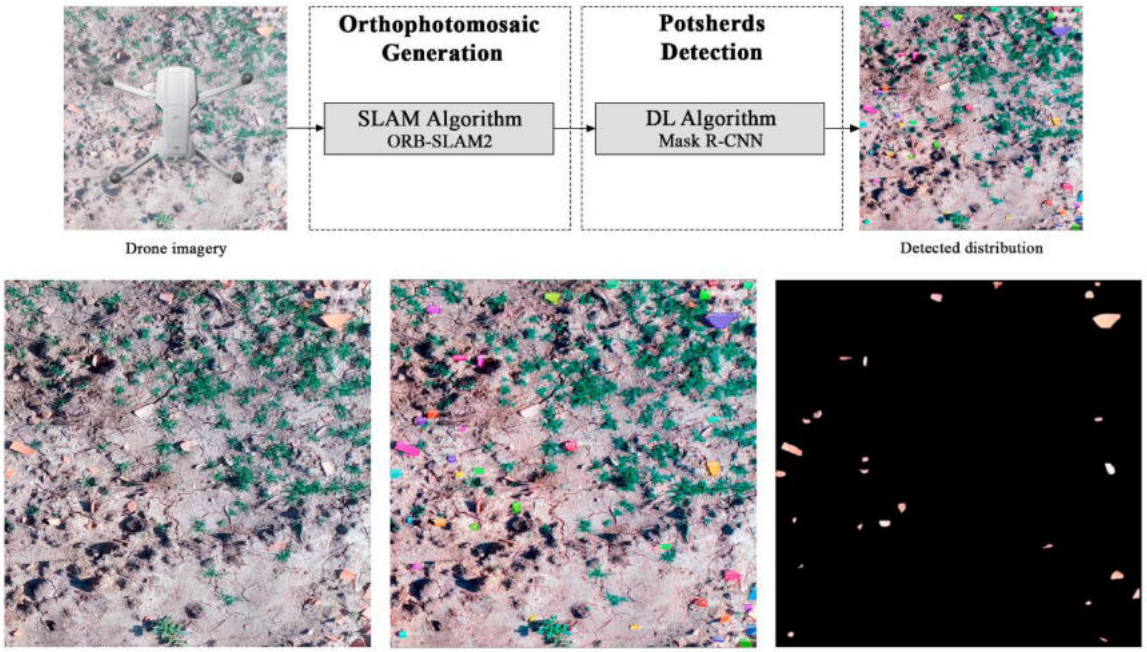


Figure 6. Graphical abstract of the fourth paper.

This selection is an incremental approach to archaeological site detection, from the visual use of lidar data to automated approaches. The pedestrian survey, the most common technique for site location, is complemented by different RS methodologies and automated by ML algorithms. This evolution begins with the use of lidar-based DTMs for the detection and visualization of archaeological features and continues with the automation of such detection using RF and DL. In the last three articles distinct region-based convolutional neural network (R-CNN) models are used, both for object detection and for mask segmentation, as well as different data sources, satellite imagery, historical cartography and drone imagery. All this, to make an approximation to the detection of sites from different data sources and through new RS and ML methods.

The application of computational methods in the field of archaeology goes beyond the development of algorithms, it is a paradigm shift within the approach to the investigation of historical remains, which will allow the ambitious goal of achieving automatic detection of archaeological sites on a global scale. A truly archaeological revolution.

SOURCES AND METHODS

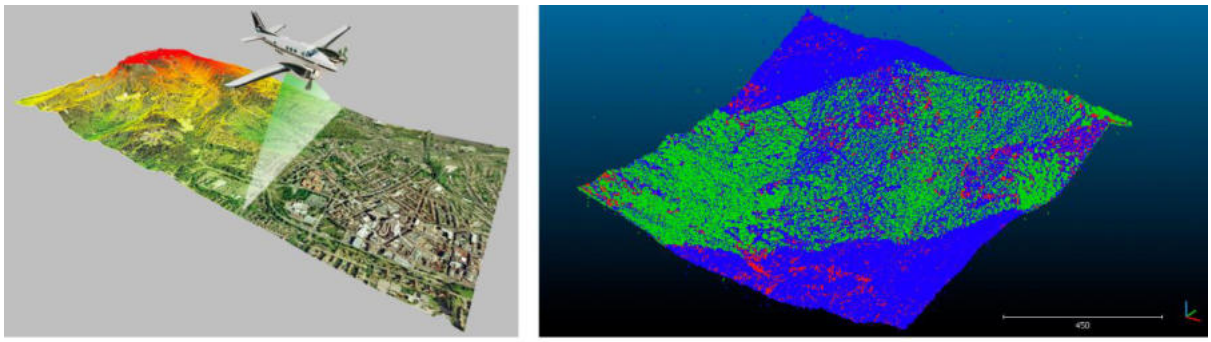
2. Sources and Methods

2.1.Sources

The case studies comprising this PhD have touched upon many sources of RS data but also upon related cartographic material with potential for the discovery of unknow archaeological sites. Indeed, the use of multiple data sources has an, still largely untapped, immense potential.

2.1.1. Lidar

A typical lidar system consists of airborne laser scanning (ALS), a GPS-derived aircraft position, an inertial measurement unit (IMU), and the measurement of the deflection angle of the transmitted laser beam (**Figure 7**). The point clouds generated by this system allow the creation of high-resolution DTMs of the landscape by filtering the points corresponding to vegetation and other elements above the surface of the terrain (Wehr and Lohr, 1999; Hesse, 2010). The remaining points corresponding to the ground surface are usually interpolated to create a continuous raster surface, the DTM. Lidar-derived technologies have been used in multiple disciplines with a geographical basis (Gutiérrez-Antuñano et al., 2017) such as engineering, geomorphology, hydrology, landscape architecture, and archaeology.



(a) (b)

Figure 7. Lidar system: (a) airborne laser scanning (GIS Blog, 2022) and (b) resulting point cloud.

In archaeology, their most common use, besides GIS-based topographic analysis, has been the detection of archaeological structures visible in the ground surface as topographic imprints. Some of these are very subtle and require specific visualisation techniques for their identification. The use of visualization algorithms can reveal important geomorphological and

cultural information (Opitz et al., 2015; Orengo and Petrie, 2018). These relief visualization tools can provide a clearer visualization of subtle topographic changes within the DTM.

The first two articles of the thesis (Berganzo-Besga et al., 2021 and 2022b) use lidar data to generate a DTM where we apply the multi-scale relief model (MSRM) visualization algorithm (Orengo and Petrie, 2018) to visualize the microreliefs and manually identify archaeological features such as walled structures in the first case, and automatically detect and analyse burial mounds and other types of tumuli through DL in the second.

2.1.2. Multispectral Satellite Imagery

In the second paper (Berganzo-Besga et al., 2021), cloud-filtered multitemporal Sentinel-2 multispectral imagery was also used to create an ML algorithm for classifying soils (**Figure 8**). Sentinel-2 incorporates 13 bands from which only the visible/near-infrared bands (VNIR B2–B8A) and the short-wave infrared bands (SWIR B11–B12) were employed. Bands B1, B9, and B10 (60 m/px each) correspond to aerosols, water vapor, and cirrus, respectively, and they were discarded in this study except for the use of the cirrus-derived cloud mask applied. Visible (B2–B4) and NIR (B8) bands provide a ground resolution of 10 m/px, while red-edge (B5-B7 and B8A) and SWIR (B11–B12) bands present a 20 m/px spatial resolution. Specifically, for this research Sentinel-2 Level 1C products representing top of atmosphere (TOA) reflectance were preferred due to the larger span of its mission (starting from June 2015). Sentinel-2 multispectral satellite images were a good compromise given their relatively high spatial and spectral resolutions and their open access policy. The use of cloud-filtered multitemporal satellite data has been successfully employed in previous research to provide long-term vegetation indices (Orengo and Petrie, 2017; Garcia-Molsosa et al., 2019), but also for the development of ML classifications (Orengo et al., 2020) as they provide images that are independent of specific environmental or land-use conditions that are particularly adequate for the development of classifications.

The use of multispectral imagery in this paper is conceived as a way to complement the information provided by lidar-derived DTMs: while lidar supplies information about the shape and 3D form of the object of interest, multispectral imagery provides information on the physical and chemical composition of the objects and their environment. This strongly contributes to obtain a better definition of the objects of interest but also of possible false positives (FPs), which might present a similar shape but be made of completely different

material. In this regard, the use of multispectral imagery in combination with lidar data provides one of the few existing examples of the use of multisensor data as an input for ML workflows (Orengo et al., 2020) and proves the enormous potential data fusion in site detection studies.

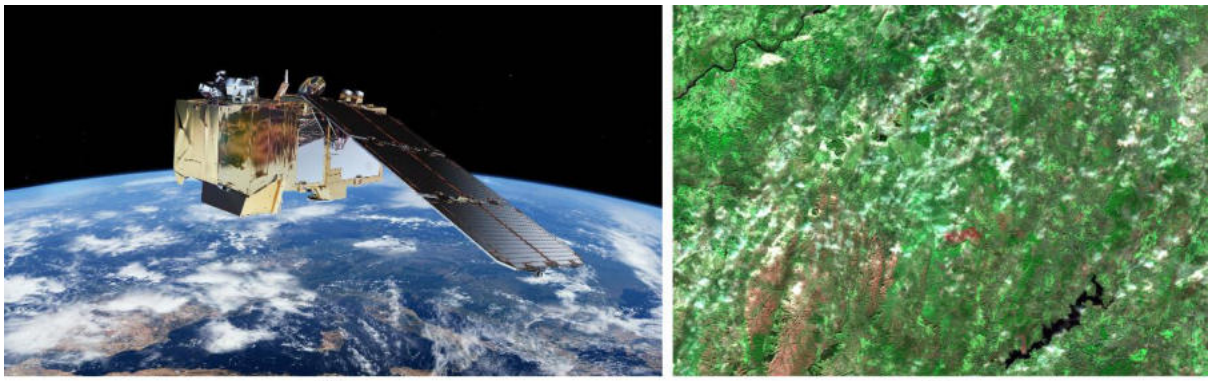


Figure 8. Sentinel-2: (a) sample satellite (ESA, 2022) and (b) resulting multispectral imagery.

2.1.3. RGB Imagery Acquired from Unmanned Aerial Vehicles

In our last article (Orengo et al., 2021), we have used 3-channel RGB images acquired from drones, unmanned aerial vehicles (UAVs) (**Figure 9**). To do this, we collaborated with HEMAV, one of the largest drone-based industries in Europe, for the design of a drone based on their popular HAR9 multirotor platform. From the moment in which the paper was written, we have moved to a modified Helix North Mapker PPK, carrying a Sony RX1r RGB sensor (42mpx, 35mm). This is a very similar model, with the same characteristics and capabilities to that proposed in the article. This Sony RX1r will has considerably increased image resolution and quality, and has allow the drone to fly much higher (and therefore faster) than common commercial drone cameras. Lower resolution cameras employed in previous tests (Orengo and Garcia-Molsosa, 2019) forced the drone to fly very low (3m above ground level) to obtain enough resolution for automated potsherd detection. The drone incorporates other characteristics directed towards archaeological survey: an altimeter that will allow the drone to fly at a constant height above ground. This will facilitate the acquisition of images with similar ground sampling distances independently of terrain variation. The drone has a flight autonomy of around 40–45 minutes depending on the weight of the selected camera (longer than most commercial and professional drone models available at the moment in which the paper was

published). Another important characteristic of the drone will be the integration of real-time kinematic (RTK) navigation with a stable base station. This will allow a positioning of centimetric precision that will be extremely useful to plan the flight and image capture and will facilitate the photogrammetric processing of the images. It will also allow the drone to continue the image acquisition at exactly the same position where it stopped when recording large areas that might require several battery changes. All in all, these features make our prototype an efficient option to conduct archaeological survey, as proven in several campaigns where the prototype has been employed during the last two years. Commercial drone options such as DJI Mavic Air 2 or Autel Evo 2 with flight times above 30 min and cameras above 40 MP do not usually include an RTK option. Despite including bottom sensors that can measure ground distance, they do not use this information to keep a constant height above ground but to avoid collision. The same is true for the DJI Phantom 4 RTK, which includes a GNSS module and ground sensor but a 20-MP camera (although with a larger CMOS sensor than the previously discussed models). The prototype cost (around €9,000, not including camera), is lower price than other drone models with similar characteristics currently available in the market.

Figure 9. Drone: (a) DJI Mavic Air 2 (DJI, 2022) and (b) resulting camera imagery.

The ground control system HEMAV Planner is a simplified version of Mission Planner for PX4 flight control software, which uses the MAVLink communication protocol (all of these being developed by successful and active open-source code projects). The ground control system is designed to operate with the drone that is controlled by a Pixhawk (PX4-compatible) autopilot and focused on the specific needs of archaeological survey. HEMAV Planner will also be configured to simplify image acquisition. The pilot will only need to draw the area that

requires survey, and the software will automatically calculate the flight height and route for optimal image resolution and image overlap. This route can be interrupted (in order to change the drone battery or the memory card) and continued at exactly the same position as many times as necessary, thus allowing the coverage of large areas with little loss of time.

2.1.4. Historical Maps

In the third article (Berganzo-Besga et al., 2023), we use historical maps, which are not a source of RS data but have long been used by archaeologists to detect archaeological and other features of historical interest (Orengo et al., 2015; Thomas and Kidd, 2017; Hammer et al., 2018; Franklin and Hammer, 2018). They hold enormous potential as they depict a landscape that, in most cases, has been strongly modified during the last 70 years or so and, therefore, they have the capacity to inform us on archaeological elements that do not longer exist. In this case, the sources employed are 1-inch topographic map series of the Survey of India (SoI) developed during the first half of the twentieth century based on topographic surveys developed during the last decades of the nineteenth century and the first decades of the twentieth century. These are particular useful to test automated detections methods as they: 1) cover a very large area including India and Pakistan, 2) the area of interest has suffered important landscape modifications and, in many cases, these are the only sources that can detect disappeared mounds (corresponding to ancient habitation sites), 3) they are well standardised and the resulting algorithms can also be potentially applied to other British colonial areas where the same type of maps were developed, 4) their scale and digital resolution is adequate for the size of the features to detect, 5) these features were represented in the maps quite systematically, 6) digital georeferenced maps have been produced and are available within as part of the MAHSA project development. MAHSA is a large project directed at the McDonald Institute of Archaeological Research (University of Cambridge) and funded by the Arcadia Fund in which ICAC is a partner. While members of GIAP are in charge of RS, ML and other site detection methods, the team at Cambridge has been in charge of collecting, scanning, and georeferencing the SoI maps available for the areas of India and Pakistan.

The large development of agriculture and urbanization during the last 60 years has caused the disappearance of many archaeological sites in India and Pakistan. For this reason, historical maps become a very relevant source of information about those lost sites. These maps allow the site detection that, given their partial or total destruction, are no longer detectable by another

type of source such as lidar or satellite imagery. The available satellite images of India date back to 1972 thanks to the Landsat satellite programme (Landsat Science, 2022), although some previous CORONA images are also available, but some of these historical maps are from the 1890s. During the rule of the British Raj in India and Pakistan (1858-1947), the SoI produced various geographical map series to record the landscape. These maps were originally intended to be geographic maps and depicted different geographic mounds which, as further research showed (Green et al., 2019), concealed archaeological sites. The ground truthing revealed there is a correlation between these mounds and proto-historical and historical sites, potentially related to landscapes of the Indus Civilization and later periods (Green et al., 2019). In this study have been used a total of 645 maps, provided by Cambridge University Library and British Library (**Figure 10**). These historical maps, drawn by the British Crown, can be classified into different periods characterized by their draughtsmen: C. Straham (1898-1899), G.C. Gore (1900-1902), F.B. Longe (1904-1907), S.G. Burrard (1912-1913) and A.R. Quraishi (1954).

The potential of these maps and other historical map series for site location studies is still largely untapped. They are a unique source where the landscape is depicted before the large-scale landscape modifications typical of the twentieth and twenty-first century and therefore provide information on site location and other features of archaeological interest that cannot be obtained using any other source including RS (Garcia-Molsosa et al., 2019).

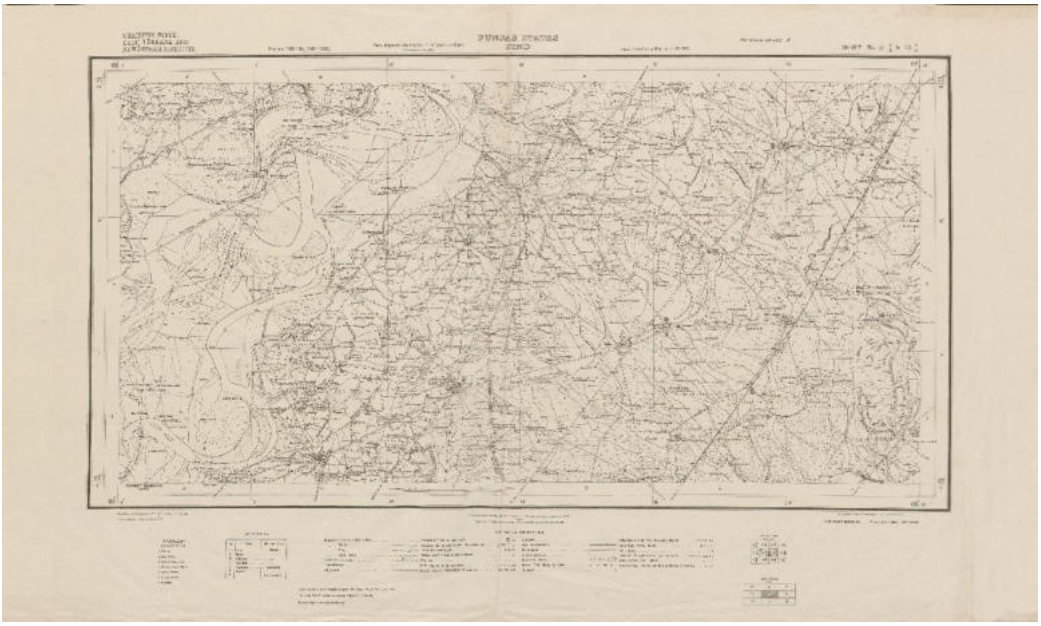


Figure 10. Historical map from the SoI.

2.2. Methods

2.2.1. Data Preparation Approaches

In all cases the sources employed needed to be treated to improve their detection potential and, in most cases, to filter out FPs. This treatment has been adapted to each specific case study. A description of the data preparation approaches employed follows, both pre- and post-processing.

2.2.1.1. Lidar Data Treatment

The first paper (Berganzo-Besga et al., 2022b) presents a method for the merging of lidar-derived point clouds of the same area taken at different moments, even when these are not co-registered. In order to obtain a higher density of points and thereby increase the resolution of the final DTM, we decided to employ all data available (in this case three coverages obtained at different moments) within a single workflow. This allowed us to obtain higher-resolution data with an average density of 6–9 points/m^2 (including all point classes). However, due to some coverages were not co-registered, we used the 3D point cloud processing software CloudCompare v2.11. alpha to align the datasets. CloudCompare is a 3D point cloud (and triangular mesh) processing software. It implements the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) to minimize the difference between the point clouds.

The next step consisted of the removal of above-surface features such as buildings and vegetation (Bewley et al., 2005). The lidar points were classified using LAStools, a collection of efficient, batch-scriptable, and multicore command line tools. LAStools also allowed us to merge the co-registered data using its *lasmerge* function and then *lasground* for bare-earth extraction, with the default step parameter of 5.0 which is adequate for forest and mountain environments. The *lasground* script classified the lidar points into ground points (*class* = 2) and non-ground points (*class* = 1). The co-registered, merged, and classified data was utilised to generate a DTM using LAStools' *las2dem* function. This tool triangulates lidar points into a triangulated irregular network (TIN), and then rasterises the TIN to create a DTM. Using QGIS projection tools we assigned the spatial reference information to each DTM.

Once the filtered high-resolution DTMs for each of the sites under study were obtained, we tested several visualization algorithms to evaluate which one provided a clearer visualization of subtle topographic changes within the DTM. For this, Digital Elevation Model (DEM) manipulation-based visualization methods (MSRM, simple local relief model (SLRM), and

slope gradient), illumination-based ones (principal component analysis of multi-azimuth shaded relief maps, which is not kernel-size dependent), and other kernel-size dependent techniques (Sky-View Factor (SVF) and Opennes) were selected (Hesse, 2010; Zakšek et al., 2011; Doneus, 2013; Kokalj and Hesse, 2017).

2.2.1.2. Multispectral Satellite Imagery Treatment

For the second paper (Berganzo-Besga et al., 2021), the use of Google Earth Engine (GEE) allowed us to access and join 1920 (at the moment of writing, which included all the available Sentinel-2 images from 23/06/2015 to 05/06/2020) in a single multi-band composite, train the classification algorithm and execute the analysis, which would have been impossible using a desktop computer. It also provided an ideal environment to join the results of the classification with that resulting of the MSRM filter of the DTM also created using GEE.

Sentinel-2 multispectral imagery was selected at level 1C, in which radiometric measurements are provided in Top of Atmosphere (TOA) reflectances, a DSM has been employed to project the image in cartographic geometry, and each image has been resampled with a constant Ground Sampling Distance. Also, we filtered images with more than a 20% of could percentage and applied a pixel cloud mask (including cirrus clouds) based on the Sentinel-2 QA band. The last treatment step was to average the 1921 images available for the study area, effectively creating a multitemporal multiband image composite, which presents ideal conditions for pixel-level ML-based classification.

2.2.1.3. Historical Maps Treatment

After the digitalisation of the maps, these were georeferenced using a minimum of 12 ground control points (GCPs) geometrically distributed within the map to achieve a good rectification. The GCPs were obtained from georeferenced high resolution RGB satellite imagery available as Web Map Services layers in standard GIS software. The georeferencing process used both second order polynomials and the adjust transformation, when enough GCPs were available for an accurate transformation. This first one was the preferred method and it was applied to most maps. These methods produced average Root Mean Square Error (RMSE) values of 0.000350º (ca. 33.7-38.8m at this latitude) using a second order polynomial and 0.000102357º (ca. 10.3m with a maximum value of ca. 26.8m) using the adjust transformation. Since the

mounds under consideration are typically much larger than these values, the georeferencing process would have resulted in mound locations within maps, which, at least, partially overlap the real locations.

The previous survey area (Green et al., 2019), which included only a small number of well-preserved mounds, showed that those smaller mounds were mostly non-archaeological sites. Likewise, previous research (Garcia-Molsosa et al., 2021) showed that many of these small mounds were mostly dunes or modern spoil from pond excavation. For this reason, it has been decided to filter, throughout the study, all those small mounds to avoid including mounds that are not likely to be archaeological sites. A second filter, using blob analysis, has been applied to remove those elongated mounds which do not usually correspond to archaeological sites. The ellipsoidal shape of each detected mound has been evaluated and all those that presented an elongation have been eliminated. Finally, in a post-processing stage, given the similarity of the mounds with the characteristic shape of mountainous peaks, a script has been applied using GEE and QGIS to filter all those mountainous areas and thus eliminate all those mounds that, although correctly identified by their drawn shape, do not correspond to possible archaeological mounds but to mountain peaks and natural topography.

2.2.1.4. Drone Imagery Treatment

The treatment of the images acquired by the drone aimed to generate a single orthophotomosaic rectifiying and merging all the available images per area. This orthophotomosaic can be then employed as an input source for the location of potsherds in the image. This process usually employs a photogrammetric processing workflow, which nowadays is almost fully automatised using commercial and open source software (Orengo, 2015). However, the photogrammetric processing of the images to create a single orthoimage that could be used for the location of potsherds was computationally costly and required specialized photogrammetry software such as Agisoft's Metashape. Although it is true that several open-source options are nowadays available for the photogrammetric processing of multiple images, these are still complex to install and use, and not all of them provide options for the output of orthomosaics. In order to address the issues, we decided to abandon photogrammetry and aim at the implementation of a simultaneous localization and mapping (SLAM) algorithm. SLAM algorithms use external data such as lidar, images and sound to geolocate the sensor and construct an approximate map of its environment as it moves. Common implementations can be found in consumer robot

vacuum cleaners, and they are an increasingly important component of autonomous vehicle navigation systems. More importantly, SLAM solutions can produce real-time data, and, although less accurate than photogrammetry, they provide an excellent alternative to the mosaicking of drone-based imagery (Bu et al., 2016; Kern et al., 2016). Therefore, we will aim to implement a SLAM algorithm, such as ORB-SLAM2 (Mur-Artal et al., 2015; Mur-Artal & Tardós, 2017), that can input information from monocular cameras while providing excellent accuracy and speed.

2.2.1.5. Data Insertion and Extraction

In archaeological survey, work is done with georeferenced data such as DTMs, historical maps and drone images, most of them in TIFF image format. This type, in addition to containing the image itself, keeps its geographic coordinates. To use these data in DL algorithms, for any type of study which requires these georeferenced coordinates, all these TIFF files have to be converted to another image type such as JPG or PNG. There is an added challenge in archaeology regarding ML algorithms, it is required the georeferenced location of all those detected archaeological elements. Once the detection is complete, the results obtained, mostly polygons of the detected archaeological features, must be transferred to the corresponding geographic coordinates for easy viewing in programs such as QGIS using a shapefile.

Also, to save computational costs, images must be cropped to sizes accessible to DL algorithms. We have used an input image size of 832x832 pixels for the detection of burial mounds, 512x512 for the maps and drones images. All cropping and georeferencing algorithms have been developed by ourselves in Python, so they can be used by other users in Google Colaboratory (Colab) to make the algorithms accessible, reproducible and replicable.

2.2.2. Machine Learning-Based Approaches

When developing ML models, it is mandatory to keep in mind the significance of selecting the correct algorithm for each project. Moreover, the archaeological perspective must be maintained to face the usual automated archaeological survey issues: the small amount of training data and the low-density of the archaeological elements to be detected.

2.2.2.1. Random Forest

This ML algorithm is a method for data classification based on decision trees (Horning, 2010). In the article on detection of burial mounds, a RF is used to classify the soil into suitable or not for the presence of tumuli, using Sentinel-2 satellite imagery. Thirteen polygons defining training areas were drawn and tagged as class 0 (areas unsuited for the presence of tumuli), which included a variety of urban features (such as roofs, roads, swimming pools, etc.), water, rock and quarries and other industrial areas. Additionally, 19 class 1 polygons were drawn within grasslands, cultivation fields and forests. Although we could have used a DL algorithm, the RF already gave results with great accuracy, so it was decided to use this traditional ML model to save computational costs.

On the other hand, a previously published RF algorithm for potsherd detection provided results more efficiently than standard pedestrian survey in terms of the number of items detected and time investment (Orengo and Garcia-Molsosa, 2019) but the algorithm was not able to detect all visible pottery sherds in the area. Another problem was the presence of FPs. Lastly, although the vectorized extraction of potsherds was a good approximation of the size of the potsherds, the shape itself was poorly defined, hindering the application of analyses based on the potsherd morphology. So, for potsherd detection was mandatory to use a DL segmentation algorithm.

2.2.2.2 Deep Learning Object Detection

Most recent research on archaeological feature detection using lidar datasets has used algorithms based on R-CNN (Trier et al., 2019; Verschoof-van der Vaart et al., 2020; Davis et al., 2021). These are object detection algorithms based on a combination of classical tools from CV and DL that has achieved significant improvements, of more than 30% in some cases, in detection metrics using reference datasets within the CV community (Girshick et al., 2014).

In the last three articles we aimed to detect certain archaeological elements in specific areas of the image, so we employed R-CNN algorithms. However, since for tumuli detection we are only interested in knowing their location and size, and not their morphology as they all have a characteristic circular shape, we used a DL object detection model like YOLO (Redmon and Farhadi, 2018), and not a segmentation one, to save computational costs. YOLO is an R-CNN-based algorithm previously employed in the field of archaeology for the detection of

inscriptions in oracle bones (Liu et al., 2020). The YOLOv3 algorithm is faster than other R-CNN methods like Faster R-CNN. Its backbone, Darknet-53, is 1.5 times faster than ResNet-101, working at 78 frames per second (Ren et al., 2016; Redmon and Farhadi, 2018). YOLOv3 predicts at three different scales, which is similar to what the feature pyramid network does (Redmon and Farhadi, 2018; Lin et al., 2017). This structure allows the detection of small objects. The bounding boxes are predicted by the anchor boxes generated using k-means clustering with an Intersection over Union (IoU) threshold of 0.5. The class prediction is made using binary cross-entropy loss and independent logistic classifiers, the latter to facilitate multilabel classification (Redmon and Farhadi, 2018). Likewise, to label the training and validation images to create our custom data, we used LabelImg, a simple graphical image annotation tool (Lin, 2021) hat allowed us to tag images directly in YOLO format.

2.2.2.3. Deep Learning Instance Segmentation

As mentioned in the previous section, in recent years, R-CNN models have become very common in archaeology, in particular segmentation algorithms such as the one we use, Mask R-CNN, and DeepLabV3+ (Landauer et al., 2022). For our two segmentation articles, we will have employed Mask R-CNN (Waleed, 2017), since we are looking for instance segmentation rather than semantic. Mask R-CNN detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance (He et al., 2017). It extends Faster R-CNN (Girshick, 2015; Ren et al., 2016) by adding a branch for predicting segmentation masks, a small fully convolutional network (FCN) (Long et al., 2015), on each region of interest (RoI), in parallel with the existing branch for classification and bounding box regression. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN (He et al., 2017). Likewise, VGG Image Annotator (VIA) from Oxford has been used to label mounds (Dutta et al., 2022).

CNNs are much more efficient than traditional ML algorithms such as RF but their training requires much more computational power and a much larger amount of training data (more than 10k examples are usually recommended). However, once the training is finalized, the algorithm can be executed with a commercial laptop within minutes. Rather than detection algorithms, we have preferred to employ segmentation procedures, both able to locate the position of objects in an image, as these are able to define the shape of the object of interest while automatically tagging the rest of the image as background.

2.2.2.4. Training Approaches

CNN-based automated detection algorithms require large amounts of data for their training, but this is not common in archaeology where the number of known archaeological features to train a ML algorithm is very low, as in our case studies. The distribution of data has been 50% for training data, to train the algorithm, 25% for validation data, to refine the training based on the accuracy metrics, and 25% for test, to make a final evaluation of the trained and validated model (Torres, 2020).

Despite having few training data, some studies like the burial mounds one did not require much due to the characteristic circular shape of the tumuli. However, the archaeological elements of the historical maps article, despite being mounds like those previously mentioned where we encountered a similar problem, are much more diverse. Since they are symbols drawn by the human hand and not images of their actual form, whether aerial or satellite, the features are much more different from each other, so little training data was not enough. Also, potsherds are so diverse that we need many examples of them. To avoid this issue we have develop a series of DA techniques such as random translation, random rotation, resizing, the so-called so-called Doppelgänger technique, refinement and creation of synthetic data; thus introducing both positive and negative new training data. Some different approaches have also been implemented like ML hybrid approaches and CL strategies (Soviany et al., 2022).

2.2.2.5. Validation Approaches

In computational archaeology, trained ML models have been shown to perform worse in areas with low-density of archaeological features than in high-density ones (Soroush et al., 2020; Verschoof-van der Vaart et al. 2022). When performing large-scale detection with few sites, many of FPs are introduced, which reduces the accuracy of the algorithm. However, the real archaeological scenarios are those where the density of elements to be detected is low. During a survey, the most common approach in archaeological case-studies is to ignore the actual density of archaeological features, so to be a useful tool, the developed of a ML algorithm must provide good results also for low-density areas. So, we have focused our validation on low-density scenarios to ensure a practical algorithm when applied to large-scale scenarios.

(1)

(2)

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

The validation metrics we have used have been the recall and precision values (Torres, 2020). The recall value is the number of elements correctly detected, true positives (TPs), out of the total present, TPs and false negatives (FNs), (1); and the precision is the number of elements correctly detected, TPs, out of the total detections, TPs and FPs, which include correct and incorrect detections (2). The most recent approaches to the detection of archaeological features are usually able to detect a high percentage of the test dataset's TPs, but they also include a large proportion of FPs. For example, Davis et al. detected 17 of 18 mounds present in their study area but they also detected 3,237 more mounds (Davis et al., 2021). After visual validation, they confirmed that from the 3,254 detected mounds, only 287 corresponded to possible burial structures (equivalent to an 8.8% success rate), pending field validation. Verschoof-van der Vaart et al. obtained a recall value of 0.796, but the precision value was 0.141 (86% of detected tumuli were FPs) (Verschoof-van der Vaart et al., 2020). Trier et al. (2021) detected 38% of known tumuli in their study area, but 89% of the detected features were identified as FPs (Trier et al., 2021). Therefore, aiming to develop practical tools, which can be applied to real archaeological survey scenarios, we have decided to focus not only on good recall but also on obtaining high precision values.

RESULTS

3. Potential of Multitemporal Lidar for the Detection of Subtle Archaeological Features under Perennial Dense Forest

Iban Berganzo-Besga[1]; Hector A. Orengo[1,2*]; Joan Canela[3] and Maria Carme Belarte[2,3]

1. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

2. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

3. Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

*Author to whom correspondence should be addressed.

Land **2022**, *11*(11), 1964; **https://doi.org/10.3390/land11111964**

Received: 23 July 2022 / Revised: 28 September 2022 / Accepted: 6 October 2022 / Published: 2 November 2022

Abstract

This paper presents a method for the merging of lidar-derived point clouds of the same area taken at different moments, even when these are not co-registered. The workflow also incorporates the filtering of vegetation allowing the classification of unclassified point clouds using the ground points of reliable coverages. The objective is to produce a digital terrain model by joining all ground points to generate a higher resolution model than would have been possible using a single coverage. The workflow is supplemented by a multi-scale relief visualisation tool that allows for better detection of archaeological micro-reliefs of variable size even in areas of complex topography. The workflow is tested in six Iberian Iron Age sites, all of them located in mountain areas with dense Mediterranean perennial forests and shrub vegetation.

Keywords:

remote sensing; lidar; archaeology; iron age; Iberian Peninsula; urbanism; Google Earth Engine

3.1. Introduction

Lidar-derived DTMs have recently become an essential tool in landscape archaeology (Bewley et al., 2005; Evans et al., 2007; Opitz et al., 2015; Parcero-Oubiña and Nión, 2022). Besides the important increase in resolution with respect to available global DSMs such as SRTM, ASTER GDEM, and ALOS, lidar allows the filtering of vegetation cover, buildings, and other human-made structures to produce high-resolution bare-soil DTMs in which archaeological remains can be detected. Raw lidar-derived DTMs only allow the detection of the most evident features and they require the use of DTM visualisation methods to detect subtler topographic changes indicative of the presence of archaeological remains. The combination of lidar-based DTMs and visualisation methods has produced important discoveries in both highly investigated archaeological landscapes and areas that, due to their dense vegetation cover, had not been previously surveyed (Devereux et al., 2005).

However, most lidar applications have been developed in areas with relatively sparse vegetation cover or in deciduous forest areas after leaves have been shed for the season. The use of lidar-derived DTMs for the detection of archaeological structures in areas covered by dense shrubs and perennial forests typical of the Mediterranean has been rather scarce and rarely produces results of equivalent quality as studies that illustrate the potential of the technique (Doneus et al., 2020). Moreover, many Mediterranean forested areas correspond to mountain environments, which tend to hide micro-reliefs when using standard relief visualisation techniques as these do not respond well to abrupt topographic changes present in mountain areas.

During recent years multitemporal RS data has started to be employed in landscape archaeology to analyse landscape patterns and visualise subtle landscape features (Orengo and Petrie, 2017; Garcia-Molsosa et al., 2019), to detect change affecting archaeological sites (Agapiou et al., 2019; Rayne et al., 2020), and to increase the efficiency of automated site detection methods (Orengo et al., 2020; Berganzo-Besga et al. 2021). However, these applications are limited to pixel-based data (mostly multispectral imagery and synthetic aperture radar) and despite the increasing availability of datasets, the combination of lidar coverages acquired during different time periods has rarely been applied for archaeological research.

This paper aims to show how the combined use of multitemporal lidar datasets and multiscale approaches to micro-relief visualisation (Orengo and Petrie, 2018; Guyot et al., 2021) can

improve the visualisation of subtle archaeological remains present as micro-reliefs in the ground surface even in complex environments such as forested mountain Mediterranean areas.

3.2. Case Studies

In order to test our approach, we produced a series of DTMs using multitemporal lidar data from a series of Iberian Iron Age settlements **(Figure 11)** from which traces of architectural remains have been detected on site but not thoroughly mapped. They all are located in low mountain forested areas. The sites under investigation are described in detail in previous work (Belarte et al. 2020), these correspond to the settlements of:

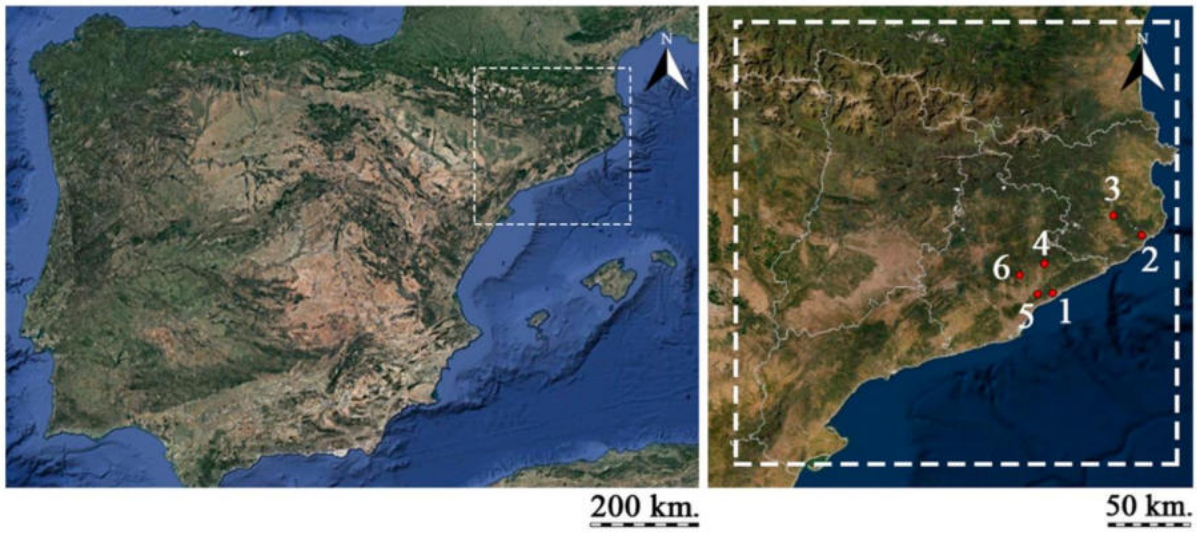


Figure 11. The location of Burriac (1), Castell Barri (2), Puig d'en Rovira (3), Puig Castell (4), Sant Miquel (5), and Torre Roja (6), the settlements under study, in the NE of the Iberian Peninsula (within the white line frame).

- Burriac (Cabrera de Mar, Maresme, Barcelona) (41°32′08″ N, 02°23′11″ E) (Barberà and Pascual, 1980). This large site of around 10 ha corresponds to a first-order settlement, which can probably be identified as ancient Ilturo;
- Castell Barri (Calonge, Baix Empordà, Girona) (41°51′19″ N, 03°02′48″ E). Surface ceramic fragments have been recovered from this second-order town, which provided an approximate dating of 4th–3rd century BC (Oliva, 1947);

- Puig d'en Rovira (La Creueta de Quart, Girona) (41°57′48″ N, 02°50′11″ E). This small settlement was occupied from the late 4th to the 2nd century BC (Riuró, 1943) with a Roman phase of use during the 1st and 2nd centuries AD;
- Puig Castell (Cànoves i Samalús, Vallès Oriental, Barcelona) (41°41′52″ N, 02°19′30″ E). This is another large (4 ha) first-order settlement that was occupied from the 6th century to the Roman period and probably corresponds to the ancient city of Lauro (Estrada, 1955; Garcia-Molsosa et al., 2015);
- Sant Miquel (Vallromanes/Montornès, Vallès Oriental, Barcelona) (41°31′44″ N, 02°16′26″ E). This middle-sized town (1.5–2.5 ha) was occupied from the second half of the 5th to the end of the 2nd century BC (Asensio and Guitart, 2010);
- Torre Roja (Caldes de Montbui, Vallès Oriental, Barcelona) (41°38′00″ N, 02°08′35″ E). This site was occupied from the late 6th century to the Middle Ages. Recent excavations have revealed an Iberian habitation area occupied from the 5/4th to the 1st centuries BC (Fortó and Maese, 2010).

The sites that were analysed correspond to birthplaces of the Iberian culture, that spread throughout the Mediterranean area of the Iberian Peninsula. The formation period of this culture corresponds to the Late Bronze Age and Early Iron Age, where a transformation process (increase in population, economic growth, trade with Greeks and Phoenician-Punics, and the introduction of iron) led to the formation of hierarchical societies, the institutionalization of hereditary inequality, and the emergence of political entities that can be considered as early states (Sanmartí, 2009). Throughout the Early Iberian period, there was a hierarchization of the settlement system, which was organized into nuclei of different sizes and functions, a system that would be clearly consolidated in the Middle Iberian Period (400–200 BC). The first stage of the Roman conquest was characterized by the abandonment and destruction of numerous Iberian settlements as a consequence of the Second Punic War (218–201 BC) military operations and the repression of uprisings by M. Porcius Cato (Sanmartí, 2009; Fernández et al. 2015).

Iberian settlements in the study area are usually classified into four categories, from dispersed rural settlements to urban sites defended by complex fortifications (5 to 10 ha); intermediate categories of sites correspond to villages and fortified sites under 1 ha (Asensio et al., 2001; Orengo et al., 2014). Common features of Iberian settlements include the preference for settling on low hills with good natural protection, the presence of defensive walls, and an urban plan with houses distributed in rows sharing walls. The main characteristics of the first and second

order of settlements, both of which possess urban characteristics, are high density of occupation, urban planning, public buildings, complex dwellings, artisan activities, storage capacity, high-quality imported goods, and evidence of administration. While some towns had some administrative and political control, other settlements had more economic or residential usage.

These settlements all share a common ecosystem and complex topography. The climate of the Catalan area under study, denoted as Csa according to the Köppen classification, is characterized by a temperate climate with dry summers and with an average temperature of above 22 °C during the warmest month. The biogeographic features that characterize the Csa climate are rivers with an irregular regime, vegetation with typical characteristics of semi-arid climates, and reddish-brown soils. The rivers have a lower flow in summer, sometimes drying up, and strong and sudden floods in autumn. The main characteristic of the vegetation is its adaptation to summer heat and dryness through short trunks, thick bark, deep roots, and small strong evergreen leaves. The variable aridity and composition of the soil give rise to four typical formations: herbaceous steppe (more arid, esparto), shrubby steppe (thuyas and dwarf palms), maquis (siliceous soils, undergrowths with heather and isolated trees such as pine trees), and garriga (limestone soils, bare trees where oak is characteristic, and aromatic plants such as thyme). The diverse soils are strongly eroded by heavy rains and deforestation (Aguilera et al., 2009).

3.3. Methods and Sources

A typical lidar system consists of airborne laser scanning (ALS), a GPS-derived aircraft position, an inertial measurement unit (IMU), and the measurement of the deflection angle of the transmitted laser beam. The point clouds generated by this system allow the creation of high-resolution DTMs of the landscape by filtering the points corresponding to vegetation and other elements above the surface of the terrain (Wehr and Lohr, 1999; Hesse, 2010). The remaining points corresponding to the ground surface are usually interpolated to create a continuous raster surface, the DTM. Lidar-derived technologies have been used in multiple disciplines with a geographical basis (Gutiérrez-Antuñano et al. 2017) such as engineering, geomorphology, hydrology, landscape architecture, and archaeology. In archaeology, their most common use, besides GIS-based topographic analysis, has been the detection of archaeological structures visible in the ground surface as topographic imprints. Some of these

are very subtle and require specific visualisation techniques for their identification. The use of visualization algorithms can reveal important geomorphological and cultural information (Orengo and Petrie, 2018).

In this study, we employed three different coverages. Initial tests employed the publicly available lidar data at the Cartographic and Geologic Institute of Catalonia (ICGC). This coverage was obtained from 2008 to 2011. It covered all of the Catalan territories with a minimum density of 0.5 points/m^2. Although in some areas it could reach 4.4 points/m^2(including all point classes), for the sites under study it ranged from 0.63 points/m^2 for Sant Miquel (acquired in 2008), 0.67 for Burriac (acquired in 2008), 0.74 for Castell Barri (acquired in 2010), 0.91 for Puig Castell (acquired in 2008), and 1.09 for Torre Roja (acquired in 2010) and Puig d'en Rovira (acquired in 2010). The data were calibrated and adjusted with multiple areas of topographic control across the Catalan territory, reaching an average Root Mean Square (RMS) of 6 cm in flat bare soil areas. Lidar point clouds were classified by the ICGC into 14 classes. For the classification of the ground points, besides the automated classification followed for all classes, these were edited manually by an expert. Although of high quality, the testing of this dataset proved the point density to be insufficient in identifying new archaeological structures in the six Iberian settlements. Consequently, a second lidar coverage was carried out specifically for this study. The new dataset was also acquired by the ICGC but with a higher resolution of 5 points/m^2. This dataset was only automatedly classified by the ICGC and incorporated many misclassified points. Several reclassification attempts were made to improve the initial automated classification until a reasonably well-filtered DTM was obtained. Although this DTM provided further insight into the architectural structure of the sites under investigation the improvements were relatively meagre. During the development of the study, the ICGC released a new lidar coverage of Catalonia which was initially acquired between 2016 and 2017, although the areas under interest were all acquired during 2016. This dataset presented similar characteristics to that of the initial publicly available dataset but with a slightly higher point density. The point densities present at the different sites under study were 0.79 for Castell Barri, 0.97 for Puig Castell, 1.05 for Burriac, 1.17 for Sant Miquel, 1.64 for Torre Roja, and 1.73 for Puig d'en Rovira. The DTMs derived from these data did not provide enough topographic information to obtain further insights into the structure of the sites under investigation.

In order to obtain a higher density of points and thereby increase the resolution of the final DTM, we decided to employ all data available within a single workflow **(Figure 12).** This

included all ICGC's coverages (both 2008–2011 and 2016–2017 coverages) and the commissioned lidar data (5 points/m^2). This allowed us to obtain higher-resolution data with an average density of 6–9 points/m^2 (including all point classes). Both public ICGC datasets were already co-registered and well-classified. Although the commissioned coverage presented the highest density of points, these were not co-registered to the publicly available ICGC datasets and their classification was inaccurate. Therefore, it was necessary to co-register this dataset with the others and filter ground points for the development of a DTM integrating all available data.

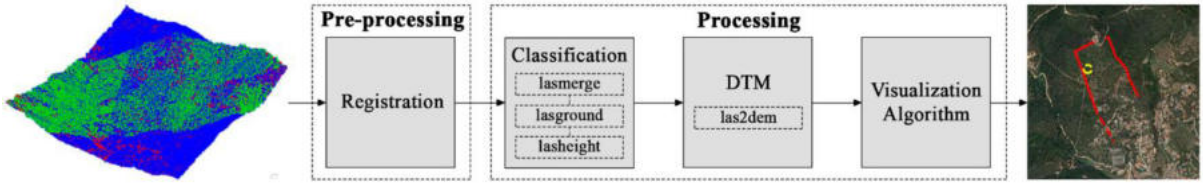


Figure 12. The implemented workflow to detect architectural features from lidar points.

We used the 3D point cloud processing software CloudCompare v2.11.alpha to align the datasets. CloudCompare is a 3D point cloud (and triangular mesh) processing software. It implements the Iterative Closest Point (ICP) algorithm (Besl and McKay, 1992) to minimize the difference between two point clouds with an RMS error difference between two iterations lower than the 1.0×10^{-5} threshold. After several co-registration attempts, it became apparent that the best results were obtained when using the combined classified ground points of the public datasets to co-register the ground points of the commissioned lidar data. Although only the ground points were utilised for the co-registration, the dataset maintained all data which was co-registered during the same process.

The next step consisted of the removal of above-surface features such as buildings and vegetation (Bewley et al., 2005). As the Csa climate vegetation is arranged in light forest formation with poor ground cover and three layers (trees, shrubs, and herbaceous cover), this step is one of the most crucial to be able to develop a DTM that can reflect small topographic variations. The lidar points were classified using LAStools (**Figure 13**), a collection of efficient, batch-scriptable, and multicore command line tools.

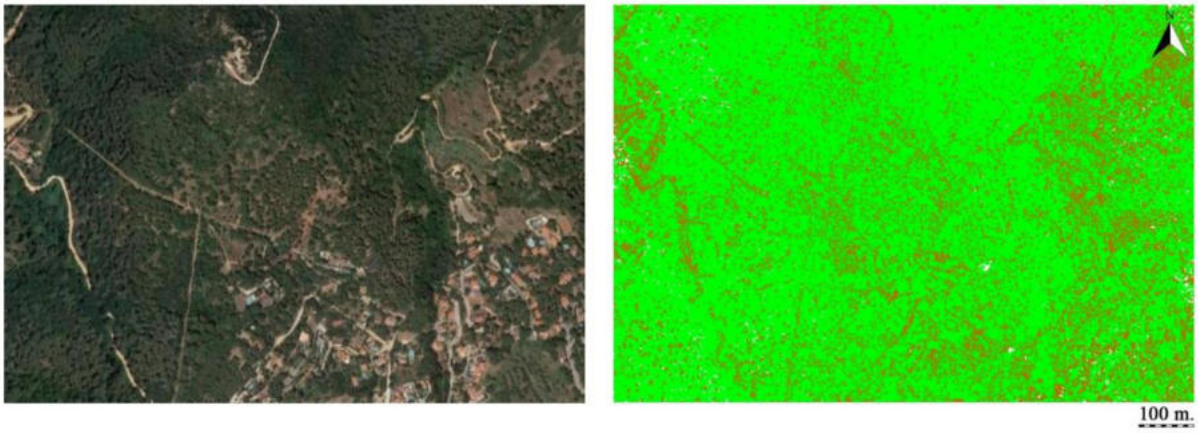


Figure 13. Satellite view (**left**) and classified lidar points (**right**) of the Burriac site.

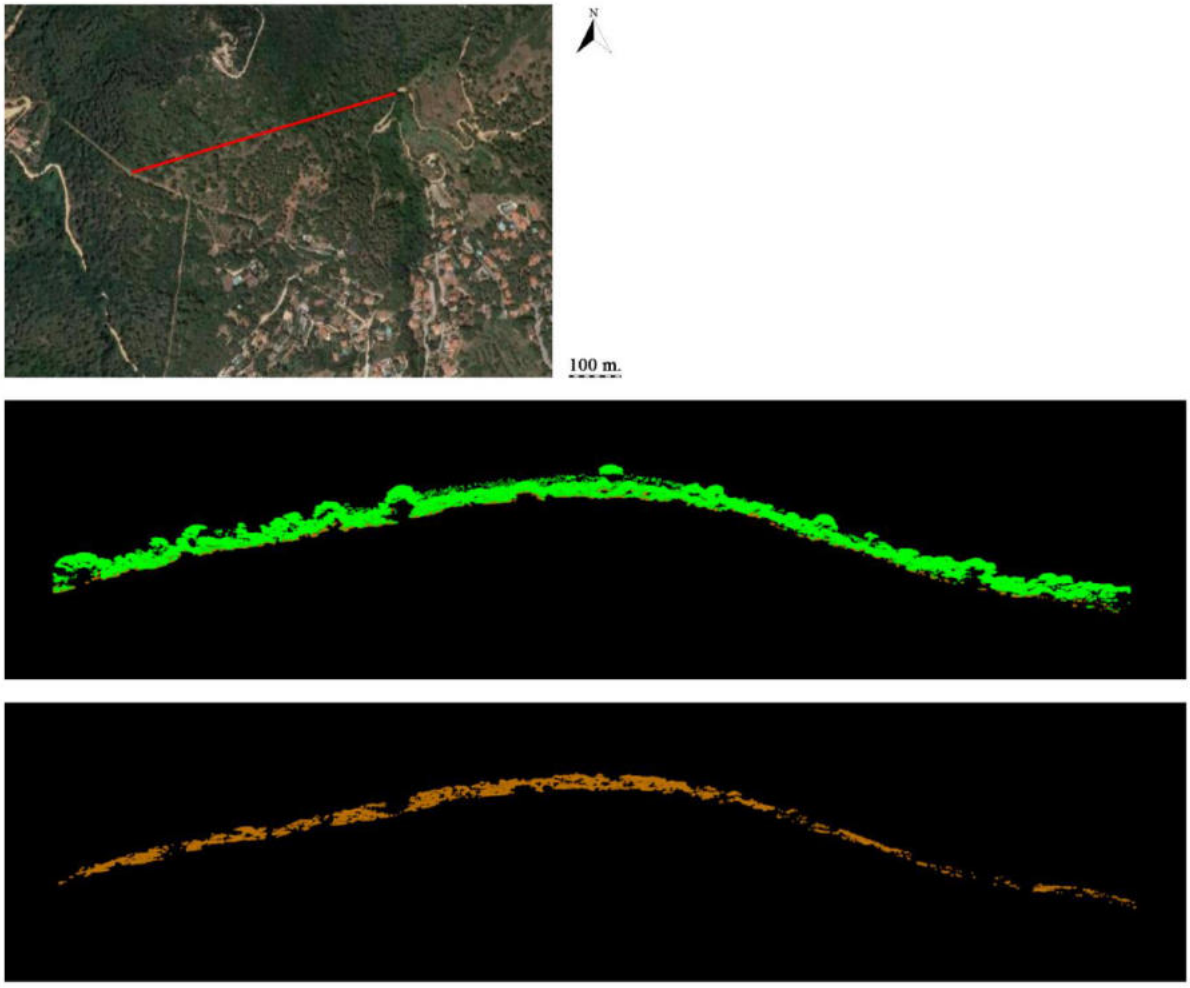


Figure 14. Profile view of the Burriac site (**centre**) and the same profile but only with the ground points (*class* = 2) (**bottom**). The cut (red) is indicated in the satellite view of the site (**top**).

LAStools also allowed us to merge the co-registered data using its *lasmerge* function and then *lasground* for bare-earth extraction, with the default step parameter of 5.0 which is adequate for forest and mountain environments. For this, the merged point clouds of ground points from

both ICGC coverages and the unclassified commissioned coverage were employed. This was the most efficient way to filter low vegetation as the *lasground* was able to employ the bare ground points of the classified coverages to better discriminate points corresponding to vegetation in the unclassified dataset. The *lasground* script classified the lidar points into ground points (*class* = 2) and non-ground points (*class* = 1). The *lasheight* function was used to keep values whose height is between −2 (*drop_below*) and 100 (*drop_above*) meters regarding previously classified terrain. As can be observed in **Figure 14**, the possible structures of the settlement cannot be seen with the naked eye in the profile view due to their very small overground topography.

The co-registered, merged, and classified data was utilised to generate a DTM using LAStools' *las2dem* function. This tool triangulates lidar points into a triangulated irregular network (TIN), and then rasterises the TIN to create a DTM. The chosen resolution of 0.25 m was chosen given the density of ground points preserved after the classification, which typically amounted to 3–5 points/m^2 in densely forested areas. Thus, choosing this ground resolution for the DTM ensured the highest possible resolution. Using QGIS v3.8 projection tools we assigned ETRS89/UTM zone 31N (EPSG:25831) spatial reference information to each DTM.

Once the filtered high-resolution DTMs for each of the sites under study were obtained, we tested several visualization algorithms to evaluate which one provided a clearer visualization of subtle topographic changes within the DTM. For this, DEM manipulation-based visualization methods (MSRM, SLRM, and slope gradient), illumination-based ones (principal component analysis of multi-azimuth shaded relief maps, which is not kernel-size dependent), and other kernel-size dependent techniques (Sky-View Factor (SVF) and Opennes) were selected (Hesse, 2010; Zakšek et al., 2011; Doneus, 2013; Kokalj and Hesse, 2017).

The best results were obtained using the MSRM (*fmn* = 0, *fmx* = 30, *x* = 1). Concerning the rest, the only one that showed results similar to the MSRM was the SLRM (*radius* = 30) (**Figure 15**). The noticeably better visualisation provided by MSRM method is probably due to its capacity to show microtopographic imprints at variable scales. Recent comparisons between multi-scale topographic visualisation, single-scale, and illumination-based methods (Berganzo-Besga et al., 2021; Guyot et al., 2021) show that multi-scale visualisation methods also produce better results for the automated detection of structures in lidar data using DL detectors. This implies a more coherent visualisation of features at slightly different scales.

GEE Code Editor, Repository and Cloud Computing Platform (Gorelick et al., 2017) was used to obtain the MSRM raster and The Relief Visualization Toolbox (RVT) for the other relief visualisation methods (Kokalj and Hesse, 2017).

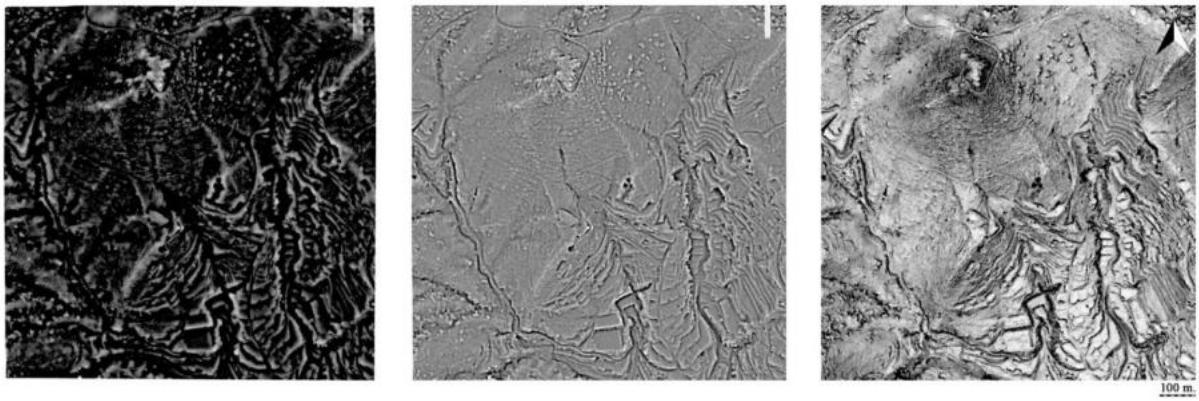


Figure 15. Comparison for the Burriac settlement of the different visualization algorithms, which have been shown to be better for the case study: MSRM, SLRM, and SVF (from left to right).

3.4. Results

Previous research (Belarte et al., 2020) based on commissioned lidar coverage presents a much lower density of points, inaccurately classified ground points, and unsophisticated visualisation methods. Unsurprisingly, comparison with the newly developed workflow shows notable improvements in the visualisation of the walled structures around five of the six archaeological sites under study (**Figure 16**). The following paragraphs will briefly describe the results obtained using the workflow described above for each of the six settlements under consideration.

The site of Puig Castell (Cànoves i Samalús, Vallès Oriental, Barcelona) presents a continuous view of the whole fortified enclosure wall. The settlement extends in a northwest–southeast direction. We did not identify any rooms inside the rampart, but we recognized one more wall in the north–south direction in the southern part of the walled area (**Figure 17**) which could have formed part of a hallway or walled access to the site.

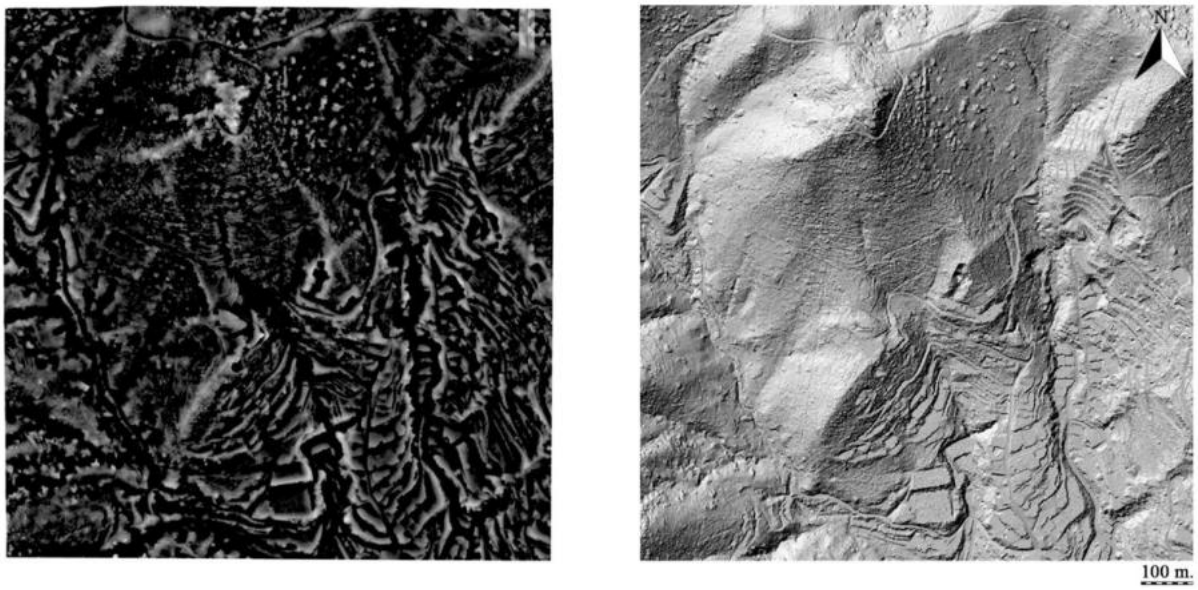


Figure 16. MSRM (**left**) and hillshade (**right**) of the Burriac site.

Figure 17. MSRM (**left**), satellite view (**centre**) and topographic map (**right**, forest in green) of the Puig Castell site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

At Burriac (Cabrera de Mar, Maresme, Barcelona) most of the walled enclosure of the northwest part has been detected. The settlement extends in a northwest–southeast direction (**Figure 18**).

Figure 18. MSRM (**left**), satellite view (**centre**) and topographic map (**right,** forest in green) of the Burriac site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

The analysis of Torre Roja (Caldes de Montbui, Vallès Oriental, Barcelona) shows a triangular walled enclosure structure. The settlement extends in a northwest–southeast direction. Moreover, an interior circular structure is observed in the northwest (**Figure 19**).

Figure 19. MSRM (**left**), satellite view (**centre**) and topographic map (**right**, forest in green) of the Torre Roja site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

The analysis of Castell Barri (Calonge, Baix Empordà, Girona) shows a large part of the perimeter wall structure. The settlement extends in a northeast–southwest direction. We were able to detect a possible additional wall in the northeast **(Figure 20).**

Figure 20. MSRM (**left**), satellite view (**centre**) and topographic map (**right**, forest in green) of the Castell Barri site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

In the case of Puig d'en Rovira (La Creueta de Quart, Girona), although the visibility of this settlement is poorer than that of the other sites, almost the entire wall structure can be observed. The settlement extends in a north–south direction. Similar to previous cases, there are no signs of interior rooms **(Figure 21).**

Figure 21. MSRM (**left**), satellite view (**centre**) and topographic map (**right**, forest in green) of the Puig d'en Rovira site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

Finally, analysis of Sant Miquel (Vallromanes/Montornès, Vallès Oriental, Barcelona) resulted in detection of several sections of wall structures. The settlement extends in a northeast–southwest direction. Two parallel walls are clearly observed in a northeast–southwest direction. In addition, the closing wall of the settlement in the northeast area is observed **(Figure 22).**

Figure 22. MSRM (**left**), satellite view (**centre**) and topographic map (**right**, forest in green) of the Sant Miquel site. In the satellite image the walls (red) and the known location of the settlement (yellow) are indicated.

3.5. Discussion

In the presented case-study, a combination of airborne lidar coverages and a multi-scale visualisation algorithm provided useful insights in all the study sites, which were located in a Mediterranean Csa environment. The results show noticeable improvement with respect to the use of a single coverage as shown in a previous study (Belarte et al., 2020). We sought to improve the visibility of archaeology-related micro-reliefs by increasing the ground point density and quality as well as employing more suitable visualization algorithms. The increase in the density of points using a purposely commissioned lidar coverage and the two public access coverages of the ICGC enabled the detection of most sites' perimeter walls and some inner large structures in the Castell Barri and Torre Roja settlements. Despite this, clear identification of smaller structures, such as the walls of houses, remains a challenge. To achieve this, a much higher point density is necessary, which will, perhaps, be achieved as more sophisticate lidar coverages become available in the future.

In any case, the delimitation of the walls and the potential extension of the archaeological sites provides important new data as:

1. It will allow for the estimation of the potential population and hierarchies between sites;
2. It will allow for the development of conscientious excavation planning, which can now focus on intramural areas and not waste efforts in potentially uninhabited sectors;

3. It will allow for more informed heritage protection measures, which will be able to take into account the site as a whole and not just the small area where remains have been detected.

The workflow presented here provides a way to co-register and filter several point clouds to increase the resolution of a DTM and, by doing so, improve the detection of subtle topographic features even in complex areas where perennial forests and shrubs combine with abrupt slopes. As many countries such as Portugal start to develop large public lidar datasets and others such as Spain, the UK, and the Netherlands have already made available multiple coverages for the same areas, the use of this and similar workflows offer a way to increase the DTM resolution and boost the detection of archaeological micro-reliefs, which are not easily identifiable in single national coverages as these rarely implement the necessary resolution.

This workflow also provides a way to combine and integrate different types of data as point clouds when a single one of those presents reliable ground points. This can be used, for example when well-classified lidar point clouds are available but these do not have enough resolution. In this case, this workflow can be implemented using the lidar ground points to register and classify photogrammetry-derived point clouds obtained, for example, with increasingly accessible UAVs or drones. Another option is the use of DGPS-derived points as the ground points which can be combined with lidar, photogrammetry-derived, or both types of point clouds. By providing a way to integrate and improve point data obtained over several periods, with different means, and at different resolutions and qualities, the proposed workflow provides an efficient way to reuse legacy data, which is widely encouraged within the discipline (e.g., Faniel et al., 2013; Opitz et al., 2018).

3.6. Conclusions

The field of landscape archaeology is rapidly developing toward the use of big structured and unstructured data. New trends such as the development of data fusion methods for multitemporal and multisource data in virtual satellite constellations (Agapiou et al., 2019; Orengo et al., 2020) and the creation of virtual data through DA processes hold much potential for the development of more holistic analysis with deeper analytical capabilities. Point cloud-data, in particular, can benefit from new approaches since many of the methods employed in archaeological surveys such as lidar, terrestrial laser scanning and total stations, differential

GPS, and photogrammetry generate cloud points. Being able to use a reliable reference dataset to structure all these diverse data through correlation, filtering, and classification methods has the potential to radically improve archaeological surveys and the detection and analysis of microtopographic archaeological features.

In this paper we presented a method for the merging and filtering of point clouds and their analysis using multiscale relief visualisation methods. Hopefully, similar workflows can unleash the emerging potential behind other sources of georeferenced point clouds.

3.7. Author Contributions

I.B.-B. and H.A.O. wrote the original paper. I.B.-B. conducted the analysis and produced the images; H.A.O. revised the text, designed the methods, and supervised the research; J.C. collected and curated the commissioned lidar data; M.C.B. contributed to the revision of the text, directed the project, and obtained the funds for the commissioned lidar data. All authors have read and agreed to the published version of the manuscript.

3.8. Funding

This research was funded by the Spanish Ministry of Economy and Competitiveness, grant number HAR2015-67946-C2-2-P…

4. Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel-2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia

Iban Berganzo-Besga[1]; Hector A. Orengo[1,2*]; Felipe Lumbreras[3]; Miguel Carrero-Pazos[4]; João Fonte[5] and Benito Vilas-Estévez[6]

1. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

2. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

3. Computer Vision Center, Computer Science Deptartment, Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellaterra, Spain

4. Institute of Archaeology, University College London, 31-34 Gordon Square, London WC1H 0PY, UK

5. Department of Archaeology, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK

6. Grupo de Estudos de Arqueoloxía, Antigüidade e Territorio, Facultade de Historia, University of Vigo, As Lagoas, s/n, 32004 Ourense, Spain

*Author to whom correspondence should be addressed.

Remote Sens. **2021**, *13*(20), 4181; **https://doi.org/10.3390/rs13204181**

Received: 21 September 2021 / Revised: 14 October 2021 / Accepted: 16 October 2021 / Published: 19 October 2021

Abstract

This paper presents an algorithm for large-scale automatic detection of burial mounds, one of the most common types of archaeological sites globally, using lidar and multispectral satellite data. Although previous attempts were able to detect a good proportion of the known mounds

in a given area, they still presented high numbers of FPs and low precision values. Our proposed approach combines random forest for soil classification using multitemporal multispectral Sentinel-2 data and a deep learning model using YOLOv3 on lidar data previously pre-processed using a multi–scale relief model. The resulting algorithm significantly improves previous attempts with a detection rate of 89.5%, an average precision of 66.75%, a recall value of 0.64 and a precision of 0.97, which allowed, with a small set of training data, the detection of 10,527 burial mounds over an area of near 30,000 km^2, the largest in which such an approach has ever been applied. The open code and platforms employed to develop the algorithm allow this method to be applied anywhere lidar data or high-resolution digital terrain models are available.

Keywords:

tumuli; mounds; archaeology; deep learning; machine learning; Sentinel-2; Google Colaboratory; Google Earth Engine

4.1. Introduction

During the last 5 years, the use of AI for the detection of archaeological sites and features has increased exponentially (Davis et al., 2021). There has been considerable diversity of approaches, which respond to the specific object of study and the sources available for its detection. Classical ML approaches such as RF to classify multispectral satellite sources have been used for the detection of mounds in Mesopotamia (Menze and Ur, 2012), Pakistan (Orengo et al., 2020) and Jordan (Liss et al., 2017), but also for the detection of material culture in drone imagery (Orengo and Garcia-Molsosa, 2019). DL algorithms, however, have been increasingly popular during the last few years, and they now comprise the bulk of archaeological applications to archaeological site detection. Although DL approaches are also diverse and include the extraction of site locations from historical maps (Garcia-Molsosa et al., 2021) and automated archaeological survey (Orengo et al., 2021), a high proportion of their application has been directed towards the detection of archaeological mounds and other topographic features in lidar datasets (e.g., Davis et al., 2021; Verschoof-van der Vaart et al., 2020; Trier et al., 2021; Trier et al. 2019; Guyot et al., 2018).

This is probably due to the common presence of tumular structures of archaeological nature across the globe but also to the simplicity of mound structures. Their characteristic tumular shape has been the primary feature for their identification on the field. They can therefore be easily identified in lidar-based topographic reconstructions presented at sufficient resolution. The simple shape of mounds or tumuli is ideal for their detection using DL approaches. DL-based methods usually require large quantities of training data (in the order of thousands of examples) to be able to produce significant results. However, the homogenously semi-hemispherical shape of tumuli, allows the training of usable detectors with a much lower quantity of training data, reducing considerably the effort required to obtain it and the significant computational resources necessary to train a CNN detector. This type of features, however, present an important drawback. Their common, simple, and regular shape is similar to many other non-archaeological features and therefore studies implementing methods for mound detection in lidar-derived and other high-resolution datasets are characterised by a very large presence of FPs (Verschoof-van der Vaart et al., 2020; Soroush et al. 2020).

Given the importance of tumuli in the archaeological literature and in that dealing with the implementation of automated detection methods in archaeology, this paper builds up from current approaches, but incorporates a series of innovations, which can be summarised as follows:

1. The use of RF ML classifier to classify Sentinel-2 data into a binary raster depicting areas where archaeological tumuli may be present or not,

2. DL approach using a relatively unexploited DL algorithm in archaeology, YOLOv3, which provides particularly efficient outputs. To boost the efficiency of the shape-detection method a series of innovations were implemented:

   - Pre-treatment of the lidar dataset with a MSRM (Orengo and Petrie, 2018), which, contrary to other methods, is usually employed to improve the visibility of features in lidar-based DTMs, considers the multi-scale nature of mounds;
   - The development of DA methods to increase the effectivity of the detector. One of them, the training of the CNN from scratch applying own pre-trained models created from simulated data;
   - The use of publicly accessible computing environments, such as GEE and Colab, which provide the necessary computational resources and assure the method's accessibility, reproducibility and reusability.

We tested this approach in the entire region of Galicia, located in the Northwest of the Iberian Peninsula. Galicia is an ideal testing area due to the following reasons: 1) its size, which allowed us to test the method under a diversity of scenarios at a very large scale (29,574 km^2, 5.8% of Spain), to our knowledge the largest area to which a CNN-based detector of archaeological features has ever been applied; 2) the presence of a very well-known Atlantic burial tradition characterised by the use of mound tombs; and 3) the availability of high-quality training and test data necessary for the successful development of the detector.

Previous research on this area has highlighted a very dense concentration of megalithic sites, mainly comprised by unexcavated mounds covered by vegetation. They present an average size of 15–20 m in diameter, and 1–1.5 m high. In some cases, the mound covers a burial chamber made of granite constituting a dolmen or passage grave (Rodríguez-Casal, 2007; Carrero-Pazos and Vilas-Estévez, 2016). The regional government (in Galician *Xunta de Galicia*) has been developing survey works since the 1980s, resulting in an official sites and monuments record. This official catalogue currently has more than 7000 records for megalithic mounds, although problems regarding its reliability have recently been pointed out (Carrero-Pazos, 2017). Another issue relates to the archaeological detection of those sites during fieldwork. The dense vegetation and forests covering a high percentage of the Galician territory and their subtle topographic nature, which makes many of them virtually invisible to the casual observer, complicates the detection of these structures even for specialised archaeologists. These problems have been identified in other Iberian and European areas (Bourgeois, 2013; Rodríguez-Del Cueto and Carrero-Pazos, 2021). The use of automatic detection methods can hugely help to validate and increase heritage catalogues' records, protect those cultural resources, and boost research on the large-scale distribution of the cultural processes that created them. Although visual approaches using lidar data have been employed for the detection and analysis of barrows in Galicia (Carrero-Pazos and Vilas-Estévez, 2016; Carrero-Pazos et al. 2015), no automatic detection of megalithic burial mounds has ever been attempted before in the area.

4.2. Materials and Methods

Most recent research on archaeological feature detection using lidar datasets has used algorithms based on R-CNN. This model is an object detection algorithm based on a combination of classical tools from CV and DL that has achieved significant improvements, of

more than 30% in some cases, in detection metrics using reference datasets within the CV community (Girshick et al., 2014).

However, the use of single-channel (or single band images) CNN-based approaches for the detection of archaeological tumuli in lidar-derived DSMs has frequently encountered strong limitations, as they cannot readily differentiate between archaeological tumuli and other features of tumular shape, such as roundabouts or rock outcrops. Initial tests solely using an R-CNN-based detection method and a filtered DTM detected hundreds of FPs corresponding to roundabouts, rock outcrops (in mountain and the coastal areas), house roofs, swimming pools but also multiple mounds in quarries, golf courses, shoot ranges, and industrial sites between others. As these presented a tumular shape, they could not have been filtered out to improve the training data without losing a large quantity of archaeological tumuli. This is a common problem in CNN-based mound detection (see, for example, Verschoof-van der Vaart et al. 2020).

To overcome this problem, a workflow combining different data types and ML approaches has been newly developed for this study:

4.2.1. Digital Terrain Model Pre-Processing

Pre-processing of the DTM is a common practice in DL-based detection. The use of micro-relief visualisation techniques in particular highlights archaeological features that are almost or completely invisible in DTMs (Berganzo-Besga et al., 2022b).

The DTM employed to conduct DL-based shape detection was obtained from the Galician Regional Government Geographical Portal (*Información Xeográfica de Galicia*) (Información Xeográfica de Galicia, 2021). The lidar-based DTM (*MDT_1m_h50*) was considered adequate due to its good quality (even in forest-filtered areas), its resolution of 1 m/px and its public availability. The DTM allowed a good visualisation of all mounds used for training data (**Figure 23**).

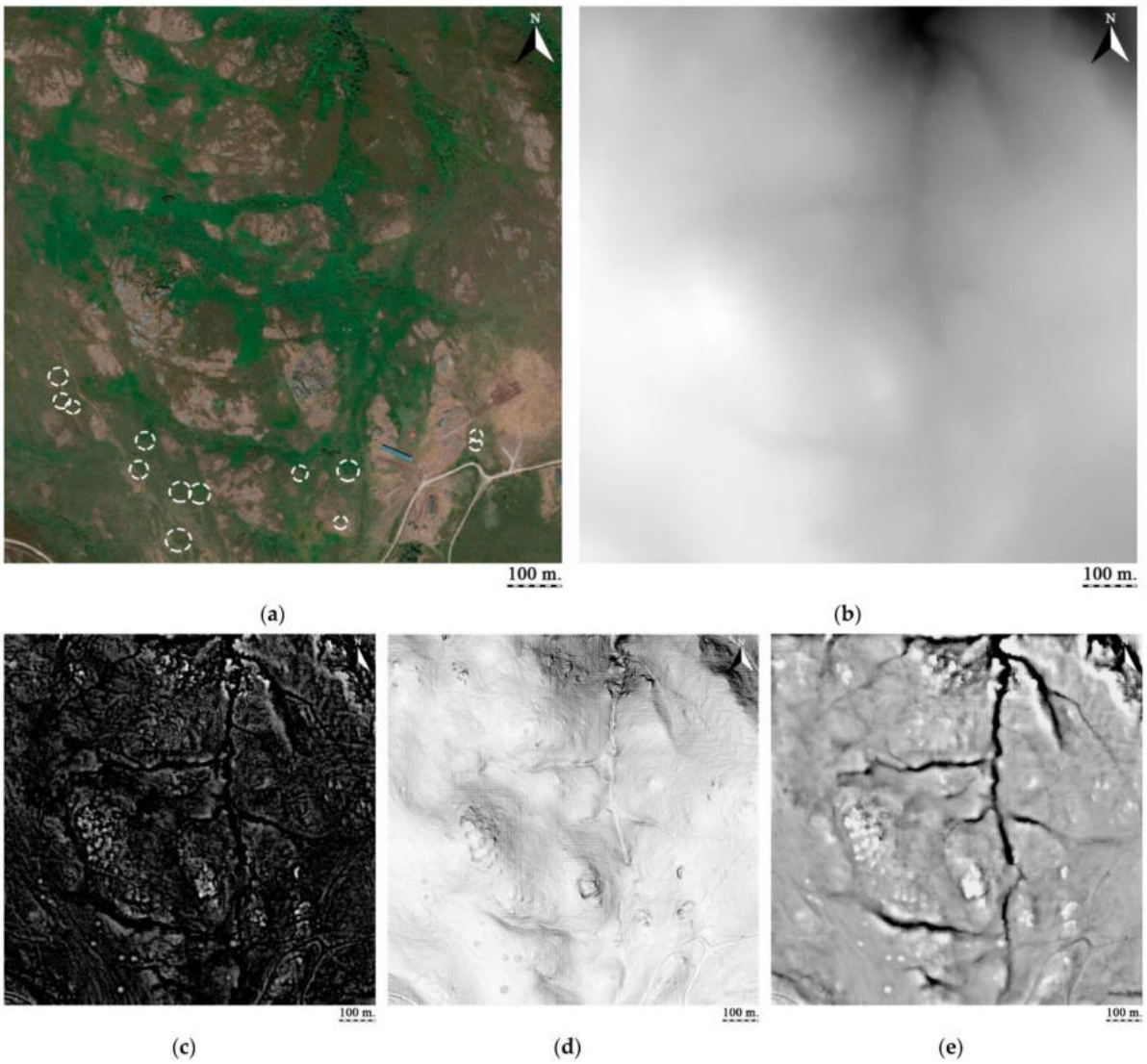


Figure 23. A comparison of a part of the used training data to evaluate which of the visualisation algorithms were better suited for the detection of burial mounds: (**a**) satellite view of that area with the known tumuli marked; (**b**) DTM; (**c**) MSRM; (**d**) slope gradient; (**e**) SLRM.

In a first approximation to mound detection using DL, we used the DTM data for algorithm training, but, as expected, an average precision (AP) of 21.81% indicated that a pre-processing stage was required on the input data. Three common relief visualization techniques were tested to improve the input data and thus facilitate the detection of burial mounds (**Figure 23**): 1. MSRM (*fmn* = 1, *fmx* = 19, *x* = 2) (Orengo and Petrie, 2018); 2. slope gradient (Doneus and Briese, 2006; Challis et al., 2011); and 3. SLRM (*radius* = 20), which is a simplified local relief model (Hesse, 2010). These constitute the most used lidar pre-processing methods for the detection of small-scale features and those in which the known burial mounds were best

observed with the naked eye. The RVT was used to obtain the slope and SLRM raster files (Kokalj and Somrak, 2019) and GEE Code Editor, Repository and Cloud Computing Platform (Gorelick et al., 2017) for the MSRM. The best results were obtained using MSRM (see the results section for details), and therefore it was the one employed for the pre-treatment of the DTM in this study.

4.2.2. Deep Learning Shape Detection

For the DTM-based shape detection we used YOLO (Redmon and Farhadi, 2018), an R-CNN-based algorithm previously employed in the field of archaeology for the detection of inscriptions in oracle bones (Liu et al., 2020). The YOLOv3 algorithm is faster than other R-CNN methods like Faster R-CNN. Its backbone, Darknet-53, is 1.5 times faster than ResNet-101, working at 78 frames per second (Ren et al., 2016; Redmon and Farhadi, 2018). YOLOv3 predicts at three different scales, which is similar to what the feature pyramid network does (Redmon and Farhadi, 2018; Lin et al., 2017). This structure allows the detection of small objects. The bounding boxes are predicted by the anchor boxes generated using k-means clustering with an Intersection over Union (IoU) threshold of 0.5. The class prediction is made using binary cross-entropy loss and independent logistic classifiers, the latter to facilitate multilabel classification (Redmon and Farhadi, 2018).

An Nvidia Titan XP graphics processing unit (GPU) with 12 GB of RAM hosted at the CVC of the Autonomous University of Barcelona (UAB) was used to run the DL algorithms. The chosen work environment was the parallel computing platform CUDA 11.2, the ML library Tensorflow 2.1.0, the DL library cuDNN 8.1.1, the software development tool CMake 3.20.2 and the CV library OpenCV 4.5.2 as recommended for YOLOv3 (Bochkovskiy, 2021).

As training and validation data, we used the current known burial mounds data obtained from the studies led by M. Carrero-Pazos and B. Vilas (Carrero-Pazos, 2017; Vilas-Estévez, 2015) in Galicia and J. Fonte in the area of Northern Portugal (**Figure 23**). The database comprised a total of 306 tumuli. From these, 200 were employed for training and 106 for validation. For the same map scale, the training, validation, and detection images size has been 1024 × 1024 pixels. The training and detection data size should not differ more than 40% to avoid FPs (Bochkovskiy, 2021). Since there is less representation of smaller diameter mounds in the training data, a DA process has been applied to add resized burial mound as new training data. DA seeks to generate more training data from our available data through a series of random

transformations to the image (Torres, 2020). As can be seen from **Figure 24**, after scaling the images to 75% and 50% of their size (DA1), practically the equivalent of the training data for small and large mounds was achieved, taking an average diameter of 18 m (Carrero-Pazos and Vilas-Estévez, 2016). DA1 added 400 more mounds for training. Likewise, to label the training and validation images to create our custom data, we used LabelImg, a simple graphical image annotation tool (Lin, 2021) that allowed us to tag images directly in YOLO format. In this step, images without burial mounds were not included.

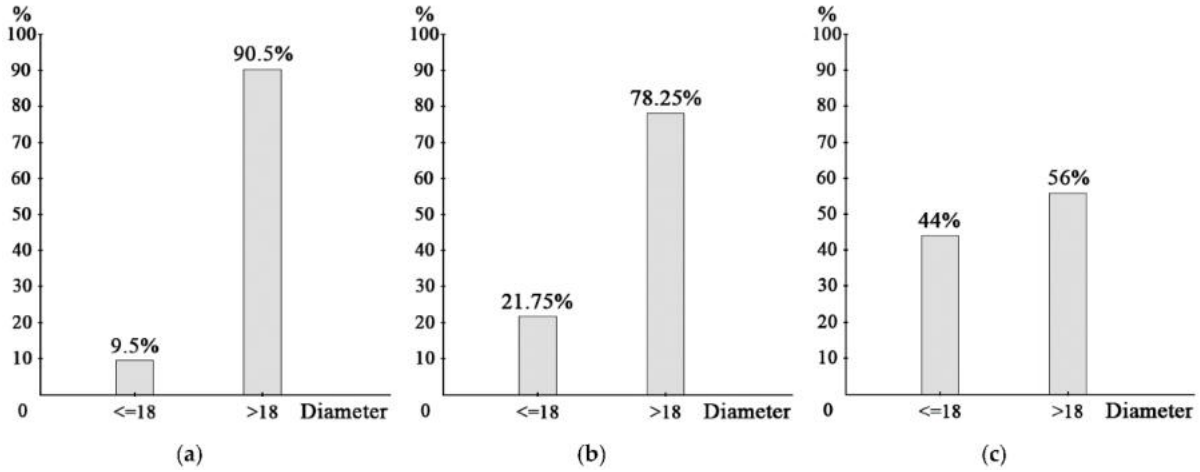


Figure 24. Number of training burial mounds with diameters of less than or equal to 18 m: (**a**) without DA; (**b**) adding the scaling to 75%; (**c**) adding the scaling to 75% and 50% (DA1).

The initially trained algorithm produced multiple FNs and FPs throughout Galicia (**Figure 25**). Additionally, some FNs were quite differently shaped from the training mounds. This issue led us to consider introducing model refinement procedures.

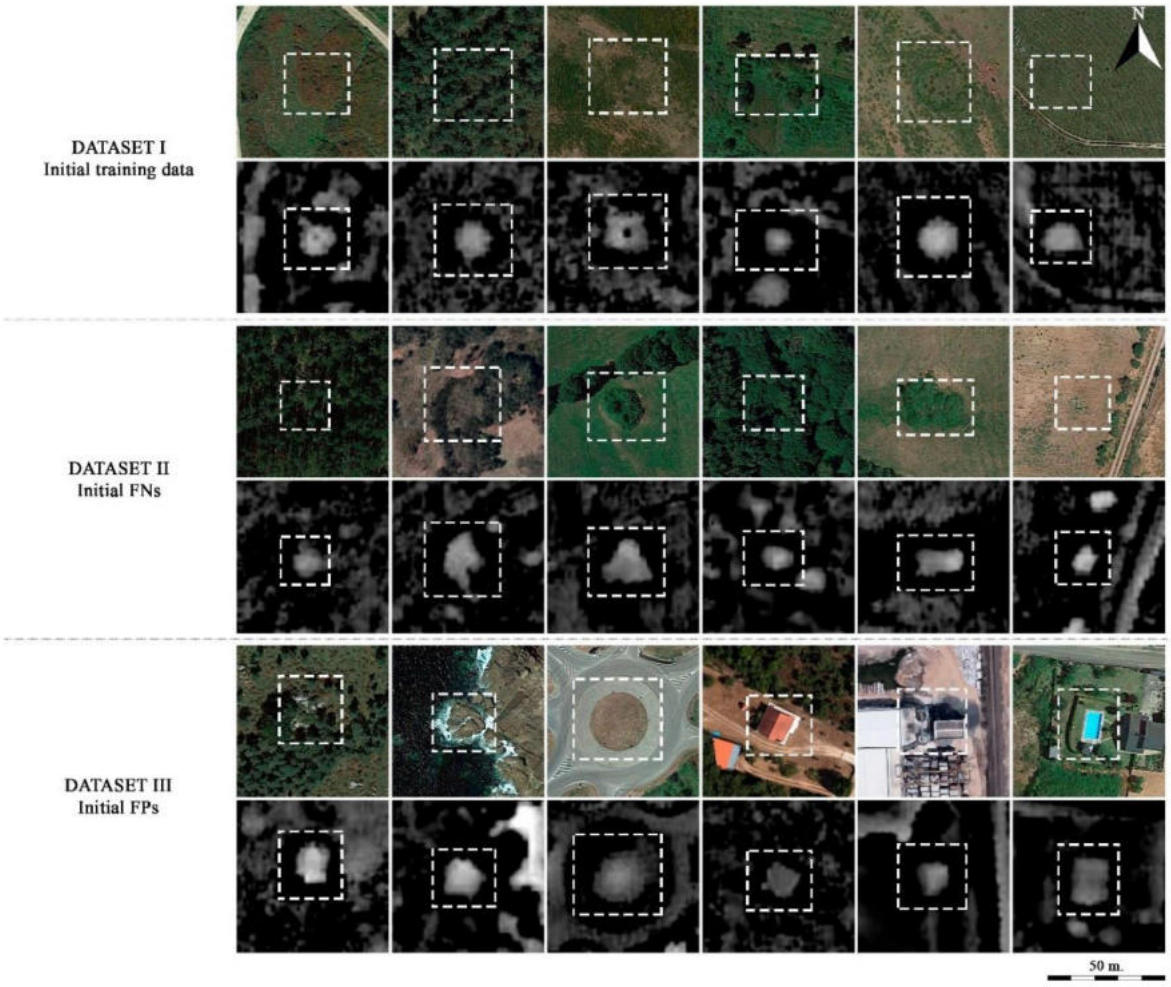


Figure 25. MSRM training data examples (Dataset I) used for the algorithm, and its FN (Dataset II) and FP examples (Dataset III) from our initial detection in Galicia. The corresponding top image for each pair is a visible satellite image, shown for the sake of visualisation, but not used in our process.

4.2.3. Model Refinement

In our initial model, the use of a CNN-based detection method and a filtered DSM detected hundreds of FPs corresponding to roundabouts, rock outcrops (also on the coast), house roofs, and swimming pools, but also multiple mounds in quarries, golf courses, shoot ranges and industrial sites between others. As many of these (particularly roundabouts) presented a tumular shape they could not have been filtered out to improve the training data without losing a large quantity of archaeological tumuli. This is a frequent problem in CNN-based mound detection. Through model refinement, we sought to reduce both FNs and FPs. In this retraining, 278 missing new burial mounds (FNs) and 88 FPs were collected from the previous training steps as new training data. From this step onwards, images without burial mounds were

included as training, obtained from the aforementioned FPs. In addition to training with FPs, a second training was proposed adding the FPs as a new extra class to find out if the algorithm was able to filter them more efficiently. To increase the burial mounds correctly detected, we proceeded to apply DA, beyond that initially developed. For this, two new models were tested. The first was a random rotation at different angles of the training features (DA2) and the second was the use of a pre-trained initial weight created specifically for the detection of circular shapes (DA3). However, these did not produce significant improvements, and were not incorporated in the final model (see discussion for details).

Despite model refinement, several FPs remained. To remove them, a filtering step using official urban, industrial and road layers was proposed. In a previous attempt to tackle this issue, Verschoof-van der Vaart et al. (2020) developed a three-level location-based ranking using the information provided by soil-type and land-use maps (Verschoof-van der Vaart et al., 2020). Instead of a ranking, such as that proposed by Verschoof-van der Vaart et al. (2020), we simply selected and eliminated the mounds detected in these areas after checking that all of them corresponded to FPs. Even though this approach eliminated most of the detected FPs in these areas, our results still included many FPs as land-use maps for the area do not classify as urban several areas in which isolated houses, swimming pools or roundabouts are present. Also, soil type maps included within the same category areas with potential archaeological mounds and FPs. For example, many archaeological mounds were located within granitic grasslands but at the same time, the specific nature and shape of granitic outcrops within these grasslands created many FPs that could not be filtered using this approach. In addition, some correct burial mounds close to the removed areas were also eliminated.

4.2.4. Random Forest Classification of Multitemporal Sentinel-2 Data

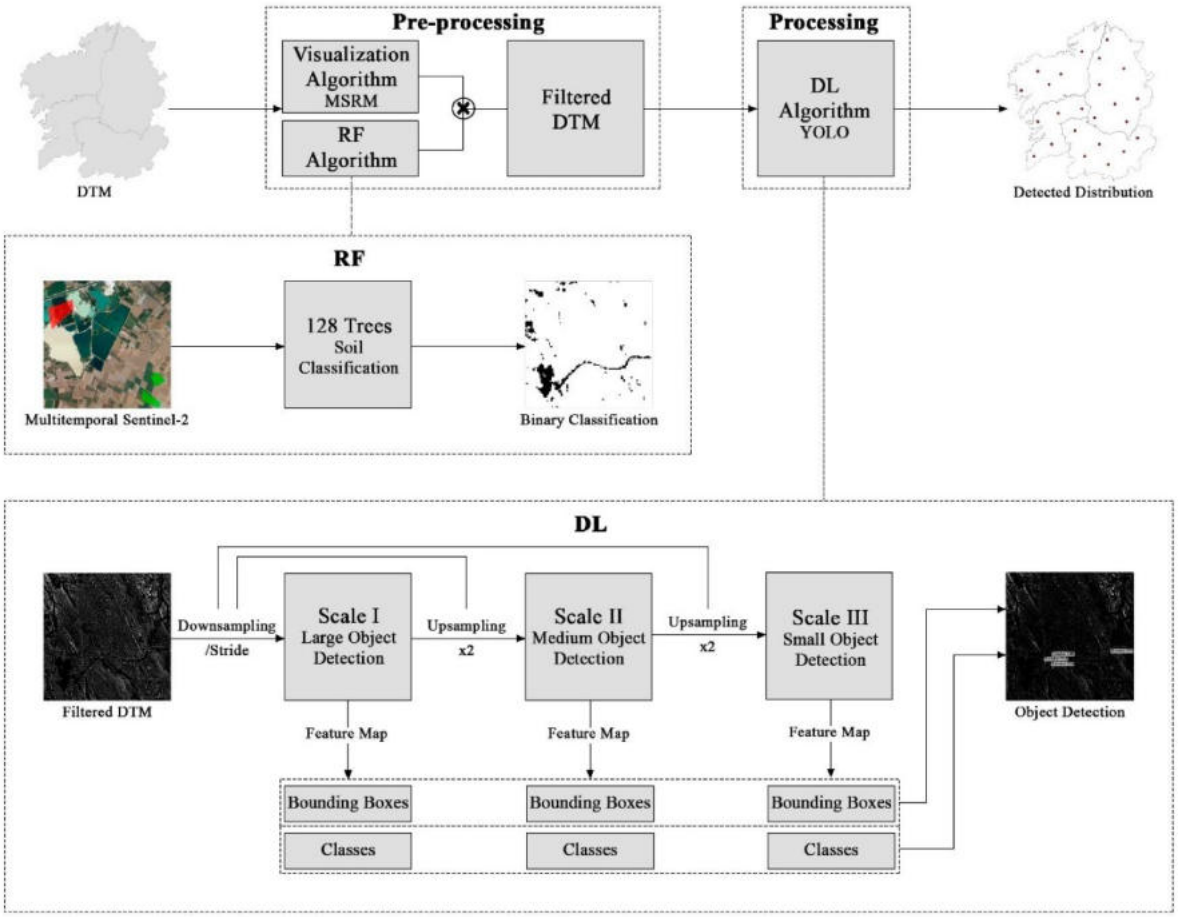
To overcome this problem, we decided to develop a binary soil classification map using GEE Code Editor, Repository and Cloud Computing Platform (Gorelick, 2017). Our objective was to eliminate those pixels that could not correspond to archaeological mounds. To reach this objective we used cloud-filtered multitemporal Sentinel-2 multispectral imagery. Sentinel-2 incorporates 13 bands from which only the visible/near-infrared bands (VNIR B2–B8A) and the short-wave infrared bands (SWIR B11–B12) were employed. Bands B1, B9, and B10 (60 m/px each) correspond to aerosols, water vapor, and cirrus, respectively, and they were not employed in this study except for the use of the cirrus-derived cloud mask applied. Visible

(B2–B4) and NIR (B8) bands provide a ground resolution of 10 m/px, while red-edge (B5-B7 and B8A) and SWIR (B11–B12) bands present a 20 m/px spatial resolution. Specifically, for this research Sentinel-2 Level 1C products representing top of atmosphere (TOA) reflectance were preferred due to the larger span of its mission (starting from June 2015). Sentinel-2 multispectral satellite images were a good compromise given their relatively high spatial and spectral resolutions and their open access policy. The use of cloud-filtered multitemporal satellite data has been successfully employed in previous research to provide long-term vegetation indices (Orengo and Petrie, 2017; Garcia-Molsosa et al., 2019), but also for the development of ML classifications (Orengo et al., 2020; Orengo and Garcia-Molsosa, 2019) as they provide images that are independent of specific environmental or land-use conditions that are particularly adequate for the development of classifications.

The use of GEE allowed us to access and join 1920 (at the moment of writing) Sentinel-2 images in a single 10-band composite, train the classification algorithm and execute the analysis, which would have been impossible using a desktop computer. It also provided an ideal environment to join the results of the classification with that resulting of the MSRM filter of the DTM also created using GEE (see previous section). Thirteen polygons defining training areas were drawn and tagged as class 0 (areas unsuited for the presence of tumuli), which included a variety of urban features (such as roofs, roads, swimming pools, etc.), water, rock and quarries and other industrial areas. Additionally, 19 class 1 polygons were drawn within grasslands, cultivation fields and forests. From these polygonal training areas, a total of 4398 sampling points corresponding to individual multispectral pixels (1832 for class 0 and 2566 for class 1) were extracted with values for all selected bands and a class identifier. These training data were employed to classify the composite raster using a RF algorithm with 128 trees, which resulted in a binary raster indicating areas where archaeological tumuli can (class 1) and cannot (class 0) be found.

4.2.5. Hybrid Machine Learning Approach

The combination of DL for shape detection and traditional ML for binary soil classification is described in **Scheme 1**. The use of GEE for the generation of both MSRM and the binary classification map made it possible to integrate both processes in a single script, which, as a last step, multiplied both outputs to produce a MSRM in which all areas not conductive to the presence of mounds had been removed.



Scheme 1. The implemented workflow for object detection with the detail of the structure and behaviour of the RF and DL algorithms.

A similar approach combining DL and traditional ML was recently published by Davis et al. (2021) (Davis et al., 2021). While we used the RF classification to eliminate areas of source of FPs for the application of the DL detector, they used the multisource multitemporal RF approach developed by Orengo et al. (2020) (Orengo et al., 2020) to evaluate the detection results from a Mask R-CNN detector. Although this approach was useful to confirm many of the detected features, it was not integrated into the detection workflow and did not contribute to reduce the large number of FPs reported.

In our case, the DL algorithm was retrained using the new raster produced by the multiplication of the MSRM and the classified binary raster. The RF removed 11 real archaeological tumuli from our initial training data and 13 from the refinement step, leaving 560 burial mounds to work with. Of those 560 mounds, 456 were employed for training and 104 for validation.

4.3. Results

4.3.1. Digital Terrain Model Pre-Processing

MSRM was the most effective DTM pre-processing method for the detection of barrows (**Table 1**), with an AP of 63.03% and higher recall and precision values. Despite showing a better result, the initial detection using MSRM presents a recall value of 0.58, which highlights the presence of a large proportion of FNs, and a precision of 0.95 indicating that some FPs were detected.

Table 1. Evaluation of the YOLOv3 models using MSRM, Slope gradient and SLRM as input data.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision
MSRM	63.03%	62	3	44	0.58	0.95
SLOPE	53.58%	49	5	57	0.46	0.91
SLRM	52.89%	44	8	62	0.42	0.85

4.3.2. Model Refinement and Data Augmentation

As said before, two different models were tested applying model refinement: a two-classes model with the FPs as the new class and one class model with the FPs as background. As shown in **Table 2**, model refinement works similarly in both cases because the background of the images is considered in the training. Although the recall and precision values have not improved significantly compared to the previous case, the key is that this result now includes the mentioned FPs and the FNs. Even though the number of FPs was reduced, several are nonetheless included.

Table 2. Evaluation of the YOLOv3 models using model refinement for one class and two classes.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision
1 class	66.77%	63	3	43	0.59	0.95
2 classes	70.30%	66	3	40	0.62	0.96

The use of DA methods provided mixed results. Although all DA methods improved the results provided by the training without DA, the resizing of the training data (DA1) proved the most

effective (**Table 3**). Even if it increased the presence of FPs it also increased the number of TPs while reducing the presence of FNs. Therefore, DA1 was implemented in the final model.

Table 3. Results of the YOLOv3 models using different types of DA.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision
None	68.31%	63	2	43	0.59	0.97
DA1	70.30%	66	3	40	0.62	0.96
DA1+DA2	67.62%	65	2	41	0.61	0.97
DA1+DA3	66.77%	66	6	40	0.62	0.92

4.3.3. Integration of Random Forest Classification

The use of the RF classification of satellite data aimed at reducing the number of FPs, by eliminating those areas with soils not conducive to the presence of burial mounds. The results of the validation (**Table 4**) show that the RF classification and filtering of the DTM improved the model in all respects. It increased the number of TPs while reducing the presence of FPs and FNs. The model trained with the classification-filtered MSRM was also able to detect 1538 tumuli more than that without the filter with a lower presence of FPs and FNs. Although a percentage of FPs are still present after using the classification to filter the MSRM (see the evaluation section for details) it was successful in eliminating all urban areas and road related infrastructure (all roundabouts were also eliminated), even those not considered as such in the official land-use maps.

Table 4. Evaluation of the YOLOv3 models using RF filtering and not using it.

Algorithm	AP@0.5	TPs	FPs	FNs	Recall	Precision	Mounds
Not RF	71.65%	66	3	38	0.63	0.96	8989
RF	66.75%	67	2	37	0.64	0.97	10,527

4.3.4. Results and Test Dataset-Based Validation

The YOLOv3 algorithm has validated the known burial mounds with an AP@0.50 of 66.75% and a loss value of 0.0592 (**Figure 26**). Furthermore, 10,527 burial mounds were detected all over Galicia with a minimum similarity of 25%, a minimum size of 7 m, a maximum size of 74 m, a mean size of 29 m and a mode of 25 m. Likewise, the locations of these detected tumuli

were indicated in order to facilitate their identification in the field. The implemented parameters were *classes* = 1, *channels* = 1, *max_batches* = 20,000, *width* = 832 px and *height* = 832 px for training configuration, and *width* = 1024 px and *height* = 1024 px for the detection one. DA1 was the DA dataset implemented.

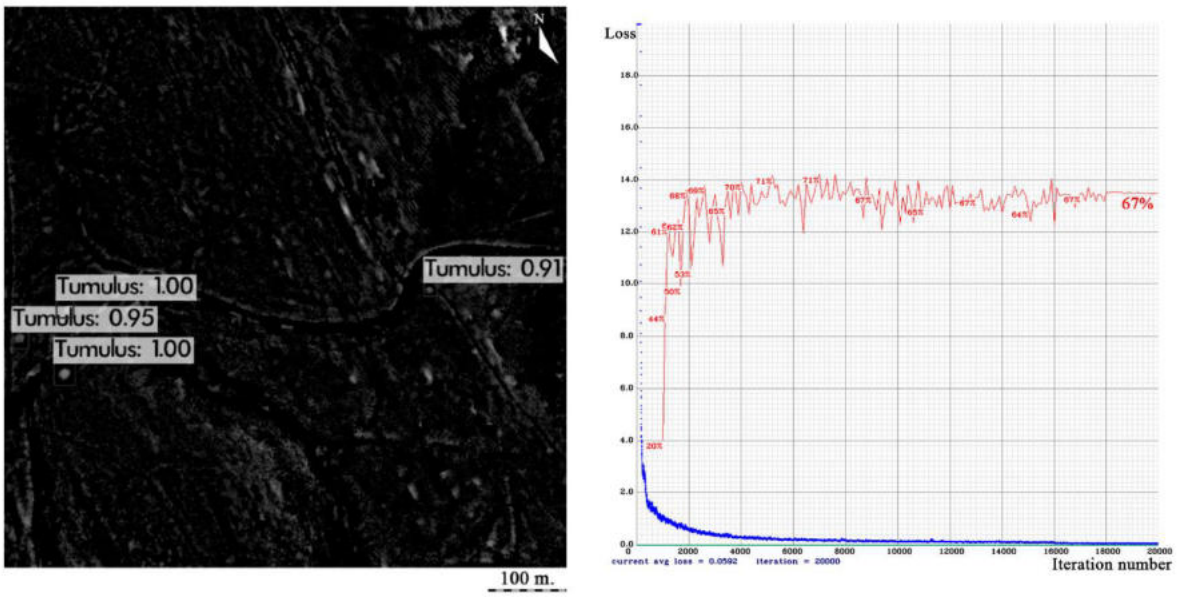


Figure 26. YOLOv3 model: (**a**) tumulus detection example; (**b**) loss (blue) and AP@0.50 (red) vs. iteration number function.

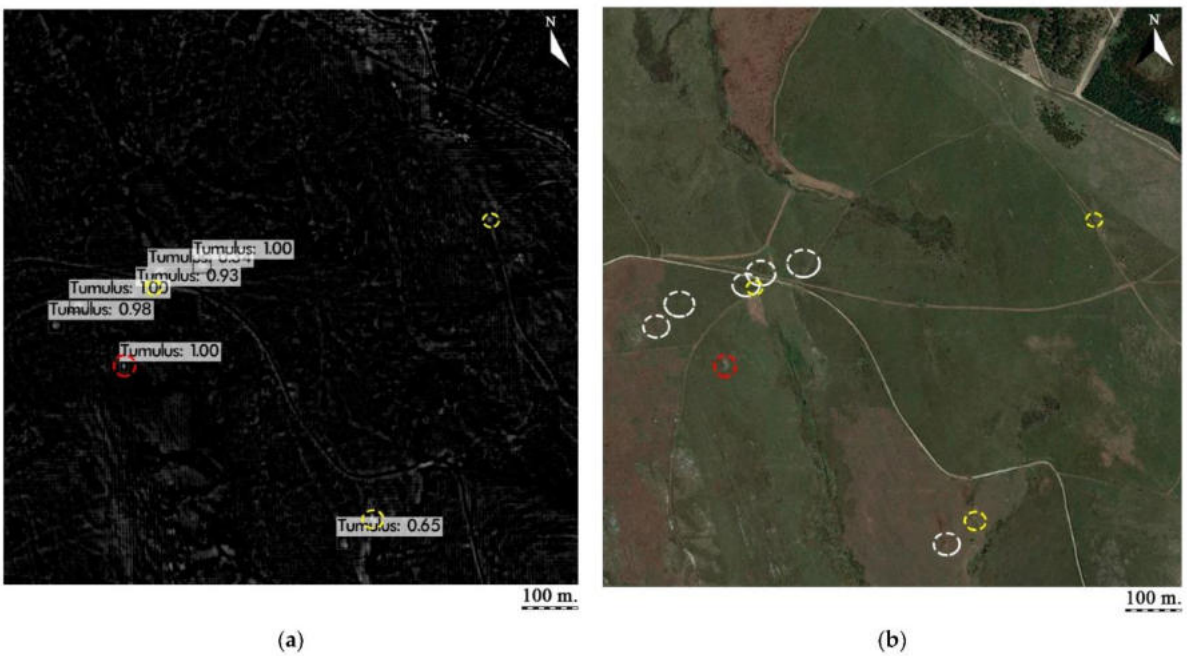


Figure 27. Tumulus detection using YOLOv3 where there were six TPs (white circles), three FNs (yellow circles) and a single FP (red circle): (**a**) output data; (**b**) satellite view.

This model proved to have similar robustness to the previous one despite having a slightly lower AP (**Table 4**). As the AP is the used area under the precision/recall curve for each recall value, it is possible that even if the precision and recall values improve, the AP may be lower. However, the AP value, calculated for an IoU threshold of 0.5, were not completely successful. On the one hand, a 0.97 precision value on the test dataset shows that the algorithm distinguishes burial mounds with high precision, but also that there were two FPs, 1.92% of the total (**Figure 27**). Both of these corresponded to small and isolated rock outcrops. On the other hand, the 0.64 recall value reveals that most of the burial mounds have been correctly validated by the algorithm, but also that there were 37 FNs, 35.58% of the total (**Figure 27**).

Finally, there were 67 correctly detected burial mounds (TPs), 64.42% of the total. This indicates that numerous burial mounds were detected in Galicia despite the aforementioned FNs (**Figure 28**), showing their large-scale distribution (**Figure 29**).

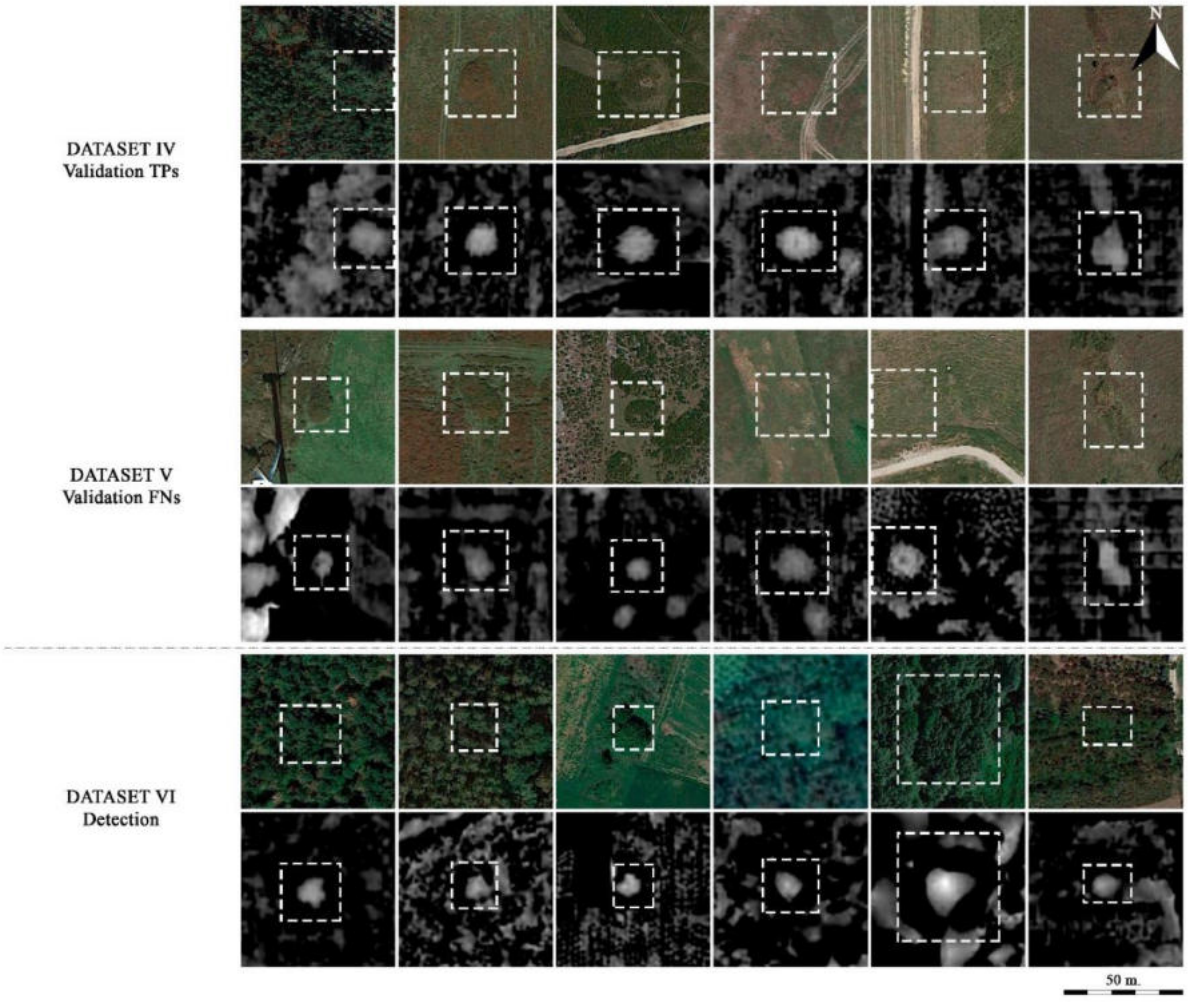


Figure 28. Validation TPs (Dataset V), FN (Dataset VI) data examples, and detections (Dataset VII). The latter were detected with a similarity of 100%, 90%, 80%, 60%, 40% and 25% (from left to right).

The corresponding top image for each pair is a visible satellite image, shown for the sake of better visualization, but not used in our process.

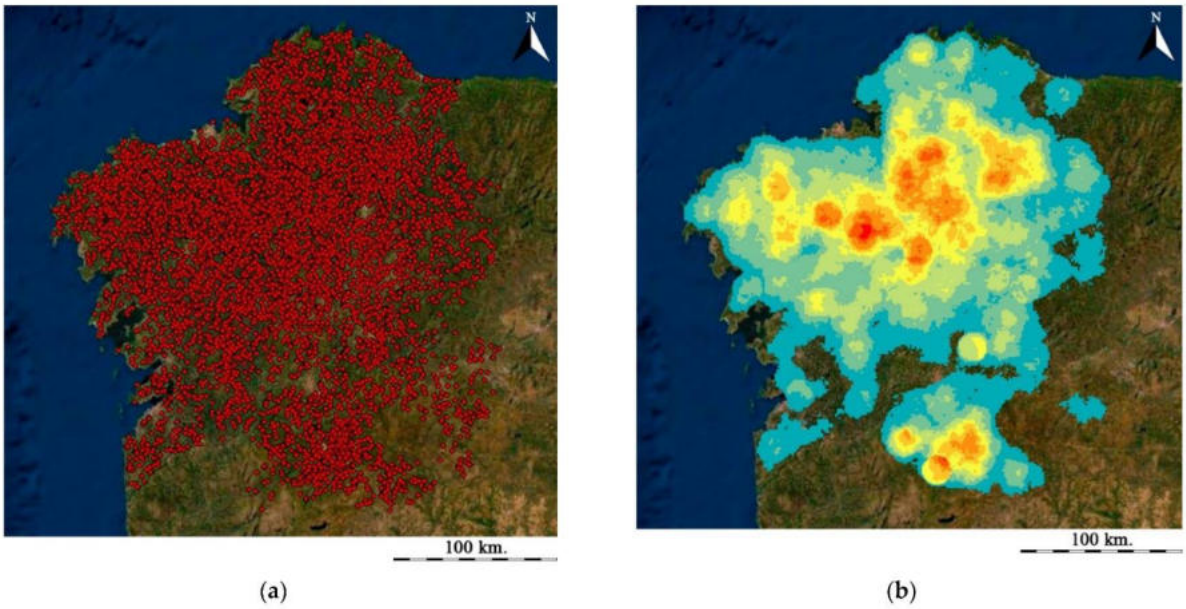


Figure 29. Detected tumuli in Galicia (Spain): (**a**) point distribution; (**b**) heat map.

4.3.5. Manual Model Validation

A last validation step consisted of manually evaluating the results. Although we extracted statistically significant performance metrics from the test dataset (see above), this dataset was extracted from a single area that did not have the variety of soil and land-use types present in the whole of the study area. As this can greatly influence the presence of FPs (e.g., areas with isolated houses could present FPs in the form of houses' roofs and eroded highland areas in the form of rock outcrops), a manual validation was considered necessary. This is a basic measure in archaeological detection studies, in particular with respect to mound detection work, as FPs tend to constitute a very high proportion of the detected features (see for example, Davis et al., 2021; Verschoof-van der Vaart et al., 2020). For the manual visual inspection of the detected features, we used three different series of high-resolution imagery provided by Google, Bing, and ESRI, accessed as XYZ Tiles, and aerial imagery at 0.25 m/px provided by SIGPAC as a WMS through QGIS. These sources cover several years before and after the acquisition of the lidar dataset and helped to evaluate the possible presence of barrows independently of specific circumstances. From the 10,527 tumuli detected, we evaluated a total of 3086 individual tumuli

in non-forested areas where the aerials allowed good visibility of ground conditions. We found that, of these, 324 corresponded to FPs, as follows: 225 were identified as rock outcrops, 33 as isolated houses' roofs, 9 as swimming pools and 57 to other mound-shaped features, most of them of anthropogenic nature. We should also note that, among this last type of FPs, some were only identified as FPs because of their context (such as mounds in golf courses) and were otherwise indistinguishable from archaeological tumuli. Mound identifications in forested areas were not considered to be FPs or TPs, as the only inspection method available for them was the lidar dataset, and this would have made it impossible for us to identify quite common occurrences such as rock outcrops.

Therefore, the manual validation indicated that 10.5% of the detected features were FPs, resulting in a detection rate of 89.5%. This suggests that, of the 10,527 tumuli detected approximately 9422 correspond to TPs. This number could be slightly higher, as approximately 23% of the tumuli are located in forested areas where, from all types of FPs, only rock outcrops (69% of the FPs) could be found.

Of course, this does not mean that all 9422 are archaeological tumuli, but their criteria did correspond to those used to identify them. Only a proper field survey and/or test pit excavations can truly document the archaeological nature of these remains, as there are many natural and human activities that could produce indistinguishable shapes in the same types of contexts.

In conjunction with the information provided by the presence of FNs (35.58% of the test data), our results suggest that the approximate number of tumular features that could correspond to archaeological tumuli in Galicia approximates 14,626 (9422 estimated TPs plus the estimation of those not detected according to the percentage of FNs).

4.4. Discussion

The automated detection of archaeological tumuli is a complex task given their common morphology. The study case presented here is particularly complex, considering the very large study area, the largest ever for this type of research. It includes multiple environmental conditions, land uses comprising urban, industrial, recreational and natural areas, and many other complex topographic settings such as granitic ranges and coasts which typically produce shapes similar to those of barrows.

Despite the complexity and scale of this study, the results are well beyond previous attempts to detect mounds using lidar data. The assessment of the test data provides a recall value of 0.64 (which means that the algorithm has detected a 64% of the known tumuli in the test area) and a precision of 0.97 (so 97% of the detections correspond to TPs). Further to that, the visual validation on randomly selected tumuli throughout the study area indicates that 89.5% of the detected features correspond to potential mounds, a total of approximately 9422 tumuli. The most recent approaches to the detection of archaeological mounds using lidar-derived data are usually able to detect a high percentage of the test dataset's true mounds, but they also include a large proportion of FPs. For example, Davis et al. (2021) detected 17 of 18 mounds present in their study area but they also detected 3237 more mounds (Davis et al., 2021). After visual validation, they confirmed that from the 3254 detected mounds, only 287 corresponded to possible burial structures (equivalent to an 8.8% success rate), pending field validation. Verschoof-van der Vaart et al. (2020) obtained a recall value of 0.796, but the precision value was 0.141 (86% of detected tumuli were FPs) (Verschoof-van der Vaart et al., 2020). Trier et al. (2021) detected 38% of known tumuli in their study area, but 89% of the detected features were identified as FPs (Trier et al., 2021).

4.4.1. Digital Terrain Model Pre-Processing

The use of MSRM instead of the most commonly used relief visualisation tools, such as LRM (Hesse, 2010) and slope gradient (Challis et al., 2011; Doneus and Briese, 2006), improved the detection rate of the algorithm. In contrast to other kernel-based methods, where the size of the feature can strongly affect its resulting shape, the multiscale nature of the MSRM produced more consistent shapes independently of the size of the tumuli. This is consistent with the results obtained by Guyot et al. (2021) (Guyot et al., 2021), which, after comparing 13 microrelief visualisation methods, concluded that multiscale approaches consistently showed better performances in CNN-based detections.

4.4.2. Model Refinement

Despite the high detection and low FP rate of our algorithm, the recall value indicates that more training data could have improved the detection rate. An increment in the burial mound's training data would increase the variability in the shape of the tumulus. In the study area, there

were only 584 known tumuli to work with (306 coming from previous works and 278 added in the refinement step), of which 478 were employed for training and 106 for validation. Theoretically, this number could be increased using DA, but given the circular nature of the mounds, the improvement with augmentation techniques such as DA2 or DA3 is very small. The obtained AP value without any DA was 68.31%, slightly lower than when implementing DA procedures (**Table 3**). This is because the validation mounds with a diameter of less than 18 m represent only 7.55% of the total. An increase of the DTM resolution to 0.5 m per pixel would allow a better detection of the smallest burial mounds, and an improvement in shape definition that would have allowed to distinguish FPs such as rock outcrops and houses' roofs.

4.4.3. Hybrid Model

Concerning the FPs, both the RF filtering and the refinement eliminated most of the possible FPs detected but we could still find some small rock outcrops, particularly those isolated and surrounded by soil types conductive to the detection of mounds. This is due to the 10-m-per-pixel ground resolution provided by Sentinel-2. In many cases, rock outcrops or other FP were located at the intersection of several 10 m pixels and the value taken was that of the positive pixel (i.e., that corresponding to a valid soil/land-use type) in some other cases the feature originating the FP was too small and the classification's pixel value was an average of the Sentinel-2 pixel footprint, resulting in the misclassification of the specific pixel. In relation to the previous point, the manual identification of FPs using high resolution imagery allowed us to identify elements that would not have been visible in lower-resolution images such as those acquired by Sentinel-2 (see the first column of **Figure 30**, where a rock outcrop is partially visible in between the tree cover). This was particularly the case for isolated small rock outcrops in grasslands, which constitute most of the FPs detected. With an improvement in the spatial resolution of the multispectral imagery employed, all those small diameter FPs in the RF classification stage would have been removed. In this study, we preferred to employ Sentinel-2 given its public nature and the fact that it is directly accessible from GEE. However, the use of imagery from SPOT, RapidEye or any of the very high-resolution satellites commercially available could importantly reduce the presence of FPs. Another problem encountered with the use of multitemporal multispectral data for the classification of soil type/land use is precisely the multitemporal nature of the data. Although it helped to achieve better classification results, in some cases, particularly recent constructions, it included data

from before the construction, producing median pixels classified as positive, which resulted in several houses' roofs being included as FPs.

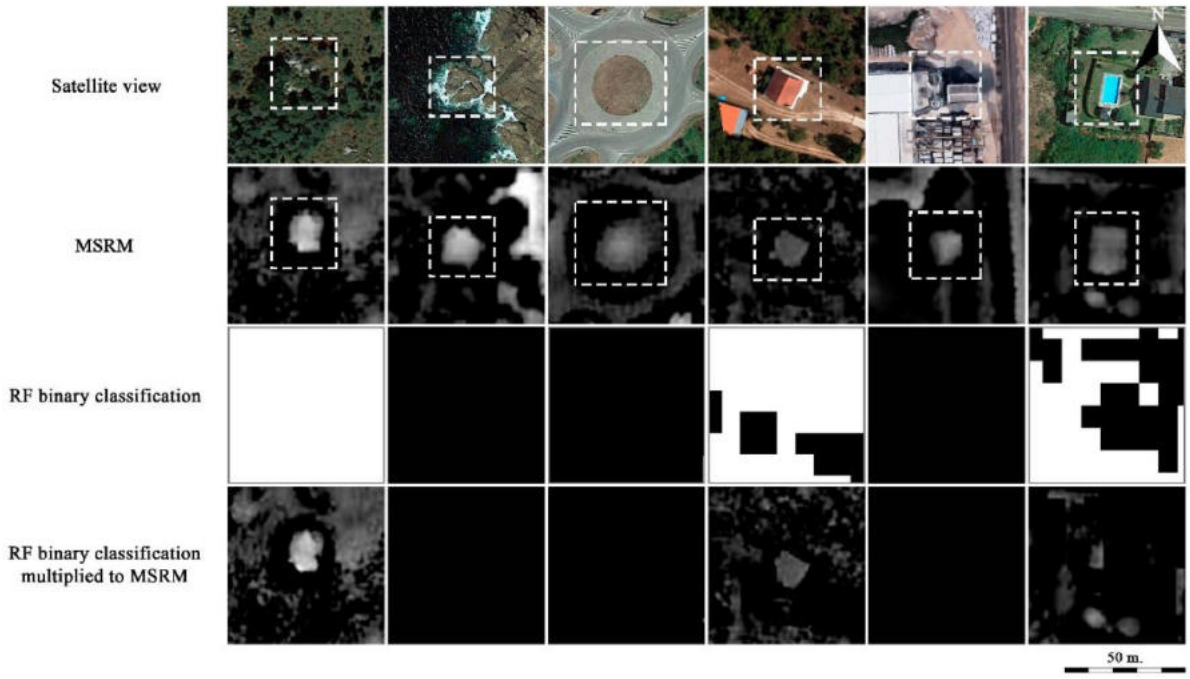


Figure 30. FP examples (Dataset III) obtained when we carried out the detection in Galicia and the same examples after applying the RF filtering. The corresponding top images are a high-resolution RGB satellite image, used for the manual validation, but not in the detection process.

It would have also been possible to include the multispectral bands as channels to the DL model (*channels* = 11) whose spectral data would have made it possible to discern the rocky areas from burial mounds and roofs from mounds in cultivation fields (which present similar values in RGB images). However, this would have resulted in a significant increase in the use of computational resources and training time.

4.4.4. Algorithm Accessibility and Reproducibility

The algorithm was designed to be accessible and reusable. All data employed are publicly available, even if the use of private imagery providers could have significantly improved the results of this study. The code is provided as **Supplementary Material** and is also available in GitHub (where future updates will be made available). Beyond evaluation and reproducibility

concerns, the code has been designed to avoid personal limitations in computing power as a personal computer with an Internet browser and an Internet connection are the only requirements to apply the whole model. GEE is used to process the MSRM and RF classification (both very costly using desktop computers), and Colab can be used to import the resulting raster and apply the YOLO algorithm seamless using a single cloud project. The design of the algorithm with a single channel source (the RF classification-filtered MSRM) instead of a costly multichannel DL approach significantly reduces computing costs allowing the detector to be applied over large areas using GEE and Colab cloud computing resources.

The results and the detected FPs are also available as Shapefiles with associated metadata. In this way, these data can be used to better understand, manage and protect the cultural heritage of Galicia.

4.5. Conclusions

The algorithm presented in this paper constitutes an important improvement over previous and current approaches to the detection of archaeological tumuli and presents, for the first time, a valid alternative to the manual detection of this very common type of archaeological structure. The comparison of the results with regional heritage databases will make it possible to validate and improve both datasets. The large number of burial mounds detected in Galicia will allow the development of future investigations on their cultural distribution, achieving a better knowledge of the Galician megalithic complex.

Future research will implement newer versions of YOLO (v4 and v5, published during the development of this study), which improve the AP and the frame rate of YOLOv3. However, given the performance of the training algorithm presented here for the detection of burial mounds, our method already constitutes a practical tool that can be applied to any other areas where tumuli are present with few modifications, thus making it a general tool for archaeological research and cultural heritage management in many areas of the world. This is also prompted by making open-access the code presented in this work.

The process could also be greatly simplified by the use of Google Cloud Projects, where GEE and Colab can be combined. GEE allows the ingestion of the user's preferred source for both lidar and satellite multispectral data (allowing to boost the results of this research with higher resolution sources without the need to modify the algorithm's code) and the training of the RF

classification algorithm can be easily achieved within GEE using its simple vector drawing tools. Colab's Jupyter notebook environment requires no configuration, runs entirely in the cloud, and allows the use of Keras, TensorFlow and PyTorch. It provides free accelerators like GPU or specialized hardware like tensor processing units, 12 GB of RAM, 68 GB of disk and a maximum of 12 h of continuous running.

4.6. Supplementary Materials

The following Supplementary Materials are available online at **https://www.mdpi.com/article/10.3390/rs13204181/s1**. Document explaining the use of the code and the scripts necessary to run it: script1.txt, script2.ipynb, JPEGtoPNG.atn, result.txt, script3.txt, resultsGIS.xlsx. Scripts can also be found in GitHub: **https://github.com/horengo/Berganzo_et_al_2021_DTM-preprocessing** (Accessed on 1 October 2021) and **https://github.com/iberganzo/darknet** (Accessed on 1 October 2021).

4.7. Author Contributions

I.B.-B. and H.A.O. wrote the paper with the collaboration of all other authors. I.B.-B. created all illustrations. M.C.-P., J.F. and B.V.-E. provided training data and input during the evaluation of the results. I.B.-B., H.A.O. and F.L. designed the algorithm. H.A.O. designed the project and obtained funding for its development. All authors have read and agreed to the published version of the manuscript.

4.8. Funding

I.B.-B.'s PhD is funded with an Ayuda a Equipos de Investigación Científica of the Fundación BBVA for the Project DIASur. H.A.O. is a Ramón y Cajal Fellow (RYC-2016-19637) of the Spanish Ministry of Science, Innovation and Universities. F.L. work is supported in part by the Spanish Ministry of Science and Innovation project BOSSS TIN2017-89723-P.M.C.-P. is funded by the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie Grant Agreement No. 886793). J.F. is funded by the European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie Grant Agreement

No. 794048). Some of the GPUs used in these experiments are a donation of Nvidia Hardware Grant Programme.

5. Curriculum Learning-based Strategy for Archaeological Mound Features Detection from Historical Maps in Low-Density Areas in India and Pakistan

Iban Berganzo-Besga[1], Hector A. Orengo[1,2*]; Felipe Lumbreras[3], Aftab Alam[4], Rosie Campbell[5], Petrus J. Gerrits[5], Jonas Gregorio de Souza[6], Afifa Khan[5], María Suárez-Moreno[5], Jack Tomaney[5], Rebecca C. Roberts[5], Cameron A. Petrie[5,7]

1. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

2. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

3. Computer Vision Center, Computer Science Deptartment, Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellaterra, Spain

4. Banaras Hindu University, Ajagara, Varanasi, Uttar Pradesh 221005, India

5. McDonald Institute for Archaeological Research, University of Cambridge, Downing St., Cambridge CB2 3ER, UK

6. Complexity and Socio-Ecological Dynamics (CaSEs) Research Group, Universitat Pompeu Fabra, Barcelona, Spain

7. Department of Archaeology, University of Cambridge, Downing St., Cambridge CB2 3DZ, UK

*Author to whom correspondence should be addressed.

Submitted: 16 February 2023

Abstract

This paper presents two algorithms for the large-scale automatic detection and instance segmentation of potential archaeological mounds on historical maps. Historical maps present a unique source of information for the reconstruction of ancient landscapes. The last 100 years have seen the unprecedented landscape modifications with the introduction and large-scale

implementation of mechanised agriculture, channel-based irrigation schemes, and urban expansion to name but a few. Historical maps offer a window onto disappearing landscapes where many historical and archaeological elements that no longer exist today are depicted. The algorithms focus on the detection and shape extraction of mound features with high probability of being archaeological settlements, mounds being one of the most commonly documented archaeological features to be found in the Survey of India historical map series, although not necessarily recognised as such at the time of surveying. Mound features with high archaeological potential are most commonly depicted through hachures or contour-equivalent form-lines, therefore, an algorithm has been designed to detect each of those features. Our proposed approach addresses two of the most common issues in archaeological automated survey, the low-density of archaeological features to be detected, and the small amount of training data available. It has been applied to all types of maps available of the historic 1'' to 1-mile series, thus increasing the complexity of the detection. Moreover, the inclusion of synthetic data, along with a Curriculum Learning strategy, has allowed the algorithm to better understand what the mound features look like. Likewise, a series of filters based on topographic setting, form, and size have been applied to improve the accuracy of the models. The resulting algorithms have a recall value of 52.61% and a precision of 82.31% for the hachure mounds, and a recall value of 70.80% and a precision of 70.29% for the form-line mounds, which allowed the detection of nearly 6,000 mound features over an area of 470,500 km², the largest such approach to have ever been applied. If we restrict our focus to the maps most similar to those used in the algorithm training, we reach recall values greater than 60% and precision values greater than 90%. This approach has shown the potential to implement an adaptive algorithm that allows, after a small amount of retraining with data detected from a new map, a better general mound feature detection in the same map.

Keywords:

historical map; mound features; archaeology; deep learning; segmentation; curriculum learning; data augmentation

5.1. Introduction

The past 100 years and, in particular, the second half of the twentieth century, have seen extensive urban growth and the large-scale implementation of mechanised agriculture and irrigated systems in India and Pakistan, causing irreversible effects on the landscape. Among

other lasting impacts, such as the implementation of large-scale irrigation systems, river avulsion and flooding, there have been systematic flattening, for cultivation and construction, of hundreds, if not thousands, of archaeological settlement mounds (Petrie et al., 2019; Garcia-Molsosa et al., 2019; Green et al., 2019). These archaeological mounds with their distinct elevation, colour and form are an indicative feature of past settlements and anthropogenic modifications of the landscape. Given their partial or total destruction, these are no longer detectable by other types of sources such as lidar or satellite imagery (Berganzo-Besga et al., 2021; Berganzo-Besga et al., 2022b). Historical maps are therefore often the only source of information about the location and size of those lost sites. Available satellite images of the Indian subcontinent date back to 1972 thanks to the Landsat satellite programme (Landsat Science, 2022), but detailed mapping of this region through triangulation dates back to 1802 and the start of the Great Trigonometrical Survey. Later, during British Crown rule in India and Pakistan (1858-1947), the SoI continued the systematic mapping of the whole subcontinent.

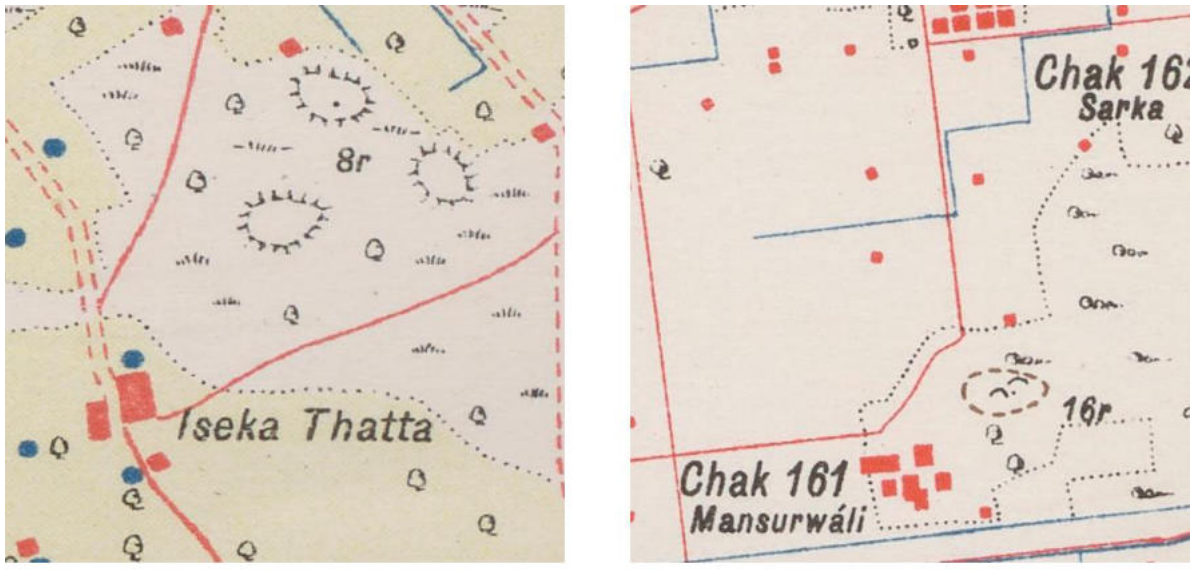
The SoI maps were originally intended to be geographic maps and depicted different topographic features including mounds, many of which, as further research has shown (Green et al., 2019), are in fact archaeological sites (**Figure 31**). Ground truthing has revealed there is a correlation between these mound features and proto-historical and historical sites dating to various periods from the period of the Indus Civilization onward (Green et al., 2019).

Figure 31. Archaeological remains found where the historical maps indicated mounds. View from an elevated mound feature in northwest India (L742) (Green et al., 2019).

DL has been widely used in recent years to aid archaeological survey by using different resources such as lidar data (Berganzo-Besga et al. 2021; Davis et al., 2021) and drone imagery (Orengo et al., 2021). This study continues the work carried out by several authors for the detection of archaeological sites using historical maps (Petrie et al., 2019; Garcia-Molsosa et al., 2019; Green et al., 2019). Previous studies made by Garcia-Molsosa focused on the present district of Multan in the Pakistani province of Punjab. The series of maps used in this study had similar production standards (Garcia-Molsosa et al., 2021). Although this previous approach produced satisfactory results it presented some drawbacks:

1. It employed a reduced series of maps of similar chronologies, depiction standards, scanning quality and preservation. This ideal situation, however, proved not to be the norm when a much larger collection of maps was assembled. Use of a larger collection of maps presented important variations in coloration, representation standards, scanning quality and preservation, which enormously complicated the large-scale application of these initial detectors (Garcia-Molsosa et al., 2021) and significantly reduced their detection capabilities.

2. The initial algorithms were designed in a proprietary web-based geospatial ML platform. The models were not available for download, analysis or free distribution and the processing was expensive, prohibitively so when considering large areas such as the one under investigation.



(a) (b)

Figure 32. The two types of mound features depictions that need to be detected in historical maps: (a) hachure [8r], and (b) form-line [16r].

The study presented in this paper uses the British historical maps produced in the late nineteenth and early twentieth century by the SoI with the aim of detecting two of the most common ways of drawing mound features (hachure and form-line, see section 2.2 for further details), which are similar to those depicted by the French in Syria and Lebanon (Garcia-Molsosa et al., 2021) (**Figure 32**). Our research seeks to develop two DL segmentation algorithms for mound feature detection, one for each mound type, extending the detection to an area of 470,500 km² (most of which corresponds to the Indus River Basin), the largest area in which such an approach has ever been applied (Berganzo-Besga et al., 2021), and to all types of maps, thus increasing the complexity of the analysis. We have employed a R-CNN segmentation algorithm as this collects information about not only the location of the mound feature, but also about its shape and extent.

Automated detection processes require large amounts of data for their training (typically in the order of tens of thousands of individual examples), but this is not common in archaeology where the number of known archaeological samples to train a ML algorithm is very low, as in this case study. Other studies with similar elements such as burial mounds (Berganzo-Besga et al., 2021), have shown that despite having limited training data, features of interest are detectable due to the characteristic circular shape of the tumuli, which presented few variations. The archaeological elements of this study, despite being mound features like those of previous studies where we encountered a similar problem, are much more diverse. Since they are symbols drawn by human hands and not images of their actual form, whether aerial or satellite, the features are noticeably divergent in style from each other. Consequently, a relatively small amount of training data was not enough to achieve meaningful results.

In computational archaeology, trained ML models have been shown to perform worse in areas with low-density of archaeological features than in high-density ones (e.g. Soroush et al., 2020; Verschoof-van der Vaart et al., 2022). When performing large-scale detection with few sites, many FPs are introduced (typically many more than the TPs), which severely reduces the accuracy of the algorithm. However, real archaeological scenarios typically present low-densities of archaeological sites that need to be detected, at least compared to other typical objects in CV studies (such as cars, trees, buildings, ships, etc.). During a survey, the actual density of archaeological features is unknown, so to be a useful tool, the developed ML algorithm must also provide good results for low-density areas.

Therefore, the use of ML approaches in archaeology entails a series of idiosyncratic challenges: including the customary small amount of archaeological data for training and the usual low-density of archaeological features. In this article we will implement a series of DA techniques and learning strategies to resolve these two issues.

The main goal of this article, besides the successful detection of mound features within acceptable parameters of precision and recall, will address these two issues by designing a workflow for the correct detection of archaeological features 1) in low-density areas and 2) with little amount of training data.

5.2. Materials and Methods

In this study, a total of 645 maps, provided by the Cambridge University Library and the British Library have been used. These historical maps were produced and distributed by the SoI, and can be classified into different periods characterized by the then current Surveyor General of the SoI, including C. Straham (1898-1899), G.C. Gore (1900-1902), F.B. Longe (1904-1907), and S.G. Burrard (1912-1913). Maps produced under A.R. Quraishi (1954) in his role of Surveyor General of the Survey of Pakistan have also been included.

5.2.1. Map Digitisation and Georeferencing

Before proceeding with the training of the DL algorithm, all 645 maps used for this study had to be scanned and georeferenced (**Figure 33**). The scanning process was done by different institutions and individuals, in different periods and using different means and resolutions as a result of the different histories, means, and the procedures of the different institutions hosting and scanning them. After the digitalisation of the maps, they were georeferenced using a minimum of 12 GCPs, and most commonly 20 GCPs with an average of 25 GCPs, geometrically distributed within the map to achieve a good distribution. The GCPs were obtained from georeferenced high resolution RGB satellite imagery available as Web Map Services layers in QGIS software. The georeferencing process mainly used second order polynomials with a minimum of 20 GCPs for an accurate transformation, which was the preferred method, and was applied to most maps. On rare occasions, where it was not possible to find enough GCPs, polynomial 1 was used. These methods produced average RMSE values

of 0.000350º (ca. 33.7-38.8m at this latitude) using a second order polynomial and 0.000102357º (ca. 10.3m with a maximum value of ca. 26.8m) using the adjust transformation. Since the mounds under consideration are typically much larger than these values, the georeferencing process results in mound feature locations, which, partially overlap the real locations (for more details on the georeferencing process see Petrie et al., 2019).

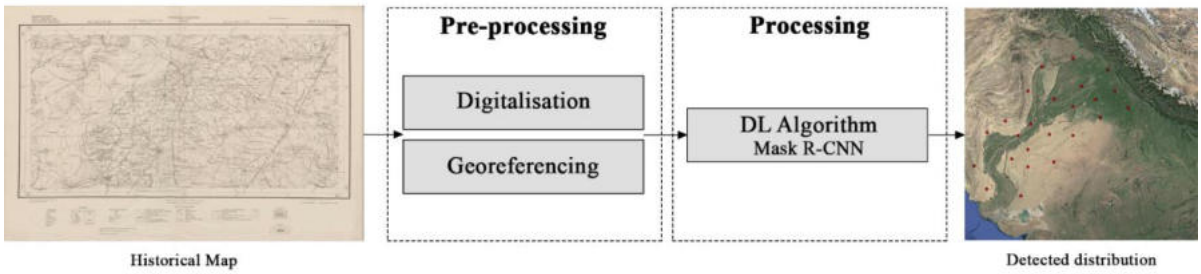


Figure 33. Scheme of the workflow for the detection of mounds in historical maps.

5.2.2. Deep Learning Model

In recent years, R-CNN models have become very common in archaeological survey, highlighting segmentation algorithms such as Mask R-CNN (Orengo et al., 2021) and DeepLabV3+ (Landauer et al., 2022). For this study, we performed two mound symbol detection DL algorithms using Mask R-CNN (Waleed, 2017), since we are looking for instance segmentation rather than semantic. Mask R-CNN detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance (He et al., 2017). It extends Faster R-CNN (Girshick, 2015; Ren et al., 2016) by adding a branch for predicting segmentation masks, a small fully convolutional network (FCN) (Long et al., 2015), on each region of interest (RoI), in parallel with the existing branch for classification and bounding box regression. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN (He et al., 2017). Likewise, VGG Image Annotator (VIA) from the university of Oxford has been used to label mound features (Dutta and Zisserman, 2019).

The digitized and georeferenced historical maps are 3-channel RGB images and we have cropped them into 512x512 pixel images to save computing costs. Of the 645 maps used, only 43 contained known mound features, which have been used for training and validation: 286 hachure and 103 form-line mound features. Of those maps, 22 were used for training, including 168 hachure and 26 form-line mound features, and 21 were used for validation, including 118 hachure and 77 form-line mound features. In addition, given the small number of known

mounds, another 21 maps, chosen randomly from the 645 original maps, were manually analysed. In this way, we have been able to create another dataset, the test dataset with 230 hachure and 137 form-line mound features, to evaluate the model obtained from training and validation for a second time.

SoI map styles, colours and symbology depended on the date the maps were produced, the team drawing them, the region and the print quality of the map (Petrie et al., 2019). Each map type also corresponds to a drawing style and, therefore, to a different mound colour, despite corresponding to the same type of mound feature. There are three typologies by which mound features are represented in the SoI maps, of which the most common ones are the hachure and the form-line mound feature. The hachure is depicted with many fragmented lines which show the orientation of the slope, whereas the form-line mound features are drawn to represent one elevation (**Figure 32**). The third type of mound feature representation on SoI historical maps is shaded-relief. Although these are also present on the maps under study, they are not included in the automated detection given the low correspondence of this type of mound feature with archaeological sites, where 86.36% of the examples visited on the ground were found not to be archaeological sites (Green et al., 2019). We have focused the form-line algorithm on detecting only its most common typology, as opposed to the hachure algorithm which detects all types of hachure depiction. This is due to the fact that other form-line mound feature types (mound feature with concentric lines, with continuous line and black ones) do not have their characteristic shape and they are similar to other typologies that have no relation to archaeological features, such as road and slope lines (**Figure 34**). Likewise, cropped mound features by the process of clipping maps to 512x512 pixels have not been detected because there are form-line and hachure-shaped features that are not closed in a circle and are not mound features.

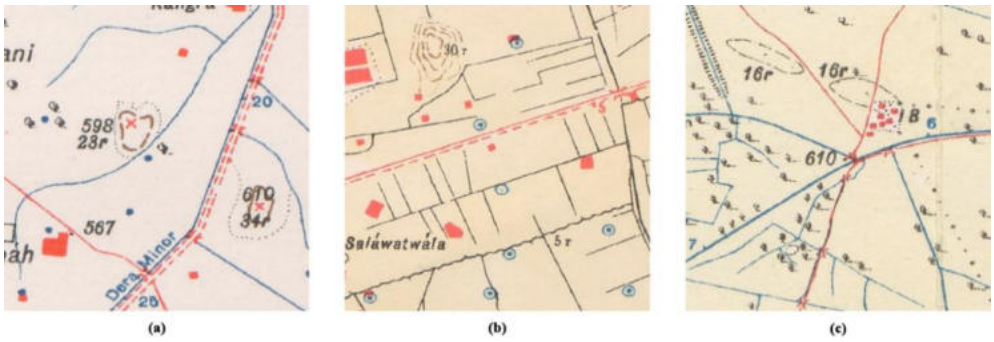


Figure 34. Different form-line typologies found in historical SoI maps: (a) dashed [23r] and solid line [34r] mound feature, (b) mound feature with concentric lines [30r], and (c) road-like black line mound feature [16r].

ML algorithms like Mask R-CNN typically evaluate their models on images that contain labelled objects and do not evaluate those without labels. Since our goal is to demonstrate the good performance of the model in low-density areas, we have created artificial mound labels on all those images without real mounds to force the analysis in them. This way, the algorithm also evaluates the presence or absence of mounds in areas of the map where we know there are no mound features to better assess its precision. The 4x4 pixel artificial mound features are placed in the upper left corner of the images and will never be detected, as our algorithm discards any detection at the edges of the images (10 pixels from the edge) to avoid FPs derived from cropped symbols. These artificial mound features will never be detected, but these areas will be analysed, allowing our model to analyse both high-density and low-density areas.

If our study had focused on areas with a high-density of mound features, our method and research could have ended here since we obtained good results after the first training for both hachure and form-line mound features. However, the majority of archaeological surveys are conducted in areas with a low-density of sites, or in places where the density of archaeological features is undetermined. Therefore, if we have looked at the reality of archaeological research and analyse the results of the first training for low-density areas, we observe that it is necessary to refine the model given the high number of FPs present in the results.

5.2.3. Model Refinement

The high number of FPs present in the first training was due to the limited number of training data available. Therefore, with the idea of introducing new training data, both positive and negative, various DA techniques have been applied. The first DA methods developed were mound feature random translation (DA1), random rotation (DA2), the so-called Doppelgänger technique (DA3).

For each type of mound feature and algorithm, 1,500 new artificial mound features were used, created randomly from the original ones used for training, and they were placed, by an automated process, randomly on all the maps used in training, implementing both DA1 and DA2. When pasting these artificial mound features at random on each of the training maps, they were emptied of any other feature than the actual mound depiction as they contained various symbols unrelated to the mound feature itself, thus avoiding possible FPs derived from the presence of these symbols, but also because the training maps had different background

colours and the inclusion of these features would have created artificial colour-related features (**Figure 35**).

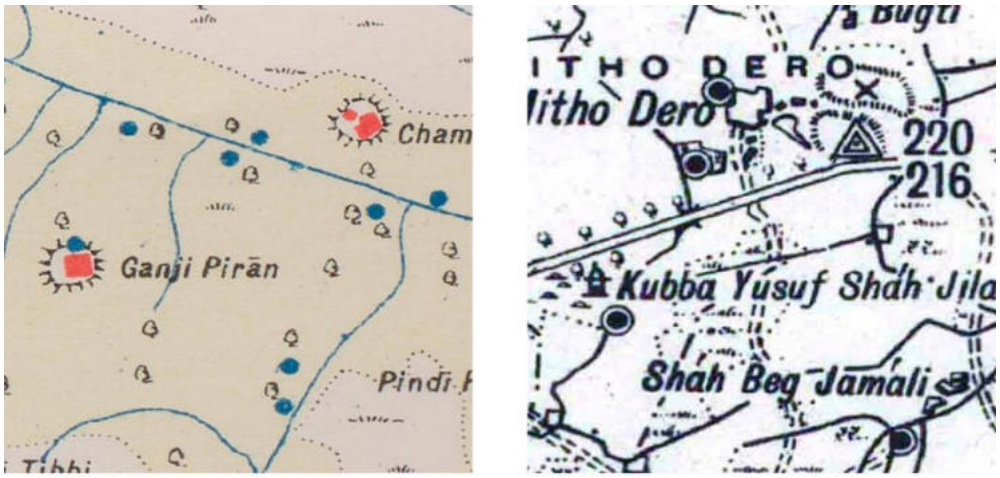


Figure 35. Some examples of hachure mound features containing different symbols inside, as well as two types of map background colour (a and b).

In order to avoid FPs due to common symbols on the maps such as roads, grass and trees where these new artificial mound features could have been placed randomly, DA3 was developed to copy the inside of each mound feature and to paste it to the outside of the mound feature so that it can be taken as negative training and just the mound feature as positive data (**Figure 36**). In this study it has been decided not to implement other possible DA techniques such as resizing, because mound features of different sizes are drawn differently than the resized mound feature itself. The hachure and form-line shapes are different for each size, increasing or decreasing the number of strokes drawn. Therefore, noise would be introduced into the algorithm. The entire DA process has been done using our own script written in Python (available as supplementary data).

Figure 36. First DA techniques used: (a) random translation (DA1), (b) random rotation (DA2) and (c) the so-called Doppelgänger technique (DA3).

After increasing the positive and negative training, the number of FPs detected was considerably reduced, but a series of specific FPs was still maintained. In order to further reduce these, a refinement stage (DA4) was included (**Figure 37**). In both cases the same correct mound features were used as continuous line circles for negative training data, so the algorithm could decide that continuous lines are not mound features. The total number of elements used as refinement is 88 for the form-line and 127 for the hachure ones, which have been placed using the DA1 and DA2 techniques up to a total of 8,800 for the form-line algorithm and 12,700 for the hachure model.

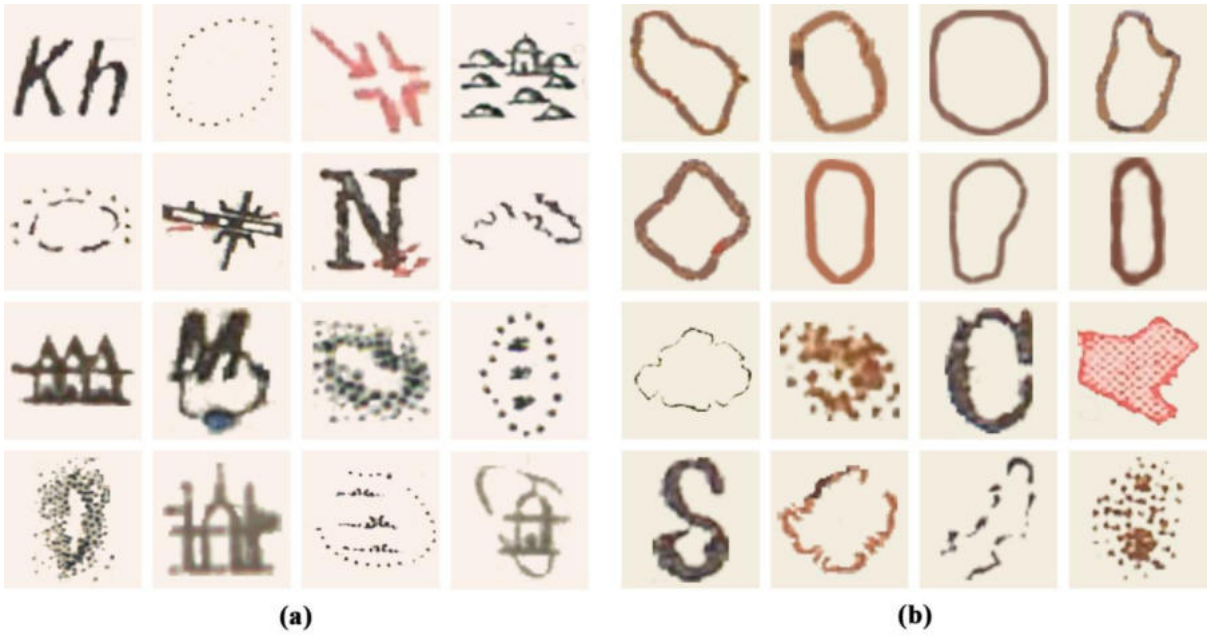


Figure 37. Some FPs used as negative training data for refinement (DA4): (a) hachure FPs and (b) form-line FPs.

5.2.4. Curriculum Learning Approach

Thanks to these DA methods we managed to reduce the number of FPs considerably, increasing the precision of the model. However, we stopped detecting some of the mound features that were initially detected, which also reduced the recall value. For this reason and with the aim of improving the accuracy metrics, it was decided to implement a CL strategy with synthetic data (DA5) (**Figure 38**).

Firstly, CL is a way to gradually introduce complexity to the model through more training phases (Soviany et al., 2022). Secondly, the lack of data forced us to create synthetic data for each mound feature class (DA5), which we have used to make the algorithm learn through a

CL strategy. In this way, the algorithm first learns the basics from the synthetic data and then more complex variations from the few known mound features in its second training, as a fine-tuning stage (**Figure 39**). A total of 75 synthetic mound features were created for each of the two types.

Figure 38. Hachure and form-line mound feature datasets for CL: (a) examples of synthetic hachure mound features (DA5), (b) examples of original hachure mound features, (c) examples of synthetic form-line mound features (DA5), (d) examples of original form-line mound features. The synthetic data (a and c) is for the first training of each of the two algorithms and the original data (b and d) for the second training also for both algorithms.

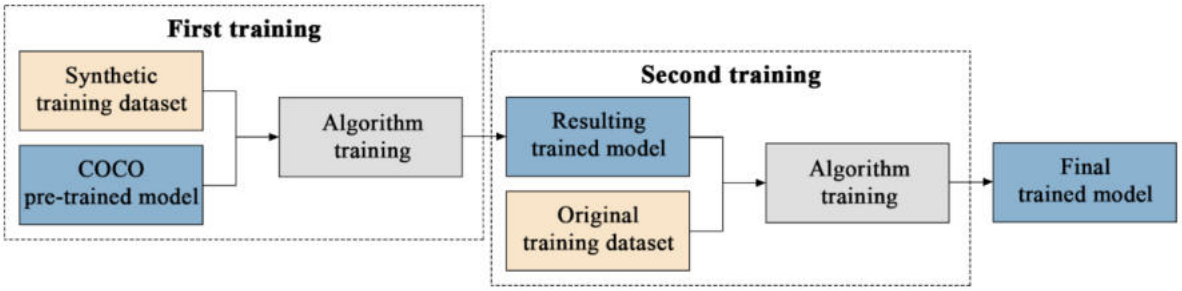


Figure 39. CL process scheme where stages with more complex aspects of the mound features are gradually included: first the synthetic dataset with DA and second the original with DA.

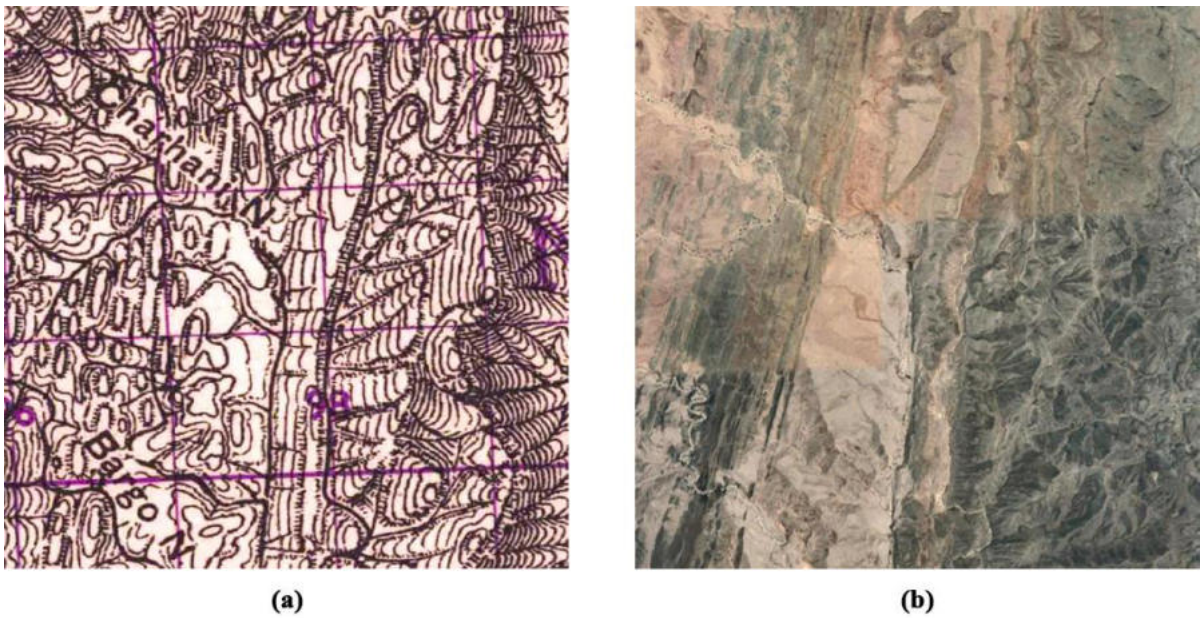
5.2.5. Model Filtering

Previous ground-truthing studies in India (Green et al., 2019), which included only a small number of well-preserved archaeological mounds, showed that those mound features smaller than 200 meters in diameter were mostly not archaeological sites, with hachure features adjacent to villages often corresponding to ponds or upcast from the creation of those ponds.

Only 7.96% of the hachure and 25.83% of the form-line mound features of less than 200 meters corresponded to archaeological sites (Green et al., 2019). Likewise, research on mound features in Pakistan showed that many of the small mound features less than 100 meters in diameter were mostly dunes or modern spoil from pond excavation (Garcia-Molsosa et al., 2021). In contrast, 56.34% of the form-line and 40% of the hachure features greater than 200 meters in diameter did correspond to sites (Green et al., 2019). For this reason, it has been decided to filter, throughout the study area, all those mound features formed by areas of less than 500 pixels, a range of 60-150 meters in diameter depending on the pixel resolution of each map, to avoid including mound features that are not likely to be archaeological sites (Filter1).

A second filter, using blob analysis, was applied to remove those elongated mound features which are not commonly archaeological sites and are mostly dunes. The ellipsoidal shape of each detected mound feature has been evaluated and all those that presented an elongation, a ratio between the largest and smallest diameter of the ellipse, greater than 3.5 were eliminated (Filter2).

Finally, in the post-processing stage, given the similarity of the mound features with the characteristic elevation shape of mountainous areas, a script was applied using GEE and QGIS to filter all those mountainous regions (Filter3), areas with a slope greater than 5 degrees (of mean value within a 7-pixel radius, equivalent in this area to 210m), and thus eliminate all mound features that, correctly identified by their drawn shape, do not correspond to possible archaeological mounds (**Figure 40**).



(a) (b)

Figure 40. Hachure mound-shaped mountain peaks on (a) historical map and (b) its satellite image.

5.2.6. Model Evaluation

Once the algorithm was trained, new mound features were detected in the remaining 581 maps for which we possessed no information on the presence of mound features. Given the diversity of the new maps compared to those used for training and validation (**Figure 41**), this evaluation was carried out differentiating the maps based on their similarity with those used in training and validation following a probability density function (**Figure 42**).

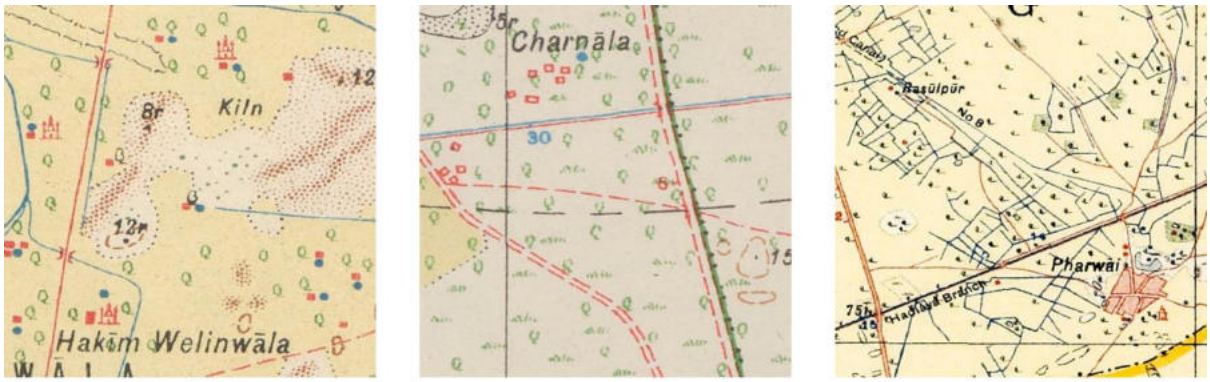


Figure 41. Similarity based on the RGB values of their backgrounds compared to the training and validation maps: (a) sample map used for training, (b) sample map used for test for a standard deviation of 0.5, and (c) sample map used for test for a standard deviation of 3.

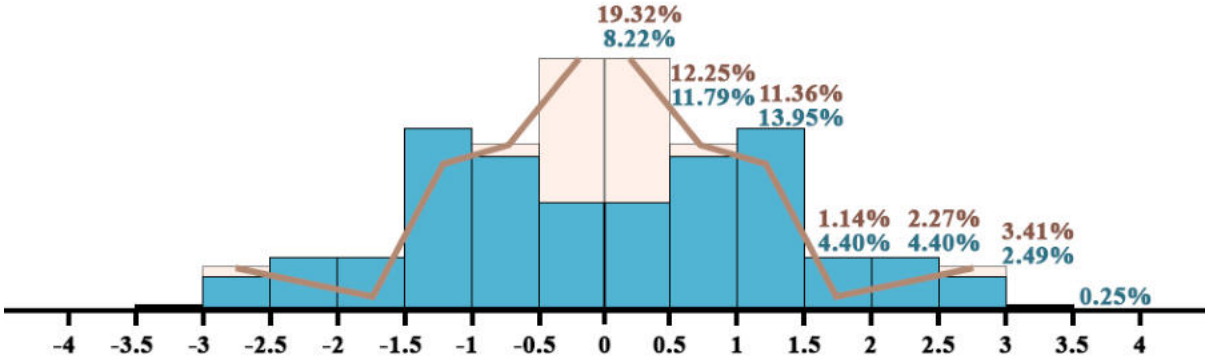


Figure 42. Percentage of maps in which new mounds are detected (blue) relative to the probability density of the maps used both in training and in validation (brown), their similarity based on the RGB values of their backgrounds.

This detection can be replicated in Colab in order to facilitate its application by other users with the aim of making this algorithm reproducible and replicable. The resulting shapefile

contains the masks of all detected mound features for easy viewing in standard GIS software such as QGIS.

5.3. Results

Below we present the results of the workflow followed for the detection of mound features in SoI historical maps. Both the initial (**Tables 5** and **6**) and the final results (**Tables 11** and **12**) of the detection of hachure and form-line mound features are presented, and only the intermediate results of the detection of hachure as an example of the evolution of the process (**Tables 7** to **10**), which was the same for both types of mound feature representations.

Table 5. Evaluation of the Mask R-CNN model in high and low-density validation datasets, average mound features per image, before the entire DA workflow for the detection of hachure mound features.

Algorithm	Density	TPs	FNs	FPs	Recall	Precision	F1
High-density	128.26%	87	21	26	80.56%	76.99%	78.73%
Low-density	2.67%	87	21	737	80.56%	10.56%	18.67%

Table 6. Evaluation of the Mask R-CNN model in high and low-density validation datasets, average mound features per image, before the entire DA workflow for the detection of form-line mound features.

Algorithm	Density	TPs	FNs	FPs	Recall	Precision	F1
High-density	95.77%	45	22	20	67.16%	69.23%	68.18%
Low-density	1.47%	45	22	1366	67.16%	3.19%	6.09%

Table 7. Evaluation of the Mask R-CNN models in low-density validation dataset using different DA techniques for the detection of hachure mound features: random translation (DA1), random rotation (DA2) and the so-called Doppelgänger technique (DA3).

Algorithm	TPs	FNs	FPs	Recall	Precision	F1
None	87	21	737	80.56%	10.56%	18.67%
DA1	71	39	37	64.55%	65.74%	65.14%
DA1+DA2	68	45	53	60.18%	56.20%	58.12%
DA1+DA2+DA3	68	44	31	60.71%	68.69%	64.45%

Table 8. Evaluation of the Mask R-CNN models in low-density validation dataset using a refinement step (DA4) for the detection of hachure mound features.

Algorithm	TPs	FNs	FPs	Recall	Precision	F1
DA1+DA2+DA3	68	44	31	60.71%	68.69%	64.45%
DA1+DA2+DA3+DA4	70	43	19	61.95%	78.65%	69.31%

Table 9. Evaluation of the Mask R-CNN models in low-density validation dataset using CL-based approach with synthetic data (DA5) for the detection of hachure mound features.

Algorithm	TPs	FNs	FPs	Recall	Precision	F1
DA1+DA2+DA3+DA4	70	43	19	61.95%	78.65%	69.31%
DA1+DA2+DA3+DA4 +DA5	77	38	11	66.96%	87.50%	75.86%

Table 10. Evaluation of area (Filter1), blob (Filter2) and slope (Filter3) filters in low-density validation dataset for the detection of hachure mound features.

Algorithm	TPs	FNs	FPs	Recall	Precision	F1
None	87	225	15	27.88%	85.29%	42.03%
Filter1	78	43	13	64.46%	85.71%	73.58%
Filter1+Filter2	77	38	11	66.96%	87.50%	75.86%
Filter1+Filter2+Filter3	77	38	10	66.96%	88.51%	76.24%

Table 11. Evaluation of the Mask R-CNN model in high and low-density validation datasets, average mound features per image, after the entire DA workflow for the detection of hachure mound features.

Algorithm	Density	TPs	FNs	FPs	Recall	Precision	F1
High-density	128.26%	77	38	3	66.96%	96.25%	78.97%
Low-density	2.67%	77	38	10	66.96%	88.51%	76.24%

Table 12. Evaluation of the Mask R-CNN model in high and low-density validation datasets, average mound features per image, after the entire DA workflow for the detection of form-line mound features.

Algorithm	Density	TPs	FNs	FPs	Recall	Precision	F1
High-density	95.77%	48	20	0	70.59%	100%	82.76%
Low-density	1.47%	48	20	4	70.59%	92.31%	80.00%

Finally, the trained model was applied to maps covering an area of 470,500 km^2 where a total of 2,802 hachure and 3,145 form-line mound features have been detected (5,947 mound features), and perfectly georeferenced by our algorithm (**Figures 43** and **44**). A manual evaluation of a series of maps of this area was performed, the aforementioned test dataset (**Tables 13** and **14**).

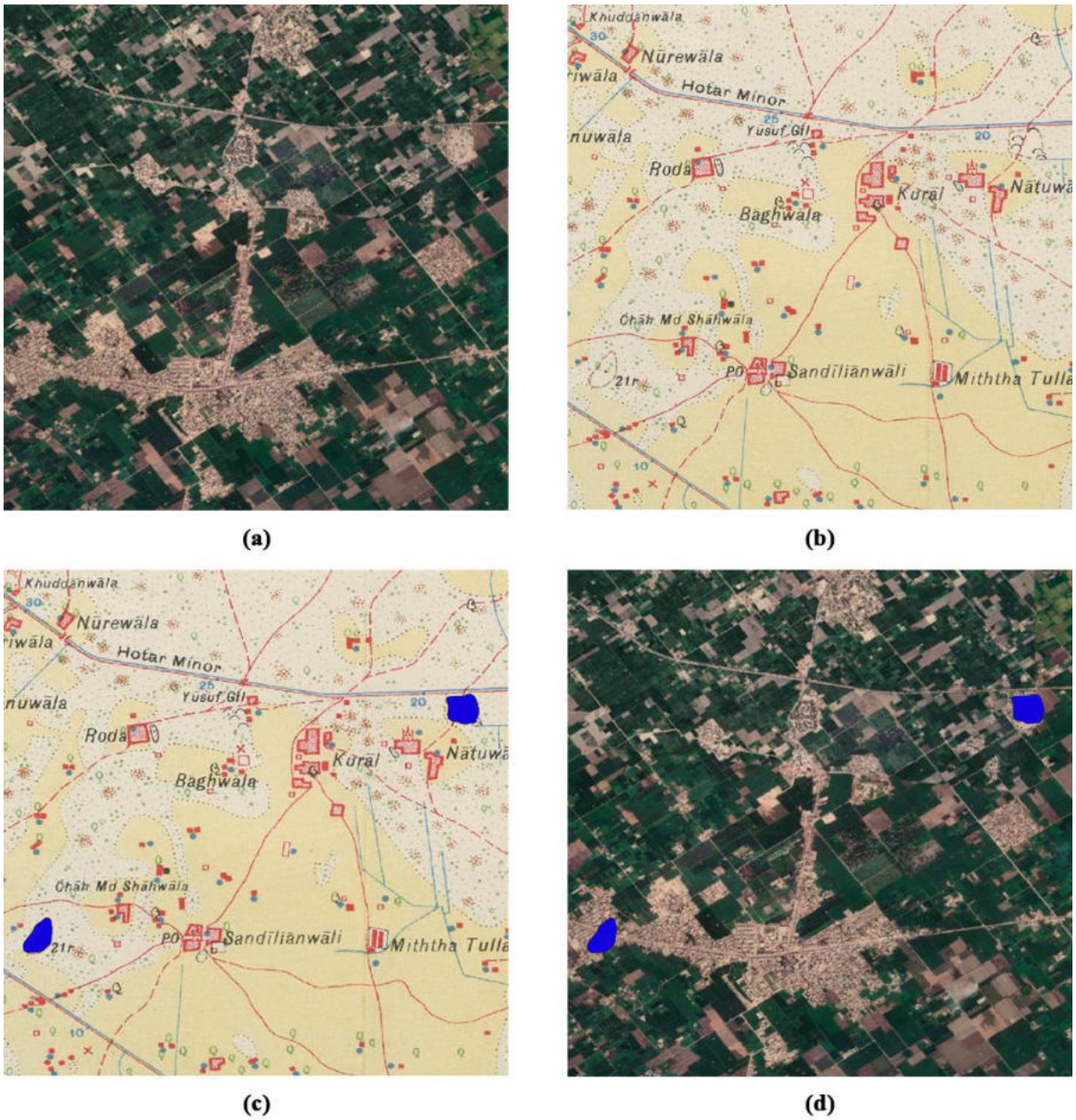


Figure 43. Detection of mound features [21r] in an area where urban and agricultural development have made those mapped mound features disappear: (a) satellite image of the area, (b) historical map of the area, (c) detection of form-line mound features (blue) on the historical map, and (d) location of the detected potential site mound features (blue) in the satellite image.

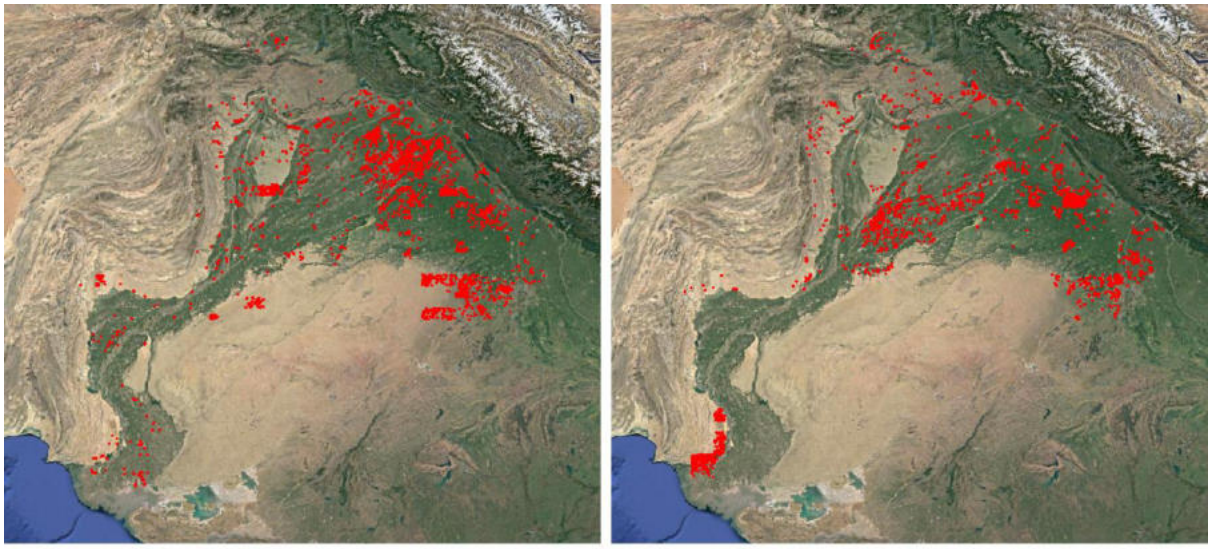


Figure 44. Distribution of detected mound features in the Indus River Basin: (a) hachure and (b) form-line mound features.

Table 13. Evaluation of the Mask R-CNN model in low-density test dataset based on its maps RGB similarity relative to training and validation ones for the detection of hachure mound features.

Similarity	TPs	FNs	FPs	Recall	Precision	F1
|0.5σ|	92	61	9	60.13%	91.09%	72.44%
|1σ|	111	89	14	55.50%	88.80%	68.31%
|2σ|	116	104	19	52.73%	85.93%	65.35%
|3σ|	121	109	26	52.61%	82.31%	64.19%

Table 14. Evaluation of the Mask R-CNN model in low-density test dataset based on its maps RGB similarity relative to training and validation ones for the detection of form-line mound features.

Similarity	TPs	FNs	FPs	Recall	Precision	F1
|0.5σ|	15	1	1	93.75%	93.75%	93.75%
|1σ|	25	6	5*	80.65%	83.33%	81.97%
|2σ|	97	40	27	70.80%	78.23%	74.33%
|3σ|	97	40	41	70.80%	70.29%	70.55%

*Four of the detected mound features were drawn in another way than the one used for training, the continuous form-line. For this reason, they have not been taken into account either as TP or as FP.

5.4. Discussion

5.4.1. Low-density Approach

In archaeology, it is common to find unsatisfactory results masked by the difference in the density of archaeological features. The density of the features must be taken into account (Soroush et al., 2020; Verschoof-van der Vaart et al. 2022) since good results in high-density areas may actually be hiding much worse results in low-density areas. The first results showed a number of FPs of up to twenty times more than the mound features present in the area (**Tables 5** and **6**). This algorithm would be useless in a large-scale survey, as it would generate a large number of FPs and an overly large dataset, which would not be of use in the planning of field validation or for archaeological analysis. These results strongly show that archaeological studies should focus their validation on low-density areas in order to avoid biased results.

During an archaeological survey, the true density of archaeological features is unknown, so algorithms must be developed to show good metrics in areas of both high and low-density of sites. Contrary to recently published discussions (Verschoof-van der Vaart et al., 2022), poor results in low-density areas due to the sparse presence of archaeological features and class imbalance are not inevitable, but these are the product of insufficient model training. The foreground-to-background imbalance as an example of class imbalance (Oksuz et al., 2019), is not the reason for poor results in the detection stage. The imbalance problem from each category for object detection in the training pipeline (Luque et al., 2019), occurs when one class heavily outnumbers the examples in the other class in the training data (Batista et al., 2004), not in the validation and test datasets. Variation in results due to the different density of archaeological features (**Tables 5** and **6**) can be resolved by different DA and CL approaches (**Tables 11** and **12**).

5.4.2. Model Refinement and Curriculum Learning Approach

The DA, with the introduction of 1,500 new mound features, significantly improves the precision by increasing the training data, both positive and negative. Both DA1 and DA2 show similar results that, despite the slight reduction in recall we have achieved a substantial improvement in precision (**Table 7**). Thanks to its negative training, the introduction of DA3 improves the precision of the model, which uses the DA4 to improve its accuracy.

The initial training data was not sufficient and resulted in a large number of FPs indicating that the model had not learned well what a mound feature looks like. The increase of the training data removed a large number of FPs, but to eliminate more specific FPs it was necessary to resort to DA3 and DA4 (**Table 8**). As shown in **Figure 37**, most of the FPs used in refinement were pointed circular and non-circular shapes for the hachure algorithm, and both continuous and dashed circular shapes for the form-line model.

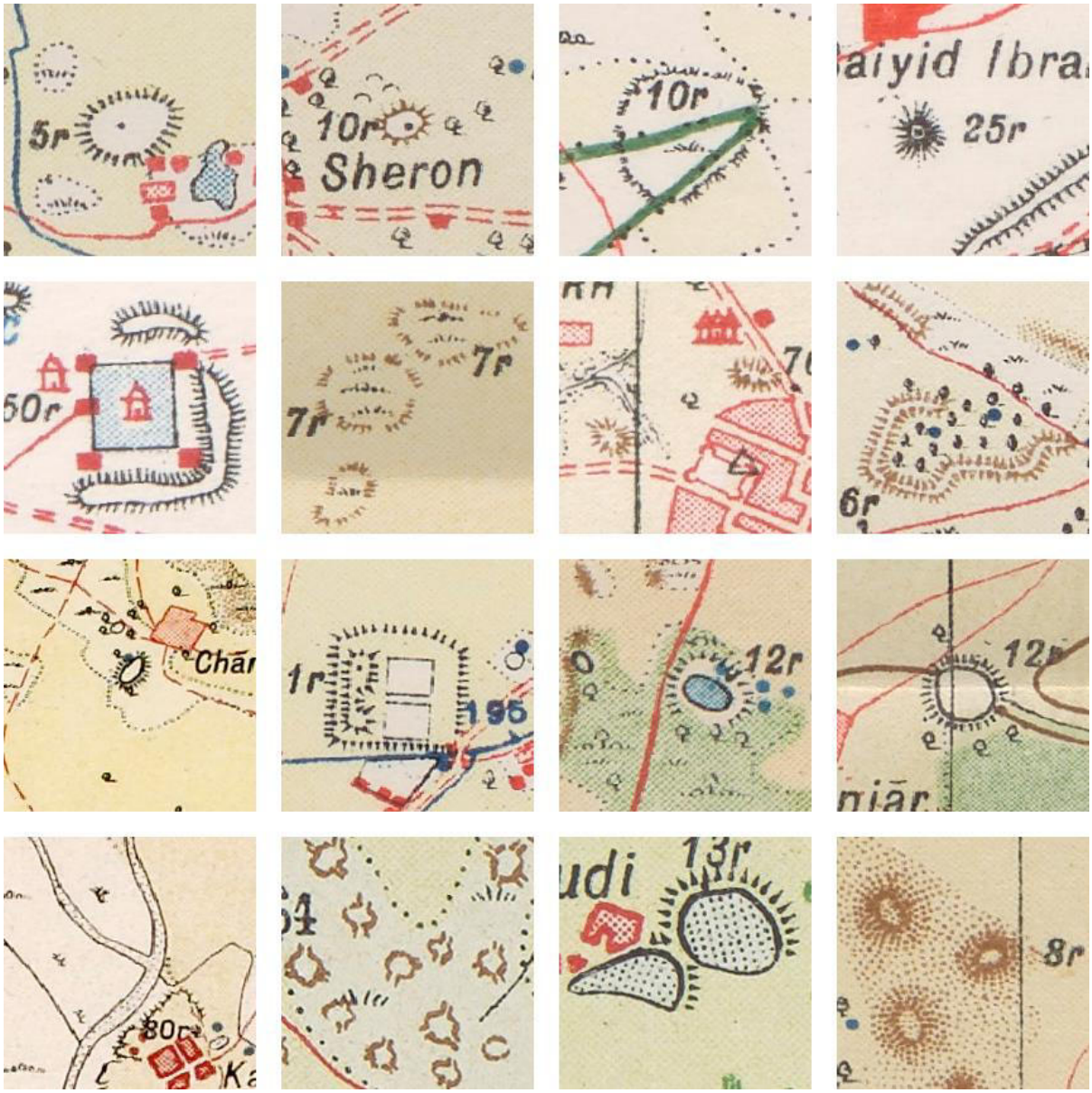


Figure 45. Different types of hachure mound features detected after applying the trained model. The last image represents the third type of mound features on the maps, the shaded relief mound features, erroneously detected as hachure but similar to them due to their characteristic pointed and circular shapes.

Likewise, as can be seen in **Figure 45**, the use of DA5 has allowed the detection of hachure shapes not included in the original training data. The inclusion of synthetic data, along with the CL strategy, has allowed the algorithm to better understand what the mound features look like. The CL using synthetic data helped to develop an algorithm from a small training dataset, which is common in archaeology. As seen in **Table 9**, both the recall value and the precision value improved noticeably.

5.4.3. Model Filtering

Smaller objects, such as mound features less than 500 pixels in area, are the most difficult for a CNN to detect, because such objects do not have enough pixels for the necessary feature extraction. That is why the recall value is so low without Filter1 but high enough when we apply it (**Table 10**). Both Filter2 and Filter3 remove many FPs, which results in an increase of the precision of the model, with fewer, but higher quality results that are more likely to be of archaeological interest.

In future work, the idea of developing new filters could be contemplated for the elimination of mound features correctly detected but not correctly classified in their type. Some hachure mound features, in addition to being detected by the hachure algorithm, have been detected by the form-line mound features algorithm. What has been detected is not the complete mound feature but only its interior, which on many occasions resembles a form-line mound feature. These misclassified mound features could easily be removed with a filter that discards the smallest duplicate detected mound feature. This can also happen with the hachure and the shaded-relief mound features. Some shaded-relief examples, as the last image of **Figure 45**, resembles a hachure mound feature. Applying the same filter mentioned above would also resolve these double detections, as well as reduce the FPs for shaded-like dunes.

5.4.4. Model Evaluation

Only 40.03% of the maps with unknown mound features, the ones used for testing, are similar to 63.64% of the maps used for training and validation (**Figure 42**), so most are substantially different. This diversity as well as its resulting metrics (**Tables 13** and **14**) indicate the need for an adaptive algorithm that allows, after a small amount of retraining with data detected from a new map, a better general mound feature detection in the same map. The more similar

the maps are to those used in training and validation, the more similar the test metrics are to the validation ones. An adaptive algorithm would improve both recall value by including different ways of drawing the mound features, only some of which have been detected thanks to the synthetic data, and precision value by including backgrounds not taken into account in the original training.

Likewise, new DA methods could be included in the training, such as random brightness jittering and random Blur/Sharpen (Berganzo-Besga et al., 2022a). Some test maps, unlike those used in training and validation, have shown darker and blurred images (**Figure 46**).

(a) (b)

Figure 46. Map samples found in the test data with different characteristics than those used in training and validation: (a) darker image and (b) blurred image.

5.4.5. Comparison to Manual Digitisation of Mound Features

The VIA was used to hand digitise 756 mound features in JSON format which were digitised using 64 random historical maps. The density of mound features is not distributed uniformly throughout each map. Instead, mound features frequently cluster together, indicating a high number of mound features on certain maps and a low number on others. This type of pattern increases the amount of labour and time necessary for manual mound feature digitising using GIS software. We predicted that manually digitising all mound features from the 645 historical maps used in this research region would take an experienced professional more than 120 work

hours based on the manually digitised mound features prepared as training data for the algorithm. The detection time, running each algorithm on a single NVIDIA A40 GPU, has been more than 6 computing hours. While 120 hours does not seem too long for this project, creating a ML-based algorithm paves the way to scale this research to the additional 2,200 historical maps covering other parts of Pakistan and India that have been scanned and are ready for analysis.

5.5. Conclusions

A workflow has been designed with different techniques and strategies that has allowed not only the detection of nearly 6,000 mound features in India and Pakistan, which will allow for a better understanding of the settlement distribution related to the Indus Civilization and later cultural periods, but has also provided solutions to common problems in archaeology such as the low-density of archaeological features in large-scale surveys and the few training data for ML models.

Historical maps constitute one of the basic sources available to both historians and archaeologists. They are housed in many archives and some series cover very large national territories using very similar symbols and conventions. This study opens the door for the large-scale automated extraction of relevant information from historical maps and, in doing so, provides a workflow and open code that has the potential to make an immensely contribute to the historical sciences.

As with other large-scale site detection methods (Berganzo-Besga et al., 2021), these DL algorithms will allow researchers to carry out studies that could not be done before given the new amount of data obtained, facilitating the task of the archaeologist. Furthermore, this model could be applied in other regions that have historical maps such as Syria and Lebanon (Garcia-Molsosa et al., 2021), but particularly those areas that were also mapped by or followed the model established by the SoI. The outputs of this study represent a powerful tool in the large-scale documentation and monitoring of archaeological heritage, with much work ahead to validate the results through remote sensing, archival work, and ground survey in collaboration with partners in India and Pakistan.

5.6.Acknowledgements

The Mapping Archaeological Heritage in South Asia (MAHSA) project is funded by Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin. The authors would like to thank Junaid Abdul Jabbar, Mou Sarmah, Ushni Dasgupta, Azadeh Vafadari, Kuili Suganya Chittiraibalan, Arnau Garcia-Molsosa and Adam Green.

5.7.Author contributions statement

IBB developed methods, executed research, wrote the initial draft of the paper, implemented corrections, produced the figures; HAO and FL, planned research, developed methods, corrected initial draft; AA, PJG, JGS, AK, RC, MSM and JT georeferenced maps; RR coordinated research; CP coordinated research, planned research and acquired funding.

6. New developments in Drone-based Automated Surface Survey: Towards a Functional and Effective Survey System

Hector A. Orengo[1,2*]; Arnau Garcia-Molsosa[2]; Iban Berganzo-Besga[2]; Juergen Landauer[3]; Paloma Aliende[2] and Sergi Tres-Martínez[4]

1. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

2. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

3. Independent researcher, Ludwigsburg, Germany

4. HEMAV Foundation, Barcelona, Spain

*Author to whom correspondence should be addressed.

Archaeol. Prospect.. **2021**, 28(4), 519-526; https://doi.org/10.1002/arp.1822

Received: 12 November 2020 / Accepted: 06 April 2021 / Published: 06 May 2021

Abstract

This paper presents new developments on drone-based automated survey for the detection of individual items or fragments of material culture visible on the ground surface. Since the publication of our original proof of concept, awarded with the Journal of Archaeological Science and Society for Archaeological Sciences Emerging Investigator Award 2019, additional funding has allowed us to implement a series of improvements to the method. These aim to improve detection capabilities and the extraction of items' shapes and increase flight autonomy, control, area covered per flight and the type of environments in which the method can be applied while reducing computing needs, processing time and expertise necessary for its application. This paper provides an account of the methods followed to achieve these objectives, their preliminary results and the current development for their implementation into a free and open-source system that can be used by the archaeological community at large.

Keywords:

drone; survey; archaeology; potsherd; SLAM; deep learning; machine learning; Google Colaboratory

6.1. Introduction

In a recent paper (Orengo and Garcia-Molsosa, 2019), we presented a proof of concept for an automated archaeological surface survey workflow combining drone-based image acquisition, photogrammetry and probabilistic ML in a cloud computing platform capable of locating and extracting potsherds from drone imagery. Since the publication of this paper, the team developing the method has received funding to move from this initial proof of concept into a fully functional system that can be routinely applied in standard archaeological surveys. This paper presents the latest and future developments of the system and explores its potential to standardize the practice of archaeological surface survey.

The workflow provided excellent results in its first field tests, proving to be more accurate and faster than standard field survey under ideal conditions. It is also unaffected by traditional survey biases such as individual surveyor experience (but see Section 3 for an account of ML's biases). In addition, the workflow provides quantifiable information such as density, size, shape and number of potsherds per unit area. This approach is also capable of providing quantifiable spatial information about visibility conditions and percentage of both FPs and undetected potsherds. This information can be used to calibrate the classification results and obtain reliable extrapolations of total ceramic densities. This new approach has the potential to change the way in which pedestrian archaeological surveys will be carried out in the future and offers, for the first time, comparable results between different surveys with a fraction of the costs and time than those of traditional survey. In this regard, the method can offer a way forward for the quantifiable large-scale use of survey data in the analysis of past human occupation.

However, the current set-up of the system presents several problems that hinder its implementation:

1. A certain amount of expertise is necessary for (a) flying or programming the drone flight to obtain photographs with enough overlap for photogrammetric reconstruction, (b) the photogrammetric processing of the acquired images and (c) the training of the

ML algorithm, which requires an experienced operator and, in its current form, needs to be retrained to adapt it to different field conditions and pottery types. In addition to this, the workflow is not integrated within a single platform requiring access to different software and hardware, strict control of the parameters for each of these and a certain level of computational expertise to combine all these processes.

2. The method requires significant amounts of computing power, which are not available in common or professional PCs. Our initial set-up proposed two intensive computational processes: the development of orthoimages from all drone-acquired photographs and the ML-based detection of pottery. The photogrammetric process can easily involve 500–1000 photographs per flight, which require 2–4 h of computation using a high-end computer (8 Intel i7 cores and 64GB RAM) and a standard photogrammetric software such as Agisoft Metashape. In a similar way, the ML process required hours of computation for the training and application of the probabilistic classifier. In order to address the latter issue, the initial workflow employed GEE geospatial analysis cloud computing platform to conduct the ML analysis, but this:
    a. Limits its use to non-commercial operations. This is an important barrier to the adoption of the method as much survey work is developed under contract archaeology.
    b. Hinders the application of DL approaches that have the potential to significantly improve the results.
3. Currently available commercial drone platforms also limit the capacity of the method to be a viable alternative to pedestrian surface survey:
    a. The method is limited to ploughsoils (in flat and unforested terrain) due to flight limitations of current commercial drone platforms, which are not capable of keeping a constant height above ground.
    b. Camera resolution requires the drone to fly at a very low altitude (3–4 m above ground) to acquire images of potsherds at enough resolution for the ML algorithm to be able to detect them. This results in a decrease of the speed at which the drone can fly. This coupled with the relatively short flight times provided by current commercial drones (around 28 min in the case of the Phantom 4 Pro V2.0, which we have previously employed), at the time of the writing of the original paper, limits the area that can be surveyed during each individual flight.

4. Although this approach is able to detect ceramic fragments, it does not differentiate between types of ceramic (which is basic to achieve chronological and typological definition of sites defined by ceramic concentrations) or between potsherds and other types of material culture.

Our new approach aims to build up on our previous research on and experience with drone-based automated archaeological survey to overcome these issues and develop a fully operational, practical and accessible system available to archaeologists in a variety of survey circumstances. In this paper, we describe a novel set of software and hardware tools specifically developed or in development for intensive archaeological survey.

6.2. Technical Improvements to the System

A series of improvements aiming to tackle the previously stated problems are currently being implemented, as outlined below.

6.2.1. Platform Developments

During the last months, we have been collaborating with HEMAV, one of the largest drone-based industries in Europe, for the development of a drone based on their popular HAR9 multirotor platform. This drone will be carrying a 40- to 60-MP high-quality camera. This camera will considerably increase image resolution and quality and will allow the drone to fly much higher (and therefore faster) than common commercial drone cameras. This will improve considerably the detection capacity and speed of our previous proof of concept. The drone incorporates other characteristics directed towards archaeological survey (**Figure 47**): an altimeter that will allow the drone to fly at a constant height above ground. This will facilitate the acquisition of images with similar ground sampling distances independently of terrain variation. Forward-facing CV cameras will allow the drone to avoid obstacles. The drone will have a flight autonomy of around 40–45 min depending on the weight of the selected camera (longer than most commercial and professional drone models available). Another important characteristic of the drone will be the integration of real-time kinematic (RTK) navigation with a stable base station. This will allow a positioning of centimetric precision that will be extremely useful to plan the flight and image capture and will facilitate the photogrammetric processing of the images. It will also allow the drone to continue the image acquisition at exactly the same position where it stopped when recording large areas that might require

several battery changes. All in all, these features make our prototype an efficient option to conduct archaeological survey. Commercial drone options such as DJI Mavic Air 2 or Autel Evo 2 with flight times above 30 min and cameras above 40 MP do not usually include an RTK option. Despite including bottom sensors that can measure ground distance, they do not use this information to keep a constant height above ground but to avoid collision. The same is true for the DJI Phantom 4 RTK, which includes a GNSS module and ground sensor but a 20-MP camera (although with a larger CMOS sensor than the previously discussed models). The prototype will be available at a price of €8000 (not including camera), which is a lower price than other drone models with similar characteristics currently available in the market.

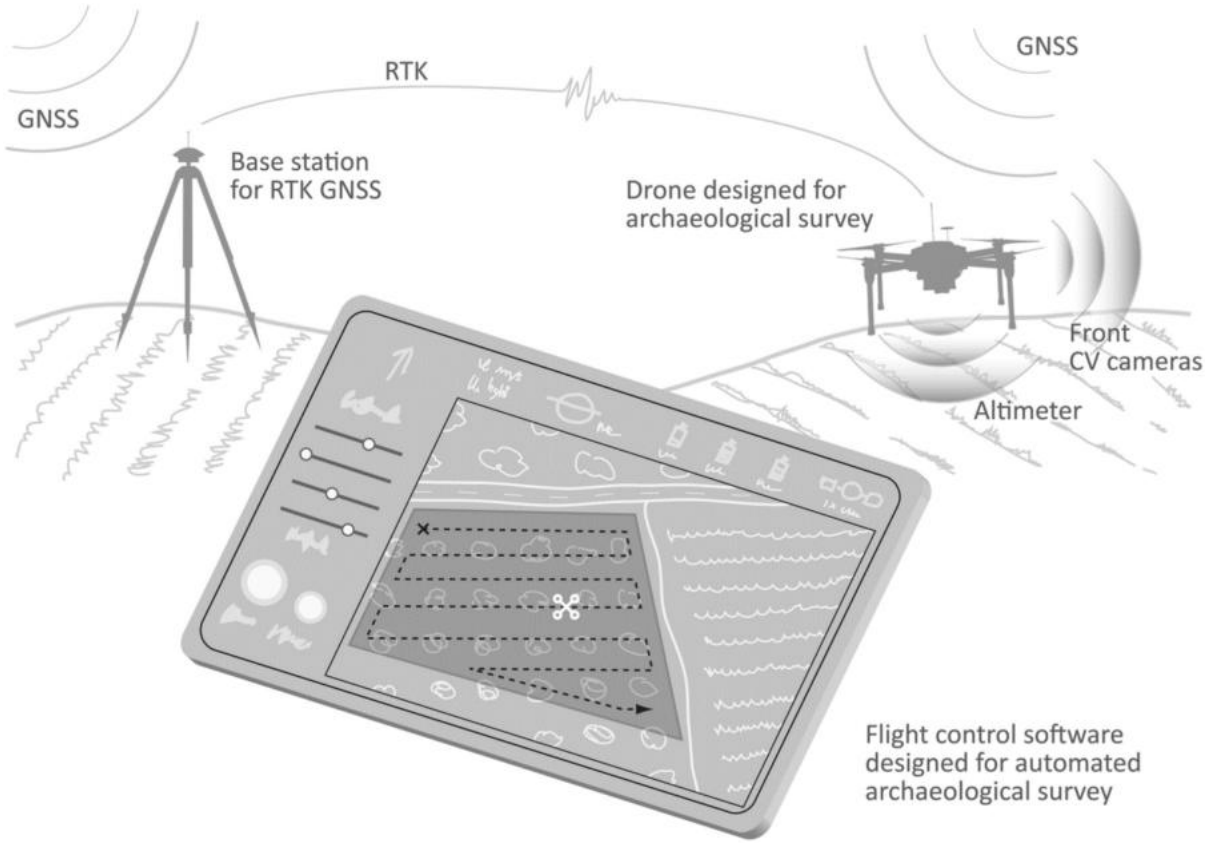


Figure 47. Field operation and sensors of the AIDAS system.

6.2.2. Ground Control System Developments

The ground control system HEMAV Planner is a simplified version of Mission Planner for PX4 flight control software, which uses the MAVLink communication protocol (all of these being developed by successful and active open-source code projects). The ground control

system is designed to operate with the drone that is controlled by a Pixhawk (PX4-compatible) autopilot and focused on the specific needs of archaeological survey. Two CV cameras able to detect objects in the drone path are installed in the front part of the drone. We chose to install only front sensors to reduce weight, battery usage and price; as a consequence, we have limited the drone to forward flight only. This will ensure the drone will always be able to detect any obstacle it its path. HEMAV Planner will also be configured to simplify image acquisition. The pilot will only need to draw the area that requires survey, and the software will automatically calculate the flight height and route for optimal image resolution and image overlap. This route can be interrupted (in order to change the drone battery or the memory card) and continued at exactly the same position as many times as necessary, thus allowing the coverage of large areas with little loss of time.

6.2.3. Orthophotomosaic Generation Process

The previous photogrammetric processing of the images to create a single orthoimage that could be used for the location of potsherds was computationally costly and required specialized photogrammetry software such as Agisoft's Metashape. Although it is true that many open-source options are nowadays available for the photogrammetric processing of multiple images, these are still complex to install and use, and not all of them provide options for the output of orthomosaics. In order to address the issues, we decided to abandon photogrammetry and aim at the implementation of a simultaneous localization and mapping (SLAM) algorithm. SLAM algorithms use external data such as lidar, images and sound to geolocate the sensor and construct an approximate map of its environment as it moves. Common implementations can be found in consumer robot vacuum cleaners, and they are an increasingly important component of autonomous vehicle navigation systems. More importantly, SLAM solutions can produce real-time data, and, although less accurate than photogrammetry, they provide an excellent alternative to the mosaicking of drone-based imagery (Bu et al., 2016; Kern et al., 2016). Therefore, we will aim to implement a SLAM algorithm, such as ORB-SLAM2 (Mur-Artal et al., 2015; Mur-Artal & Tardós, 2017), that can input information from monocular cameras while providing excellent accuracy and speed.

We are now developing tests to decide the most efficient way of SLAM implementation. This can take several forms: the generation of an orthoimage from the acquired imagery using an ORB-SLAM-based system such as Map 2DFusion (Bu et al., 2016), which will be later

subjected to potsherd extraction or the direct georeferencing of the potsherds detected in each of the individual images using the output of SLAM location data. This last method has the advantage of using the original images instead of a composed image that can introduce deformations and saving the computation necessary to generate the orthoimage and the drive space it will require to be stored. However, it can incorporate duplicate detections due to the same pottery sherds being present in different overlapping images. Duplicates can be removed at a later stage based on the spatial coordinates of their bounding boxes, but this might cause problems with overlapping potsherds.

6.2.4. Machine Learning Algorithms

Our previously published RF ML algorithm provided results more efficiently than standard pedestrian survey in terms of the number of items detected and time investment (Orengo and Garcia-Molsosa, 2019). However, the algorithm was only able to locate from 1/3 to 3/4 (depending on the field) of the potsherds counted by experienced surveyors using a very intensive total count method. Another problem was the presence of FPs, which could account for a small part of the total potsherd count. Lastly, although the vectorized extraction of potsherds (see **Figure 48**) was a good approximation of the size of the potsherds, the shape itself was poorly defined, hindering the application of analyses based on the potsherd morphology.

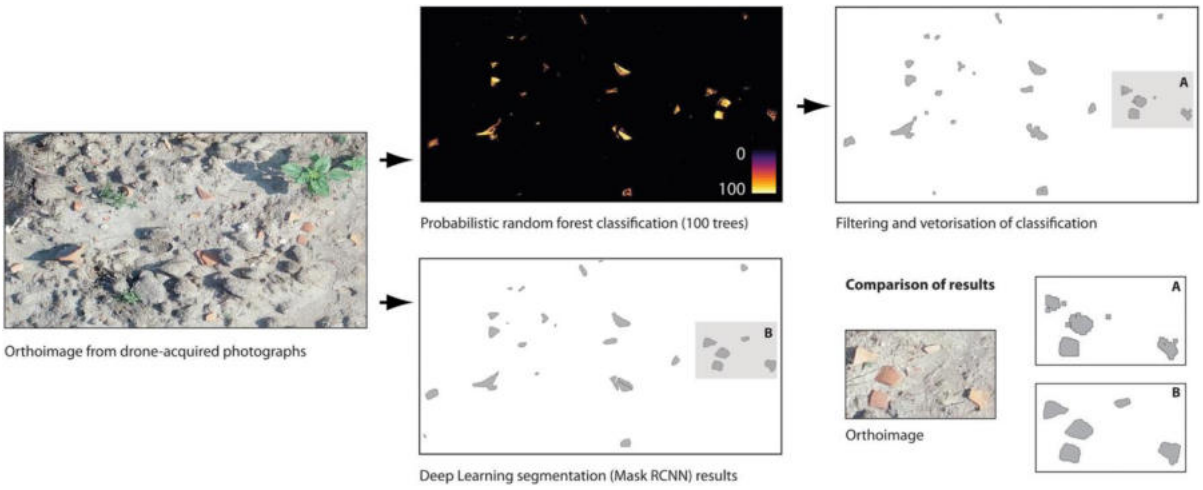


Figure 48. Comparison between the RF (a) and the mask R-CNN algorithms (b).

We are testing and implementing new segmentation algorithms based on CNNs, in particular Mask R-CNN (Waleed, 2017). CNNs are much more efficient than traditional ML algorithms such as RF, which we used in our previous work (see **Figure 48** for a comparison), but their training requires much more computational power and a much larger amount of training data (more than 10 k examples are usually recommended). However, once the training is finalized, the algorithm can be executed with a commercial laptop within minutes. Rather than detection algorithms, we have preferred to employ segmentation procedures, both able to locate the position of objects in an image, as these are able to define the shape of the object of interest while automatically tagging the rest of the image as background (i.e. all other types of objects that are not potsherds and have not been delineated as such), which is adequate to improve the algorithm through several interactions.

R-CNN is an object detection algorithm based on a combination of classical tools from CV and DL, which achieved a mean average precision (mAP) of 53.7%, more than 30% relative to the previous best result (Girshick et al., 2014). The fundamental difference between a CNN and fully connected (FC) neural networks is that the former is capable of detecting local patterns versus global patterns that the latter detects (Torres, 2020). Mask R-CNN detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance (He et al., 2017). It extends Faster R-CNN (Girshick, 2015; Ren et al., 2016) by adding a branch for predicting segmentation masks, a small fully convolutional network (FCN) (Long et al., 2015), on each region of interest (RoI), in parallel with the existing branch for classification and bounding box regression. Mask R-CNN is simple to train and adds only a small overhead to Faster R-CNN (He et al., 2017).

The chosen work environment is Colab environment, a Jupyter notebook environment that requires no configuration and runs entirely in the cloud. It allows the use of Keras, TensorFlow and PyTorch. The main motivation for its choice was that it provides free accelerators like graphics processing units or specialized hardware like tensor processing units.

Preliminary results present a modest improvement (see **Figure 50**) as the Mask R-CNN algorithm has managed to detect a larger number of potsherd than those previously detected in the same field (1698 compared with 1597 for our previously published RF-based algorithm). Despite these encouraging results, we aim to improve the accuracy of the algorithm using more training data at higher resolution and increase the use of DA methods. DA processes are able to multiply the training data, creating new synthetic data through the modification of exiting

available data. Usual approaches in CV include simple image techniques such as scaling, rotating, flipping, cropping and changes in contrast, saturation, brightness, hue and so on. They constitute a popular option in DL models as these usually require large amounts of training data that are not always available.

Initial tests using segmentation-based approaches suggest an increase of 6% in the detection rate and a decrease of 11% in the presence of FPs. The RF classifier seemed to outperform the DL segmentation only when it comes to the detection of fragments typically smaller than 1.2–1.5 cm^2 (for a ground sampling distance of approximately 1.34 mm^2); the convolution cannot properly identify the texture and edge of the object. For these small fragments, the more colour-focused approach of the RF algorithm can still identify the pixels as belonging to ceramic fragments. All in all, the enhancement in potsherd detection together with a notable improvement in shape extraction (see **Figure 48** for a comparison between these two approaches) that the use of the CNN-based segmentation approach offers widely justifies its implementation in the new system.

However, RF potsherd detection can still be useful in specific survey situations and needs (see **Figure 51**) such as extensive sampling-based potsherd detection where the drone is flying too high (and therefore faster) for the convolutional approaches to work properly given the reduction in image resolution. Depending on the strategy adopted (see **Figure 49**), the surveyor might prefer not to acquire overlapping images and move into a regular sampling of the area where every image will be analysed independently to provide a measure of the presence of potsherds in that specific image. However, this approach will have to be more sensitive to the environmental conditions in which survey is carried out as the colour of the potsherds will play a more important role in their detection and high contrast between soil surface and potsherd can significantly increase detection chances. Also, although this approach will be able to provide fast results for a relatively large area, this might need a specific training of the algorithm.

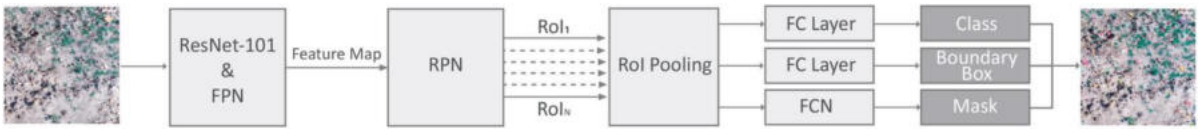


Figure 49. The implemented Mask R-CNN workflow for instance segmentation. The first step is to obtain the feature map of the input image using 101-layer residual network (ResNet-101) and feature pyramid network (FPN). Then, through region proposal network (RPN), the region of interest (RoI) is obtained. After a pooling process, the fully connected (FC) layer is applied to obtain the class and the boundary box, and the fully convolutional network (FCN) to get the mask. With these three elements, we obtain the resulting image with the detected potsherds.

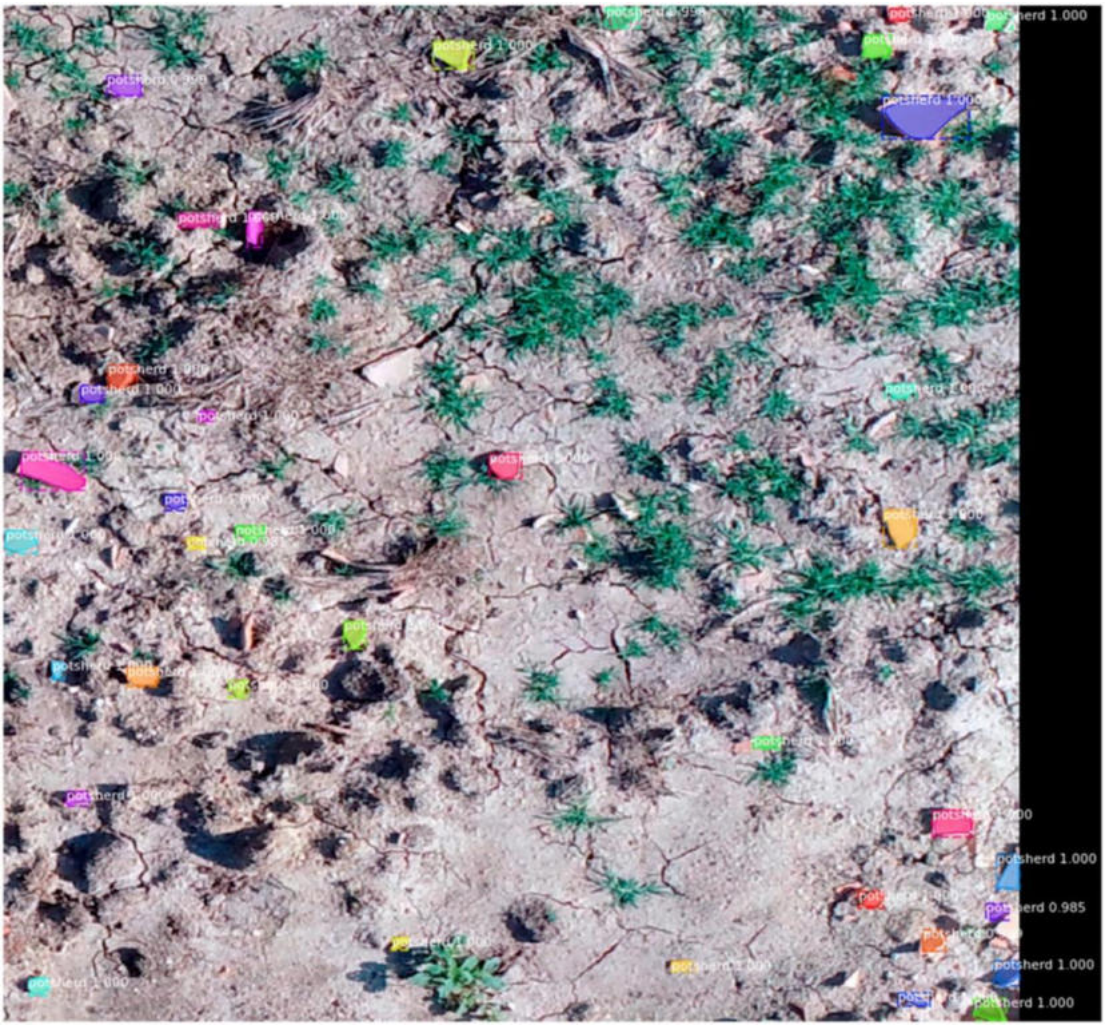
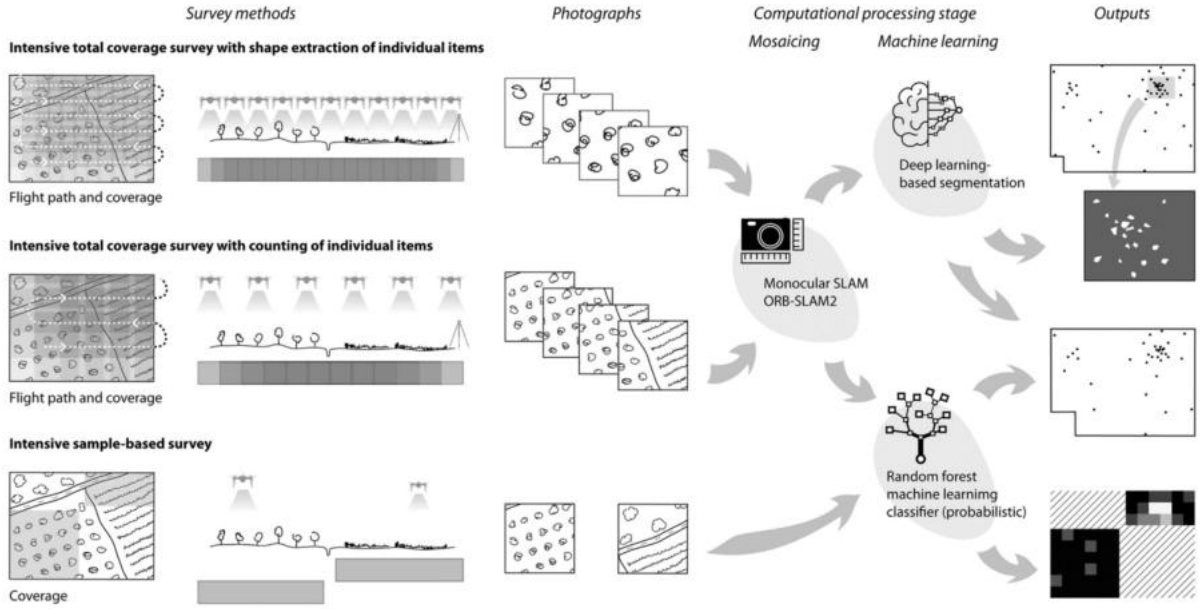


Figure 50. Potsherd detection using Mask R-CNN. The used environments were TensorFlow 1.15.2 and Keras 2.1.0. The implemented parameters were epoch = 160 and Fliplr(0.5) for DA. The results were a loss value of 0.0531, an mAP of 57.33% and a number of potsherds detected of 37.



Type of survey	Aprox. altitute	Flight speed	Area covered	Comput. processing	Accuracy of ids	Cost	Capabilities	Pros	Cons
Intensive total coverage survey with shape extraction of individual items	6-10 m	★★★☆☆	★★☆☆☆	★★☆☆☆	★★★★★	★★★☆☆	Intensive archaeological survey and recording identification of shape, size, orientation, etc.	Extreme accuracy and capabilities beyond current pedestrian survey. Allows comparison between surveys. Advanced analysis of artefacts' distribution.	Small area coverage, highest need of computational resources, highest time investment.
Intensive total coverage survey with counting of individual items	10-20 m	★★★★★	★★★☆☆	★★★★☆ Rand. forest ★★★☆☆ CNNs	★★★☆☆	★★★★☆ Rand. forest ★★★☆☆ CNNs	Intensive archaeological survey, fragment indentification and counting	Good detection capabilities equivalent to current total coverage intensive pedestrain survey. High speed.	Fragment shape is lost or not well defined. Counts are only approxi-mate. Smaller fragments may be ignored.
Intensive sample-based survey	10-30 m (48 Mpx camera)	★★★★★	★★★★★	★★★★★	★★★☆☆	★★★★★	Discontinuous recor-ding from number of fragmetns to density per area depending on height. Sampling.	Very fast, ideal for large areas. It requires less computational resources, and less drone components.	Recording does not reflect the real fragment number but a sample. Might require specific algortihm training. Sites might be lost.

Figure 51. Workflow development from data capture to output generation depending on survey type.

The combination of approaches allows flexibility to carry out surveys with different objectives, from rapidly sampling a large area with a minimum of computational resources to very intensive survey, which includes the extraction of the shape of each fragment of material culture detected.

6.2.5. Automated Survey Software

An important component of the system will be stand-alone software where archaeologists in charge of survey campaigns can introduce the images captured by the drone or have them directly delivered while the drone flies. This is an important element of the system as it will

not just allow the user to evaluate the performance and take survey decisions while on the field based on the results obtained but will make the system independent of an Internet connection, which was necessary to execute the original workflow. The software will automatically mosaic the images (which the drone flight control system will have acquired at suitable distances for correct overlapping and potsherd identification) and apply the potsherd detection algorithm. We are working towards the integration of an automated method to calculate field visibility conditions using the images. At the moment, our efforts in this regard have resulted in the automated production of green pixel maps that can be used to evaluate the distribution of material culture. However, the range of surface vegetation colours is very wide, and special specific types of vegetation might need to be accounted with purposely trained algorithms. In this regard, the software package will also provide options for the training of new segmentation algorithms for specific survey needs. Besides the identification of specific types of vegetation, this will serve several other purposes: (a) to allow the development of specific detectors to adapt the algorithm to particular survey conditions that might have not been taken into account with our initial training and/or specific types of potsherds, (b) to enable the system to detect specific types of material culture beyond potsherds (lithics, bone, stone and so on) and (c) to facilitate the adoption of automated survey procedures in other disciplines with similar interests. It is important to note that although the system has been originally designed for archaeological survey, there are multiple disciplines that could benefit from such a survey system. Detection of plagues, wildlife monitoring (in particular small species), sedimentology, forensic work and precision agriculture are just a few of these for which this system might be useful.

The software should be able to run the SLAM and segmentation algorithms in a laptop computer preferably with NVIDIA graphics hardware. These are of the order of €1000 to 1500 and should provide enough resources to run all the algorithms within 0.5–2 h, depending on the data size. However, for the training of new algorithms (for specific pottery types or objects), the software will need to access much larger computational resources such as cloud computing services or computing clusters. In this regard, we are considering the option to offer a computation service in a high-performance server to cover this need. The training will employ the acquired images to identify specific types of material culture or any other object of interest. The selected objects will undergo a process of DA before the training of the algorithm, and the resulting detector will be evaluated and improved through several interactions until a satisfactory result will be achieved. The resulting detector will then be incorporated into the

software to be used on the field. We aim to simplify the training process through the use of a web-based frontend equipped with interactive vector drawing tools.

6.3. Discussion

The new implementations of the method are directed to improve our initial workflow so it can be applied to a wide range of field conditions. The ultimate objective is to develop a system that can become standard practice in archaeological survey. In this regard, many of the concerns detailed in our initial proof of concept have been at least partially addressed:

1. With the implementation of SLAM and DL algorithms, the computing resources necessary to carry out the workflow are drastically reduced, making it available for almost real-time on-site analysis without the need of an Internet connection to access cloud computing services or hours of computing.
2. The new drone has an extended flight time and can carry a high-quality camera, which will allow it to fly at higher altitudes and faster, covering a much larger area per photograph and therefore increasing the distance between flight lines that will also significantly reduce the length of the flight necessary to cover the same space. This will increase considerably the area covered in each flight. The RTK GNSS connection with the base station will allow the flight to be interrupted for battery change and retrieval of photographs and continued at exactly the same position when surveying large areas.
3. The ground control system and analysis software will ease access to the method to surveyors with little experience with drones or computational methods, significantly reducing the level of expertise necessary to apply it.
4. The increase in image quality and the use of DL methods will increase significantly the detection rate and shape extraction while reducing the presence of FPs. Not being so dependent of RGB colour combinations, CNN-based approaches will also improve the detection capabilities for diverging ceramics while reducing the effect of shadows and different light conditions.

However, several of the problems noted in our initial paper are still unsolved and will remain so for some time to come:

The method is still less efficient than very intensive total count pedestrian survey as executed by experienced surveyors. Despite the increased efficiency with respect to previous RF

approaches, very small-sized fragments (<1.2 cm^2) are not so efficiently detected when using the images captured with the previous drone camera. This size will be significantly decreased when employing the new drone prototype with a higher resolution camera; however, this is something to take into account when flying the drone at higher altitudes for the fast covering of large areas when CNN-based approaches might be disadvantaged with respect to a simpler RF purposely trained.

Future drone survey campaigns, currently postponed due to the Covid-19 crisis, will increase the amount of potsherd (and other items of material culture) data available to train the algorithm. However, it will prove impossible to include in our training data all items of material culture, particularly those unique finds of higher value in account of their rarity such as figurines and written tablets.

It is also important to note that the method will not be able to detect more pottery fragments than any knowledgeable human conscious effort to record all pottery fragments in a given area. The notorious 1-s rule (Ng, 2016), although not applicable to many scenarios, offers a useful guide here. It states that anything than takes an average human less than 1 s can be automated using ML. In this regard, to identify a potsherd will certainly take a surveyor much less than 1 s, and therefore, the system is well suited to the task at hand. However, the simple fact of being in the field instead of looking at photographs offers the surveyor the possibility to improve the data in which the analysis is based. For example, if in doubt, the surveyor can always get closer to the potsherd, change the angle of vision to avoid vegetation and ultimately take the potsherd to have a closer look to its surface and fabrics to make sure it is indeed a pottery fragment. Given a certain amount of expertise, if possible, a cultural identification will be obtained. This is something that will be very difficult to achieve with RS, and, therefore, an equal performance to that of human surveyors will not be reached in the near future. The method assumes a kind of big data approach in which the quantity of data is more important than the quality of the data. The sheer quantities of a typical pedestrian survey project joined with information of location (*x*and *y* coordinates), size and shape make this a very useful approach. It will be possible for a person to do this task, and indeed, this person would do the task much better than the ML algorithm, but to geolocate and draw each fragment will cost such an amount of time that it would be impractical given the relative importance of the data gathered. In this regard, many of the opinions received after the publication of our initial paper showed concern about the possibility to substitute traditional pedestrian survey for our automated approach. This is clearly not possible as collection and laboratory-based analysis of

selected sherds in which important information, such as the pottery date or function, will always be necessary. However, this is not to say that future iterations of this method cannot eventually substitute the bulk of surface survey campaigns and leave the collection and analysis of artefacts in the hands of survey and material culture specialists rather than student workforce.

Despite the improvements designed in the drone, there will always be environments where the application of this method will not be possible, such as forested areas or areas where the lack of visibility does not allow to conduct traditional pedestrian survey.

A large hole in the future application of this method is the legal restrictions on the use of drones that different countries might have put in place. Whereas in many countries drone flight projects require of a licensed drone pilot and a detailed project describing the activity and all relevant information that needs to be approved by the competent authorities, other countries do not have legislation in place, or the use of drones for archaeology relies on the discretion of archaeological and heritage authorities. Many areas such as airport surroundings are restricted, and special permits are required to be able to fly within.

Another limitation might be the price of the drone we have developed. Even though the pricing of the prototype described in this paper is very competitive compared with similar solutions, it might be still higher than some archaeological projects can afford to pay. Although we remain convinced that such as investment would repay itself with a minimum use, new commercial drones, such as the Autel Robotics Evo 2 (in production at the time of writing) and DJI Mavic Air 2, can be considered as affordable (€850–1300) viable alternatives at the cost of reduction in flight times and image quality but still noticeably superior to the commercial solution we used in our original proof of concept (Orengo and Garcia-Molsosa, 2019). At the moment of writing the revised version of the manuscript, DJI has already launched the Matrice 300 RTK, which promises to outperform all previously discussed unmanned aerial vehicles (UAVs) and incorporates terrain-follow flight capabilities. Unfortunately, prices for the drone and cameras are still not available. It seems clear to us that drone technology is rapidly catching up with the method requirements, and there is little doubt that this will result, during the next few years, in an increase in the quality and efficiency of automated survey methods and the data they will produce.

6.4. Conclusions

Initial testing of the hardware and algorithms composing the system makes us optimistic in expecting an important boost of the method's applicability and efficiency. The higher quality camera will allow obtaining images at higher resolution and with better contrast, which will increase the number of TPs and the segmentation quality, while allowing the drone to fly higher and faster. This combined with the longer flight times and constant height will allow covering much larger areas and flight over fruit trees and irregular terrain.

The combination of SLAM and DL algorithms will bring an important step change in the applicability of the system, allowing the algorithm to be executed on the field while survey is being conducted and obtaining results in almost real time using a standard laptop computer.

Lastly, we would like to conclude this paper with a call to surveyors, computational archaeologists and software engineers for collaborative work in the development of the method. We are open to discussion, criticisms, opinions and suggestions but also to new collaborations for its development. We believe that it is important for the technique to be developed as a community effort so it can be accessed by and be useful to as many of us as possible.

6.5. Author Contributions

H.A.O., A.G.-M., I.B.-B. and J.L. wrote the original paper. H.A.O. and A.G.-M. revised the text, designed the methods, and supervised the research; I.B.-B. conducted the DL analysis; J.L. conducted the SLAM analysis; and P.A. collaborated with the data curation. All authors have read and agreed to the published version of the manuscript.

6.6. Funding

I.B.-B.’s PhD is funded with an Ayuda a Equipos de Investigación Científica of the Fundación BBVA for the Project DIASur. H.A.O. is a Ramón y Cajal Fellow (RYC-2016-19637) of the Spanish Ministry of Science, Innovation and Universities. Other funding: Marie Skłodowska-Curie Actions and Agència de Gestió d'Ajuts Universitaris i de Recerca.

DISCUSSION

7. Discussion

7.1. Achievements and Limitations

Computational archaeology and, in particular, its application to automated large-scale site detection requires, as seen in this thesis, georeferenced Big Data such as satellite imagery, and training datasets. Due to data size and processing complexity, the processing of these data required the use of online cloud-computing platforms such as GEE and Colab. The training of the DL algorithms developed in this thesis exceeded the freely available resources of both online platforms, so we had to resort to our own GPU servers, first in collaboration with the CVC and later with the development of our own computational laboratory at the ICAC. The exponential growth of computing needs and capabilities has been strongly related to the use of neural networks running on graphics processing units (GPUs), where one of its first discoverers were Krizhevsky et al. with their AlexNet network (Krizhevsky et al., 2012). GPUs, originally designed for visualization, but mainly used for gaming (and lately for cryptocurrency mining), have enabled the development of faster and more accurate ML models, such as DL. The setup of the Computational Archaeology Lab is comprised of an HPC server with the latest generation AMD EPYC Milan processors (€24,200) and several NVIDIA A40 GPUs (€5,000 each). Also, this server, due to its large RAM memory (500GBs), helps with other computational issues in archaeology such as photogrammetry (Orengo et al., 2015) and large-scale computational processes for which much RAM and many CPU cores are requiered. In the next few years, research groups and units that develop research in computational archaeology with an emphasis on DL applications will require a similar setup.

7.2. Adequate Algorithm Selection

As mentioned above, there is a wide variety of ML algorithms, all of which have previously been used in archaeology (Menze and Ur, 2012; Bicho et al., 2017; Trier et al., 2019; Mikhailova et al., 2019). Therefore, it is mandatory to analyse each task in detail to decide which is the most suitable algorithm to solve it.

First, we need to decide whether we should use remote data or ML approaches. In our case, the combination of both was necessary. Regarding RS, it is important to check which visualization

algorithm is best for our case. According to Guyot et al., who compares 13 microrelief visualization methods, each case may require a different one. Local dominance and SLRM detrending techniques worked well for small scale-terrain variation, however their limits where apparent for smaller and larger radius than that used for analysis. But the multiscale MSTP and e2MSTP, with an average of 65%, had the best performances regardless of their size or morphology (Guyot et al., 2021). So, multiscale approaches, such as the MSRM, employed in our two first papers, consistently showed better performances in CNN-based detections.

Every AI algorithm used in this thesis sought to classify a new dataset for which there were previous records manually classified by professional archaeologists. Therefore, all the ML algorithms used are supervised classification models. The second article combines a RF algorithm for pixel classification, one of the classic ML algorithms, with a CNN DL model for object detection. In relation to the use of the RF algorithm, Sentinel-2 satellite multispectral data is used for soil classification. In particular, the study area was classified in areas appropriate for the location of mounds and those that could not host tumuli. Since this approach was a pixel-based approach that allowed the use of CPU-based parallel computing, a traditional ML approach was selected. A DL algorithm could have also been used for this classification task, but since the RF's accuracy is already very high, it was not considered necessary to apply a DL algorithm that requires many more computational resources. This same paper and the last two articles, incorporate DL algorithms to detect several types of archaeological features with specific repetitive shapes (burial mounds, mounds, and potsherds, which could be defined as objects in CV terms) within several visual data sources (DTMs, maps, and drone images, respectively). Neural networks, specifically R-CNNs, make it possible to highlight the edges of those objects that are intended to be detected in the mentioned images. Given that in the case of mounds, and potsherds (3rd and 4th papers), in addition to detecting objects, we are interested in knowing their exact shape, the CNN algorithm needs to incorporate a segmentation process, in our case the employed one was Mask R-CNN. Since all burial mounds (2nd paper) are almost circular in shape, applying a masks algorithm is not necessary and an object detection algorithm is enough. In our case, the one used was a detection algorithm, in this case a YOLO algorithm.

7.3. Algorithm Combination for Refining

This thesis has also shown the importance of combining algorithms and sources of different nature, not only to detect settlements and other features of archaeological interest but also to

complement and increase the information that can be obtained from these features. It is necessary to adapt the treatment of the sources to each specific case study by combining ML with other geospatial approaches to refine the results and eliminate FPs.

In the second paper, an RF algorithm trained with Sentinel-2 multispectral imagery was used to filter unsuitable soils for burial mounds. In this way, in addition to finding circular patterns such as those of tumuli using the DL algorithm, we guarantee that those burial mound shapes that were correctly detected in unsuitable areas are discarded by the algorithm combination. Furthermore, in the third article, a global DSM (ALOS World 3D) was employed to filter out FP mound shapes corresponding to mountain peaks using a combination of slope-based classification and median convolutions. Although still not implemented, this same project aims to employ multispectral images from Sentinel-2 to further discard those mounds correctly detected with DL in the maps but which correspond to dunes. These detected dunes are made up of sandier soil, making it easier for Sentinel-2 to filter them out and to leave the potential archaeological sites as the only feature detected. Although it has not been implemented in the last article due to the good accuracy obtained, it is mentioned heat maps can be created using specialized thermal cameras to filter out possible FPs such as rocks that can be confused with pottery. Further to this thermal imagery and multispectral drone-mounted images can be used to discern other types of material culture such as lithics and metal objects. Many commercial drones allow payloads of up to 10 kg and multispectral and thermal cameras (and combinations of these) are available for many popular professional drone-models. Thus, we remark how the combination of algorithms allows not only to filter FPs but also to ensure better detection of new sites and archaeological features.

7.4. Data Engineering

This thesis introduces new concepts little used in archaeology but widely used in CV: refinement (with its positive and negative training concepts, and different methods for DA and the use of synthetic data), instance segmentation instead of semantic one (more common in archaeology with DeepLabv3+ algorithm), CL and blob analysis. All of them have been shown to improve the accuracy of detections.

In addition, developed Python-based code, which can be executed on different platforms such as Colab, has been developed for Big Data processing both for the creation of datasets

according to the requirements of the DL algorithms and for the development of DA techniques to develop training data that can generate a successful detection. Given the low amount of data in archaeology due to its own idiosyncrasy, the use of DA can be considered as almost mandatory. During this thesis, the following DA techniques adapted to each case have been developed: rotation, scaling, translation, the so-called Doppelgänger technique and the creation of synthetic data. Outside the papers presented in this PhD, others techniques for DA such as blurring, and colour jittering (Berganzo-Besga et al., 2022a) have been employed. Only by understanding the nature of archaeological data and their challenges, it is possible to select adequate DA techniques to obtain good detection accuracy.

7.5. Archaeological Approach to Validation Metrics

This thesis has showcased the potential of computational archaeology as a prospection technique for archaeological research. As such, it can yield as many site location hypotheses as a pedestrian survey would, and even more because it has non-pedestrian technologies (satellite imagery) and other ones that allow the visualisation of sites that are not visible or even no longer exist and to access hard-to-reach places like mountainous and forested landscapes. In addition, these hypothetical locations can be obtained in record time and on a large scale with very exhaustive mathematical validation thanks to ML. There is neither money nor time to survey and excavate all potential sites. This is where computational archaeology proves its worth.

The analysis of the results of our first work (Berganzo-Besga et al., 2022b) allows, without a real excavation, to determine the fortified walls of a series of Iberian sites. The lidar-based DTMs visualised using the MSRM, are clear enough to confirm that the walls that have been detected are underground in the settlement. Another issue is that excavation is still necessary to determine the age of these ruins, since we can only make hypotheses based on the data obtained and the studied archaeological documentation. Likewise, in the following two articles (Berganzo-Besga et al., 2021 and 2023) possible archaeological funerary tumuli and habitation mounds are correctly detected as a pedestrian reconnaissance would or even better, with the only difference being the time it takes to perform the work. As an example, the detection of near 10K archaeological burial mounds in Galicia, within an area of 29,574 km^2 (5.8% of Spain), only took 30 minutes using Colab. In this case, excavations would be necessary to confirm if these are real archaeological mounds and to determine their age. But as before, based

on the archaeological documentation and the data obtained, we can already draw important hypotheses about their origin. In addition, multispectral imagery allows differentiating artificial from natural tumuli in a similar way to how it would be done in situ. The latest article (Orengo et al., 2020) makes it possible to determine archaeological sites from found pottery with greater precision than humans could in a reasonable amount of time. Finally, large-scale studies allow the development of analyses that were previously impossible in terms of time, resources, and scale. This could include, for example, the distribution of past population over large territories. All of them statistically significant without excavation.

However, this computational investigation must be explicit in its conception, application and validation to avoid misinterpretation. All found sites are detected with a series of validation parameters (recall and precision) that must be correctly understood for a correct interpretation and use. Computational algorithms allow the detection of thousands of sites in a short time but they add FPs too. So, it is important, from an archaeological perspective, to have a high precision value, low FPs compared to TPs. If the precision is low, the number of detected sites does not matter because they will be hidden by many other possible sites that are, in reality FPs. This would result in a similar or even higher amount of resources being necessary for the ground validation than those required for a pedestrian survey. Therefore, it is necessary to balance the detection capacity with the presence of FPs. In our studies, we have always sought to have a precision value of at least 80%.

7.6. Machine Learning Approaches

The results obtained in this PhD clearly show that automated and ML-based approaches to the detection of archaeological features can be successfully applied in many case studies. However, most applications are still in their infancy as seen in the problems present in many archaeological applications: 1) high percentage of FPs, 2) reuse of algorithms and pre-trained networks without adapting them to specific case studies, and 3) lack of treatment for the sources employed. Therefore, this thesis has sought to offer alternatives and solutions to these common problems in archaeology.

The latest advances in ML have been related to the improvement of DA techniques given the usual lack of data, at least in the field of archaeology, and to the improvement in the algorithms' speed (from YOLOv3 to YOLOv7 in the four years the work on this thesis has been

developing). The future of AI, from a global perspective, is its arrival at the user-level, to be of practical use to users with no particular training or experience, thanks to the speed of its algorithms at the training and detection stage, and its adaptability to different situations such as DALL·E and ChatGPT (DALL·E, 2022; ChatGPT, 2022). The future of ML-based applications in archaeology will look to: 1) data generation in a way that can be machine-readable and usable as ML training data, 2) increased involvement of administrations in terms of investment in new data sources such as lidar and site protection, these methods have revealed new sites which will need new protection, 3) exploration of new sources and ML categories and methodologies, such as reinforcement learning (Angourakis et al., 2015), regression (Bicho et al., 2017) and other DL algorithms like generative adversarial networks (GANs) which holds much potential for reconstruction of ceramic pots (Navarro et al., 2022), and 4) new types of applications beyond site detection. AI algorithms are not only for site-detection, but also encompass any other archaeological methodology such as the professional interpretation of archaeologists as shown in phytolith classification (Berganzo-Besga et al., 2022a). The AI algorithms used here are used in other fields like engineering or medicine, the same algorithm used for mound detection on maps can be used for self-driving cars or to detect cancer cells in the human body (Venkadesh et al., 2021). Therefore, funding invested in this type of research holds the potential to boost the advancement of multiple fields.

7.7. Computational Algorithms as a Tool for Archaeological Research

“The Analytical Engine (...) can do whatever we know how to order it to perform”, Ada A. Lovelace found the key to the most common problems when applying computational algorithms in archaeology: it is necessary to understand how and with which data the algorithm is working to interpret correctly the results it shows. This PhD has tried to show different perspectives on how to deal with the use of computational algorithms for similar and distinct archaeological approaches.

In summary, this thesis has developed a probabilistic framework for the detection of archaeological sites. That is, the archaeologist cannot just take the results of an algorithm or detection method but: 1) it is necessary to combine approaches (not just ML-based detection) to achieve the best possible results, 2) ML needs to be employed with a probabilistic mind frame necessary to understand the process and validation in order to provide useful interpretations, it provides data that needs to be interpreted within a particular context, and 3)

ML should not substitute other types of site detection but be employed in conjunction with these, adding to the probabilities that a particular detected feature corresponds to an archaeological site.

7.8. Open Paradigm

This PhD has published its four articles in open access with open source to guarantee the accessibility, reproducibility and reusability of its research, advances and algorithms. Also, the code can be found on the online Github repository at the following link:

https://github.com/iberganzo

Nowadays, most articles are not yet open access, therefore, they have an economic barrier that prevents the knowledge of said research, its reproduction and the possible continuation or improvement by other researchers. Likewise, most of the current research in computational archaeology is focused on local studies which, despite being of great interest, prevents the reuse of its algorithms on a global scale or even in other countries. The data from our models follows the FAIR principles (GO FAIR, 2022): these data can be found in prestigious international scientific journals and Github, it is accessible due to the open access of the articles, it is interoperable because the code is written in common programming languages like Python and it is reusable in the cloud-computing platforms such as GEE and Colab for studies in other regions since they are algorithms designed with a large-scale approach.

CONCLUSIONS

8. Conclusions

8.1. General Conclusions

This thesis has shown, with the development of new methods, that computational archaeology is an incipient but powerful tool for site detection. The methods described and applied as part of this doctoral thesis allow the 1) knowledge of potential sites prior to excavation, 2) better management of future surveys and excavations, 3) reduction of cost, both economic and time, and 4) greater protection of archaeological sites still unknown. Also, it remarks the potential and importance of developing new computational archaeology laboratories in those institutions and universities with an interest on archaeological research.

8.2. A New Way to Understand and Protect Archaeological Heritage

Automated methods, and ML in particular, can provide accurate ways of locating previously unknown archaeological sites so they can be included in heritage catalogues and protected (Bewley et al. 2005). Likewise, these low-cost systems can be easily implemented in measures for the protection of regional or national heritage (Rayne et al. 2020). For this, a good knowledge of the method and the interpretation of its results is necessary but also the use of standardised validation procedures.

Likewise, an extensive catalogue of sites with different probabilities of existence should be a basic for any heritage administration, mostly taking into account that we only know a very minor percentage of the existing sites. They will also be useful and are necessary for automated monitoring methods (Conesa et al., 2023), which will have a very important impact in the future.

8.3. Beyond the Thesis Proposal, Practical Site Detections

As the methodologies developed have been applied to specific cases of ongoing projects the detection of multiple archaeological sites has been achieved improving the former knowledge and methodologies. Thanks to this thesis, the walled enclosures of six Iberian Iron Age sites in the North-Eastern of the Iberian Peninsula have been detailed, nearly 10,000 archaeological

tumuli have been detected in North-Western Iberia and nearly 6,000 potential mounds have been detected which would correspond to settlements of the Indus civilization in India and Pakistan. The algorithm developed for the detection of ceramics, which made a detailed map of the potsherds from the Abdera site, will be implemented both in Greece, Catalonia and other study areas in the Mediterranean and beyond for the detection and monitoring of archaeological sites based on the detection and analysis of surface pottery.

8.4. Beyond Sites, the Potential of Automated Methods in Archaeology

AI in archaeology, and in particular automated site detection, has grown exponentially in recent years. But archaeological sites are not the only element for which ML-based automation can be employed. Archaeology includes multiple subdisciplines in which automated procedures can have an important impact. In particular all those activities in which a large amount of individuals needs to be located, counted, measured and/or classified. However, we must not assume that these tools only allow automating processes. These ML algorithms have the ability to learn and look for patterns hidden in plain sight. That is why their full potential must be maximized, not only by automating detections, but also by helping archaeology professionals to find these patterns (visible to us through tools like Grad-CAM in DL (Selvaraju et al., 2019)). One of the clearest examples for this can be found in the automated classification of coins (Schlag and Arandjelovic´, 2017), ceramics or glass fragments (Mikhailova et al., 2019) and phytoliths (Berganzo-Besga et al., 2022a). Regarding pytholiths, previous works (Hošková et al., 2021; Díez-Pastor et al., 2020; PHYTAID, 2019; Cai and Ge, 2017) have shown potential of automated classifications but focused on single-cell phytoliths, which are relatively easy to identify for CNN-based classification algorithms given their unique and non-repetitive patterns and their nuclear shape, which allows the algorithm's convolutions use the edge shape to identify them. Nonetheless, this multi-cell method has the potential to revolutionise all these fields by allowing not just the development of much larger analytical datasets in a fraction of the time than was previously feasible but also by allowing the incorporation of new measurements and analysis methods (such as fragmentation patterns, phytolith size, etc.), assuring consistency in phytolith identification, and increasing the validity of sample analysis by moving from statistical estimations to total phytolith counts (Berganzo-Besga et al., 2022a).

8.5. Understanding and Adapting to Automatization in Archaeology

The ML algorithms developed throughout this thesis have demonstrated 1) the capacity of ML to detect large amounts of unknown archaeological sites, 2) the need to combine technologies, both between RS and ML and among themselves, for the detection of new sites and information retrieval, 3) a better comprehension of the data obtained and a better archaeological interpretation of the same thanks to the correct understanding of the validation metrics, and 4) the need to adapt the different algorithms and techniques such as DA, developing our own procedures, due to the different challenges posed by archaeology, which is very rich in its variety and heterogeneity but poor in spatial density.

The results of a process that can be automated are but the more mechanical part of archaeological research. The use of ML approaches to automate or, simply, facilitate research can be considered as an implicit substitution of human expertise. However, this PhD has demonstrated that archaeological expertise needs to be involved in all parts of the automation process. Without it the results will rarely be satisfactory and, more importantly, correct interpretations will rarely be achieved. It is necessary that the archaeological expert understand the process followed and results obtained by ML algorithms to reach conscious interpretations.

The development of these algorithms needs to be understood as a way to reduce time and funds invested in the obtention of basic research data. Perhaps the results will not be as accurate as with manual processes but these have the potential to multiply the amount of data and increase statistical quality of the data (in terms of quantity, reproducibility, and standardisation). Therefore, the use of these approaches marks a turn toward a more reproductible, data-conscious, and statistics-based research.

8.6. Future Work

All the algorithms developed can be applied to other regions of interest and can be improved by introducing new data sources such as lidar intensity values and using the latest generation DL algorithms which appeared during the development of this thesis such as EfficientDet (Tan et al., 2020). However, their most important aspect is that they raise the possibility of developing new types of analyses given the amount of consistent data obtained by the large-scale automation process or, at least, make context analyses much more statistically significant.

Ideally, archaeologist should not just be involved in the development of these algorithms but lead it. Future generations of archaeologists trained in computational methods will be able to bridge this gap and move forward the discipline with understandable and conscious automation processes. The most important part of the archaeological process, interpretation, can only be reached by trained experts that can understand the results provided by ML processes. This should provide these experts with the time and data necessary to conduct their core expertise and not the technical process. In a few years, if not sooner, similar approaches to those developed and validated in this PhD will form the basis of much new research focused on computational archaeology.

As a final remark, it is worth to remember that the basis of AI and all the computational methods which have been developed within this thesis is the domain expertise and experience, which develops the data from which these systems learn, the data we feed the algorithm to show us hidden patterns or to automate arduous tasks. Therefore, it is essential to remember the statement by Albert Einstein within On the Method of Theoretical Physics (1933): “All knowledge of reality starts from experience and ends in it”.

REFERENCES

9. References

Agapiou, A.; Alexakis, D.D.; Hadjimitsis, D.G. Potential of virtual earth observation constellations in archaeological research. *Sensors* **2019**, *19*, 4066. [Google Scholar] [CrossRef]

Aguilera, M.J.; Borderías, M.P.; González, M.P.; Santos, J.M. *Geografía General I (Geografía Física)*, 1st ed.; Editorial Universidad Nacional de Educación a Distancia: Madrid, Spain, 2009; pp. 266–271. [Google Scholar]

Angourakis, A.; Santos, J.I.; Galán, J.M.; Balbo, A.L. Food for all: An agent-based model to explore the emergence and implications of cooperation for food storage. *Environ. Archaeol.* **2015**, *20(4)*. [Google Scholar] [CrossRef]

Argyrou, A.; Agapiou, A. A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. *Remote Sens.* **2022**, *14*, 6000. [Google Scholar] [CrossRef]

Asensio, D.; Guitart, J. *El Jaciment ibèric de la Muntanya de Sant Miquel. Montornès del Vallès i Vallromanes. Recull de Documentació i assaig d'interpretació*, 1st ed.; Diputació de Barcelona: Barcelona, Spain, 2010. [Google Scholar]

Asensio, D.; Morer, J.; Rigo, A.; Sanmartí, J. Les formes d'organització social i económica a la Cossetània ibèrica: Noves dades sobre l'evolució i tipología dels assentaments entre els segles VII-I a.C. In *Territori Polític i Territori Rural Durant L'edat del Ferro a la Mediterrània Occidental. Actes de la Taula Rodona celebrada a Ullastret*, 1st ed.; Martí, A., Plana, R., Eds.; Museu d'Arqueologia de Catalunya-Ullastret: Girona, Spain, 2001; Volume 1, pp. 253–271. [Google Scholar]

Barberà, J.; Pascual, R. Burriac, un yacimiento protohistórico de la costa catalana (Cabrera de Mar, Barcelona). *Ampurias* **1980**, *41–42*, 203–242. [Google Scholar]

Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newslett.* **2004**, *6(1)*, 20-29. [Google Scholar] [CrossRef]

Belarte, M.C.; Canela, J.; Orengo, H.A.; Berganzo-Besga, I. Using Lidar to detect architectural features in urban sites in the coast of Northern Iberia (6th–3rd centuries BC). In *Urbanization in Iberia and Mediterranean Gaul in the First Millennium BC*, 1st ed.; Treballs de la

Mediterrània Antiga, 7, Belarte, M.C., Noguera, J., Plana-Mallart, R., Sanmartí, J., Eds.; ICAC: Tarragona, Spain, 2020; Volume 1, pp. 137–148. [Google Scholar] [Handle]

Berganzo-Besga, I.; Orengo, H.A.; Lumbreras, F.; Alam, A.; Gerrit, P.J.; Khan, A.; Campbell, R.; Suárez-Moreno, M.; Tomaney, J.; Roberts, R., Petrie, C. Curriculum learning-based strategy for Archaeological Mound Features detection from Historical Maps in low-density areas in India and Pakistan. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain. 2023, *manuscript to be submitted*.

Berganzo-Besga, I.; Orengo, H.A.; Lumbreras, F.; Aliende, P.; Ramsey, M.N. Automated detection and classification of multi-cell Phytoliths using Deep Learning-Based Algorithms. *J. Archaeol. Sci.* **2022a**, *148*, 105654. [Google Scholar] [CrossRef]

Berganzo-Besga, I.; Orengo, H.A.; Canela, J.; Belarte, M.C. Potential of Multitemporal Lidar for the Detection of Subtle Archaeological Features under Perennial Dense Forest. *Land* **2022b**, *11*, 1964. [Google Scholar] [CrossRef]

Berganzo-Besga, I.; Orengo, H.A.; Lumbreras, F.; Carrero-Pazos, M.; Fonte, J.; Vilas-Estévez, B. Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel-2-Based Machine Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. *Remote Sens.* **2021**, *13*, 4181. [Google Scholar] [CrossRef]

Besl, P.J.; McKay, N.D. A method for registration of 3-D shapes. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE Computer Society: Washington, DC, USA, 1992; Volume 14, pp. 239–256. [Google Scholar] [CrossRef]

Bewley, R.; Crutchley, S.; Shell, C. New light on an ancient landscape: LIDAR survey in the Stonehenge World Heritage Site. *Antiquity* **2005**, *79*, 636–647. [Google Scholar] [CrossRef]

Bicho, N.; Cascalheira, J.; Gonçalves, C. Early Upper Paleolithic colonization across Europe: Time and mode of the Gravettian diffusion. *PLoS ONE* **2017**, *12(5)*: e0178506. [Google Scholar] [CrossRef].

Bochkovskiy, A. YOLOv3. *GitHub Repository*. Available online: https://github.com/AlexeyAB/darknet (accessed on 25/01/2021).

Bourgeois, Q.P.J. Monuments on the Horizon: The Formation of the Barrow Landscape throughout the 3rd and the 2nd Millennium BCE. Ph.D. Thesis, University of Leiden, Leiden, The Netherlands, 2013. [Google Scholar]

Bu, S.; Zhao, Y.; Wan G.; Liu, Z. Map 2DFusion: Real-time incremental UAV image mosaicing based on monocular SLAM. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, 2016, 4564– 4571. [Google Scholar] [CrossRef]

Cai, Z., Ge, S. Machine learning algorithms improve the power of phytolith analysis: a case study of the tribe Oryzeae (Poaceae). J. *Systemat. Evol*. **2017**, *55(4)*, 377–384. [CrossRef] [CrossRef]

Carrero-Pazos, M. Density, intensity and clustering patterns in the spatial distribution of Galician megaliths (NW Iberian Peninsula). *Archaeol. Anthropol. Sci.* **2019**, *11*, 2097-2108. [Google Scholar] [CrossRef]

Carrero-Pazos, M. El Fenómeno Tumular y Megalítico en Galicia. Aportaciones Desde los Sistemas de Información Geográfica y la Estadística Especial para el Estudio de los Patrones de Localización. Ph.D. Thesis, University of Santiago de Compostela, Santiago de Compostela, Spain, 2017. [Google Scholar]

Carrero-Pazos, M.; Vilas-Estévez, B. The possibilities of the aerial LiDAR for the detection of Galician megalithic mounds (NW of the Iberian Peninsula). The case of Monte de Santa Mariña (Lugo). In *CAA2015. Keep the Revolution Going, Proceedings of the 43rd Annual Conference on Computer Applications and Quantitative Methods in Archaeology*, 1st ed.; Campana, S., Scopigno, R., Carpentiero, G., Eds.; Archaeopress: Oxford, UK, 2016; pp. 901–908. [Google Scholar]

Carrero-Pazos, M.; Vilas-Estévez, B.; Romaní-Fariña, E.; Rodríguez-Casal, A.A. La necropolis del Monte de Santa Mariña revisitada: Aportaciones del LIDAR aéreo para la cartografía megalítica de Galicia. *Gallaecia* **2015**, *33*, 39–57. [Google Scholar] [CrossRef]

Challis, K.; Forlin, P.; Kincey, M. A Generic Toolkit for the Visualization of Archaeological Features on Airborne LiDAR Elevation Data. *Archaeol. Prospect.* **2011**, *18*, 279–289. [Google Scholar] [CrossRef]

ChatGPT. Available online: https://chat.openai.com/chat (accessed on 16/12/2022).

Conesa, F.C.; Orengo, H.A.; Lobo, A.; Petrie, C.A. An Algorithm to Detect Endangered Cultural Heritage by Agricultural Expansion in Drylands at a Global Scale. *Remote Sens.* **2023**, *15*, 53. [Google Scholar] [CrossRef]

Conyers, L.B. Ground-Penetrating Radar Techniques to Discover and Map Historic Graves. *Hist. Archaeol.* **2006**, *40*, pp. 64–73. [Google Scholar] [CrossRef]

Crawford, O.G.S. Air Survey and Archaeology. *The Geographical Journal* 1923, 61(5), pp. 342-360. [Google Scholar] [CrossRef]

DALL·E. Available online: https://labs.openai.com (accessed on 01/12/2022).

Davis, D.S.; Gaspari, G.; Lipo, C.P.; Sanger, M.C. Deep learning reveals extent of Archaic Native American shell-ring building practices. *J. Archaeol. Sci.* **2021**, *132*, 105433. [Google Scholar] [CrossRef]

Devereux, B.J.; Amable, G.S.; Crow, P.; Cliff, A.D. The potential of airborne LiDAR for detection of archaeological features under woodland canopies. *Antiquity* **2005**, *79*, 648–660. [Google Scholar] [CrossRef]

Díez-Pastor, J.F., Latorre-Carmona, P., Arnaiz-González, Á., Ruiz-Pérez, J., Zurro, D. You Are Not My Type": An Evaluation of Classification Methods for Automatic Phytolith Identification. *Microsc. Microanal.* **2020** *26(6)*, 1158–1167. [Google Scholar] [CrossRef]

Dimensions. Available online: https://app.dimensions.ai/discover/publication (accessed on 20/12/2022).

DJI. Available online: https://www.dji.com/ (accessed on 20/12/2022).

Doneus, M.; Mandlburger, G.; Doneus, N. Archaeological Ground Point Filtering of Airborne Laser Scan Derived Point-Clouds in a Difficult Mediterranean Environment. *JCAA* **2020**, *3*, 92–108. [Google Scholar] [CrossRef]

Doneus, M. Openness as Visualization Technique for Interpretative Mapping of Airborne Lidar Derived Digital Terrain Models. *Remote Sens.* **2013**, *5*, 6427–6442. [Google Scholar] [CrossRef]

Doneus, M.; Briese, C. Full-waveform airborne laser scanning as a tool for archaeological reconnaissance. In *From Space to Place. Proceedings of the 2nd International Conference on*

Remote Sensing in Archaeology. BAR International Series 1568, 1st ed.; Campana, S., Forte, M., Eds.; Archaeopress: Oxford, UK, 2006; pp. 99–106. [Google Scholar] [Handle]

Dutta, A; Zisserman, A. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, October 21–25, **2019**, Nice, France. ACM, New York, NY, USA, 4 pages. [Google Scholar] [CrossRef]

ESA. Available online: https://www.esa.int/ (accessed on 20/12/2022).

Estrada, J. *Síntesis Arqueológica de Granollers y sus Alrededores*, 1st ed.; Museo de Granollers: Granollers, Spain, 1955. [Google Scholar]

Evans, D.; Pottier, C.; Fletcher, R.; Hensley, S.; Tapley, I.; Milne, A.; Barbetti, M. A comprehensive archaeological map of the world's largest preindustrial settlement complex at Angkor, Cambodia. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 14277–14282. [Google Scholar] [CrossRef] [PubMed]

Faniel, I.; Kansa, E.; Whitcher Kansa, S.; Barrera-Gomez, J.; Yakel, E. The challenges of digging data: A study of context in archaeological data reuse. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, USA, 22–26 July 2013; ACM: New York, NY, USA, 2013; pp. 295–304. [Google Scholar]

Fernández, A. (Coord); Hernando, A.; Maillo, J.M.; Muñoz, F.J.; Quesada, J.M.; Ripoll, S. *Prehistoria II. Las sociedades metalúrgicas*, 2nd ed.; Editorial Universitaria Ramón Areces: Madrid, Spain, 2015; pp. 665–689. [Google Scholar]

Fortó, A.; Maese, X. La Torre Roja: Un jaciment ibèric i medieval (Caldes de Montbui, Vallès Oriental; Sentmenat, Vallès Occidental). *Trib. D'arqueologia* **2010**, *2009–2010*, 113–152. [Google Scholar]

Franklin, K.; Hammer, E. Untangling palimpsest landscapes in conflict zones: A "remote survey" in Spin Boldak, Southeast Afghanistan. J. *Field Archaeol.* **2018**, *43*, 58–73. [Google Scholar] [CrossRef]

Garcia-Molsosa, A.; Orengo, H.A.; Lawrence, D.; Philip, G.; Hopper, K.; Petrie, C.A. Potential of deep learning segmentation for the extraction of archaeological features from historical map series. *Archaeol. Prospect.* **2021**, *28*, 187–199. [Google Scholar] [CrossRef]

Garcia-Molsosa, A.; Orengo, H.A.; Conesa, F.C.; Green, A.S.; Petrie, C.A. Remote Sensing and Historical Morphodynamics of Alluvial Plains. The 1909 Indus Flood and the City of Dera

Ghazi Khan (Province of Punjab, Pakistan). *Geosciences* **2019**, *9*, 21. [Google Scholar] [CrossRef]

Garcia-Molsosa, A.; Flórez Santasusana, M.; Palet Martínez, J.M. Arqueología del paisaje en el entorno de «Lauro»: Una aproximación microrregional a la construcción del territorio romano en el noreste de la Península Ibérica. *Zephyrus* **2015**, *76*, 99–119. [Google Scholar] [CrossRef]

Girshick, R. Fast r-cnn. In *2015 Proceedings of the IEEE international conference on computer vision*, 2015, 1440– 1448. [Google Scholar] [CrossRef]

Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1st ed.; IEEE: New York, NY, USA, 2014; pp. 580–587. [Google Scholar]

GIS Blog. Available online: https://paulblgis.wordpress.com/ (accessed on 20/12/2022).

Go FAIR. Available online: https://www.go-fair.org/fair-principles/ (accessed on 20/12/2022).

Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [Google Scholar] [CrossRef]

Green, A.S.; Orengo, H.A.; Alam, A.; Garcia-Molsosa, A.; Green, L.M.; Conesa, F.; Ranjan, A.; Singh, R.N.; Petrie, C.A. Re-Discovering Ancient Landscapes: Archaeological Survey of Mound Features from Historical Maps in Northwest India and Implications for Investigating the Large-Scale Distribution of Cultural Heritage Sites in South Asia. *Remote Sens.* **2019**, *11*, 2089. [Google Scholar] [CrossRef]

Gutiérrez-Antuñano, M.A.; Tiana-Alsina, J.; Rocadenbosch, F.; Sospedra, J.; Aghabi, R.; González-Marco, D. A wind-lidar buoy for offshore wind measurements: First commissioning test-phase results. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 1st ed.; IEEE: Fort Worth, TX, USA, 2017; pp. 1607–1610. [Google Scholar] [CrossRef]

Guyot, A.; Lennon, M.; Hubert-Moy, L. Objective comparison of relief visualization techniques with deep CNN for archaeology. *J. Archaeol. Sci. Rep.* **2021**, *38*, 103027. [Google Scholar] [CrossRef]

Guyot, A.; Hubert-Moy, L.; Lorho, T. Detecting Neolithic Burial Mounds from LiDAR-Derived Elevation Data Using a Multi-Scale Approach and Machine Learning Techniques. *Remote Sens.* **2018**, *10*, 225. [Google Scholar] [CrossRef] [Green Version]

Hammer, E.; Seifried, R.; Franklin, K.; Lauricellaca, A. Remote assessments of the archaeological heritage situation in Afghanistan. J. *Cult. Herit.* **2018**, *33*, 125–144. [Google Scholar] [CrossRef]

He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017, 2961– 2969. [Google Scholar] [CrossRef]

Hesse, R. LiDAR-derived Local Relief Models—A new tool for archaeological prospection. *Archaeol. Prospect.* **2010**, *17*, 67–72. [Google Scholar] [CrossRef]

Hošková, K., Pokorná, A., Neustupa, J., Pokorný, P. Inter- and intraspecific variation in grass phytolith shape and size: a geometric morphometrics perspective. *Ann. Bot.* **2021** *127(2)*, 191–201. [Google Scholar] [CrossRef]

Información Xeográfica de Galicia. Available online: http://mapas.xunta.gal/portada (accessed on 01/09/2021).

Kern, A.; Bobbe, M.; Bestmann, U. Towards a real-time aerial image mosaicing solution. In *International Micro-Air Vehicle Conference and Competition (IMAV)*, Beijing (China), 2016. [Google Scholar]

Kokalj, Ž.; Somrak, M. Why Not a Single Image? Combining Visualizations to Facilitate Fieldwork and On-Screen Mapping. *Remote Sens.* **2019**, *11*, 747. [Google Scholar] [CrossRef] [Green Version]

Kokalj, Ž.; Hesse, R. *Airborne Laser Scanning Raster Data Visualization: A Guide to Good Practice*, 1st ed.; Založba ZRC: Ljubljana, Slovenia, 2017. [Google Scholar]

Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1sd ed.; Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; 25. [Google Scholar] [CrossRef]

Landauer, J.; Hoppenstedt, B.; Allgaier, J. Image Segmentation to Locate Ancient Maya Architectures Using Deep Learning. In *Discover the Mysteries of the Maya: Selected Contributions from the Machine Learning Challenge & The Discovery Challenge Workshop at*

ECML PKDD 2021, 2nd ed.; Kocev, D., Simidjievski, N., Kostovska, A., Dimitrovski, I., Kokalj, Ž., Eds.; arXiv: Ithaca, NY, USA, 2022; arXiv:2208.03163, pp. 7–12. [Google Scholar] [CrossRef]

Landsat Science. Landsat1 Available online: https://landsat.gsfc.nasa.gov/satellites/landsat-1/ (accessed on 01/12/2022).

Liss, B.; Howland, M.D.; Levy, T.E. Testing Google Earth Engine for the automatic identification and vectorization of archaeological features: A case study from Faynan, Jordan. *J. Archaeol. Sci.* **2017**, *15*, 299–304. [Google Scholar] [CrossRef] [Green Version]

Liu, G.; Xing, J.; Xiong, J. Spatial Pyramid Block for Oracle Bone Inscription Detection. In *ICSCA 2020, Proceedings of the 2020 9th International Conference on Software and Computer Applications*, 1st ed.; Association for Computing Machinery: New York, NY, USA, 2020; pp. 133–140. [Google Scholar] [CrossRef]

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1st ed.; IEEE: New York, NY, USA, 2017; pp. 936–944. [Google Scholar] [CrossRef]

Lin, T. LabelImg. *GitHub Repository*. Available online: https://github.com/tzutalin/labelImg (accessed on 25/01/2021).

Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, 3431– 3440. [Google Scholar] [CrossRef]

Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216-231. [Google Scholar] [CrossRef]

Mayor, A. Gods and Robots: Myths, Machines, and Ancient Dreams of Technology, 1st ed.; Princeton University Press: Princeton, USA, 2018. [Google Scholar] [CrossRef]

Menze, B.H.; Ur, J.A. Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E778–E787. [Google Scholar] [CrossRef] [PubMed] [Green Version]

Mikhailova, M.; Mikhailova, E.; Grafeeva, N. The Application of Clustering Techniques to Group Archaeological Artifacts. In *New Knowledge in Information Systems and Technologies*, 1st ed.; Rocha, A., Adeli, H., Reis, L.P., Costanzo, S., Springer: Cham, Switzerland, 2019; Volume 1, pp. 50-57. [Google Scholar] [CrossRef]

Mur-Artal, R.; Tardós, J. D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. In *IEEE Transactions on Robotics*, **33**(5), 2017, 1255– 1262. [Google Scholar] [CrossRef]

Mur-Artal, R.; Montiel, J.M.M.; Tardos, J. D. ORB-SLAM: A versatile and accurate monocular SLAM system. In *IEEE Transactions on Robotics*, **31**(5), 2015, 1147– 1163. [Google Scholar] [CrossRef]

Navarro, P.; Cintas, C.; Lucena, M.; Fuertes, J.M.; Segura, R.; Delrieux, C.; González-José, R. Reconstruction of Iberian ceramic potteries using generative adversarial networks. *Sci. Rep.* **2022**, *12*, 10644. [Google Scholar] [CrossRef]

Ng, A. (2016). What artificial intelligence can and can't do right now. *Harvard Business Review*. Available online: https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now (accessed on 12/11/2020).

Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance Problems in Object Detection: A Review. *arXiv* **2020**, arXiv:1909.00169. [Google Scholar] [CrossRef]

Oliva, M. El poblado ibérico de Castell Barri. *Ampurias* **1947**, *9–10*, 288–293. [Google Scholar]

Opitz, R.; Herrmann, J. Recent Trends and Long-standing Problems in Archaeological Remote Sensing. *J. Comput. Appl. Archaeol.* **2018**, *1*, 19–41. [Google Scholar] [CrossRef]

Opitz, R.S.; Ryzewski, K.; Cherry, J.F.; Malone, B. Using airborne LiDAR Survey to explore historic-era archaeological landscapes of Montserrat in the eastern Caribbean. *J. Field Archaeol.* **2015**, *40*, 523–541. [Google Scholar] [CrossRef]

Orengo, H.A.; Garcia-Molsosa, A.; Berganzo-Besga, I.; Landauer, J.; Aliende, P.; Tres-Martínez, S. New developments in drone-based automated surface survey: Towards a functional and effective survey system. *Archaeol. Prospect.* **2021**, 1–8. [Google Scholar] [CrossRef]

Orengo, H.A.; Conesa, F.C.; Garcia-Molsosa, A.; Lobo, A.; Green, A.S.; Madella, M.; Petrie, C.A. Automated detection of archaeological mounds using machine learning classification of multi-sensor and multi-temporal satellite data. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 18240–18250. [Google Scholar] [CrossRef] [PubMed]

Orengo, H.A.; Garcia-Molsosa, A. A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery. *J. Archaeol. Sci.* **2019**, *112*, 105013. [Google Scholar] [CrossRef]

Orengo, H.A.; Petrie, C.A. Multi-scale relief model (MSRM): A new algorithm for the visualization of subtle topographic change of variable size in digital elevation models. *Earth Surf. Process. Landf.* **2018**, *43*, 1361–1369. [Google Scholar] [CrossRef] [PubMed]

Orengo, H.A.; Petrie, C.A. Large-scale, multi-temporal remote sensing of palaeo-river networks: A case study from northwest India and its implications for the Indus Civilisation. *Remote Sens.* **2017**, *9*, 735. [Google Scholar] [CrossRef]

Orengo, H.A. Open Source GIS Geospatial Software for Archaeology: Towards its Integration into Everyday Archaeological Practice. In Open Source Archaeology Ethics and Practice, 1st ed.; Wilson, A.T., Edwards, B., Eds.; Walter de Gruyter GmbH & Co KG, 2015; pp. 64–82. [Google Scholar]

Orengo, H.A.; Krahtopoulou, A.; Garcia-Molsosa, A.; Palaiochoritis, K.; Stamati, A. Photogrammetric re-discovery of the hidden long-term landscapes of western Thessaly, central Greece. *J. Archaeol. Sci.* **2015**, *64*, 100–109. [Google Scholar] [CrossRef]

Orengo, H.A.; Ejarque, A.; Albiach, R. Water management and land-use practices from the Iron-Age to the Roman period in Eastern Iberia. *J. Archaeol. Sci.* **2014**, *49*, 265–275. [Google Scholar] [CrossRef]

Parcero-Oubiña, C.; Nión, S. At Last! Remote-Sensing Discovery of Archaeological Features through Aerial Imagery and Lidar in Galician Hillforts. *AARGnews* **2022**, *64*, 22–33. [Google Scholar] [Handle]

Peréx, M.J. (Coord.) *Métodos y Técnicas de Investigación Histórica I*, 1st ed.; Editorial UNED: Madrid, Spain, 2012. pp. 153-225. [Google Scholar]

Petrie, C.A.; Orengo, H.A.; Green, A.S.; Walker, J.R.; Garcia, A.; Conesa, F.; Knox, J.R.; Singh, R.N. Mapping Archaeology While Mapping an Empire: Using Historical Maps to

Reconstruct Ancient Settlement Landscapes in Modern India and Pakistan. *Geosciences* **2019**, *9*, 11. [Google Scholar] [CrossRef]

PHYTAID, 2019. Phytolith Automatic Identification System" v.0.2. Facilitated by the University of Washington and The Burke Museum of Natural History and Culture. Published on the Internet. Available online: http://www.phytaid.burkemuseum.org/ (accessed on 30/12/2021).

Raczkowski, W. Science and/or art: aerial photographs in archaeological discourse. *Archaeologia Polona* **2001**. [Google Scholar]

Rayne, L.; Gatto, M.C.; Abdulaati, L.; Al-Haddad, M.; Sterry, M.; Sheldrick, N.; Mattingly, D. Detecting Change at Archaeological Sites in North Africa Using Open-Source Satellite Imagery. *Remote Sens.* **2020**, *12*, 3694. [Google Scholar] [CrossRef]

Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [Google Scholar] [CrossRef]

Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [Google Scholar] [CrossRef] [PubMed] [Green Version]

Riuró, F. El poblado de La Creueta. *Ampurias* **1943**, *5*, 117–131. [Google Scholar]

Rodríguez-Casal, A.A. El fenómeno tumular y megalítico en Galicia: Caracterización general, problemas y perspectivas. In *Proceedings of the International Congress on the Study of Megaliths and Other Contemporary Burials in a Social, Economic and Cultural Context*, 1st ed.; Fernández-Eraso, J., Mujika-Alustiza, J.A., Eds.; Sociedad de Ciencias Aranzadi: Donostia-San Sebastián, Spain, 2007; pp. 58–93. [Google Scholar]

Rodríguez-Del Cueto, F.; Carrero-Pazos, M. Límites y posibilidades de los análisis Lidar aplicados al megalitismo asturiano. Revisión de cuatro conjuntos tumulares prehistóricos en el concejo de Salas (España). *Veleia* **2021**, *38*, 9–31. [Google Scholar] [CrossRef]

Sanmartí, J. From the archaic states to romanization: A historical and evolutionary perspective on the Iberians. *Catalan Hist. Rev.* **2009**, *2*, 9–32. [Google Scholar] [Handle]

Schlag, I.; Arandjelovic, O. Ancient Roman Coin Recognition in the Wild Using Deep Learning Based Recognition of Artistically Depicted Face Profiles. In *Proceedings of the IEEE*

International Conference on Computer Vision (ICCV), **2017**, pp. 2898-2906. [Google Scholar] [CrossRef]

Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv* **2019**, arXiv:1610.02391. [Google Scholar] [CrossRef]

Soroush, M.; Mehrtash, A.; Khazraee, E.; Ur, J.A. Deep Learning in Archaeological Remote Sensing: Automated Qanat Detection in Kurdistan Region of Iraq. *Remote Sens.* **2020**, *12*, 500. [Google Scholar] [CrossRef] [Green Version]

Soviany, P.; Ionescu, R.T.; Rota, P.; Sebe, N. Curriculum Learning: A Survey. *arXiv* **2022**, arXiv:2101.10382. [Google Scholar] [CrossRef]

Thomas, D.C.; Kidd, F.J. On the margins: Enduring pre-modern water management strategies in and around the Registan desert, Afghanistan. J. *Field Archaeol.* **2017**, **41**, 29–42. [Google Scholar] [CrossRef]

Torres, J. *Python Deep Learning. Introducción Práctica con Keras y TensorFlow2*, 1st ed.; Marcombo: Barcelona, Spain, 2020; pp. 231–253. [Google Scholar]

Trier, Ø.D.; Reksten, J.H.; Løseth, K. Automated mapping of cultural heritage in Norway from airborne lidar data using faster R-CNN. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *95*, 102241. [Google Scholar] [CrossRef]

Trier, Ø.D.; Cowley, D.C.; Waldeland, A.U. Using deep neural networks on airborne laser scanning data: Results from a case study of semi-automatic mapping of archaeological topography on Arran, Scotland. *Archaeol. Prospect.* **2019**, *26*, 165–175. [Google Scholar] [CrossRef]

Venkadesh, K.V.; Setio, A.A.A.; Schreuder, A.; Scholten, E.T.; Chung, K.; Wille, M.M.W.; Saghir, Z.; van Ginneken, B.; Prokop, M.; Jacobs, C. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT. *Radiology*, **2021**, 30:2. [Google Scholar] [CrossRef]

Verschoof van der Vaart, W.; Bonhage, A.; Schneider, A.; Ouimet, W.; Raab, T. Automated large-scale mapping and analysis of relict charcoal hearths in Connecticut (USA) using a Deep Learning YOLOv4 framework. *Archaeol. Prospect*. **2022**, 1–16. [Google Scholar] [CrossRef]

Verschoof-van der Vaart, W.B.; Lambers, K.; Kowalczyk, W.; Bourgeois, Q.P.J. Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 293. [Google Scholar] [CrossRef]

Vilas-Estévez, B. Estudio de las Orientaciones y Emplazamientos de los Túmulos de la Necrópolis de la Serra do Laboreiro en base a la Arqueología del Paisaje y la Arqueoastronomía. Master's Thesis, University of Santiago de Compostela, Santiago de Compostela, Spain, 2015. [Google Scholar]

Waleed, A. (2017). Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. *GitHub repository*. Available online: https://github.com/matterport/Mask_RCNN (accessed on 14/07/2020).

Willey, G.R. A pattern of diffusion-acculturation. *Southwestern Journal of Anthropology*, **1953**, *9*. [Google Scholar]

Wehr, A.; Lohr, U. Airborne laser scanning—An introduction and overview. *ISPRS J. Photogramm. Remote Sens.* **1999**, *54*, 68–82. [Google Scholar] [CrossRef]

Zakšek, K.; Oštir, K.; Kokalj, Ž. Sky-View Factor as a Relief Visualization Technique. *Remote Sens.* **2011**, *3*, 398–415. [Google Scholar] [CrossRef]

ANNEXES

10. Annexes

10.1. Using Lidar to Detect Architectural Features in Urban Sites in the Coast of Northern Iberia (6th – 3rd Centuries BC). Preliminary Results

Maria Carme Belarte[1,2*]; Joan Canela[2]; Hector A. Orengo[1,3] and Iban Berganzo-Besga[3]

1. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

2. Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

3. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

*Author to whom correspondence should be addressed.

Belarte, M. C.; Noguera, J.; Plana-Mallart, R.; Sanmartí, J., *Urbanization in Iberia and Mediterranean Gaul in the first millennium BC*, *Treballs de la Mediterrània Antiga,* 7, ICAC, Tarragona, p. 137-148.; http://hdl.handle.net/2072/417717

First edition: December 2019 / Published online: 2020

Abstract

We present here the first results of an ongoing research project aimed at improving our knowledge of the urban settlements of the north-eastern Iberian Peninsula during the Iron Age. In the 4th-3rd centuries BC, and probably as early as the 6th-5th centuries BC, we detect a strongly hierarchical settlement pattern in this area. It was composed of settlement types that were differentiated by their size and function. The urban sites at the top of the hierarchy are the least known, as their excavation and study present several difficulties, such as the large areas they cover (around 10 hectares) and the fact that most of them lie under dense forest cover that obscures the archaeological remains. This last factor makes it difficult to apply certain non-invasive methods, including geophysical prospection. They are, however, suitable for study by remote sensing techniques. In this paper we discuss the efficiency of those techniques, more specifically the use of lidar data as a method of detecting architectural features in these settlements.

Keywords:

remote sensing; lidar; iron age; Iberian Peninsula; urbanism

1. Introduction

The study of the Iron Age settlements in the north-eastern Iberian Peninsula (present-day Catalonia) has achieved important results during the last four decades. As discussed elsewhere (see Sanmartí et al. in this volume), Iberian communities were organized into different political entities, each with its own territory. In what is now Catalonia, at least in the coastal areas, each territory contained settlements of different categories and with diverse functions. There were various sizes of towns, villages, fortified sites and rural settlements. They were organized under a hierarchical pattern and a proto-state structure with the main towns acting as the capitals of each territory. This organization is attested at least by the Middle Iberian period (4^{th}-3^{rd} century BC) (Asensio et al., 1998; Sanmartí, 2002 and 2004; Sanmartí, 2014, 462) and probably began during the 6^{th}-5^{th} centuries BC (Sanmartí et al., 2006, 153). Urban sites were at the top of the settlement hierarchy. We can distinguish between first-order towns, which could be as large as 10 ha or even more, and second-order towns of between 2 and 4 ha. They were typically located on hilltops and their urban area may also have extended over one or more of the hillsides, sometimes following a terraced pattern. They were usually protected by defensive walls, towers and moats (Sanmartí and Santacana ,1994; Asensio et al., 1998). In spite of their importance and interest, these large sites are the least known settlement type. Recent fieldwork has mainly focused on smaller sites and, except for the case of Ullastret (see Codina, Plana and Prado in this volume), none of the urban sites has been extensively excavated. This is because their excavation and study present several challenges: they cover large areas (about 10 ha) and most have overlapping occupations from the Roman, medieval and modern periods. As for the possibilities of non-destructive survey methods, most of these sites lie under a dense Mediterranean forest cover consisting of evergreens (mostly pine) and dense bushes that obscure the archaeological remains and hinder attempts to carry out geophysical surveys. Given the characteristics of Iberian towns, RS techniques appear to be suitable for their study, as they allow large areas of land to be analysed. Among the different RS possibilities, we decided to carry out a study using lidar (Light Detection and Ranging) data. The object of this was two-fold: firstly, to gain information about the main building features and urban planning of these

settlements, and secondly to test the suitability of the methodology. In this paper we present the first results of the study.

2. Methodology

The methodology for the study of the selected settlements was, as stated above, the use of RS techniques, more specifically lidar data. This method is based on the emission of laser light and the measurement of the reflected pulses to generate a point cloud representing the earth's surface (DSM). The model includes every feature of the relief, including the vegetation. We can then select the elements of the topography, filtering out the vegetation cover, and obtain a high resolution DTM. This allows small reliefs created by anthropic action to be highlighted through specific filters, the use of GIS and specific software. Lidar technology has been used in archaeology since the beginning of the century (Bewley and Raczkowski, 2002). It has been applied to various types of archaeological site presenting very different challenges. It has been used to study archaeological sites in forest areas, such as the First Iron Age settlement of Purbach in Austria (Doneus et al., 2008) and the prehistoric settlement of Welshbury Hill in Gloucestershire, England (Devereux et al., 2005). It has also been applied to landscape studies in Germany, specifically at Rastatt (Sittler 2004), and the Mayan site of Caracol in Belize (Chase et al., 2011). This technology has also proved to be efficient at sites not obscured by vegetation, for example, the high-resolution DTMs that highlight archaeological microtopography, such as in the area around Stonehenge, England (Bewley, Crutchley and Shell, 2005). Lidar-derived DTMs are today commonly used as a data source for the automated detection of cultural heritage elements (Davis, Lipo and Sanger, 2019; Schneider et al., 2015; Trier, Zortea and Tonning, 2015). On the Iberian Peninsula, lidar technology came later to archaeological studies. It is worth highlighting the research carried out to locate the Roman military camps in the north-west (Menéndez et al., 2017; Fonte and Costa-García, 2017), as well as the modern-era fortifications on the Miño River in Galicia and Portugal (Blanco-Rotea et al., 2016). It has also been used in more specific cases, such as the First Iron Age settlement of Irueña in Fuenteguinaldo, Salamanca province (Berrocal-Rangel et al., 2017), or to describe hydraulic engineering techniques in Roman mining works in north-western Spain (Fernández-Lozano, Gutiérrez-Alonso and Fernández-Morán, 2015).

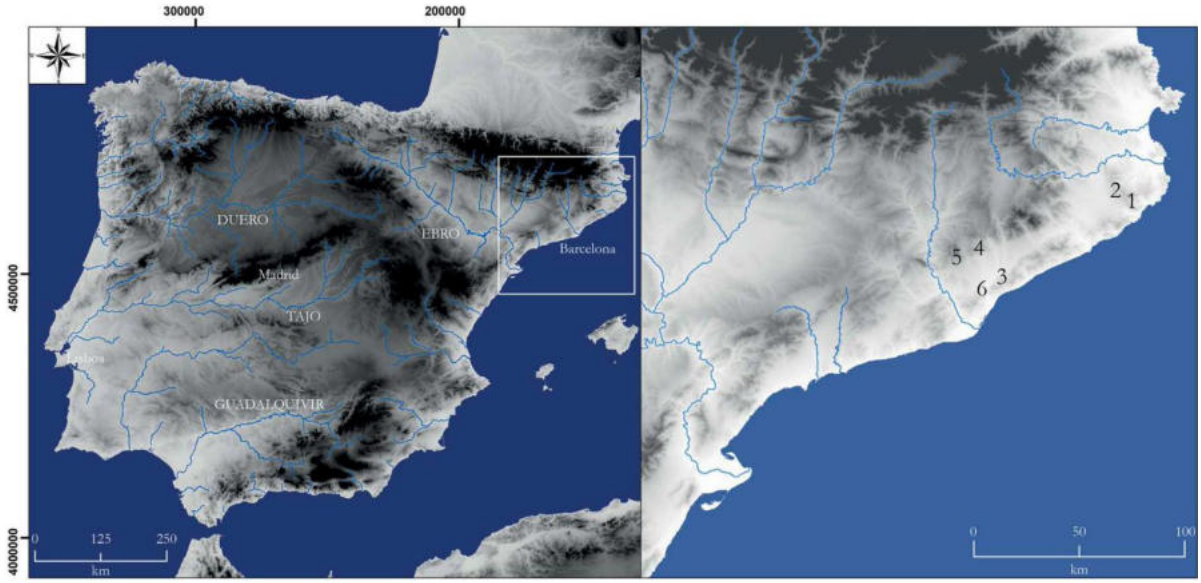


Figure 1. Study area within the Iberian Peninsula and the locations of the studied sites: 1) Castell Barri (Calonge, Girona); 2) Puig d'en Rovira (La Creueta de Quart, Girona); 3) Burriac (Cabrera de Mar, Barcelona); 4) Puig Castell (Cànoves i Samalús, Barcelona); 5) La Torre Roja (Caldes de Montbui, Barcelona); 6) Sant Miquel (Montornès/Vallromanes, Barcelona). Map background: DTM SRTM 90 m (CGIAR-CSI).

For this research we selected six settlements in which only limited fieldwork has been carried out to date (**Figure 1**). They all lie under dense forest cover (**Figure 2**). Our research was carried out in two phases and the results were of differing quality. Firstly, we worked with the available resolution data from the Catalan Cartographic and Geographic Institute (ICGC) lidar data, with a medium resolution of 2p/m^2 (unclassified). We obtained satisfactory results for only two sites: Puig Castell (Cànoves i Samalús) and Burriac (Cabrera de Mar). In order to increase the number of settlements with positive results, we attempted a second research phase with the acquisition of higher resolution lidar data through a new survey (5 unclassified points per m^2), also from the ICGC, but carried out specifically for the project. Our hypothesis was that a greater number of points per m^2 would allow some of these to penetrate the vegetation cover and facilitate more ground returns, giving us more information about the weakest topographic anomalies that could be signs of smaller archaeological structures. The results obtained in the first phase were not uniform, as the density of the vegetation cover did not allow enough ground returns to filter it out and generate a DTM with sufficient resolution. Likewise, at some sites the points classification provided by the ICGC may have led to errors, such as identifying possible topographical anomalies in the understory as low vegetation. In these

cases, where the LAS files are used to isolate ground returns and produce DTMs, those topographic anomalies that may indicate the presence of archaeological structures in the subsoil are not visible. We processed the lidar files using QGIS 2.12.0- Lyon and LAStools. More specifically, we worked with the module lasground tool, which allows you to extract and work only with the points corresponding to the terrain. These data were later refined and, after examining several possibilities, we observed that the best results were obtained by establishing "nature" as the type of land and "fine" as the pre-processing. The resulting .las file, which only corresponded to the points classified as "ground", was transformed into a DTM using the lasAdem tool. A hillshade was later applied to the resulting DTM, allowing initial verification of the quality of the DTM and a first review of the presence or absence of topographical anomalies that could be interpreted as archaeological.

3. Case studies: location and previous research

The six settlements of our sample are located in the northern and central coastal regions of Catalonia and correspond to the ancient Iberian territories of the Indiketes (in the northern area, organised around the town of Ullastret) and the Laeetani (in the central area, organised around Burriac). Burriac is a first-order settlement, while the other five correspond to smaller urban settlements, considered as second-order towns. From north to south, the site of Castell Barri (Calonge, Baix Empordà, Girona) is located on a hill at 297 m a.s.l., very near the sea and in a densely built-up area (**Figure 2.1**). Very little is known about it. In the mid-20th century, Miquel Oliva (1947) visited it and described the remains of the western and north-western stretches of a defensive wall 1.80-1.90 m wide that would have delimited the town, although he did not record any visible architectural remains inside the wall. No plan or photographs of the architectural features accompanied this description. Oliva also collected some archaeological surface finds that suggest a possible dating of between the 4th and the 3rd centuries BC (Oliva, 1947, 291). The site has not been excavated, except for a small area in the 1990s (Caravaca et al., 1996). More recently, Aurora Martin described it as a lookout point to monitor the surrounding territory (Martin, 2005, 333). We interpret it as a probable secondary town of the Indiketan polity. We selected a second settlement in the same area, Puig d'en Rovira (La Creueta de Quart, Girona) (**Figure 2.2**). Located on a wooded hill at 150 m a.s.l., it is about 30 kilometres from the coast, near the modern city of Girona and therefore in a densely built-up area. Part of the site was destroyed by a modern quarry. The settlement was occupied at least from the late 4th century BC (Martin 2005, 328). Excavations carried out in the 1940s revealed

part of the defensive wall, the remains of habitation walls, some of them superimposed and corresponding to different phases, and abundant archaeological finds (Riuró 1943, 119). All these data suggest that this settlement would have been either a small town or a third-order settlement. The other four sites in our sample are in the Laeetanian area. Among them, Burriac (Cabrera de Mar, Maresme, Barcelona) is a first-order settlement, probably the capital of the Laeetani (**Figure 2.3**). It is located on the southern slope of a hill on the Catalan coast at a maximum height of some 400 m a.s.l. It has been proposed that the occupied area covered approximately 10 ha (Zamora 2007, 325), whereas the remaining habitation sites in this region do not exceed 4 ha. The archaeological excavations at the site were carried out mainly between the 1930s and 1980s (Barberà and Pascual, 1979-1980; Benito et al., 1982-1983 and 1985; Garcia and Zamora, 1994; 1986; Ribas, 1931, 1952 and 1964; Ribas and Lladó, 1977-1978). However, only a small part of the fortified enclosure is known and there is practically no information regarding the internal urban layout. We only know of the existence of several rows of houses organised on terraces at different levels, following a wellknown pattern of urbanism in the Iberian culture. A larger construction attached to the eastern wall was initially interpreted as a public building (Barberà and Pascual, 1979-1980, 212-222), although more recently as part of a large dwelling (Zamora, 2007). The fieldwork allowed the site to be dated to between the 6th and the 1st centuries BC. The centrality of this town is also confirmed by the existence in its surroundings of three funerary areas dating from the 3rd century BC: Turó dels dos Pins, Can Rodon and Can Ros (Zamora, 2007). This is significant given that the few necropolises documented in the study area are always associated with the first-order towns (see Sanmartí et al. in this volume). Finally, groups of rural buildings and silos are located in the town's peri-urban area (Zamora, 2012), an aspect also characteristic of the main settlements. A second site in this territory is Puig Castell (Cànoves i Samalús, Vallès Oriental, Barcelona) (**Figure 2.4**). It is located in the Catalan Pre-Coastal Mountains, approximately 30 km from the coast and built on a hilltop at 631 m a.s.l. Discovered by Josep Estrada in the nineteen-fifties (Estrada, 1955; Estrada and Villaronga, 1967), it was studied by Arnau Garcia-Molsosa (2015) for his PhD. It is in fact the best-known settlement in our selection, thanks to excavations carried out in recent years by Marc Guàrdia. These have brought to light a large part of the rampart, a structure two metres in width built in the 4th century BC, as well as part of the towers in the same defensive system. The archaeological site is located in the Montseny Natural Park where deforestation is not permitted, making open-area excavations very complicated (**Figure 3**). As in other cases, practically no fieldwork has been undertaken in the interior of the town, and its urban layout is therefore barely known.

It covers an area of approximately 4 ha and would have been an important town in the Laeetani territory (Garcia-Molsosa, 2015; Guàrdia, 2015 and 2016). Still in this area, Torre Roja (Caldes de Montbui, Vallès Oriental, Barcelona) is also located in the Catalan Pre-Coastal Mountains and is built on an upper slope 375-400 m a.s.l. (**Figure 2.5**). The site was occupied from the late 6th century to the mediaeval era. Recent excavations have revealed a habitation area dated to the Iberian period (5th-4th to 1st centuries BC), with several buildings initially laid out in regular rows and separated by a central street, as well as later modifications that transformed the urban layout into a more complex system. The defensive wall has also been partially identified (Fortó and Maese, 2009-2010).

However, the available information is insufficient to define the total area of the settlement. The last site is Sant Miquel (Vallromanes/Montornès, Vallès Oriental, Barcelona) (**Figure 2.6**). Located on an inland hill at 410 m a.s.l. in the Catalan Coastal Range, this Iberian settlement occupied the top and two slopes of the hill. Limited fieldwork was carried out in the late 1960s when the remains of the rear wall and domestic buildings were discovered, as well as finds from the 3rd century BC (Barberà and Pascual, 1969-1970), although little is known about the urban planning of the settlement. Although no later excavations have been undertaken, a recent review of the archaeological finds allows us to date the occupation period of the settlement from at least the second half of the 5th century to the end of the 3rd century or the beginning of the 2nd century BC (Asensio and Guitart, 2010, 70). An updated topography of the settlement has allowed an area of between 1.5 and 2.5 ha to be proposed (Asensio and Guitart, 2010, 63). It probably corresponds to a second order town.

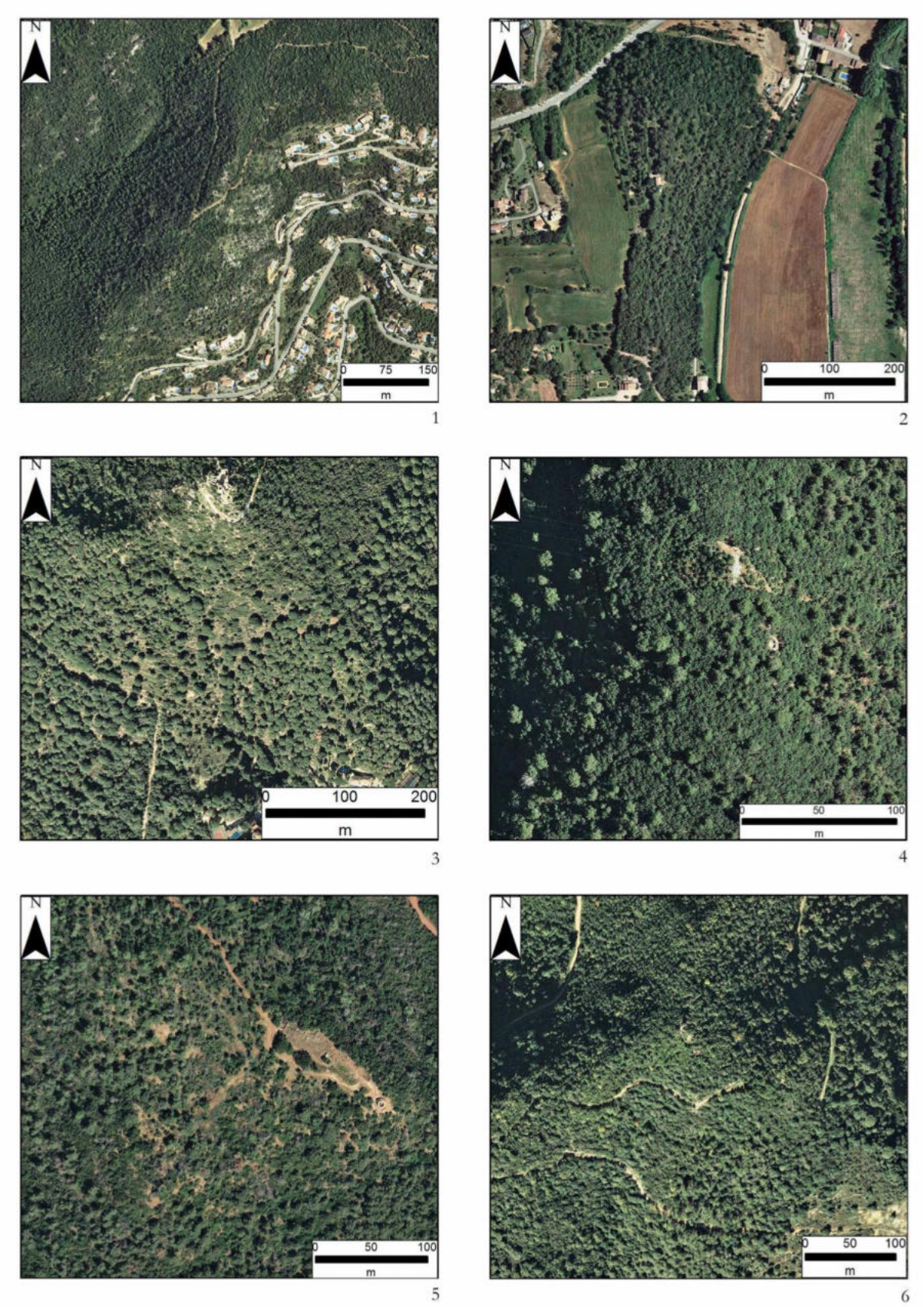


Figure 2. An aerial perspective that allows us to appreciate that the analysed sites are covered by the forest. 1) Castell Barri (Calonge, Girona); 2) Puig d'en Rovira (La Creueta de Quart, Girona); 3) Burriac

(Cabrera de Mar, Barcelona); 4) Puig Castell (Cànoves i Samalús, Barcelona); 5) La Torre Roja (Caldes de Montbui, Barcelona); 6) Sant Miquel (Montornès/Vallromanes, Barcelona). Map background: Orthophotomap 1:2.500 ICGC.

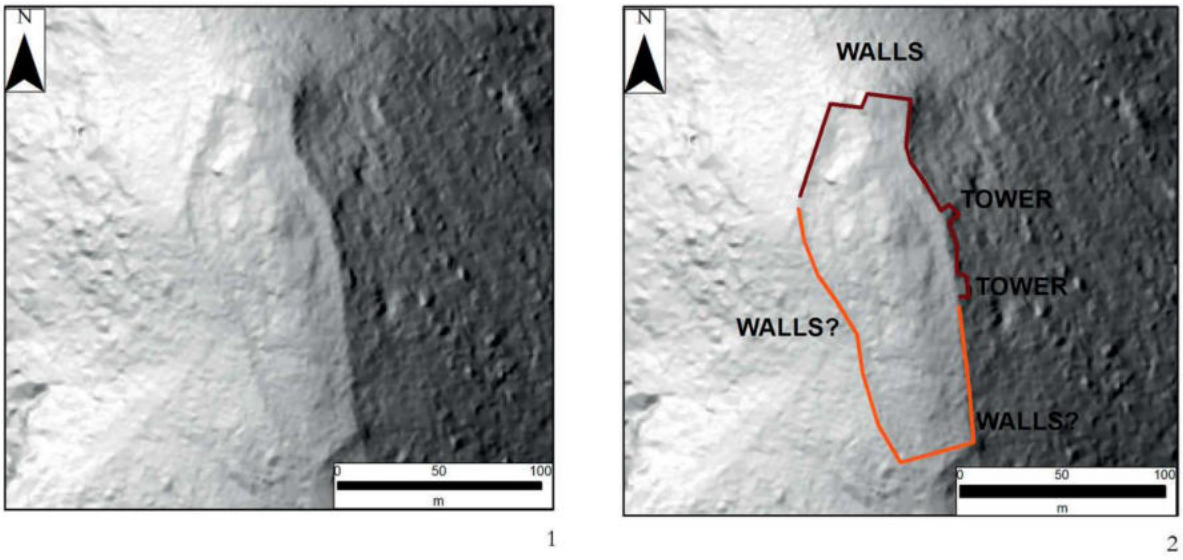


Figure 3. Puig Castell: The DTM shows the entire fortified enclosure of the site (wall and several towers). Map background: Orthophotomap/Lidar ICGC.

4. Results and discussion

The best results were obtained at Puig Castell (Cànoves i Samalús), where almost the whole fortified enclosure was detected, as well as some of the defensive towers. Its general outline coincides with that obtained from the excavations (Guàrdia, 2015, 54, **Figure 2**). However, we did not identify any topographical anomalies inside the ramparts (**Figure 4**).

At Burriac (Cabrera de Mar), the analysis revealed two anomalies to the east and west of the site that could correspond to two stretches of the wall. No anomalies were observed inside the site's enclosure and there is no trace of the southern stretch of the fortification (**Figure 5**). During the second phase, the most satisfactory results continued to be from Burriac and Puig Castell. The analysis of the rest of the sites gave minor results, with the exception of Torre Roja (**Figure 8**), where the rampart surrounding the settlement perimeter was also detected.

Figure 4. Detail of the rampart of Puig Castell), after Guàrdia, 2015, 55.

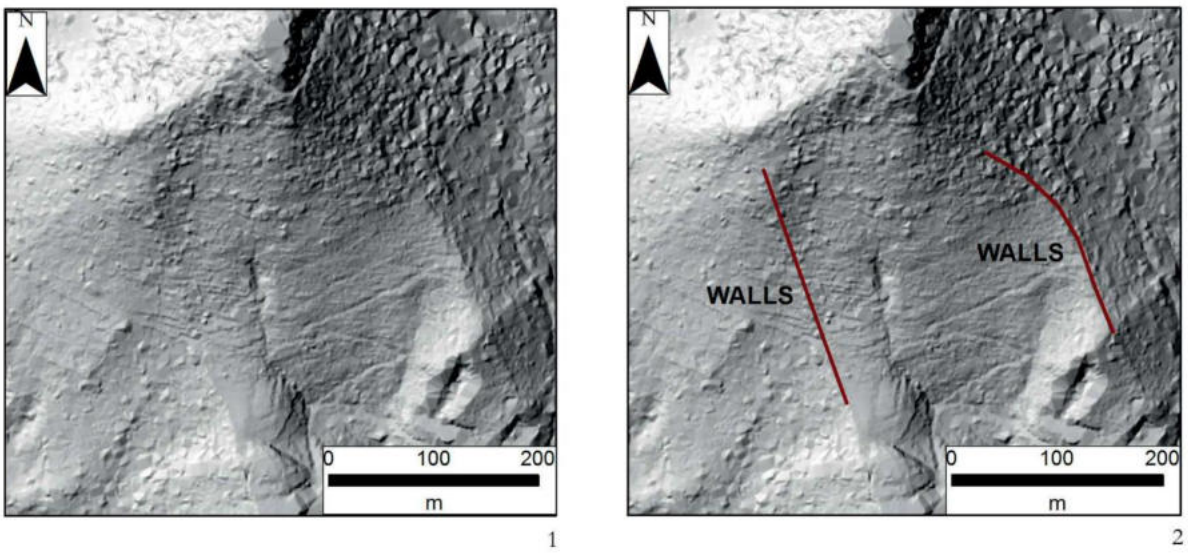


Figure 5. Burriac: The DTM allows us to appreciate the eastern and western stretches of the wall. Map background: Orthophotomap/Lidar ICGC.

In the case of Castell Barri (Calonge), a possible anomaly was detected on the upper part of the southern hill slope; this could correspond to part of the eastern, southern and western stretches of the town wall, with a total length of 283 m. They define a partial area of 1 ha, although the settlement probably extended northwards and thus undoubtedly occupied a larger area (**Figure 6**).

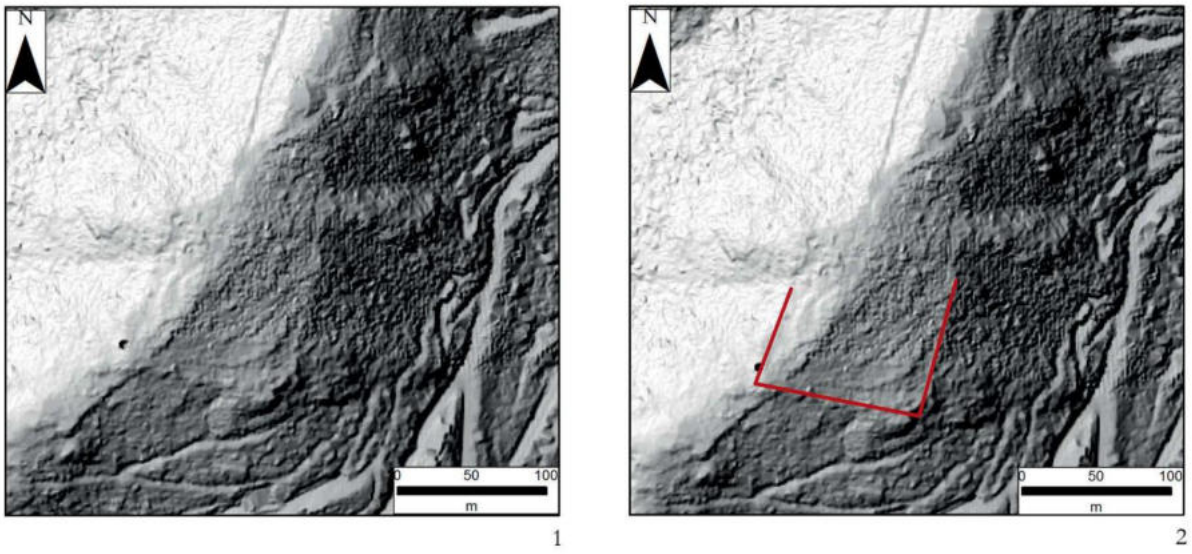


Figure 6. Castellbarri: The analysis allows us to locate a part of the hypothetical fortified enclosure, specifically part of the eastern, southern and western stretches of the wall. Map background: Orthophotomap/Lidar ICGC.

In the analysis of Puig d'en Rovira (La Creueta, Quart), it can be clearly observed that the hill has been considerably affected by a modern quarry and has marked terracing on the eastern slope. On the northern slope, however, there are anomalies that could be of an archaeological nature, corresponding more precisely to a possible 126-metre stretch of wall that follows the relief of the hill (**Figure 7**).

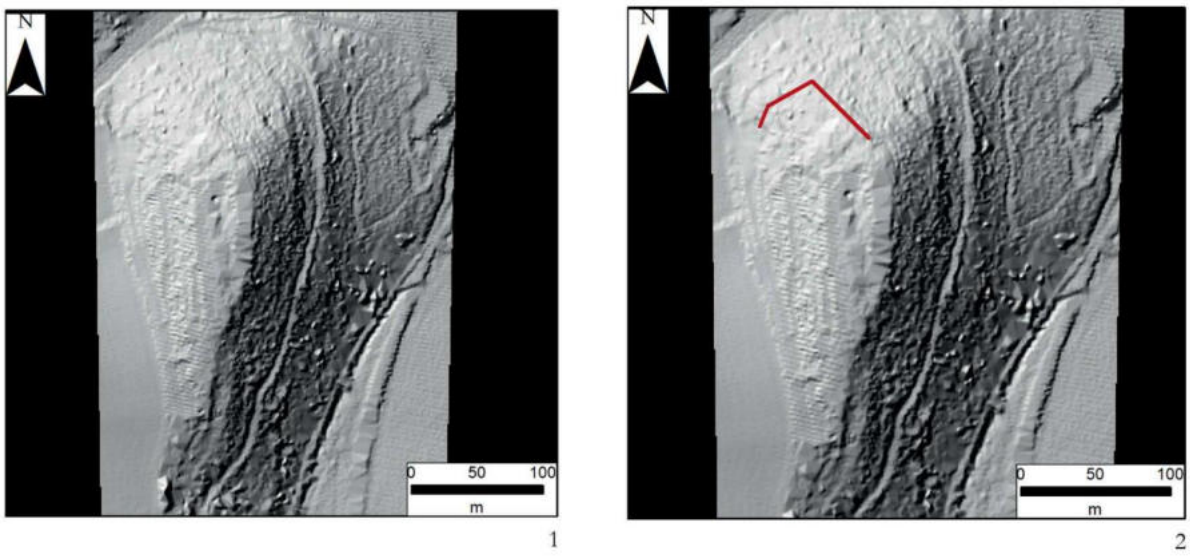


Figure 7. La Creueta: Observation of DTM allows to follow part of the possible layout of the wall to the north of the hill. Map background: Orthophotomap/Lidar ICGC.

In the case of Torre Roja (Caldes de Montbui, Barcelona), most of the perimeter wall of the settlement appears to be defined. On the upper part of the hill, as well as on its south-eastern slope, several topographic anomalies are detected that appear to mark the northern, southern and eastern stretches of the wall; they have a total length of 676 m, giving a possible size for the settlement of around 3 ha (**Figure 8**).

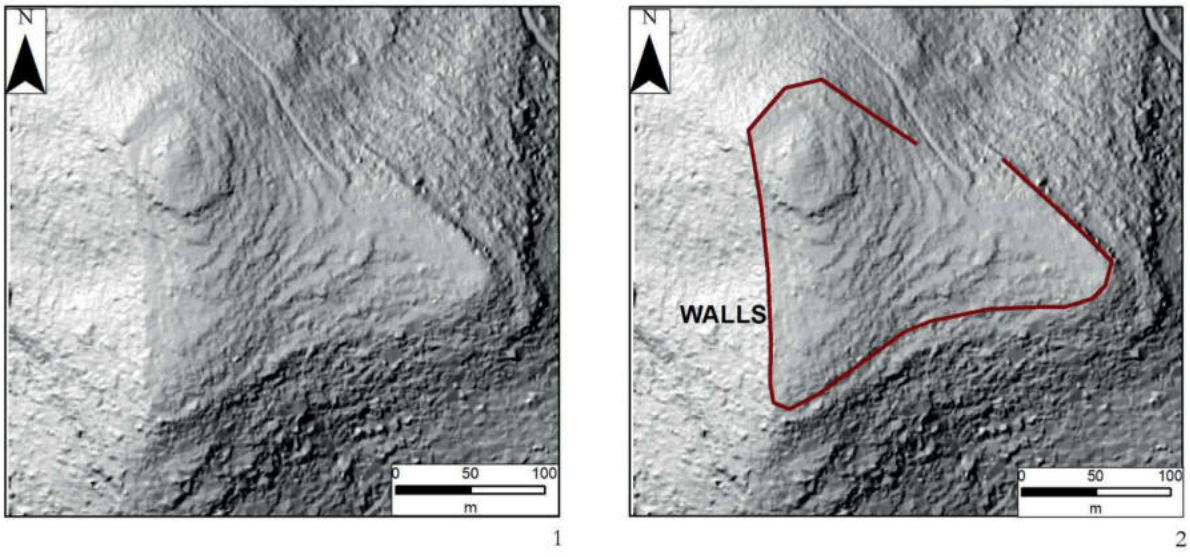


Figure 8. Torre Roja: The lidar data processing allows us to locate almost all the probable layout of the wall. Map background: Orthophotomap/Lidar ICGC.

Finally, the study of Sant Miquel (Montornès/ Vallromanes) presents greater difficulty. The dense vegetation cover on the eastern, northern and western slopes makes it difficult to record the points corresponding to the understory. Nevertheless, two possible anomalies can be identified on the southern slope of the hill, probably corresponding to stretches of the defensive wall of 120 m and 70 m respectively. However, there is also a mediaeval fortification on this site (known as Sant Miquel castle) and it cannot be ruled out that the elements identified are related to this later construction (**Figure 9**).

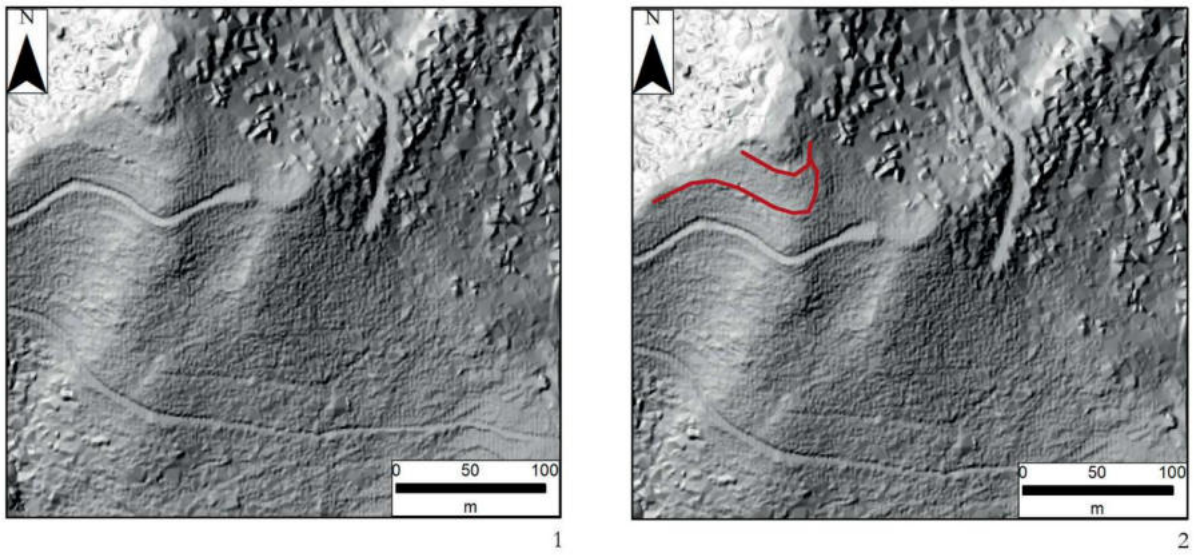


Figure 9. Sant Miquel de Vallromanes: Two topographical anomalies can be identified just south of the hilltop, although they could also be related to the medieval fortification. Map background: Orthophotomap/ Lidar ICGC.

By and large, the structures detected through lidar coincide with those identified by archaeological excavations, where they have been carried out. The latter are sometimes still visible on site, although in the case of earlier excavations (Sant Miquel de Vallromanes, Puig d'en Rovira or Burriac) they are not discernible due to the vegetation.

5. Conclusions and perspectives

Preliminary results show that this technology is useful for the study of large areas with dense vegetation, at least for the identification of large defensive structures such as towers, walls and

moats. This is a major step forward if we take into account that for most of the cases under study we do not even know the extent of the sites. The perimeter walls detected in our survey surrounded settlements with areas of between 2 and 4 hectares, a figure that corresponds to the proposed size of second-order sites. An exception would be Burriac, the only first-order nucleus in the sample, which is much larger, probably reaching around 10 ha. To date, however, this method has not allowed us to clearly identify smaller structures, such as house walls. With the preliminary results of the second phase we have been able to confirm that a higher resolution of points per m^2 allows us to obtain incipient results at sites that had not previously revealed possible archaeological structures. Despite this, it has not been possible to improve the results already obtained for Burriac and Puig Castell. It is entirely possible that the habitation structures within the perimeter wall do not present any topographic imprint on the surface. This will be tested during a third phase of the project by means of a high resolution topographic survey using a differential GNSS system. In some cases, it becomes apparent that the points density is not high enough to provide the ground returns necessary to produce a high-resolution DTM, given the dense tree vegetation and the undergrowth (Sant Miquel). This is particularly problematic as evergreen vegetation maintains a similar leaf density thorough the year and ground visibility conditions do not improve at any specific time. A second challenge is the steep slope across which part of these sites is built. Slopes are a challenge as they increase the illuminated ground area but not the point density, which can result in a substantial reduction in ground return density. The slopes in the area also tend to concentrate forest and shrub vegetation, which results in a further decrease in ground returns per area. Lastly, hillshading, the visualisation method currently employed, tends to produce unsatisfactory results. Our research is now entering its final phase and the aim is to achieve the highest possible quality DTMs in order to extract as much information as possible from these settlements and to define to what extent there may be other types of impediment that do not allow us to obtain more results.

6. Future work

The ICGC recently released a second version of the Catalan lidar point cloud. After initial testing, it became apparent that new version presents a similar point density to the first, although with a different distribution. This opens up new avenues for increasing the ground returns at the different sites. In the coming months we aim to correlate the two versions of the ICGC's cloud point data with that acquired specially for the project. This will involve the use

of automatic cloud point correlation algorithms. Once all ground heights from the three cloud points have been perfectly correlated, they will be classified using algorithms that can better discriminate ground returns on slopes. The classified ground returns will be used to generate a DTM through interpolation procedures. Lastly, the surface will be subjected to microtopographic analysis using a MSRM. This method of identifying microtopographic imprints on the terrain produces superior results to previous methods such as hillshading (Orengo and Petrie, 2018), but it has yet to be tested in complex topographic settings. This will be an excellent opportunity to do so.

7. Funding

This research has been funded by the Spanish Ministry of Economy and Competitiveness (Project HAR2015-67946- C2-2-P).

8. References

Asensio, D.; Belarte, M.C.; Sanmartí, J.; Santacana, J. Paisatges ibèrics. Tipus d'assentaments i formes d'ocupació del territori a la costa central de Catalunya durant el període ibèric ple. In *Los Iberos. Príncipes de Occidente*. Actas del Congreso Internacional, Fundación la Caixa, Barcelona, 1998, pp. 373-385. [Google Scholar]

Asensio, D.; Guitart, J. *El Jaciment ibèric de la Muntanya de Sant Miquel. Montornès del Vallès i Vallromanes. Recull de Documentació i assaig d'interpretació*, 1st ed.; Diputació de Barcelona: Barcelona, Spain, 2010. [Google Scholar]

Barberà, J.; Pascual, R. Burriac, un yacimiento protohistórico de la costa catalana (Cabrera de Mar, Barcelona). *Ampurias* **1980**, *41–42*, 203–242. [Google Scholar]

Barberà, J.; Pascual, R. El poblado prerromano de la Muntanya de Sant Miquel, en Vallromanes - Montornès (Barcelona). *Ampurias* **1970**, *31-32*, 273-283. [Google Scholar]

Benito, N.; Burjachs, F.; Espadaler, M.M.; Defaus, J.M.; Molina, M. Les excavacions al poblat ibèric de Burriac (Cabrera de Mar, el Maresme) durant l'any 1984. Resultats preliminars i noves dades estratigràfiques", *Tribuna d'Arqueologia* **1985**, *1984-1985*, 15-23. [Google Scholar]

Benito, N.; Burjachs, F.; Espadaler, M.M.; Defaus, J.M.; Molina, M. Les excavacions al poblat ibèric de Burriac (Cabrera de Mar, el Maresme). *Laietània* **1983**, *2-3*, 42-45. [Google Scholar]

Berrocal-Rangel, L.; Paniego, P.; Ruano, L.; Manglano, G. Aplicaciones LiDAR a la topografía arqueológica: El Castro de Irueña (Fuenteguinaldo, Salamanca). *Cuadernos de Prehistoria y Arqueología*, Universidad Autónoma de Madrid, 2017, *43*, 195-215. [Google Scholar] [CrossRef]

Bewley, R.; Crutchley, S.; Shell, C. New light on an ancient landscape: LIDAR survey in the Stonehenge World Heritage Site. *Antiquity* **2005**, *79*, 636–647. [Google Scholar] [CrossRef]

Bewley, R.H.; Raczkowski, W. Aerial Archaeology. Developing Future Practice, IOS Press: Amsterdam, Netherlands, 2002. [Google Scholar]

Blanco-Rotea, R.; Fonte, J.; Güimil-Fariña, A.; Mañana-Borrazás, P. Using airborne laser scanning and historical aerial photos to identify modern age fortifications in the Minho Valley, Northwest Iberian Peninsula. In *The three dimensions of archaeology. Proceedings of the XVII UISPP World Congress*; Kamermans, H., De Neef, W., Piccoli, C., Posluschny, A., Scopigno, R., Eds.; Archaeopress Archaeology 7, Oxford, 2016, 111-120. [Google Scholar]

Caravaca, J.; Codina, F.; Margall, J.; Prado, G. De; Punseti, D. Excavacions al poblat ibèric de Castell Barri, Calonge (Baix Empordà). In *Terceres Jornades d'Arqueologia de les Comarques Gironines*; Centre d'Investigacions Arqueològiques de Girona, Santa Coloma de Farners, 1996, 109-118. [Google Scholar] [Zenon]

Chase, A.; Chase, D.; Weishampel, J.; Drake, J.; Shrestha, R.; Slatton, C.; Awe, J.; Carter, W. Airborne LIDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. *J. Archaeol. Sci.* **2011**, *38*, 387-398. [Google Scholar] [CrossRef]

Davis, D.S.; Gaspari, G.; Lipo, C.P.; Sanger, M.C. Deep learning reveals extent of Archaic Native American shell-ring building practices. *J. Archaeol. Sci.* **2021**, *132*, 105433. [Google Scholar] [CrossRef]

Devereux, B.J.; Amable, G.S.; Crow, P.; Cliff, A.D. The potential of airborne LiDAR for detection of archaeological features under woodland canopies. *Antiquity* **2005**, *79*, 648–660. [Google Scholar] [CrossRef]

Doneus, M.; Briese, C. Archaeological prospection of forested areas using full-waveform airborne laser scanning. *J. Archaeol. Sci.* **2008**, *35*, 882-893. [Google Scholar] [CrossRef]

Estrada, J. *Síntesis Arqueológica de Granollers y sus Alrededores*, 1st ed.; Museo de Granollers: Granollers, Spain, 1955. [Google Scholar]

Estrada, J.; Villaronga, L. La Lauro monetal y el hallazgo de Cànoves (Barcelona), Diputació de Barcelona, Barcelona, 1967. [Google Scholar]

Fernández-Lozano, J.; Gutiérrez-Alonso, G.; Fernández-Morán, M. Using Airborne LiDAR Sensing Technology and Aerial Orthoimages to Unravel Roman Water Supply Systems and Gold Works in NW Spain (Eria Valley, León). *J. Archaeol. Sci.* **2015**, *53*, 356-73. [Google Scholar] [CrossRef]

Fonte, J.; Costa-García, J. Scope and limitations of airborne LiDAR technology for the detection and analysis of Roman military settlements in Northwest Iberia. In *Archaeology and Geomatics: Harvesting the benefits of 10 years of training in the Iberian Peninsula (2006-2015)*; Mayoral, V., Parcero-Oubiña, C., Fábrega-Álvarez, P., Eds Sidestone Press: Amsterdam, Netherlands, 2017, 55-71. [Google Scholar] [Handle]

Fortó, A.; Maese, X. La Torre Roja: un jaciment ibèric i medieval (Caldes de Montbui, Vallès Oriental; Sentmenat, Vallès Occidental). *Tribuna d'Arqueologia* **2010**, *2009-2010*, 113- 152. [Google Scholar]

Garcia-Molsosa, A. Arqueologia dels paisatges culturals del Massís del Montseny. Dinàmiques històriques de la prehistòria a l'edat mitjana. PhD Dissertation. Catalan Institute of Classical Archaeology / University Rovira i Virgili, Tarragona, 2015. [Google Scholar] [Handle]

Garcia, J.; Zamora, D. La Vall de Cabrera de Mar. Un model d'ocupació del territori a la Laietània ibèrica, El poblament ibèric a Catalunya (Actes), Laietània 8. Actes del seminari El poblament ibèric a Catalunya, Mataró, 1994, 147-179. [Google Scholar]

Guàrdia, M.A. les portes de Lauro: el poblat ibèric del Puig del Castell de Samalús (Cànoves i Samalús), Ponències. *Revista del Centre d'Estudis de Granollers* **2015** *19*, 73-108. [Google Scholar]

Guàrdia, M.A. Lauro y el poblado ibérico del Puig del Castell de Samalús (Cànoves i Samalús, Barcelona), hacia una nueva propuesta de la localización de la ceca. In *XV Congreso Nacional de Numismática*; Grañedo, P., Eds., Museo Arqueológico Nacional, Madrid, 2016, 863-886. [Google Scholar]

Martin, A. Territori i hàbitat al nord-est català en època ibérica. In *Món Ibèric als Països Catalans, Homenatge a Josep Barberà Farràs, XIII Col·loqui Internacional d'Arqueologia de Puigcerdà*; Mercadal, O., Eds., Institut d'Estudis Ceretans: Puigcerdà, 2005, 323-345. [Google Scholar]

Menéndez, A.; González-Álvarez, D.; Costa-García, J. M.; Fonte, J.; Gago, M.; Álvarez, M. Seguindo os passos do exército romano: uma proposta metodológica para a detecção de assentamentos militares romanos no Noroeste Peninsular. In *Genius Loci: lugares e significados / places and meanings*; Rosas, L., Sousa, A.C., Barreira, H., Eds., CITCEM – Centro de Investigação Transdisciplinar Cultura, Espaço e Memória 2: Porto, Portugal, 2017, 67-79. [Google Scholar] [Handle]

Oliva, M. El poblado ibérico de Castell Barri. *Ampurias* **1947**, *9–10*, 288–293. [Google Scholar]

Orengo, H.A.; Petrie, C.A. Multi-scale relief model (MSRM): A new algorithm for the visualization of subtle topographic change of variable size in digital elevation models. *Earth Surf. Process. Landf.* **2018**, *43*, 1361–1369. [Google Scholar] [CrossRef] [PubMed]

Ribas, M. El poblat ibèric de Burriac i la seva necròpolis, Foment d'Estudis de la Maresma, Mataró, 1931. [Google Scholar]

Ribas, M. El poblament d'Ilduro. Estudi arqueològic i topogràfic des dels temps prehistòrics fins a la destrucció d'Iluro, Societat Catalana d'Estudis Històrics, Barcelona, 1952. [Google Scholar]

Ribas, M. El poblado ibérico de Ilduro, Excavaciones Arqueológicas en España 30, Ministerio de Educación Nacional, Madrid, 1964. [Google Scholar]

Ribas, M.; Lladó, J. Excavació d'unes habitacions pre-romanes a Burriac (Cabrera de Mataró). *Pyrenae* **1978**, *13-14*, 153-180. [Google Scholar]

Riuró, F. El poblado de La Creueta. *Ampurias* **1943**, *5*, 117–131. [Google Scholar]

Sanmartí, J. Les territoires politiques et la formation des états ibériques sur la côte de Catalogne (IVe – IIIe av. J.-C.). In *Territoires celtiques. Espaces ethniques et territoires des agglomérations protohistoriques d'Europe occidentale. Actes du XXIVe colloque international de l'AFEAF*; Garcia, D.; Verdin, F., Eds., Éditions Errance: Paris, France, 2002, 30-36. [Google Scholar]

Sanmartí, J. From local groups to early states, *Pyrenae* **2004**, *35(1)*, 7-41. [Google Scholar]

Sanmartí, J. Long-term social change in Iron Age Northern Iberia (ca. 700–200 BC). In *The Cambridge Prehistory of the Bronze and Iron Age Mediterranean*; Knapp, A.B., Van Dommelen, P., Eds., Cambridge University Press: Cambridge, UK, 2014, 454-470. [Google Scholar] [CrossRef]

Sanmartí, J.; Asensio, D.; Belarte, M. C.; Martin, A.; Santacana, J. La iberització a la Catalunya costanera i central. In *De les comunitats locals als estats arcaics: la formació de les societats complexes a la costa del Mediterrani occidental. Homenatge a Miquel Cura, Actes de la III Reunió Internacional d'Arqueologia de Calafell, Arqueo Mediterrània 9*; Belarte M. C.; Sanmartí, J., Eds., Universitat de Barcelona – Institut Català d'Arqueologia Clàssica: Barcelona, 2006, 145-163. [Google Scholar]

Sanmartí, J.; Santacana, J. L'urbanisme protohistòric a la costa de Catalunya, *Cota Zero* **1994**, *10*, 27-37. [Google Scholar]

Schneider, A.; Takla, M.; Nicolay, A.; Raab, A.; Raab, T. A Template-matching Approach Combining Morphometric Variables for Automated Mapping of Charcoal Kiln Sites. *Archaeol. Prospect.* **2015**, *22*, 45-62. [Google Scholar] [CrossRef]

Sittler, B. Revealing historical landscapes by using airborne laser scanning. A 3D model of ridge and furrow in forests near Rastatt (Germany). *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **2004**, *36*, 258-262. [Google Scholar]

Trier, Ø. D.; Zortea, M.; Tonning, C. Automatic detection of mound structures in airborne laser scanning data. *J. Archaeol. Sci.* **2015**, *2*, 69-79. [Google Scholar] [CrossRef]

Zamora, D. L'oppidum de Burriac. Centre de poder polític de la Laietània ibèrica. *Laietània* **2007**, *suplement 17*, Mataró. [Google Scholar]

Zamora, D. L'espai periurbà de l'oppidum laietà de Burriac. De l'ibèric ple a la romanització. In *El paisatge periurbà a la Mediterrània Occidental durant la Protohistòria i l'Antiguitat /Le paysage périurbain en Méditerranée Occidentale pendant la Protohistoire et l'Antiquité. Actes du colloque international de Tarragone, Documenta 26*; Belarte, M. C.; Plana Mallart, R., Eds., Institut Català d'Arqueologia Clàssica: Tarragona, 2012, 149-164. [Google Scholar]

10.2. Automated Detection and Classification of Multi-Cell Phytoliths Using Deep Learning-Based Algorithms

Iban Berganzo-Besga[1]; Hector A. Orengo[1,2*]; Felipe Lumbreras[3]; Paloma Aliende[1] and Monica N. Ramsey[4,5]

1. Landscape Archaeology Research Group (GIAP), Catalan Institute of Classical Archaeology (ICAC), Pl. Rovellat s/n, 43003 Tarragona, Spain

2. Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

3. Computer Vision Center, Computer Science Deptartment, Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellaterra, Spain

4. McDonald Institute for Archaeological Research, University of Cambridge, Downing Street, Cambridge, CB2 3ER, UK

5. Department of Anthropology, Health Sciences Complex, University of Toronto Missisauga, Mississauga, ON, L5L 1C6, Canada

*Author to whom correspondence should be addressed.

J. Archaeol. Sci. **2022**, 148, 105654; https://doi.org/10.1016/j.jas.2022.105654

Received: 19 May 2022 / Revised: 27 July 2022 / Accepted: 28 July 2022 / Published: 09 October 2022

Abstract

This paper presents an algorithm for automated detection and classification of multi-cell phytoliths, one of the major components of many archaeological and paleoenvironmental deposits. This identification, based on phytolith wave pattern, is made using a pretrained VGG19 deep learning model. This approach has been tested in three key phytolith genera for the study of agricultural origins in Near East archaeology: *Avena*, *Hordeum* and *Triticum*. Also, this classification has been validated at species-level using *Triticum boeoticum* and *Avena* images. Due to the diversity of microscopes, cameras and chemical treatments that can influence images of phytolith slides, three types of data augmentation techniques have been implemented: rotation of the images at 45-degree angles, random colour and brightness

jittering, and random blur/sharpen. The implemented workflow has resulted in an overall accuracy of 93.68% for phytolith genera, improving previous attempts. The algorithm has also demonstrated its potential to automatize the classification of phytoliths species with an overall accuracy of 100%. The open code and platforms employed to develop the algorithm assure the method's accessibility, reproducibility and reusability.

Keywords:

Phytoliths; Computational archaeology; Deep learning; Machine learning; Google colaboratory

1. Introduction

The automated identification of multi-cell phytoliths has the potential to revolutionise phytolith research. Phytoliths are opaline microscopic structures made of silica, they are formed in the cellular and intercellular spaces in plants from the uptake of soluble silica (SiO2-nH2O) in the ground water. Formed as individual cells (single-cells) or as conjoined cells (multi-cells), phytoliths vary according to their anatomical origin (leaves, culm, husks, etc.), with specific forms having varying diagnostic potential (Rapp and Mulholland, 1992). Phytoliths are especially prominent in the grass family. Grass husks, composed of the glume, lemma and palea, surrounding the grain, have diagnostic multi-cell forms (Rapp and Mulholland, 1992; Rosen, 1992; Weisskopf and Lee, 2016; Ge et al., 2018 and 2019; Lu et al., 2009; Hošková et al., 2021). Phytoliths are extremely robust and are a major component of many archaeological and paleoenvironmental deposits (Piperno, 2006). However, phytolith analysis is prohibitively time consuming, leaving specialists to work with very small datasets over long spans of time. An automated detection and classification process could radically reduce work time while increasing the quantity of processed data, the quality and quantity of information retrieved (including information such as fragmentation patterns and shape and achieving total slide identification, see e.g. Strömberg, 2009 regarding minimum count issues), and assuring consistent identification across all datasets. The aim of this paper is to demonstrate that it is possible to automate phytolith identification using DL-based algorithms and, in order to do so, we present a workflow to implement these algorithms in microscope slide images.

Previous studies carried out in relation to the automated detection and classification of objects such as mounds (Berganzo-Besga et al., 2021) or potsherds (Anichini et al., 2020; Orengo et al., 2021) in images have demonstrated the potential of DL in the field of archaeology. Studies with pollen (Battiato et al., 2021) or cells (Waleed, 2017) have also shown their potential in bioarcheology. Regarding phytoliths, previous works (Hošková et al., 2021; Díez-Pastor et al., 2020; PHYTAID, 2019; Cai and Ge, 2017) have shown potential of automated classifications but focused on single-cell phytoliths, which are relatively easy to identify for CNN-based classification algorithms given their unique and non-repetitive patterns and their nuclear shape, which allows the algorithm's convolutions use the edge shape to identify them. In this article we address the classification of multi-cell phytoliths, specifically three key grass husk multi-cell types: wheat (wild einkorn, *Triticum boeoticum*, wild emmer, *Triticum dicoccoides*), barley (wild, *Hordeum spontaneum*) and oat (domestic common oat, *Avena sativa*) (**Figure 1**). Multi-cell phytoliths, particularly grass husks provide more specific genera level identifications and are therefore critical to the archaeological application of phytolith analysis. Also, given the complexity of forms that multi-cells present, and the similarity between these forms, these identifications can be time consuming and challenge even experienced phytolith analysts. The method for identifying Near East husk multi-cells was originally outlined by Rosen, 1992 where she describes the characteristics of wheat, barley, oat and rye grass according to: 1) wave pattern, the negative space between dendritic long-cells (**Figure 1B**, **b1**); 2) papillae size, shape and regularity (**Figure 1B**, **b2**); and 3) the number of pits around the papillae (**Figure 1B**, **b2**). This method is still employed for the identification of these Near Eastern grasses.

2. Materials and Methods

For both phytolith classification and detection we have a single dataset of 378 microscope images (**Figure 1**). The comparative plant material comes from M.N. Ramsey's comparative collections, including the Mount Scopus Collection, collected over four seasons by trained botanists in Israel, and the Hillman Collection, sampled from UCL's comparative material originally collected by Prof. Gordon Hillman in Turkey. The comparative material was prepared by first cleaning with distilled water to remove dust. Once dry, the samples of the different plant parts (leaf, culm, husk, awn) were separated. Organic material was burned off in a muffle furnace for 3 h at 500 ◦C, the remaining ash was then directly mounted on slides in Entellan (Weisskopf and Lee, 2016). Images for *Avena*, *Hordeum* and *Triticum* were taken

from multiple accessions (multiple slides of different plants from the same taxa), ensuring that our results are not biased by the unique characteristics of one plant or slide preparation. The images were taken on a Leica DM500 and a GXCAM-U3-5 digital camera with ToupeLite software. The resulting 2560 × 1922-pixel images dataset is composed of 121 images of *Avena sativa* (32.01% of the total), 90 of *Hordeum spontaneum* (23.81% of the total), 142 of *Triticum boeoticum* (37.57% of the total) and 25 of *Triticum dicoccoides* (6.61% of the total). This paper will use this dataset to try to classify the phytoliths visible in the images in these 3 genera.

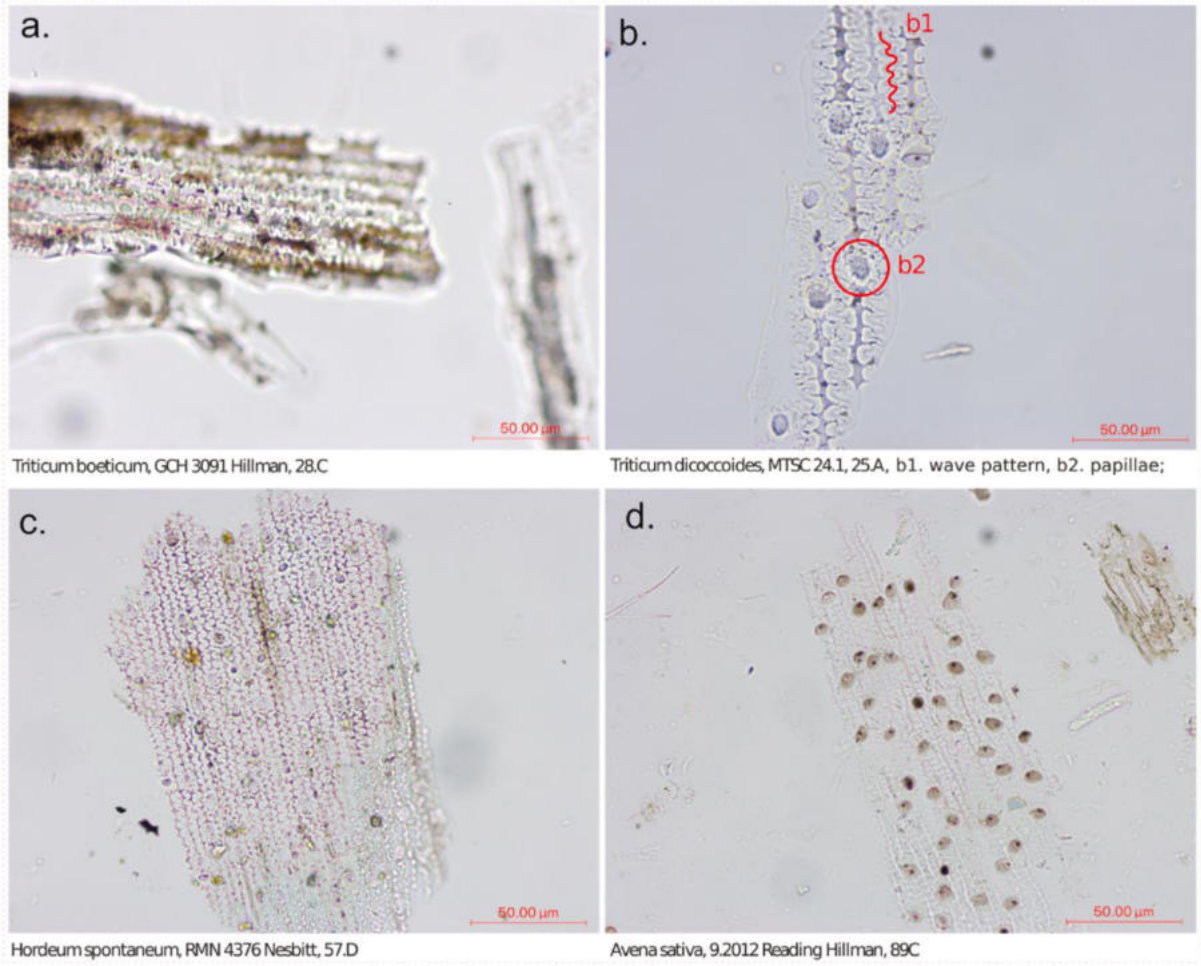


Figure 1. Selection of phytolith images from the original dataset featuring in detail some of its dendritic long-cells in wave patterns (b1) and papillae (b2).

From this initial dataset two new datasets have been created: one with 1470 phytoliths with no background (dataset 1), extracted from the original images, as many of these images included several instances of phytoliths; and other with 585 images of 256 × 256 pixels divided from the original images (dataset 2).

Five different approaches have been tested to investigate their detection capacity: 1. automated classification of phytolith genera without background; 2. automated classification of phytolith genera with background; 3. automated classification of phytolith genera in microscope images; 4. automated detection and classification of phytolith genera in microscope images, which would constitute the most practical application as they can be directly applied to digitised microscope images or directly implemented in digital microscopes; and 5. automated classification of phytolith species with background, which would show how far we can go with phytolith classification. While the previous four approaches have been aimed to classify phytolith genera, this last proof of concept aims to explore to what extent it is possible to differentiate between phytoliths species and thus assess the potential of the technique to go beyond current manual approaches to phytolith identification.

2.1. Automated classification of phytolith genera without background

In a first stage we have sought to demonstrate the possibility of classifying different types of phytoliths using DL. The elimination of the background allowed us to explore the potential of the detector to classify phytoliths when other elements, which can affect the results, are not present. For this demonstration we used dataset 1 (**Figure 2**), which includes: 490 images of *Avena sativa* (33.33% of the total), 490 of *Hordeum spontaneum* (33.33% of the total) and 490 of *Triticum boeoticum/dicoccoides* (33.33% of the total). Likewise, we have divided this dataset into 735 images for training (50.00% of the total), 369 for validation (25.10% of the total) and 366 for testing (24.90% of the total).

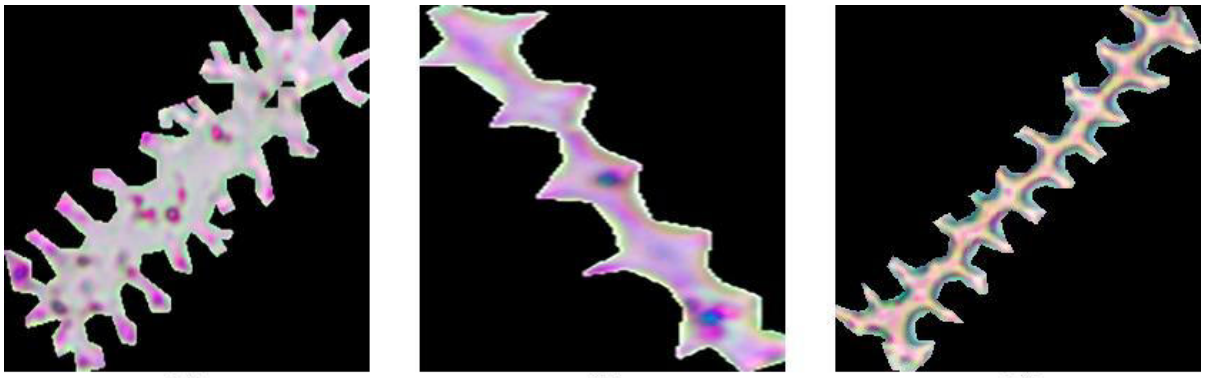


Figure 2. *Avena* (a), *Hordeum* (b) and *Triticum* (c) dendritic long-cell phytoliths from the dataset 1.

The manual classification of multi-cell grass husk phytoliths is primarily based on distinguishing the negative 'wave pattern' or the distance between their cells (Rosen, 1992). Therefore, with the aim of automating this classification, a CNN algorithm has been implemented due to their ability to extract features from images and their capacity to detect local patterns. In order to ensure correct classification based on the dendritic long-cells and not on phytolith colour (Rosen, 1992), image processing has been carried out. Likewise, with the same objective and to increase the number of phytoliths in the training, a DA process has been applied. DA seeks to multiply the training data available through a series of random transformations of the images (Torres, 2020).

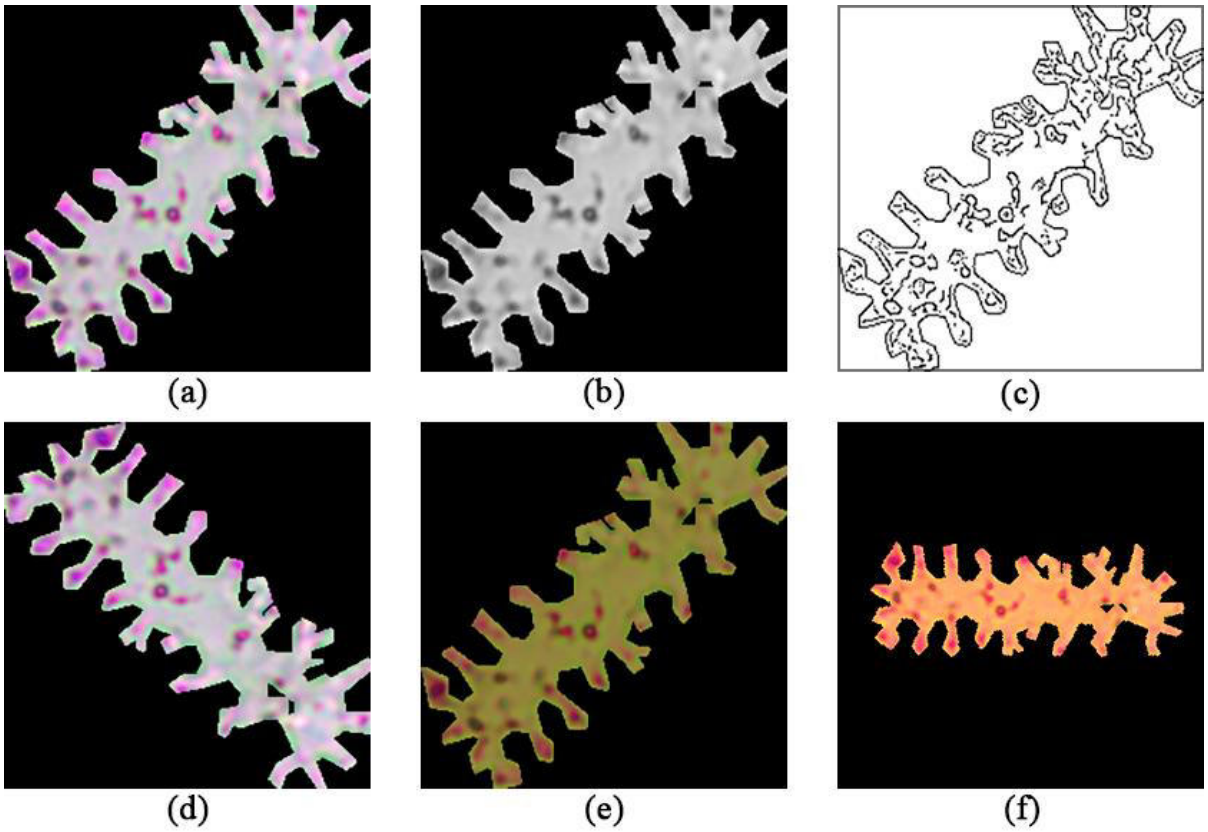


Figure 3. *Avena* phytolith image from dataset 1 processed with different image and DA techniques: RGB (a), greyscale (b), binary edges (c), RGB with DA1 (d), RGB with DA2 (e), and RGB with DA1 and DA2 (f).

In this case, we have employed two types of DA techniques: the rotation of the images at 45-degree angles (DA1), and a random colour and brightness jittering (DA2). The brightness and colour jittering were added to multiply the colour variation naturally occurring in the images and to make sure the algorithm would not use colour to differentiate between genera as this

factor can be quite variable depending on the microscope, mounting media, camera, or chemical treatment, which makes colour dependent on context and methodology. Multiplying the possible colour variations, we make sure the resulting classifier will be applicable to phytolith images acquired by other means as colour is not conditioning our analysis. This process added 58,065 new images for training. Dataset 1 was treated with several image processing techniques to check which method provided the most accurate classification: RGB with the presence and absence of the two types of DA (3 channel), greyscale (1 channel) and binary edges (1 channel) (**Figure 3**).

Four neural networks architectures have been tested in order to check their efficiency: VGG19 and ResNet50V2 both non-pretrained and pretrained with ImageNet weights. To prevent exceeding the GPU memory limit the input to the algorithm have been rescaled to 256 × 256-pixel images.

2.2. Automated classification of phytolith genera with background

Just as the classification of the phytoliths without background from the dataset 1, we tested the same classification but with the phytoliths from the dataset 2. The difference between the phytoliths in this dataset from the previous ones is that in these new images with background multiple phytoliths are shown together and many of them are cropped, as found in the microscope images (**Figure 4**).

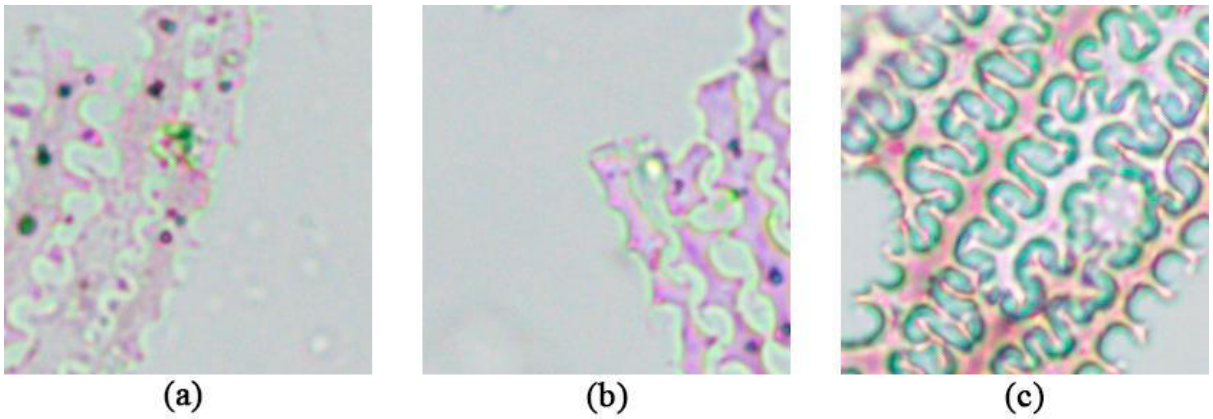


Figure 4. *Avena* (a), *Hordeum* (b) and *Triticum* (c) phytoliths from the dataset 2.

For the training of this algorithm, we used 585 images of 256 × 256 pixels with multiple phytoliths from the dataset 2: 195 of *Avena sativa* (33.33% of the total), 195 of *Hordeum*

spontaneum (33.33% of the total) and 195 of *Triticum boeoticum/dicoccoides* (33.33% of the total). Likewise, we have divided this dataset into 294 images for training (50.25% of the total), 147 for validation (25.13% of the total) and 144 for testing (24.62% of the total). The previous analysis has shown that the colour of the phytoliths is not decisive for the class to which it belongs, so no image treatment has been implemented on this occasion. However, the previous two types of DA have been maintained and a new one has been added, random Blur/Sharpen (DA3), which means adding 23,226 new images for training. This new DA is included to simulate common blurring present in microscope images, which due to the 3D nature of slides and the phytoliths within some parts of the phytoliths can be slightly out of focus (**Figure 5**). For this algorithm we also evaluated the performance of pretrained and non-pretrained VGG19 and ResNet50V2 neural network architectures.

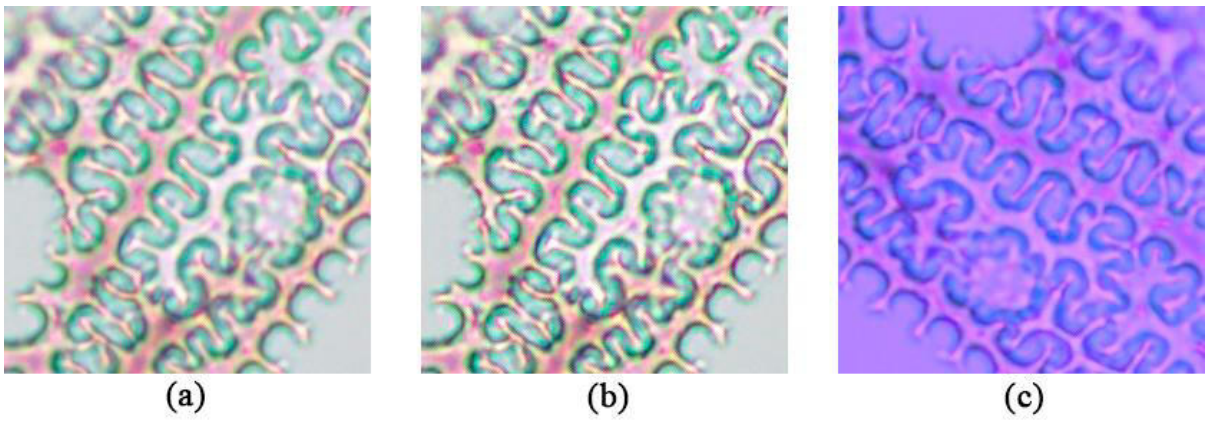


Figure 5. *Triticum* phytolith image from dataset 2 processed with different DA techniques: RGB with no DA (a), RGB with DA3 (b), and RGB with DA1, DA2 and DA3 (c).

2.3. Automated classification of phytolith genera in microscope images

To automate the identification process and avoid adding a new detection phase by background subtraction or phytolith segmentation that reduces the overall accuracy, a new algorithm has been trained that, in addition to distinguishing phytolith classes, also differentiates the background from the original image. The microscope imaging process delivers images with phytoliths but also with different artifacts due to breakage of other phytoliths but also other elements that the slide preparation has not been able to eliminate and are found surrounding the phytoliths (**Figure 6**).

Figure 6. A *Triticum* phytoliths (white outline) microscope image from the initial dataset showing some artifacts around them.

This forced us to include two extra classes in addition to those used for each type of phytolith, one for these artifacts and another one for the background. As we sought to classify *Avena*, *Hordeum* and *Triticum* type phytoliths we have trained a DL algorithm with 5 classes. For this training we add to the previous dataset 2 the two new classes obtaining a dataset of 975 images of 256 × 256 pixels with multiple phytoliths, artifacts and background (**Figure 7**): 195 of *Avena sativa* (20.00% of the total), 195 of *Hordeum spontaneum* (20.00% of the total), 195 of *Triticum boeoticum/dicoccoides* (20.00% of the total), 195 of artifacts (20.00% of the total) and 195 of background (20.00% of the total). Likewise, we have divided this dataset into 490 images for training (50.25% of the total), 245 for validation (25.13% of the total) and 240 for testing (24.62% of the total).

The objective was to classify the objects within the images in these 5 classes. The implemented neural network and DA techniques have been the same as with the previous training.

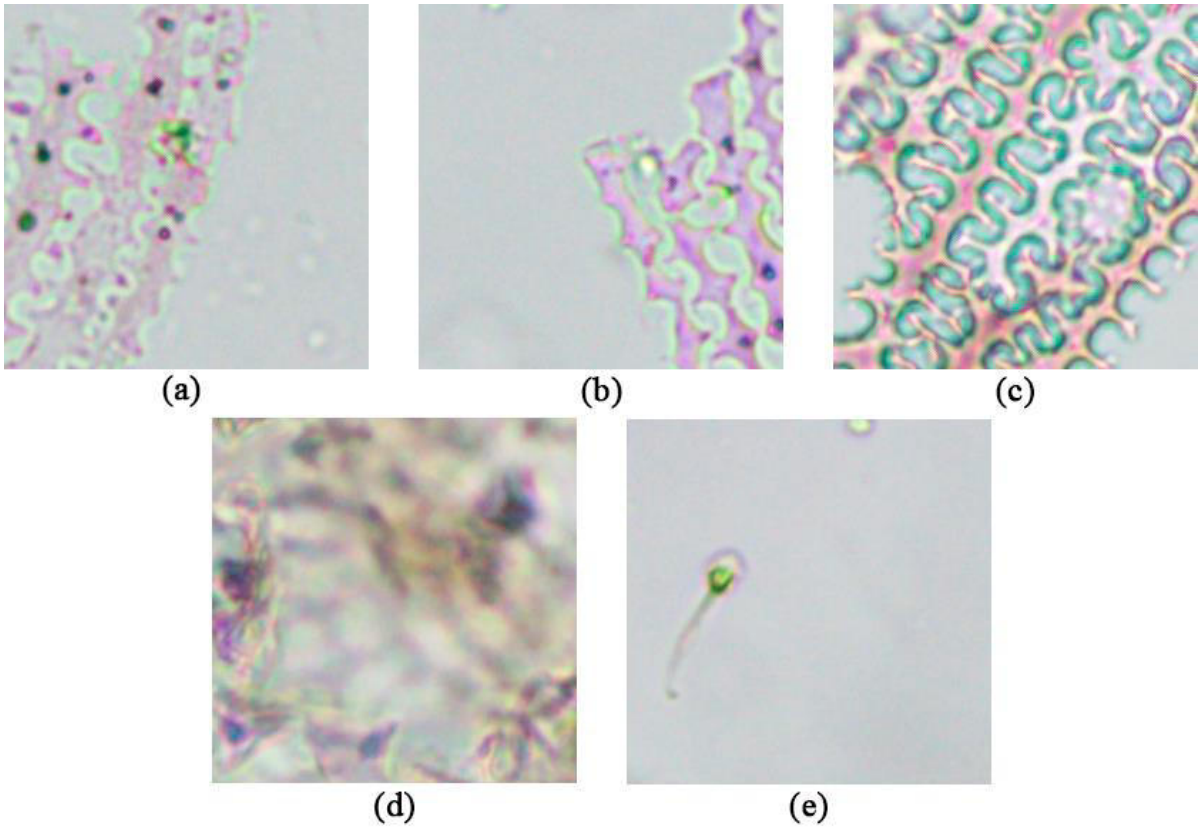


Figure 7. *Avena* (a), *Hordeum* (b), *Triticum* (c), artifacts (d) and background (e) from the dataset 2.

2.4. Automated detection and classification of phytolith genera in microscope images

Finally, we designed a workflow using the automated classification of phytolith genera in microscope images described in the previous section for the identification of images with phytoliths obtained from the microscope (**Figure 8**). In the first part, the original image will be divided into 256 × 256-pixel RGB images. Note than in the previous sections the images used for the detection were not cropped but the training data consisted of 256 × 256-pixel images with which the VGG19 model was trained. After the classification of the divided images, the class of the entire original image is determined based on the highest number of subimages detected relative to a class. For the purposes of evaluation, if a tie occurs, the general classification is considered correct if the correct class is among the tied ones. To evaluate and validate this model 190 microscope images (50.26% of the total) not used in any previous training parts have been taken from the original dataset: 64 of *Avena sativa* (33.68% of the total), 62 of *Hordeum spontaneum* (32.64% of the total) and 64 of *Triticum boeoticum/dicoccoides* (33.68% of the total).

Likewise, in the resulting image the areas with classified phytoliths are drawn to allow later statistical calculations, such as the study of the area of the identified phytoliths.

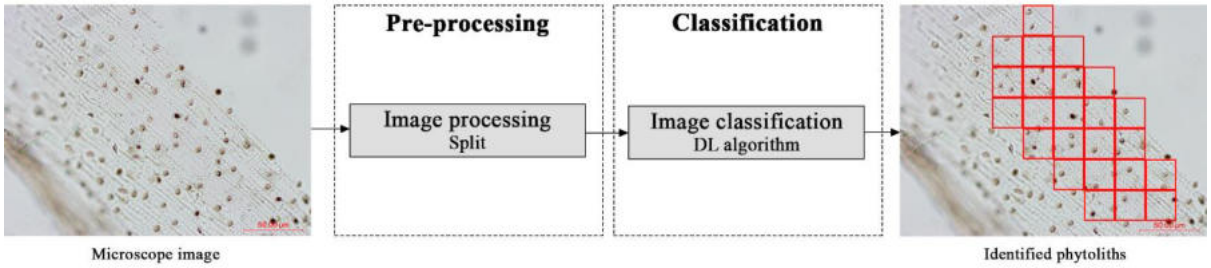


Figure 8. The implemented workflow for phytolith identification using a DL algorithm.

A new refinement stage has been included as, in this stage, the use of full images with more information could reduce the efficiency of the algorithm. The previous classification algorithm has been retrained, including new training data extracted from the original images added to dataset 2 (**Figure 9**) resulting in a total number of: 272 *Avena sativa* images (20.00% of the total), 272 *Hordeum spontaneum* images (20.00% of the total) and 272 *Triticum boeoticum/dicoccoides* images (20.00% of the total), 272 artifact images (20.00% of the total) and 272 background images (20.00% of the total). Likewise, we have divided this dataset into 875 images for training (64.34% of the total), 245 for validation (18.01% of the total) and 240 for testing (17.65% of the total).

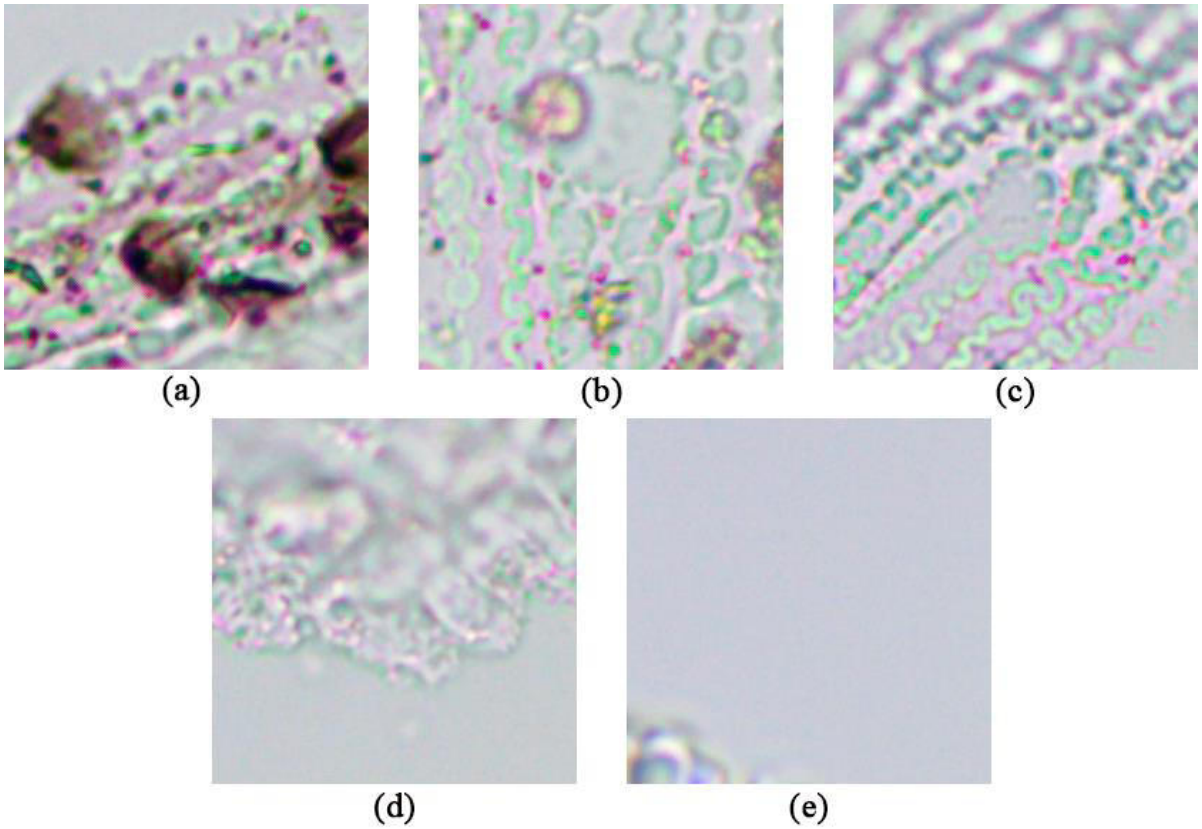


Figure 9. *Avena* (a), *Hordeum* (b), *Triticum* (c), artifacts (d), and background (e) from the dataset 2 used for refinement.

2.5. Automated classification of phytolith species with background

Once the automatic detection and classification of phytolith genera in microscopic images has been demonstrated, we have proceeded to evaluate whether this classification can be carried further, to the species level. For this, 244 new images of the *Triticum* genus extracted from the initial dataset were added to dataset 2 (**Figure 10**), obtaining 258 images of the *boeoticum* (50.00% of the total) and 258 of the *dicoccoides* (50.00% of the total) species. We have divided these species-level images into 322 for training (62.40% of the total), 98 for validation (18.99% of the total) and 96 for testing (18.61% of the total). As before, the implemented neural network has been the pre-trained VGG19, and the DA techniques have been DA1, DA2 and DA3. An Nvidia GeForce RTX 3090 graphics processing unit (GPU) with 24 GB of RAM was used to run the DL algorithms. The chosen work environment was the parallel computing platform CUDA 11.2, the ML library Tensorflow 2.1.0, the DL library cuDNN 8.1.1, the software development tool CMake 3.20.2 and the CV library OpenCV 4.5.2 as recommended for DL. Ten evaluations for each classification have been processed to assure consistent results.

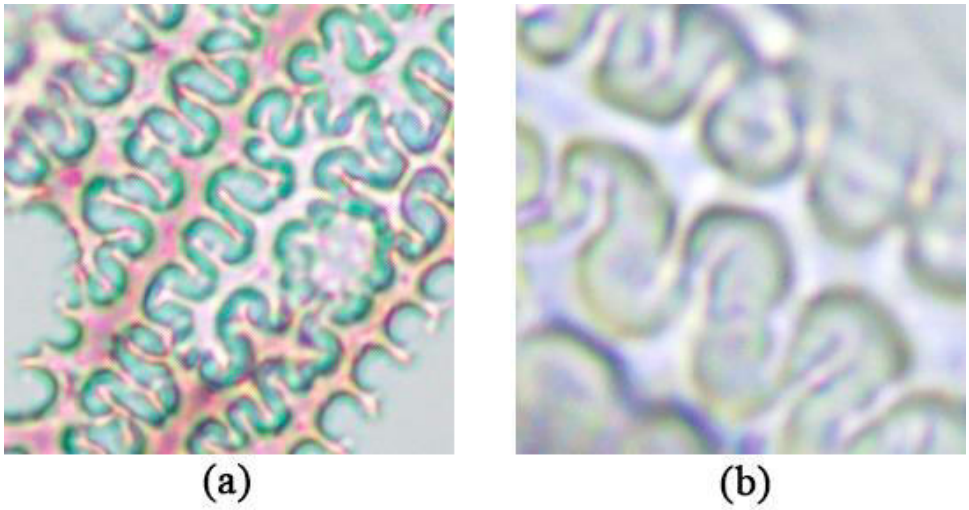


Figure 10. *Triticum boeoticum* (a) and *Triticum dicoccoides* (b) phytoliths from the dataset 2 used for species-level classification.

3. Results

3.1. Automated classification of phytolith genera without background

The results of the different techniques employed for the automated classification of phytolith genera without background are shown in **Table 1**. These show that image processing, which includes greyscaling and binarization to remove colour and intensity, reduce accuracy while

DA1, including image rotation, improve it. Also, colour jittering (DA2) has shown that the algorithm does not use colour as a decision factor in classification with similar accuracy around 89% and standard deviation (for ten evaluations) lower than 1% for DA2 and non-DA2 cases. (**Table 1**).

Table 1. Phytolith genera image processing model evaluation from dataset 1 for pretrained VGG19 architecture and test dataset.

#	Model	DA	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.
1	RGB	DA1 & 2	89.23%	0.79%	90.90%	84.02%	92.78%
2	RGB	DA2	85.63%	1.81%	84.10%	82.30%	90.49%
3	RGB	DA1	89.67%	0.85%	91.64%	85.09%	92.29%
4	RGB	No	85.93%	0.92%	85.49%	82.30%	90.00%
5	Greyscale	DA1	87.19%	0.95%	89.26%	79.67%	92.70%
6	Greyscale	No	83.74%	1.30%	81.72%	80.90%	88.60%
7	Binary	DA1	85.03%	0.92%	89.02%	77.71%	88.36%
8	Binary	No	80.41%	1.12%	76.56%	83.12%	81.56%

Between the two evaluated CNN architectures VGG19 has proved to be the best for individual phytolith classification and the use of DA has significantly improved the results. The pretrained models have better accuracy than the non-pretrained ones with the VGG19 model being the best with an overall accuracy of 89.67% for the test dataset. Also, the use of DA is a key factor (**Table 2**). All these models have been trained and evaluated with no image processing, since the previous results (**Table 1**) show the use of RGB images provide the best results.

Table 2. Phytolith genera model architecture evaluation for RGB images from dataset 1 and test dataset.

#	Model	Pretr.	DA	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.
1	VGG19	Yes	DA1	89.67%	0.85%	91.64%	85.09%	92.29%
2	VGG19	Yes	No	85.93%	0.92%	85.49%	82.30%	90.00%
3	VGG19	No	DA1	83.03%	1.99%	83.03%	83.94%	82.13%
4	VGG19	No	No	56.34%	6.98%	61.56%	65.16%	42.30%
5	ResNet50V2	Yes	DA1	88.91%	1.07%	92.07%	81.23%	93.44%
6	ResNet50V2	Yes	No	84.59%	1.10%	82.30%	81.07%	90.41%
7	ResNet50V2	No	DA1	88.06%	1.43%	86.97%	89.51%	87.79%
8	ResNet50V2	No	No	60.58%	2.10%	73.03%	64.02%	44.67%

3.2. Automated classification of phytolith genera with background

Table 3 provides an overview of the results of the tested models. All three types of DA increase the accuracy of the results, with DA1 being the most prominent. As before, colour jittering has shown that the algorithm does not use colour as a decision factor in classification. The combination of DA2 and DA3 slightly worsens the results but always within the range of the standard deviation. For this reason, together with the fact that the simulation of the microscope using the DA3 has improved the results in the rest of the cases, we have decided to continue the study with the resulting weights from the application of the three types of DA (**Table 3**).

Table 3. Phytolith genera image processing model evaluation from dataset 2 for pretrained VGG19 architecture and test dataset.

#	Model	DA	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.
1	RGB	DA1, 2 & 3	93.06%	0.98%	96.67%	85.42%	97.08%
2	RGB	DA2 & 3	87.15%	2.00%	92.50%	77.08%	91.88%
3	RGB	DA1 & 3	91.88%	1.32%	95.21%	83.13%	97.29%
4	RGB	DA1 & 2	93.82%	0.79%	96.88%	87.29%	97.29%
5	RGB	DA3	84.72%	3.73%	92.50%	72.08%	89.58%
6	RGB	DA2	87.50%	2.06%	94.58%	76.46%	91.46%
7	RGB	DA1	90.70%	0.89%	92.92%	82.08%	97.08%
8	RGB	No	84.24%	2.35%	91.04%	70.21%	91.67%

As in the previous section, the pretrained VGG19 with DA has proved to be the best model to classify three classes in multiple phytoliths images with an overall accuracy of 93.06% and a standard deviation lower than 1% (**Table 4**).

Table 4. Phytolith genera 3 class model architecture evaluation for RGB images from dataset 2 and test dataset.

#	Model	Pretr.	DA	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.
1	VGG19	Yes	DA1, 2 & 3	93.06%	0.98%	96.67%	85.42%	97.08%
2	VGG19	Yes	No	84.24%	2.35%	91.04%	70.21%	91.67%
3	VGG19	No	DA1, 2 & 3	78.89%	2.90%	87.50%	59.17%	90.00%
4	VGG19	No	No	61.39%	3.72%	57.92%	52.29%	74.17%
5	ResNet50 V2	Yes	DA1, 2 & 3	90.21%	1.30%	94.58%	84.58%	91.46%
6	ResNet50 V2	Yes	No	82.50%	2.00%	93.13%	70.63%	83.75%

7	ResNet50 V2	No	DA1, 2 & 3	45.07%	5.33%	77.71%	52.08%	5.63%
8	ResNet50 V2	No	No	68.82%	1.95%	78.96%	53.13%	74.17%

3.3. Automated classification of phytolith genera in microscope images

The best model for classifying five classes multiple phytoliths images is that incorporating all types of DA as in the previous classification tests. The overall accuracy has been 90.79% with a standard deviation lower than 1% (**Table 5**).

Table 5. Phytolith genera 5 class model architecture evaluation for RGB images from dataset 2 and test dataset.

#	DA	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.	Artifacts Acc.	Background Acc.
1	DA1, 2 & 3	90.79%	0.73%	96.46%	81.04%	95.00%	81.67%	99.79%
2	No	84.79%	1.27%	91.25%	64.58%	91.46%	79.38%	97.29%

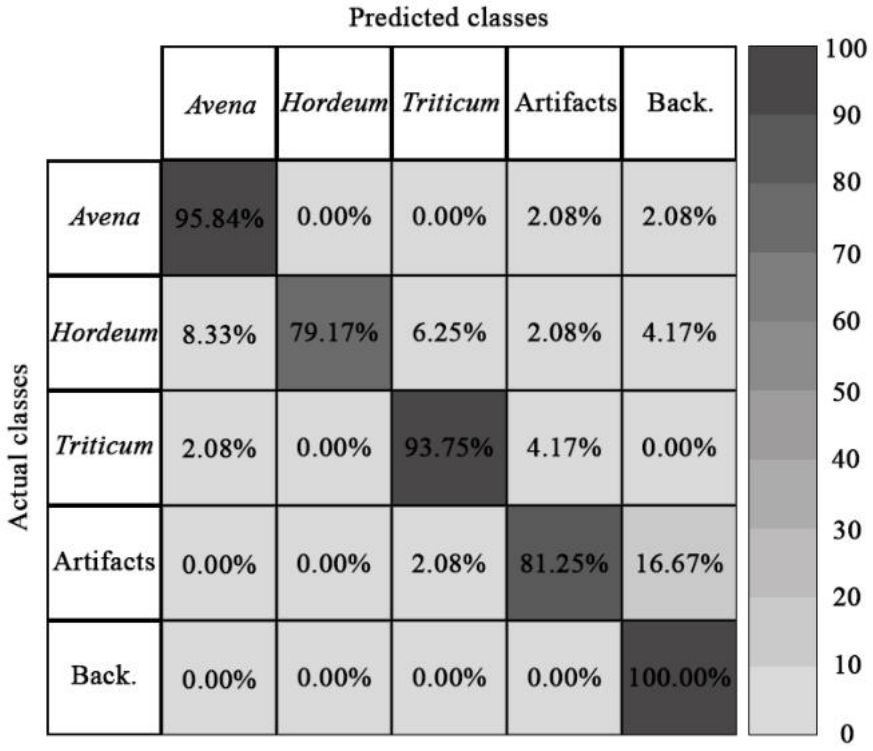


Figure 11. Phytolith genera 5 class model confusion matrix generated from the refinement best results of Table 6.

The refinement of the algorithm with new and different training data has worsened previous results (**Table 6**). This does not necessarily reflect a decrease in the quality of the classification but an increase in the diversity of the data used to train it. This is also suggested by the confusion matrix, which shows the quality of the results (**Figure 11**).

Table 6. Phytolith genera 5 class model refinement evaluation for RGB images from dataset 2, pretrained VGG19 architecture and test dataset.

#	Refin.	Overall Acc.	Overall Dev.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.	Artifacts Acc.	Back. Acc.
1	Yes	88.04%	1.14%	94.37%	75.00%	94.37%	76.46%	100.00%
2	No	90.79%	0.73%	96.46%	81.04%	95.00%	81.67%	99.79%

3.4. Automated detection and classification of phytoliths in microscope images

The implemented workflow has resulted in an overall accuracy of 93.68%, achieved after the refinement step (**Table 7**). There has only been a tie between classes in 4.21% of the images. It shows resulting images as the one from **Figure 12**.

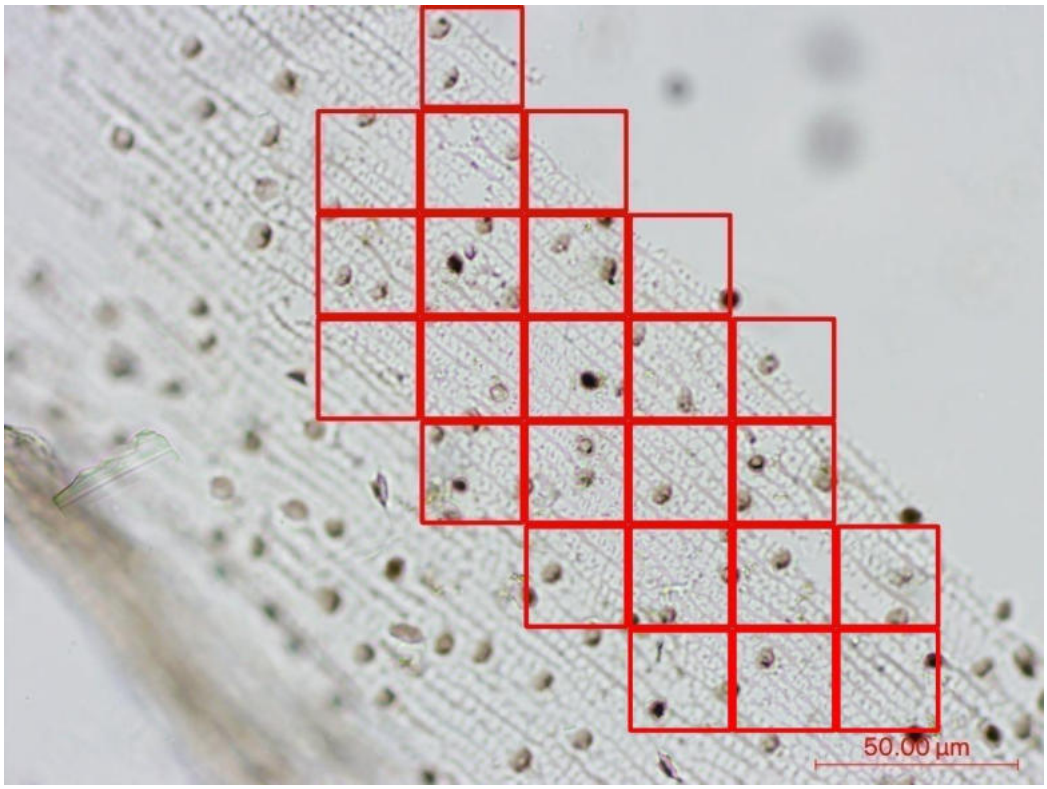


Figure 12. *Avena* phytolith (red) from the original dataset correctly classified.

Table 7. Microscope image model evaluation for the test dataset.

#	Refinement	Overall Acc.	*Avena* Acc.	*Hordeum* Acc.	*Triticum* Acc.
1	Yes	93.68%	98.44%	93.55%	89.06%
2	No	88.95%	93.75%	91.94%	81.25%

3.5. Automated classification of phytolith species with background

While previous algorithms classified phytolith genera, this is focused on the classification of phytolith species of the same genus. The results show that it is not only possible to classify phytoliths by genus but also by species. The algorithm has classified *Triticum* species with an overall accuracy of 100% (**Table 8**). The DA has remained relevant.

Table 8. *Triticum* species 2 class model architecture evaluation for RGB images from dataset 2, pretrained VGG19 architecture and test dataset.

#	DA	Overall Acc.	OverallDev.	*Triticum boeoticum* Acc.	*Triticum dicoccoides* Acc.
1	DA1, 2 & 3	100.00%	0.00%	100.00%	100.00%
2	No	94.38%	14.82%	100.00%	88.75%

4. Discussion

Díez-Pastor et al. obtained particularly satisfactory results (about 88% of accuracy using traditional ML methods) achieving their automatic classification objective for the eight classes used (Díez-Pastor et al., 2020). The same goes for the study by Cai and Ge (2017). The difference with these previous studies lies in the fact that the single-cell phytolith classes they considered were quite different from each other and individually distinctive, therefore, their classification could only be applied to distinctive phytoliths that do not pose a complex challenge for machine or expert classification, leaving the potential of DL for more complex classifications of multi-cell phytoliths unexplored. The average accuracies for their DL algorithms are below 80% in both cases with some genus less than 65%, while the number of images used per class is relatively low (about 50). These paper presents accuracies for each multi-cell genus higher than 89% for the complete microscope images and over 85% for the cropped ones, with an overall accuracy of 93% for both cases. Much better have been our results in the classification of phytoliths by species, reaching the 100% of accuracy. However,

the relatively small training dataset available for *Triticum dicoccoides* prompts us to be cautious with these results until more species-level training data are available for further testing.

Our results show that the application of multiple image DA techniques together show better results than each one separately (**Table 3**) allowing with a trained CNN model like VGG19 classify phytolith with a high accuracy. The refinement step has also proven to be necessary (**Table 7**) due to the diversity of phytoliths and artifacts found in real-life scenarios. All approaches tested here have been able to produce results that are above 88% overall accuracy in the correct identification of phytolith classes. In particular, the automated detection and classification of phytoliths in microscope images, with an overall accuracy of 93.68% shows enormous potential to be implemented as an automated stand-alone tool within digital microscope workflows. This can be particularly useful when combined with the use of complex digital microscopes able to produce focussed images by combining multiple digital images at different focus. These can significantly improve the workflow by producing crisp images that will reduce the misidentifications related to unfocussed parts of phytoliths. Also, given the fact that some of these microscopes can incorporate DL algorithms it should be possible for these to produce phytolith types counts automatedly near instantaneously, a radical improvement on the current analysis speeds (manual methods produce speeds of between 1 and 10 slides a day, contingent on analyst experience and slide composition). Beside its incorporation in digital microscope workflows, the algorithm is designed to be employed by other interested parties using freely available computational resources. The supplementary information includes the training algorithm and the detection/classification algorithm in Python prepared for their use within Colab and freely accessible Jupyter notebook and computing platform. In this way the algorithms can be tested applied and improved with the user's own data.

Although costly, the continuous improvement and incorporation of new training datasets for phytolith genera and species and the most common phytolith morphotypes (ICPT, 2019) could result in a viable automatization of phytolith identification. This would produce a revolution in phytolith studies, which would exponentially increase the data available in a fraction of the time and reach much more significant conclusions at larger scales of analysis. At the same time, it would allow researchers working with phytolith data to ask new archaeologically significant questions, and then concentrate on interpretation and data analysis.

In addition, this workflow's outputs may provide much finer or more specific identifications, for example, it may be that there are other single-cell or multi-cell phytoliths from specific taxa that are currently not considered unique and identifiable to the human-eye, but using this workflow are diagnostic – improving our ability to identify phytoliths of different genera or even species, including wild versus domestic taxa. Moreover, this workflow may also provide complementary data such as the level of phytolith fragmentation, area, and shape that can offer important insights during the interpretation phase. Such data could potentially shed light on plant growing conditions and phytolith formation, which could help answer questions related to climate and irrigation for example. The implementation of such methods will also allow researchers to move from estimated phytolith counts (Strömberg, 2009) to actual totals. Indeed, by identifying all phytoliths within the entire slide, versus a small portion (current practice is based on minimum counts of 300 single-cell and 50 multi-cells per slide) it will drastically improve the statistical significance and accuracy of the data.

5. Conclusion

This study has shown that it is possible to automate the classification of phytoliths with a high-level of accuracy. Furthermore, a workflow for identifying phytoliths on microscope images has been presented that can be used for three of the most common genera present in archaeological contexts and two species as demonstrated, and in the future, with the incorporation of more training data, for as many classes of phytoliths as necessary, both at the genus and species level.

The incorporation of ML-based workflows in archaeology, while still little explored beyond site detection studies, presents significant potential within archaeological research. By automatising technical tasks, it will be possible to increase the scale of analysis to achieve quantitatively solid interpretations while reducing the time and effort necessary for the development of these tasks. The incorporation of new methods and automated detection and classification algorithms should ultimately allow archaeologists to concentrate their efforts into the historical and sociocultural interpretations that make archaeological insight unique and necessary.

6. Author contributions

Iban Berganzo-Besga: formal analysis, investigation, methodology, validation, software, data curation, writing of the original draft, visualisation. Felipe Lumbreras: methodology, resources, writing, review and editing, supervision. Monica N. Ramsey: conceptualisation, data curation, writing of the original draft, project administration, funding acquisition. Hector A. Orengo: conceptualisation, methodology, resources, writing, review and editing, supervision, project administration, funding acquisition. Paloma Aliende: data curation.

7. Funding

M.N.R is a Leverhulme Early Career Fellow (EFC-2020-318) and was awarded a D M McDonald Research Grant from the McDonald Institute for Archaeological Research (Deep Origins: AI Deep Learning ID of Plant Phytoliths for the Origins of Agriculture) which partly funded I.B–B's analysis. H.A.O. is a Ramo´n y Cajal Fellow (RYC-2016-19637) of the Spanish Ministry of Science, Innovation and Universities. F.L. work is supported in part by the Spanish Ministry of Science and Innovation project BOSSS TIN2017-89723-P. Some of the GPUs used in these experiments are a donation of Nvidia Hardware Grant Programme.

8. References

Anichini, F., Banterle, F., Buxeda i Garrigo´s, J., Callieri, M., Dershowitz, N., Dubbini, N., Lucendo, D., Evans, T., Gattiglia, G., Green, K., Gualandi, M.L., Hervas, M.A., Itkin, B., Madrid i Fernandez, M., Miguel, E., Remmy, M., Richards, J., Scopigno, R., Vila, L., Wolf, L., Wright, H., Zallocco, M. Developing the ArchAIDE Application: a digital workflow for identifying, organising and sharing archaeological pottery using automated image recognition. *Internet Archaeol.* **2020**, 52. [CrossRef] [Handle]

Battiato, S., Guarnera, F., Ortis, A., Trenta, F., Ascari, L., Siniscalco, C., De Gregorio, T., Suárez, E. Pollen grain classification challenge 2020. In *Pattern Recognition. ICPR International Workshops and Challenges*, 1st ed.; Del Bimbo, A., cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R., Eds., Springer International Publishing: Cham, Switzerland, 2021, pp. 469–479. [Google Scholar] [CrossRef]

Berganzo-Besga, I.; Orengo, H.A.; Lumbreras, F.; Carrero-Pazos, M.; Fonte, J.; Vilas-Estévez, B. Hybrid MSRM-Based Deep Learning and Multitemporal Sentinel-2-Based Machine

Learning Algorithm Detects Near 10k Archaeological Tumuli in North-Western Iberia. *Remote Sens.* **2021**, *13*, 4181. [Google Scholar] [CrossRef]

Cai, Z., Ge, S. Machine learning algorithms improve the power of phytolith analysis: a case study of the tribe Oryzeae (Poaceae). J. *Systemat. Evol*. **2017**, *55(4)*, 377–384. [CrossRef] [CrossRef]

Díez-Pastor, J.F., Latorre-Carmona, P., Arnaiz-González, Á., Ruiz-Pérez, J., Zurro, D. You Are Not My Type": An Evaluation of Classification Methods for Automatic Phytolith Identification. *Microsc. Microanal.* **2020** *26(6)*, 1158–1167. [Google Scholar] [CrossRef]

Ge, Y., Lu, H., Zhang, J., Wang, C., He, K., Huan, X. Phytolith analysis for the identification of barnyard millet (*Echinochloa* sp.) and its implications. *Archaeol. Anthropol. Sci.* **2018**, *10*, 61–73. [Google Scholar] [CrossRef]

Ge, Y.; Lu, H.; Zhang, J.; Wang, C.; Gao, X. Phytoliths in inflorescence bracts: preliminary results of an investigation on common panicoideae plants in China. *Front. Plant Sci.* **2019**, *10*, 1736. [Google Scholar] [CrossRef]

Hošková, K., Pokorná, A., Neustupa, J., Pokorný, P. Inter- and intraspecific variation in grass phytolith shape and size: a geometric morphometrics perspective. *Ann. Bot.* **2021** *127(2)*, 191–201. [Google Scholar] [CrossRef]

International Committee for Phytolith Taxonomy (ICPT). International code for phytolith nomenclature (ICPN) 2.0. *Ann. Bot.* **2019**, *124(2)*, 189–199. [Google Scholar] [CrossRef]

Lu, H., Zhang, J., Wu, N., Liu, K., Xu, D., Li, Q. Phytoliths analysis for the discrimination of foxtail millet (*Setaria italica*) and common millet (*Panicum miliaceum*). *PLoS One* **2009**, *4(2)*, e4448. [Google Scholar] [CrossRef]

Orengo, H.A.; Garcia-Molsosa, A.; Berganzo-Besga, I.; Landauer, J.; Aliende, P.; Tres-Martínez, S. New developments in drone-based automated surface survey: Towards a functional and effective survey system. *Archaeol. Prospect.* **2021**, 1–8. [Google Scholar] [CrossRef]

PHYTAID, 2019. Phytolith Automatic Identification System" v.0.2. Facilitated by the University of Washington and The Burke Museum of Natural History and Culture. Published on the Internet. Available online: http://www.phytaid.burkemuseum.org/ (accessed on 30/12/2021).

Piperno, D.R. *Phytoliths: A Comprehensive Guide for Archaeologists and Paleoecologists*, 1st ed., AltaMira Press: Lanham, USA, 2006. [Google Scholar]

Rapp, G.R., Mulholland, S.C. *Phytolith Systematics: Emerging Issues. Advances in Archaeological and Museum Science*, 1st ed., Springer: Boston, USA, 1992. [CrossRef]

Rosen, A.M. Preliminary identification of silica skeletons from near eastern archaeological sites: an anatomical approach. In *Phytolith Systematics: Emerging Issues. Advances in Archaeological and Museum Science*, 1st ed. Rapp, G.R., Mulholland, S.C., Eds., Springer: Boston, USA, 1992, pp. 129–147. [Google Scholar] [CrossRef]

Strömberg, C.A.E. Methodological concerns for analysis of phytolith assemblages: does count size matter? *Quat. Int.* **2009**, *193(1–2)*, 124–140. [Google Scholar] [CrossRef]

Torres, J. *Python Deep Learning. Introducción Práctica con Keras y TensorFlow2*, 1st ed.; Marcombo: Barcelona, Spain, 2020; pp. 231–253. [Google Scholar]

Waleed, A. (2017). Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. *GitHub repository*. Available online: https://github.com/matterport/Mask_RCNN (accessed on 14/07/2020).

Weisskopf, A.R., Lee, G.A. Phytolith identification criteria for foxtail and broomcorn millets: a new approach to calculating crop ratios. *Archaeol. Anthropol. Sci.* **2016**, *8(1)*, 29–42. [Google Scholar] [CrossRef]

UNIVERSITAT
ROVIRA i VIRGILI