An IEEE Societies and Technical Councils Publication
IEEE Data Descriptions

Received 3 March 2025; revised 6 April 2025 and 13 May 2025; accepted 13 May 2025; Date of publication 16 May 2025; date of current version 12 June 2025.

Digital Object Identifier 10.1109/IEEEDATA.2025.3571025

Descriptor: *Archaeological Cropmark Synthetic Signatures (ACSS)*

ELIAS GRAVANIS AND ATHOS AGAPIOU

Department of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol 3603, Cyprus

CORRESPONDING AUTHOR: Elias Gravanis (e-mail: elias.gravanis@cut.ac.cy).

This work was supported in part by the European Commission (HORIZON-WIDERA-2021-ACCESS-03, Twinning Call, ENGINEER PROJECT) under Grant 101079377, and in part by the DEPLOYED project, Cyprus University of Technology Internal Inter-Disciplinary Research Program. The authors contributed equally to this article.

ABSTRACT With the advent of artificial intelligence and machine learning algorithms, research in the domain of remote sensing archaeology may be greatly assisted by new tools available. Nevertheless, a major barrier relies on the scarcity of existing ground truthing remote sensing data, which eventually hinders training effective models, acquiring general understanding, and performing predictive modeling. In this article, we introduce an extensive dataset of synthetic hyperspectral signatures (ACSS)—of 1 nm interval—simulating cropmarks spectral measurements, which can be used as a proxy indicator for the detection of shallow buried archaeological remains. Simulation generation involved: 1) encoding the observed hyperspectral ground truth signatures, collected over a barley test field, into leaf and canopy bio-physical parameter values, via inversion of the radiative transfer model PROSAIL; and 2) modeling the statistics of the physical parameters by a joint pdf, sampling this distribution, and then running PROSAIL forward to generate two types of simulated signatures, i.e., over buried remains (cropmarks) and over pure soil. The dataset includes 10 000 synthetic signatures from each type in comma-separated value (csv) file format, as well as a Python generator script. To the best of the authors' knowledge, this is the first time that such an extensive spectral cube has been published, which can be further reused by the scientific community to better understand the formation of cropmarks and optimize archaeological proxies through the development of targeted remote sensing archaeological AI models and architectures. We therefore anticipate that ACSS can be a starting point for the development of targeted data resources in the field of remote sensing archaeology that will facilitate model building toward addressing important questions, which are otherwise hard to answer.

IEEE SOCIETY/COUNCIL Geoscience and Remote Sensing

DATA TYPE/LOCATION Measured and synthetic hyperspectral signatures over crops; Cyprus

DATA DOI/PID 10.17632/465vpcsd84.1

INDEX TERMS Archaeological prospection, archaeological proxies, cropmarks, hyperspectral signatures, inverse modeling, machine learning, PROSAIL, physically based, radiative transfer models, synthetic signatures.

BACKGROUND

Optical remote sensing has been widely used and adopted for identifying shallow buried archaeological remains, known in the literature as archaeological proxies [1], [2], [3], [4]. These proxies are essential for supporting archaeological prospection surveys before excavation, as they can provide insights into extensive areas.

A key indicator of archaeological proxies is the phenomenon known as cropmarks [5], [6], [7]. Cropmarks occur when the growth of cultivated crops is affected by shallow buried archaeological remains. These effects can appear as either negative cropmarks, where the crops are stressed, or positive cropmarks, where crop growth is enhanced [8], [9], [10]. These variations are caused by differences in

the subsoil's chemical and physical properties due to the buried archaeological features, which influence the growth conditions of the crops [11].

Optical remote sensing technologies have proven to be a primary tool for detecting cropmarks and are a valuable source of archaeological information [12]. Their detection is carried out by analyzing the spectral signatures in the visible and the near-infrared (NIR) part of the spectrum (around 400–900 nm). Several examples can be found where scholars and researchers have used satellite sensors for their needs. For instance, satellite multispectral satellite sensors have been used in tropical environments on the Indian Ocean islands of Madagascar, Mauritius, and Zanzibar using PlanetScope satellite imageries [13]. At the same time, high-resolution WorldView images were elaborated using deep learning to support "AI-assisted surveys" in the southern Peruvian highlands. Moreover, in the case of [14], Quickbird-2 and IKONOS-2 images were processed to analyze different archaeological environments. Other examples can also be found using hyperspectral sensors, which are limited by spatial resolution. As pointed out in [15], the low spatial resolution of hyperspectral sensors such as HYPERION and the weakly reflective bands in the NIR limit the utility of the results.

In recent years, the evolution of unmanned aerial vehicles and the increased potential of camera sensors have populated the applications for the detection of cropmarks. In [16], a detailed recent review is provided in this domain, including a range of helpful vegetation indices that can be applied to high-resolution orthophoto products. Examples of these technologies can be found in the case studies of the fortified settlements in the northwestern Iberian Peninsula [17], while short-wave infrared (SWIR) cameras have also recently been introduced [18].

Nevertheless, ground hyperspectral (optical) sensors such as ground spectroradiometers remain limited in the literature. Although other ground remote technologies, such as the geophysical prospection sensors, have been adopted throughout the last decades with significant outcomes, the use of ground spectral signatures and the processing of hyperspectral signatures remains limited. A few studies carried out since 2010 have demonstrated the potential use of ground sampling for supporting various steps of satellite processing and enhancement of cropmarks, e.g., [19]. Recent research [20], [21] shows that spectral libraries collected throughout a crop's complete phenological cycle can be used to develop physically based models. Combined with machine learning classification techniques, these models provide promising results for detecting cropmarks. The dataset constructed by the means and methods presented here was used in [20].

COLLECTION METHODS AND DESIGN

Raw Data Collection

Ground spectral signatures were systematically collected from a test field established near the village of Alampra in Cyprus. Detailed information about this test field and additional results on processing the collected spectral signatures are available in [22]. As described in [22], a 5×5 m square area was constructed to simulate "tombs" located 25 cm below the surface. A key priority during the construction of this controlled field was to preserve the soil's natural stratigraphy, ensuring that the original surface layer remained intact. At the time of the study, the surrounding area was cultivated with barley crops, and the precise location of the "tombs" was identified using global navigation satellite systems (GNSSs) (Leica Viva Pro GS15, Leica Geosystems AG, Heerbrugg, Switzerland).

Spectroradiometric measurements of the ground were performed during each campaign, covering both the simulated "archaeological" area and a non-archaeological control area. A handheld GER 1500 spectroradiometer (Geophysical and Environmental Research Corporation, University of Edinburgh, Scotland) was used for these measurements. With 512 distinct channels and a full width at half maximum (FWHM) of approximately 1.5 nm, this instrument can measure electromagnetic radiation within the visible to NIR spectrum (400–1050 nm). All measurements taken over the crops were calibrated using a Lambertian spectralon panel (Geophysical and Environmental Research Corporation, University of Edinburgh, Scotland) to account for incoming solar radiation. This panel reflects nearly 100% (99.996%) of incoming radiation uniformly in all directions. The spectroradiometer's field of view (FOV) was set to 4°, corresponding to an area of roughly 0.02 m^2 when positioned 1.2 m above the ground. Incoming radiance was first determined using the spectralon panel as a reference, followed by measurements of the target areas. This process was repeated across all campaigns to ensure consistency and calibration of the spectral signatures, which ranged from 0% to 100%.

Multi-temporal ground spectroradiometric campaigns were conducted over the simulated test field from October 2011 to April 2012. These campaigns spanned 16 observation days, with the number of measurements varying daily. Table I details the number of reflectance measurements (ρ) taken on specific dates. Measurements labeled "A" were collected over the simulated "archaeological" subsurface features, while those labeled "H" were taken over "healthy" vegetation (crops). Photos from sample measurements can be found in the `Photos` subfolder of the present dataset.

Synthetic Data Generation

The procedure of generating data cubes of synthetic A and H signatures is detailed in section II of [20]. Here, we summarize certain main points. Only the dates in bold in Table I are used, i.e., the period January, February, and March (JFM), in deducing physical parameter values to be used in the simulation generation via the radiative transfer model PROSAIL [23]. The procedure of synthetic signature generation can be briefly described in the following steps.

TABLE I. Dates and Numbers of Measurements per Day

Date	# of A	# of H	Date	# of A	# of H
17/10/2011	19	19	20/12/2011	9	9
26/10/2011	18	18	**03/01/2012**	**9**	**9**
31/10/2011	9	9	**11/02/2012**	**9**	**9**
09/11/2011	9	9	**21/02/2012**	**9**	**9**
16/11/2011	5	5	**04/03/2012**	**9**	**9**
23/11/2011	9	9	**17/03/2012**	**8**	**8**
28/11/2011	9	9	**29/03/2012**	**9**	**9**
13/12/2011	9	9	17/04/2012	9	9

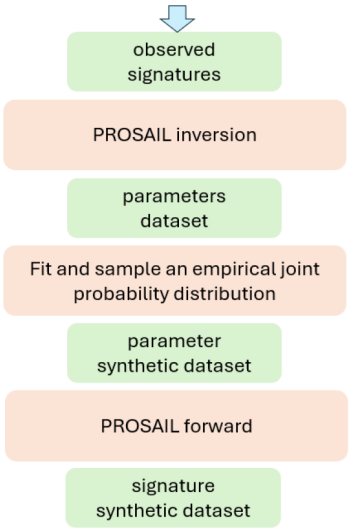

FIG. 1. Flowchart of the basic steps of the synthetic signature dataset generation.

1) The 53 A and 53 H measured signatures from the JFM period are turned to leaf and canopy (collectively called "physical") parameters by inversion of the PROSAIL model. Inversion is done by standard Python language optimizers [25], which employ least-squares minimization. The physical parameters and the ranges or values used are given in Table II, which correspond to generic ranges of grain crops [23].
2) The statistics of the physical parameter values of the A and H signatures are modeled by an empirical joint probability distribution. Simulations of physical parameters for A and H signatures are obtained by sampling this distribution.
3) Running PROSAIL forward, we generate type A and H signatures.

A flowchart of the procedure is given in Fig. 1.

Steps 2 and 3 are implemented in the simulations generating script included in this dataset.

The physical parameters values obtained in step 1 and the associated PROSAIL generated signatures are given in the dataset. Each estimated seven-parameter group and signature are labeled by A or H. The quality of the fit is illustrated in Fig. 2, where we plot the indicated percentile curves and median line of modeled-to-observed reflectance difference across the given range of wavelengths. In Table III, we also provide statistics (including mean value, standard deviation, and 5th and 95th percentile) for the root-mean-square (RMSE) and the Pearson R^2 values (see also [20]). We observe that differences are in general smaller than the radiometric uncertainty of widely used satellite sensors.

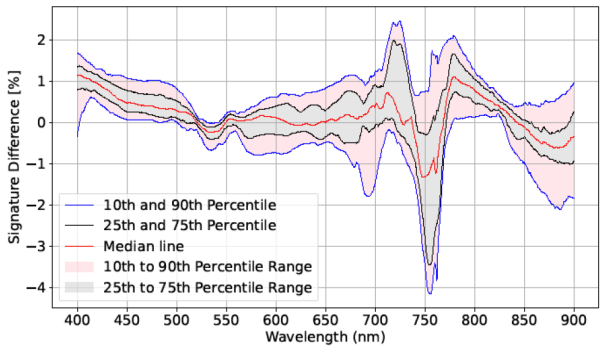

FIG. 2. Percentile curves of the modeled-to-observed difference in units 0–100 across the 400–900 nm wavelength range.

TABLE II. Ranges of the PROSAIL Input Parameters

Parameter	Value
Leaf structure index, N (-)	1.5
Leaf chlorophyll content, Cab ($\mu g\ cm^{-2}$)	**0–100**
Leaf carotenoid content, Car ($\mu g\ cm^{-2}$)	**0–30**
Leaf dry matter content, Cm ($g\ cm^{-2}$)	**0.002**
Equivalent water thickness, Cw (cm)	**0.02**
Brown pigment content, Cbrown (-)	0
Leaf anthocyanin content, Cant ($\mu g\ cm^{-2}$)	**0–4**
Leaf area index, LAI ($m^2\ m^{-2}$)	**0.5–8**
Average leaf angle, lidfa (deg)	**30–89**
Hot-spot size parameter, hspot ($m\ m^{-1}$)	**0.25–0.75**
Soil reflectance, psoil (-)	**0–1.0**
Sun zenith angle (deg)	0
View zenith angle (deg)	0
Relative azimuth angle (deg)	0

TABLE III. Statistics of Goodness-of-Fit Measures

Measure	Mean	St. Dev.	p5	p95
RMSE	0.0083	0.0030	0.0038	0.0128
R^2	0.9973	0.0030	0.9915	0.9993

In step 2, the seven estimated parameters, namely Cab, Car, lidfa, hspot, LAI, psoil, and ant, which are associated with each type (A and H) of signature, are treated as distinct variables, carrying information about the type. The statistics of these fourteen variables is encoded in their covariance matrix and marginal distributions (histograms). From this information, an empirical joint pdf is deduced and sampled for the simulated signatures generating part in step 3. One

should note that the measured signatures are not given spatial reference; hence, each type (A or H) is treated as a single sample.

VALIDATION AND QUALITY

To independently check the quality of the synthetic A and H signatures, we added a validation step. (Note that this is not part of [20]). To this end, we generate 1000 synthetic signatures from each kind, A and H. Then, dimensionality reduction methods with variable number of components, and classifying algorithms, are employed in a pipeline to model the synthetic dataset as effectively as possible. The obtained models are then tested on the original (measured) signatures dataset, which has not been used in the construction of the models.

In particular, we use five dimensionality reduction methods including principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA), singular value decomposition (SVD), and partial least squares regression (PLS). After initial experimentation with different classifiers, including logistic regression, k-nearest neighbors, random forests, support vector machines, and neural networks (NNs), we concluded that NN classifiers are the most effective in our case. The NN classifiers are simply configured with the default parameters in Python 3.9. In particular, NN is an MLP classifier, with one hidden layer of 100 neurons, ReLU activation function, Adam optimizer, a learning rate of 0.001, and is trained for a maximum of 200 iterations.

We employ fivefold cross validation (cv $=$ 5). We use accuracy as a performance measure of the models. We use a range of components of the dimensionality reduction methods from 5 to 8. We observed that with higher number of components, the models tend to overfit the synthetic dataset and return relatively poor scores when evaluated on the test (measured signatures) set. We may note that this range is consistent with the number of variable physical parameters used in the physically based construction of the synthetic dataset, i.e., 7. We selected the top 10 performing models whose accuracy is 99%. Below, each model is designated by the name of the classifier, the dimensionality reduction transformer, and the number of components of the latter. The results are shown in Table IV, where we include the mean value and standard deviation of accuracy of the corresponding cross-validation fold.

The performance of these models on the measured signatures set is summarized in Table V. We may recall that the measured set size is only 106. One observed that all models return consistently high performance, with the possible exception of NN_FA_7, which performs less well. The average value of each classification score across the top 10 modes is 95%. We may interpret these results as arguing strongly in favor of the good quality of the synthetic dataset.

TABLE IV. Top 10 Performing Models

Model	Mean Accuracy	Standard Dev.
NN_ICA_8	0.9939	0.0014
NN_FA_8	0.9938	0.0015
NN_PLS_8	0.9934	0.0017
NN_PLS_6	0.9933	0.0012
NN_PLS_7	0.9932	0.0011
NN_ICA_6	0.9931	0.0018
NN_ICA_7	0.9931	0.0015
NN_FA_7	0.9929	0.0013
NN_PLS_5	0.9921	0.0018
NN_FA_6	0.9896	0.0024

TABLE V. Performance of Models on the Original Dataset

Model	Accuracy	Precision	Recall	F1
NN_ICA_8	0.9717	0.9464	1.0000	0.9725
NN_FA_8	0.9623	0.9455	0.9811	0.9630
NN_PLS_8	0.9340	0.9600	0.9057	0.9320
NN_PLS_6	0.9811	0.9636	1.0000	0.9815
NN_PLS_7	0.9811	0.9636	1.0000	0.9815
NN_ICA_6	0.9811	0.9811	0.9811	0.9811
NN_ICA_7	0.9811	1.0000	0.9623	0.9808
NN_FA_7	0.8302	0.8571	0.7925	0.8235
NN_PLS_5	0.9340	0.9792	0.8868	0.9307
NN_FA_6	0.9434	0.9796	0.9057	0.9412

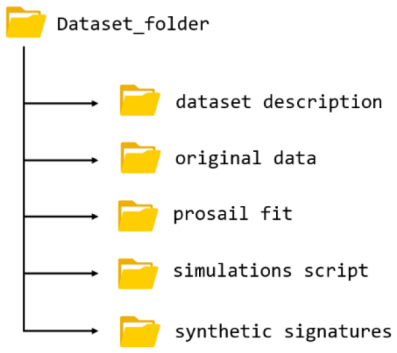

FIG. 3. Structure of the dataset folder.

RECORDS AND STORAGE

The structure of the dataset folder is given in Fig. 3. The `original data` folder contains the original (measured) signatures and photos of sample measurements. The `prosail fit` folder contains the values of the physical parameters obtained by PROSAIL inversion and the corresponding PROSAIL generated (fitted) signatures. The `simulations script` folder contains `simulations_generation.ipynb` (Jupyter notebook) and two additional files are required for it to run.

The variable `n_samples` is to be set equal to the desired number of A and H signatures to be generated. The output dataset size is equal to `n_samples` of each

type of signature in the wavelength range of 400–900 nm, stored in csv files (`simulated_A_{n_samples}`, `simulated_H_{n_samples}`). Finally, the `synthetic signatures` folder contains 10 000 ready-made synthetic signatures of each type (A and H) in the form of csv files. The signatures are given in the 400–900 nm wavelength range.

INSIGHTS AND NOTES

ACSS dataset aims to promote further research in collecting and processing ground truth hyperspectral signatures over buried archaeological remains by optimizing the detection of cropmarks. A complete spectral cube of 10 000 synthetic signatures in the range of 400–900 nm is provided in this dataset that can serve as a valuable tool for future research and practical applications in this domain. To the best of the authors' knowledge, this is the first-ever synthetic spectral cube generated using physically based models such as PROSAIL, and its extent provides enough data for testing various AI models and architectures in the domain of remote sensing archaeology, targeted for the detection of archaeological proxies. An example of the application of these synthetic signatures is given in [20]. A limitation of ACSS is the absence of spatial resolution of the measured signatures. This conceals the spatial component of the correlation between the spectral and physical data, resulting in a coarse-grained synthetic dataset. This is consistent with the relative smallness of the test field area. Hence, the constructed dataset (ACSS) primarily serves as encoding the physical differences between the cropmark and healthy crop signatures. Future developments along the lines of the present work may include: greater time-span than a single year, different crops, geographic locations, and physical conditions, including soil type.

SOURCE CODE AND SCRIPTS

The ACSS dataset is stored and publicly available in the open-source cloud-based data-sharing platform, Mendeley Data, under the name *Archaeological Cropmark Synthetic Signatures* (ACSS) [24]. The dataset can be directly accessed using the following URL: https://doi:10.17632/465vpcsd84.1.

ACKNOWLEDGMENT

The authors have declared no conflicts of interest.

REFERENCES

[1] N. Masini, C. Marzo, P. Manzari, A. Belmonte, C. Sabia, and R. Lasaponara, "On the characterization of temporal and spatial patterns of archaeological crop-marks," *J. Cult. Heritage*, vol. 32, pp. 124–132, 2018.

[2] A. Agapiou, A. Hegyi, and A. Stavilă, "Observations of archaeological proxies through phenological analysis over the Megafort of Csanádpalota-Juhász T. Tanya in Hungary using Sentinel-2 images," *Remote Sens.*, vol. 15, no. 2, p. 464, 2023.

[3] D. Stott, D. S. Boyd, A. Beck, and A. G. Cohn, "Airborne LiDAR for the detection of archaeological vegetation marks using biomass as a proxy," *Remote Sens.*, vol. 7, no. 2, pp. 1594–1618, 2015.

[4] D. Ronchi, M. Limongiello, and S. Barba, "Correlation among earthwork and cropmark anomalies within archaeological landscape investigation by using LiDAR and multispectral technologies from UAV," *Drones*, vol. 4, no. 4, p. 72, 2020.

[5] N. Masini and R. Lasaponara, "Sensing the past from space: Approaches to site detection," in *Sensing the Past: From Artifact to Historical Site.* Cham, Switzerland: Springer, vol. 16, 2017, pp. 23–60.

[6] A. R. Beck, "Archaeological site detection: The importance of contrast," in *Proc. Annu. Conf. Remote Sens. Photogramm. Soc.*, Newcastle, U.K.: Newcastle Univ., Sep. 2007, pp. 307–312.

[7] D. C. Cowley, "In with the new, out with the old? Auto-extraction for remote sensing archaeology," in *Proc. Remote Sens. Ocean, Sea Ice, Coastal Waters, Large Water Regions*, vol. 8532, Edinburgh, UK: SPIE, Oct. 2012, pp. 37–45.

[8] M. Gojda and M. Hejcman, "Cropmarks in main field crops enable the identification of a wide spectrum of buried features on archaeological sites in Central Europe," *J. Archaeol. Sci.*, vol. 39, no. 6, pp. 1655–1664, 2012.

[9] P. G. Deshbhandari and A. Shetty, "Archaeological investigations of Honnavar Fort in Uttara Kannada District of Karnataka, India using geospatial technology," *J. Indian Soc. Remote Sens.*, vol. 52, no. 5, pp. 1031–1043, 2024.

[10] A. J. O. Villarejo and L. M. Gutiérrez, "Combining Sentinel 2A timeseries, Lidar data and archaeological prospection for the study of ancient roads and poblational pattern in right side of Guadalén River, (Vilches, Spain)," *Preprints*, 2024.

[11] M. Limongiello, L. Gujski, and C. De Vita, 2020, "Analysis of RGB images to enhance archaeological cropmark detection: The case study of Nuceriola, Italy," in *Proc. 42th Int. Conf. Represent. Discip. Teachers Congr. Unione Italiana per il Disegno*, Milan, Italy: Franco Angeli, pp. 2360–2368.

[12] C. Moriarty, D. C. Cowley, T. Wade, and C. J. Nichol, "Deploying multispectral remote sensing for multi−temporal analysis of archaeological crop stress at Ravenshall, Fife, Scotland," *Archaeol. Prospect.*, vol. 26, no. 1, pp. 33–46, 2019.

[13] W. Alders, D. S. Davis, and J. J. Haines, "Archaeology in the fourth dimension: Studying landscapes with multitemporal PlanetScope satellite data," *J. Archaeol. Method Theory*, vol. 31, no. 4, pp. 1588–1621, 2024.

[14] H. Guo, X. Wang, F. Chen, and C. Wang, 2024, "Spatial cognition for information of world cultural heritage," in *Introduction to Space Archaeology.* Singapore: Springer Nature, pp. 195–257.

[15] S. H. Savage, T. E. Levy, and I. W. Jones, "Prospects and problems in the use of hyperspectral imagery for archaeological remote sensing: A case study from the Faynan copper mining district, Jordan," *J. Archaeol. Sci.*, vol. 39, no. 2, pp. 407–420, 2012.

[16] D. Kaimaris, "Aerial remote sensing archaeology—A short review and applications," *Land*, vol. 13, no. 7, p. 997, 2024.

[17] S. Peña-Villasenín, M. Gil-Docampo, and J. Ortiz-Sanz, "Hidden archaeological remains in heterogeneous vegetation: A crop marks study in fortified settlements of Northwestern Iberian Peninsula," *Remote Sens.*, vol. 16, no. 21, p. 3923, 2024.

[18] J. Casana and C. Ferwerda, "Drone-acquired short-wave infrared (SWIR) imagery in landscape archaeology: An experimental approach," *Remote Sens.*, vol. 16, no. 10, p. 1671, 2024.

[19] A. Agapiou, "Optimal spatial resolution for the detection and discrimination of archaeological proxies in areas with spectral heterogeneity," *Remote Sens.*, vol. 12, no. 1, p. 136, 2020.

[20] E. Gravanis and A. Agapiou, "Physically-based detection algorithm of buried archaeological remains using spectral signatures," *IEEE Access*, vol. 12, pp. 197217–197232, 2024.

[21] A. Agapiou and E. Gravanis, "A machine-learning-assisted classification algorithm for the detection of archaeological proxies (cropmarks) based on reflectance signatures," *Remote Sens.*, vol. 16, no. 10, p. 1705, 2024.

[22] A. Agapiou, "Development of a novel methodology for the detection of buried archaeological remains using remote sensing techniques," Ph.D. thesis, Cyprus Univ. of Technol., Limassol, Cyprus, 2013. Accessed: May 8, 2024. [Online]. Available: https://hdl.handle.net/20.500.14279/877

[23] Python Programming Language. Accessed: Mar. 14, 2024. [Online]. Available: https://www.python.org/

[24] E. Gravanis and A. Agapiou, "Archaeological cropmark synthetic signatures (ACSS)," *Mendeley Data*, vol. V1, 2025, doi: 10.17632/465vpcsd84.1.