Machine learning for rapid mapping of archaeological structures made of dry stones – Example of burial monuments from the Khirgisuur culture, Mongolia –

Fabrice Monna, Jérôme Magail, Tanguy Rolland, Nicolas Navarro, Josef Wilczek, Jamiyan-Ombo Gantulga, Yury Esin, Ludovic Granjon, Anne-Caroline Allard, Carmela Chateau-Smith

Machine learning for rapid mapping of archaeological structures made of dry stones

- Example of burial monuments from the *Khirgisuur* culture, Mongolia -

Fabrice Monna[a,*], Jérôme Magail[b], Tanguy Rolland[a], Nicolas Navarro[c,d], Josef Wilczek[a,e,f], Jamiyan-Ombo Gantulga[g], Yury Esin[h], Ludovic Granjon[i], Anne-Caroline Allard[j], Carmela Chateau-Smith[k]

[a] ARTEHIS, UMR CNRS 6298, Université de Bourgogne–Franche Comté, 6 Boulevard Gabriel, Bat. Gabriel, 21000 Dijon, France

[b] Musée d'anthropologie préhistorique de Monaco, 56, boulevard du Jardin exotique, 98000 MC, Monaco

[c] EPHE, PSL Research University, 21000 Dijon, France

[d] Biogéosciences UMR CNRS 6282, Université Bourgogne Franche-Comté, 6, boulevard Gabriel, Bat. Gabriel, 21000 Dijon, France

[e] Ústav archeologie a muzeologie, Masarykova univerzita, Arna Nováka 1, 602 00 Brno, Czech Republic

[f] Katedra archeologie, Univerzita Hradec Králové, Rokitanského 62, Czech Republic (present address)

[g] Bronze and Early Iron Age Department, Institute of History and Archaeology, Mongolian Academy of Sciences, Jucov street-77, Ulaanbaatar-51, Mongolia

[h] Khakassian Research Institute for Language, Literature, and History, 23, Shchetinkin Street, 655017 Abakan, Republic of Khakassia, Russia

[i] MSH de Dijon, USR CNRS 3516, Université Bourgogne Franche-Comté, 6, esplanade Erasme, 21066 Dijon, France

j Institut d'Art et d'Archéologie, Université Paris IV Sorbonne, 3, rue Michelet, 75006 Paris, France

[k] CPTC, Université de Bourgogne, 4, boulevard Gabriel, 21000 Dijon, France

∗ Corresponding author. E-mail address: Fabrice.Monna@u-bourgogne.fr (F. Monna), phone: +33 (0)3 80 39 63 60

Abstract: The present study proposes a workflow to extract from orthomosaics the enormous amount of dry stones used by past societies to construct funeral complexes in the Mongolian steppes. Several different machine learning algorithms for binary pixel classification (i.e. stone *vs* non-stone) were evaluated. Input features were extracted from high-resolution orthomosaics and digital elevation models (both derived from aerial imaging). Comparative analysis used two colour spaces (RGB and HSV), texture features (contrast, homogeneity and entropy raster maps), and the topographic position index, combined with nine supervised learning algorithms (nearest centroid, naive Bayes, *k*-nearest neighbours, logistic regression, linear and quadratic discriminant analyses, support vector machine, random forest, and artificial neural network). When features are processed together, excellent output maps, very close to or outperforming current standards in archaeology, are observed for almost all classifiers. The size of the training set can be drastically reduced (to ca. 300 samples) by majority voting, while maintaining performance at the highest level (about 99.5% for all performance scores). Note, however, that if the training set is inadequate or not fully representative, the classification results are poor. That said, the methods applied and tested here are extremely rapid. Extensive mapping, which would have been difficult with traditional, manual, or semi-automatic delineation of stones using a vector graphics editor, now becomes possible. This workflow generally surpasses pedestrian surveys using differential GPS or a total station.

Keywords: pixel classification, grey level co-occurrence matrix, RGB colour space, texture, topographic position index, photogrammetry, burial complex planigraphy, Mongolia, Bronze Age, Iron Age.

1. Introduction

Soon after the start of aerial photography, which became fully operational during WW1, archaeologists realized the potential of this technique for discovering new sites, apprehending large complexes in a new way, and understanding the spatial organization of archaeological structures [1]. For a long time, greyscale pictures were captured in low sunlight to reveal elevation anomalies as shadow marks [2]. Later, the introduction of colour photography became an important asset to identify subtle colour variations, which may occur in field crops, either because soils above buried walls generally retain less moisture, or because water may accumulate close to the structures [3]. Nowadays, the range of acquisition techniques in aerial archaeological investigation is huge, from satellites to small unmanned aerial vehicles (UAVs), and images are produced in both in visible and invisible spectra (e.g. [4]). However, because of their low cost and versatility, UAVs are often privileged by archaeologists over other solutions. High-definition surveys over several km^2 are possible with UAVs, by dividing the area of interest into several smaller tiles (depending on UAV flight capacity). In contrast with pioneering works, photographs are now often acquired not only to provide aerial images, but also high-definition, georeferenced orthomosaics, and digital elevation models, also known as DEMs [5]. Landscape representations are then reconstructed with the help of photogrammetry, a technique that is about to become the new standard in archaeology for field documentation [6,7]. Scientists dealing with such a massive information flow may, however, encounter serious difficulties in producing documentation suitable for further exploitation in a reasonable amount of time [8], particularly when stones or archaeological structures are delineated manually with a vector graphics editor. The burial structures in the Arkhangai province (Mongolia) are a perfect example of this type of bottleneck. This region is extremely rich in funeral monuments dating from the Bronze Age and the Iron Age. Funeral complexes from the Iron Age are composed of decametric dry-stone arrangements [9-10], sometimes encompassing several km^2. Even for a skilled expert, it is extremely difficult to identify any clear organisation of the monument from the ground, mainly due to the vast area covered. By contrast, orthomosaics provide extremely valuable

information [11]. Almost every individual stone can be distinguished in the open steppes, especially when the grass is low due to livestock grazing. In this specific example, as in many other situations of this kind, a rapid and accurate procedure to extract, at least semi-automatically, stone boundaries from the data acquired by UAV would be a valuable improvement in the acquisition speed of archaeological documentation. The problem, in a nutshell, consists of a binary pixel classification (stone *vs* non-stone), and solutions can be sought in the numerous machine learning algorithms increasingly used in archaeology [12-15]. Colour information is an obvious candidate for input data, as stones are clearly visible in the images. Other variables, such as those related to the spatial arrangement of the tonal information (also known as textural features) may also be highly relevant, as demonstrated in various fields, such as remote sensing, ecology, etc. [16-17]. Local altitudinal variations may also be useful as input data for classification [18].

2. Research aim

To treat the set of input features mentioned above, supervised learning should be privileged in the present case, because the operator determines the number of output classes (here two) in a very specific way, in conformity with the objectives, unlike unsupervised learning [19]. Although this approach inevitably introduces a manual and relatively time-consuming training step, outcomes should be much closer to documentation directly exploitable from an archaeological point of view. Here, our objective is to compare several solutions for binary pixel classification, by combining different input features: colour, textural parameters, and topography, with nine supervised machine learning algorithms: nearest centroid, naive Bayes, *k*-nearest neighbours, logistic regression, linear and quadratic discriminant analyses, support vector machine, random forest, and artificial neural network. For the first example, the famous 9-ha site of Jargalant, descriptive features were progressively introduced into the algorithms, and classification performance was investigated statistically and then empirically, by assessing the archaeological potential of the maps obtained.

Once the best approach had been selected, it was applied to the larger area of Tsatsiin Ereg, and the results were evaluated.

3. Material and methods

3.1. The sites

The photogrammetric campaign took place during three summer periods (2016 to 2018), within the framework of the "Joint Monaco-Mongolian Archaeological Mission". It focused mainly on the site of Tsatsiin Ereg, in the Khoid Tamir valley, located about 500 km southwest of Ulaanbaatar, but also on other smaller sites, either in the vicinity of Tsatsiin Ereg, or in Jargalant, a site famous for its deer stones, about 80 km away [20]. The site of Tsatsiin Ereg is characterized by a remarkable concentration of large, well-preserved complexes, including barrows, satellite quadrangles, circles, enclosures, and stone alignments formed by the accumulation of dry stones. The most sophisticated complexes, which can extend over several square kilometres, date from the late 2nd to early 1st millennium BC [21-23]. The plethora of structures composing these funeral complexes makes it difficult to understand the precise chronology of the building phases. However, the repetition of certain elements, related to funeral practices, sacrificial rituals, and artistic style, indicates cultural coherence shared over a large area by past nomadic societies, which should be studied to better apprehend cultural interactions.

3.2. Orthomosaic and DEM production

Pictures were captured by an unmanned aerial vehicle (UAV), a DJI Phantom 3 PRO equipped with a GPS and a 12 Mpix camera. The lens was a 20 mm (equivalent 35 mm) f/2.8, producing a diagonal field of view of 94°. The flight plan was programmed *via* a free Android application (Altizure App,

https://www.altizure.com/), where target area positioning is facilitated by displaying a satellite image as background. The operator sets the height flight above ground, the size and orientation of the region of interest, as well as the capture density, by choosing the amount of forward and side overlap (typically 75-80%). The UAV can then automatically follow a zigzag pattern, taking a series of photographs in the nadir direction. As sensor definition and focal of the lens are fixed, the distance between the centres of two consecutive pixels at ground level, also known as the ground sample distance (GSD), depends only on the height of flight [24]. At this point, it is worth recalling that GSD should be at least half the size of the smallest details to be captured optically. If the smallest stones of interest measure approximately 20 cm, they should be recognisable at a flight height of 100-150 m (GSD of 4.3-6.5 cm/px, theoretically). In practice, the areas studied were divided into smaller square regions of interest of approximately 9-10 ha, which can be covered by UAV without battery replacement at a flight height of 100-150 m, a value typically used. Ground control points were placed before acquisition, and the distances between them were measured with a laser telemeter Leica Disto D510, able to work in sunlight up to 200 m, with a precision of ca 1 mm. Typically 80-110 pictures were captured per tile, taking 15-20 min in the field. In the laboratory, orthomosaics and DEMs were produced by the Photoscan PRO software, v. 1.4.3 from Agisoft. Picture alignment and subsequent sparse cloud construction were strongly constrained by the distances between GCPs. This step, which helps to structure the 3D model, was useful here, as wind causing movement of long grass (which dominates the landscape studied) may lead to the accumulation of slight alignment errors between pictures. The workflow then consisted in densifying the cloud, producing height field models, DEMs, and orthomosaics, at a fixed resolution of 5 cm / pixel for Tsatsiin Ereg and 8 cm/pixel for Jargalant (see [22, 25] for more details about the photogrammetrical workflow). It is worth mentioning that the use of GCPs precisely georeferenced with differential GPS, for example, would have been optimal for accurate positioning of the maps produced. Although georeferencing here only derives from the GPS embarked into the drone because of logistic constraints, the relative error

of the models produced does not exceed 10-20 cm, while absolute error, assessed by projecting orthomosaics on Google Earth (considered as true reference), does not exceed 1-3 m.

3.3. Feature inputs

In order to evaluate which input feature (or combination of features) is the most pertinent for binary classification (i.e. stone *vs* non-stone), several image representations were obtained from original orthomosaics and DEMs.

RGB colour space. In its most common version: 24-bit encoding, the image is composed of one channel for each of the three primary colours processed by cameras and computers: red, green, and blue. Each is encoded on 8 bits, producing 256 possible discrete values per channel, and a palette of 16 777 216 discrete combinations. This colour space uses additive colour mixing to compose the final image. In the following, images are split into 3 channels (namely R, G, and B), and each colour channel is treated separately as a single feature (Fig. 1).

*HSV colour space***.** Colour transformations into non-RGB colour space have sometimes been shown to enhance classification performance [26]. Similarly to the RGB colour model, the HSV colour space is composed of three channels (for hue, saturation, and value), denoting colour property, perceived colour intensity, and perception of brightness (Fig. 1) [27]. The HSV channels are obtained from those composing the RGB.

Grey Level Co-Occurrence Matrix (GLCM) and texture parameters. Developed by Haralick et al. (1973) [28], GLCM texture parameters belong to the family of statistics describing the arrangement of pixels separated by a certain distance, in different directions. Originally 14 parameters were proposed as image texture features, but only three of them are used in the following: contrast (CON), homogeneity (HOM), and entropy (ENT), because they have been recognized to enhance classification accuracy [29,30]. Their calculation is a two-step process: the computation of the grey

level co-occurrence matrix from an image with *g* grey levels (obtained from the RGB image), and the calculation of the descriptors from this matrix (see [31] for details). The contrast descriptor, calculated for each pixel, illustrates the local variation of pixel intensity within a certain spatial range, while homogeneity and entropy describe the local sameness of grey levels of pixels, in other words the tonal variations in space. Three parameters must therefore be tuned: spatial scale (i.e. window size), the number of grey levels in the image to be processed, and the directions for which GLCM is computed (Note that calculations covered all directions using 32 grey levels and a window size of 9 x 9 pixels, i.e. representing typically 0.2-0.5 m^2. See Fig. 1 for map examples).

Topographic position index (TPI). This index, widely used for automatic landform classification [32], is simply defined as the difference in altitude between a central pixel and the mean of the surrounding cells in the DEM [33]. The TPI depends only on topography and the size of the search window defining neighbours. In the present case, this size should be larger than the elements of interest, to highlight them as positive or negative anomalies (In the following, a square window of 201 x 201 pixels, covering about 100-250 m^2, was found appropriate. See Fig. 1, where the DEM is coloured and hill-shaded for better understanding).

3.4. Machine learning algorithms

Underlying idea. Here, the aim is to predict, for each pixel, the presence or absence of stone, from a set of features selected among those enumerated above (i.e. colour channels, texture maps, and TPI). Let Y be the class ensemble composed, in our case, of two categories: y_c, (with $c \in [0,1]$), and **x** a vector describing the set of *n* features, $\mathbf{x} = \{x_1,\ldots,x_n\}$. In supervised learning, the operator first teaches the mathematical model, labelling by hand a set of pixels with and without stones. From that training set or a subset of it, patterns are sought in **x** to predict class labels, Y. The classifier tries to find a mapping function (i.e. a decision rule), $f(\mathbf{x}) \rightarrow Y$, which is then used to map new, unseen data. Nine popular classifiers listed below were tested in this study. As they are extensively explained in many

textbooks [34-36], they are only briefly described, to facilitate clearer understanding for readers unfamiliar with machine learning, and a simplified pipeline is provided in Fig. 2.

Naive Bayes (NB). This classifier is one of the simplest algorithms used for binary classification. As indicated by its name, it is based on the Bayes' theorem, and is naïve as it assumes that every pair of features is independent, a situation rarely met in real-world data. Continuous variables are assumed to follow Gaussian distributions within each class, allowing the calculation of conditional probabilities, $p(x_i|y_c)$, from the training set [37]. Combined with the independence assumption, a class label, $\hat{y}_{NB}$, corresponding to the most probable class (that with maximum *a posteriori* probability) is assigned following:

$$\hat{y}_{NB} = \underset{c}{\operatorname{argmax}}\ p(\mathrm{y}_c) \prod_{\mathrm{i}=1}^{\mathrm{n}} p(x_i|y_c)$$

Nearest centroid (NC). It simply compares the position of a sample in the feature space to the centroid of each class determined from the training set, and labels it with the class where the mean, $\boldsymbol{\mu}_c$, is the closest (see Supplementary Material S1a for an illustration):

$$\hat{y}_{NC} = \underset{c}{\operatorname{argmin}}\ \|\boldsymbol{\mu}_c - \boldsymbol{x}\|$$

*k-nearest neighbours (*KNN*).* For the classification task, the *KNN* algorithm identifies the *k* nearest neighbours in the feature space of the training samples, $T = \{S_1^{NN}, \ldots, S_k^{NN}\}$, and proceeds by majority voting to assign a class label, $\hat{y}_{KNN}$ (Supplementary Material S1b):

$$\hat{y}_{KNN} = \underset{c}{\operatorname{argmax}} \sum_{S_i \in \mathrm{T}} I(y_c = y_i^{NN})$$

with $y_i^{NN}$the class label of the *i*-th neighbour among the *k* nearest neighbours, and $I(y_c = y_i^{NN})$ equal to 1 if the classes of S_1^{NN} and y_c are the same, and 0 otherwise. Despite its simplicity, this algorithm often provides competitive results, but the value of *k* (usually odd) needs to be set by the operator, or optimized [38]. The operator may also choose the metric used for distance, e.g.

Euclidian, Manhattan, and possibly a weight, inversely proportional to the distance of each of the *k* neighbours, which may be useful when classes are clearly unbalanced.

Logistic regression (LR). A new variable, *z*, is first built as input from a linear combination of weights and sample features: $z=\mathbf{w}^{\mathrm{T}}\mathbf{x}$. Then, the logistic function, also known as a sigmoid cumulative logistic distribution, quashes the range of possible outputs within the [0, 1] range that can be interpreted in terms of probabilities (Supplementary Material S1c). From these results, a class label is assigned as follows [35]:

$$\hat{y}_{LR} = \underset{c}{\operatorname{argmax}}\ p(\mathrm{y}_c|\boldsymbol{x}, \boldsymbol{w})$$

Several different strategies exist to optimize the weights, **w**, and to perform a regularization step, to handle collinearity among features, as well as to prevent overfitting. Logistic regression performs well for classes that are linearly separable (details can be found in [39]).

Linear and quadratic discriminant analyses (LDA & QDA). These models assume a Gaussian density for each class. Bayes' rule is applied to calculate conditional probabilities, $p(\mathrm{y}_c|\boldsymbol{x})$, and hence to predict classes, by choosing *c*, which maximizes $p(\mathrm{y}_c|\boldsymbol{x})$. For LDA, all classes are supposed to have the same covariance matrix, and decision boundaries are linear (Supplementary Material S1d), but not for QDA (Supplementary Material S1e), which makes decision boundaries more flexible [40].

Support-vector machine (SVM). This very popular algorithm for classification seeks to maximize the margin between the decision boundary hyperplane and the closest training samples from this hyperplane, which are called support vectors [41]. The idea behind this procedure is to produce a clear gap between samples belonging to both categories (Supplementary Material S1f). Interestingly, SVM can also be used for data not linearly separable, after the application of a mapping function, which transforms the input data, in a higher dimensional space where classes become linearly separable. This step takes advantage of the so-called "kernel trick" for computation, by applying, most often, a radial basis function (RBF), which is in fact a Gaussian kernel [42].

Random forest (RF). It belongs to the class of ensemble methods, capable of both classification and regression, in the case of linear and non-linear problems. The algorithm proceeds by aggregating a bunch of classification trees built randomly: instead of constructing splits on the basis of feature importance, the best feature is sought among a random subset of input variables (Supplementary Material S2a). The idea behind the algorithm is that predictions made by individual decision trees may not be correct but, once they have been combined, label predictions will be more accurate and stable [43]. Several parameters must be tuned for forest construction (e.g. number of trees, number of levels in each decision tree, etc.), and concerning the method used for sampling data points.

Artificial neural network (ANN). This algorithm is vaguely inspired by the human brain [44]. Several hidden layers composed of several nodes are placed between input (i.e. features) and output (class labels) layers. Each node receives input values from the previous layer. Values are pondered by weights and biased, and then passed through an activation function, used to determine whether and to what extent the signal moves to the next layer (Supplementary Material S2b). Weight and bias values are optimized by iterating the following steps: (i) predicted output calculation (feedforward step), and (ii) update of weight and bias (backpropagation step) [36]. The operator must set the number of hidden layers and units, the learning rate, the type of activation function, etc.

Hard voting. Hard voting consists in aggregating predictions made by each individual classifier, or a subset of them, and then predicting the class by simple majority voting (Supplementary Material S2c). The underlying idea is that several models are probably more reliable than just one.

3.5. Hyperparameter tuning and metrics for model evaluation

Basically, a good practice to evaluate the capacity of the models produced to generalize to unseen data is to split the supervised data into two groups, one for training (here 70% of the dataset), with the remainder (30%) as a test dataset [36]. However, depending on learning strategies, as mentioned above, several model hyperparameters (some of them tackling overfitting *via* regularization) have to

be fine-tuned. This optimization step is operated by a brute force search on a grid of possible hyperparameter values, using inner cross-validation on the training set. For final model evaluation, two strategies are possible. The first simply applies the model to the test dataset. This method provides a single evaluation value, which is unbiased, as these data were not used to build the models [44]. Note, however, that results may depend on how the data were split into training and test sets. The second strategy computes an outer cross-validation by splitting the data into *k* folds, applying the model to *k*-1 folds, keeping the remainder for performance evaluation, and repeating the operation *k* times [45]. Results are almost unbiased [46] and, interestingly, may be expressed in terms of confidence intervals. Both strategies were used here.

Several metrics are available to evaluate the models, including precision, recall, F1-score, and accuracy [36]. Let TP, TN, FP, and FN be the True Positive, True Negative, False Positive and False Negative, respectively (with positive being a stone).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score deals with both precision and recall. It is the harmonic of both scores:

$$\text{F1} - \text{Score} \quad = 2.\frac{\text{Precision} \,.\, \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is defined as the percentage of correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Most archaeologists prefer to save all the stones if possible, even if they need to perform a minor *a posteriori* cleaning of false positives. As a consequence, hyperparameters were tuned to maximize recall.

3.6. Practical implementation

Computation of GLCM and TPI used the `glcm` package and a homemade script, both written for the free R software (https://www.r-project.org/). Colour manipulation, application of machine learning algorithms to standardized data, and georeferencing of the maps produced used the `scikit-learn 0.20.3`, `opencv 4.1.1`, `rasterio 1.0.22`, `gdal 2.4.2`, and `scipy 1.2.1` libraries for Python 3.7.1 (https://www.python.org/). Final results, expressed as georeferenced polygon vector layers, were integrated using the free QGIS software (https://www.qgis.org). Point pickup to produce the training dataset used ImageJ (https://imagej.nih.gov/ij/), or a homemade snippet based on OpenCV.

4. Results and discussion

4.1. Feature engineering and choice of machine learning algorithm

Tests to select the best combination of features and algorithm were conducted, for the Jargalant site, on images extending over 5043 x 4546 pixels. Six feature combinations: (i) R + G + B, (ii) H + S + V, (iii) H + S, (iv) CON + HOM + ENT, (v) R + G + B + CON + HOM + ENT, and (vi) R + G + B + CON + HOM + ENT + TPI were evaluated with each of the 9 algorithms listed above. Such a progressive scheme of feature selection aims at identifying the most relevant features in the classification process, thus building a better classifier without computational overload. Results, expressed as precision, recall, F1-Score, and accuracy using an inner cross-validation on the training set are summarized for each combination in Table 1 (see also Supplementary Material SM3 for scores from test set). Pixels were manually picked out on the orthomosaic to train the model with 363 positive and 1010 negative cases. Such a dataset may appear oversized and time-consuming for a procedure requiring manual operation. This is especially true when repeated several times for large sites, covering several square kilometres. However, the objective here was to evaluate the influence of the size of the training set, in terms of performance, which therefore required a vast number of samples.

RGB as input features. Results appear to be good, most of the time > 93%, up to 98.5%, whatever the algorithm applied to the RGB information alone (Table 1). These good scores are all the more remarkable in that they include some rather basic procedures, such as naive Bayes, nearest centroid, and, to a lesser extent, linear discriminant analysis, which is among the four best results (Table 1). It is worth mentioning that the passage of clouds during aerial picture acquisition resulted in colour inhomogeneity, especially in the northern half of the orthomosaic (cf. Fig. 1). This could have affected the classification performance based on colour, but an appropriate training set circumvented this problem. The SVM procedure was finally selected, because it slightly outperformed the other procedures. Another reason for this choice is the capacity of SVM, in comparison with the other algorithms, to train relatively quickly at this level of quality, although convergence between the training and cross-validation scores was not reached, even when 1000 samples were processed for training (Fig. 3a). In terms of archaeological output, the map produced by SVM is acceptable (Fig. 3b), but several bare soil areas, corresponding to car tracks, were erroneously classified as dry stones. As a result, serious manual cleaning would be necessary before producing usable documentation.

HSV or HS as input features. When HSV channels were applied as alternative input features to RGB, results were, at best, similar to those obtained by RGB (e.g. SVM, RF, ANN), or exhibited worse performances (e.g. NB, LR, LDA, QDA) (Table 1). Using H and S alone could be a better option, because of the gain in robustness (for H and S) against illumination changes (V) [47], whereas the RGB colour space is generally sensitive to this parameter (Sural et al. 2002). This was not the case here, as the results obtained from H and S (not shown) decreased dramatically in comparison with those based on HSV, similarly to results already observed in other circumstances [26]. As a consequence, the RGB colour space will be preferred in the following.

Texture features alone. The combination of contrast, homogeneity, and entropy yields an excellent classification for all algorithms, always surpassing RGB by a few percent, which is noticeable at this

level of performance (Table 1). The logistic regression model was selected for its efficiency, and because convergence is reached quickly, after 300-400 total samples, at a level of accuracy close to 99.3% (Fig. 3c). This procedure provides a map where car tracks have now disappeared, demonstrating the power of texture features for classification (Fig. 3d).

Combining both texture features and RGB. Such a combination should take into account both spatial arrangement of the tonal information, and spectral distribution of light; two sets of features which are not fully independent because the former is, in part, deduced from the latter, but which should describe two different sides of the image. This might push classification capabilities a little further [48]. Table 1 confirms this expectation, as performances increased slightly, almost systematically (note however that these improvements cannot be statistically demonstrated, considering confidence intervals). The classifier trained quickly (Fig. 3e), and several isolated false positives have vanished from the resulting map, although such an improvement is difficult to see at the scale of the document (Fig. 3f).

Combining texture features, RGB and TPI. Stone mounds exhibit positive relief with respect to their immediate surroundings, while hollows are expected for furrows resulting from agricultural activities, or car tracks in the steppes. As a consequence, information derived from topography might efficiently enrich the feature set. When TPI is introduced as a new input variable, together with RGB colour and texture, performance scores tend to improve by a further few tenths of a percent (Table 1). Random forest, logistic regression, and SVM slightly surpass other approaches, and produce scores that generally exceed 99.5%. However, attention must be paid to training speed in addition to performance scores, because the more rapidly the model learns, the less time-consuming will be the supervision step. Here, even for the most favourable case, at least 400 samples were necessary to reach convergence between the training and cross-validation scores (not shown). Hard voting was therefore tested, using the results of the KNN, SVM, LR, LDA, RF and ANN classifiers, in the hope of reducing this number, if possible, without impacting score quality. This ensemble-based method

reached correct convergence for only 200 training samples (Fig. 3g), while maintaining scores at their highest levels: > 99.5% for precision, F1-Score, and accuracy (Fig. 3h displays the resulting map). Note that gradient boosting and adaboost (i.e. adaptive boosting), which seek to transform a set of weak learners into strong learners [49], were also tested. Both methods produced scores comparable to those of hard voting, but at a lower learning speed. As a consequence, hard voting was preferred.

4.2. Visual evaluation of the Jargalant output map

Introducing complexity into a model to improve classification performance by only a few percent or tenths of a percent in comparison with the use of RGB alone might appear, at first glance, not really relevant, but purely academic. It must, however, be kept in mind that a gain of only 0.1% in terms of accuracy corresponds to more than 20 000 pixels in an image of more than 20 Mpix. Any improvement, even minimal, may therefore save a considerable amount of time during post-processing, so that efforts must be made in this direction. Fig. 4a displays the original orthomosaic of Jargalant, overlain by a vector layer corresponding to a polygonised black and white map, obtained by hard voting (i.e. Fig. 3h), and a close-up of two particular areas (Fig. 4bc). At this scale, the quality of the output map is undeniable. All archaeological structures composed of accumulations of dry stones were precisely delineated, while bare soils were not misclassified as stones (see Fig. 4bc), except in very few instances in the north-east (arrow 1), on the car track in the south-east (arrow 2), and for the livestock enclosure in the west of the orthomosaic (arrow 3). Some positive cases were also delineated in the fields, but they correspond to actual stones raised to the surface by ploughing.

4.3. Performance of the method for a large site and evaluation of the operational framework

At Tsatsiin Ereg, archaeological structures cover several square kilometres. However, only nine tiles, of about 10 ha each, corresponding to the B10 complex are presented here (Fig. 5a), because this

area has already been the object of a precise survey by two topographers, equipped with a total station during 2 missions, each lasting 1 month (Fig. 5b). Each of the 9 orthomosaics was produced from about 100 pictures acquired in 2016, and hard voting was applied to RBG + texture + TPI, following the procedure described above. About 200-250 samples per tile for both positive and negative cases were selected manually, so that, without the test subset, the total number of training samples finally used for learning was around 300; a value sufficient to reach an acceptable convergence between training and cross-validation scores.

During the past 3000 years, rain erosion has almost certainly led to the accumulation of a thin slope-wash layer made of granitic arena, but as it is only 5-10 cm thick, even small stones can be recognized. Dry stone structures are therefore perfectly visible in the steppes and can be correctly delineated. The method proposed clearly outperforms pedestrian surveys. It appears more precise, partly because human error during topographic surveys of an area very dense in anthropogenic structures can be avoided. From flight programming to photo processing, the gain in working time for the operator is considerable (cf. Table 2). Only 15-20 minutes of flight are necessary to cover an area of 400 x 400 m^2, which is remarkable, considering that time spent in the field is the main limiting factor for massive recording. Interestingly, the operator can also evaluate orthomosaics the same day, to detect possible technical problems, and so perform the operation again if necessary. Note that several isolated rocks are recorded. Depending on the final objective, these positive cases could be easily removed, either by hand (taking typically less than 30 min), or automatically using an algorithm, taking into account the local density of positive cases, and/or the distance to the nearest positive neighbours.

One final test was performed to examine if the time spent acquiring the training dataset could be significantly reduced. Tiles n°5 and n°3 were treated with their specific training dataset, but also with that acquired to process tile n°4 (Fig. 6). Good results were anticipated for n°5, since the images were captured the same day, approximately at the same time, and thus in the same lightning and topsoil

moisture conditions, but there was some doubt about tile n°3, which exhibits greener images (like tile n°6), probably because the images were taken after a rainy event. In fact, both treatments produce comparable outcomes (Fig. 6ab and Fig. 6cd), because texture variables are less prone to variation than colour variables. However, it must be noticed that, when training used samples from tile n°4, more false positives were observed for tile n°5, while a few stones were missed on tile n°3. Again, depending on the final objective, the operator will be able to decide between optimal accuracy, but a relatively time-consuming training step, or simplified training by picking pixels from only a single tile.

5. Conclusion

In the context of large archaeological sites, covering several hectares, with little vegetation, acquiring images by drone, with treatment by appropriate methods, is a very effective solution for further automatic archaeological mapping. The low cost and simple logistics, especially in remote field conditions, undeniably argue for this type of aerial photogrammetry. The method proposed for treating the data generally surpasses pedestrian surveys , as it is almost fully automatic, rapid, and accurate, while a traditional record by GPS or total station is time-consuming, and may lead to errors difficult to avoid when archaeological structures are small and numerous. Acquisition speed is a strong asset, as one of the most limiting factors is the time spent in the field, especially for studies undertaken overseas, where field campaigns are often time-constrained. With the increasingly high resolution of images, and technological progress making it possible to collect hundreds of images for each flight, applying machine learning algorithms becomes indispensable. The operator intervenes significantly during two crucial steps. The first is the manual selection of the training dataset. In the examples presented, it was very important to sample different kinds of stones, but also grass, car tracks, bare soil, etc., because if the training data is inadequate or not adequately representative, poor classification results are to be expected. The second step concerns feature engineering. While the selection of the best learning model and hyperparameter tuning can be performed almost

automatically, the workflow consisting in reformatting, processing, enriching, calibrating, and finally selecting features requires some experience. In the present case, the use of colour information in the RGB colour space, three texture parameters among those available, and one feature derived from the topography produced suitable outputs for almost all classifiers tested. That is probably a good start for undertaking such mapping in other circumstances, but it is likely that some adjustments will be necessary to attain optimal results. Depending on the final objective, it may be interesting (or not) to keep isolated stones. In our examples, their position might be useful for cultural heritage preservation, by better understanding of stone displacements caused by livestock perambulation or water runoff. By contrast, for studying the spatial organisation of archaeological structures, these isolated stones can be removed manually, or alternatively by using an appropriate algorithm. For the sake of completeness, it must however be mentioned that this study takes place in a part of the world where conditions are optimal: the archaeological structures are not masked by vegetation, and they have almost never been buried or disturbed since their construction. Such a situation is rarely met in other environments, probably making the application of the proposed workflow more difficult. Finally, although binary pixel classification was here proved to perform well, other extremely powerful approaches, such as deep learning for object detection, should also be tested in the near future.

6. Acknowledgements

This research was funded by the Join mission Mongolia – Monaco, and the project ROSAS (uB-FC and RNMSH). We are grateful for helpful comments by an anonymous reviewer and the editor, which have greatly improved the manuscript.

7. References

[1] J. Bourgeois, M. Meganck, Aerial photography and archaeology 2003. A century of information. Archaeological Reports Ghent University Ghent: Academia Press. 4 (2005).

[2] D.N. Riley, The technique of air-archaeology, Archaeol. J. 101 (1946) 1-16.

[3] R.S. Solecki, Practical aerial photography for archaeologists, Am. Antiq. 22 (1957) 337-351.

[4] N.G. Smith, L. Passone, S. al-Said, M. al-Farhan, T.E. Levy, Drones in archaeology: integrated data capture, processing, and dissemination in the al-Ula Valley, Saudi Arabia, Near East. Archaeol. 77 (2014) 176-181.

[5] D.C. Cowley, C. Moriarty, G. Geddes, G.L. Brown, T. Wade, C.J. Nichol, UAVs in context: archaeological airborne recording in a national body of survey and record, Drones 2 (2018) 2.

[6] I. Aicardi, F. Chiabrando, A. Lingua, F. Noardo, Recent trends in cultural heritage 3D survey: The photogrammetric computer vision approach. J. Cult. Herit. 32 (2018) 257–266.

[7] S. Campana, Drones in archaeology. State-of-the-art and future perspectives, Archaeol. Prospect. 24 (2017) 275-296.

[8] A. Traviglia, A. Torsello, Landscape pattern detection in archaeological remote sensing, Geosci. 7 (2017) 128.

[9] J. Magail, Tsatsiin Ereg, site majeur du début du 1er millénaire en Mongolie, Bull. Musee Anthrop. Prehist. Monaco 48 (2008) 107–120.

[10] W.W. Fitzhugh, The Mongolian deer stone-khirigsuur complex: dating and organization of a Late Bronze Age menagerie. In: J. Bemmann, H. Parzinger, E. Pohl & D. Tseveendorzh (Eds.) Current Archaeological Research in Mongolia, (Bonn Contributions to Asian Archaeology 4.) Bonn: vfgarch.press uni-bonn, 2009 183–99.

[11] J. Magail, F. Monna, Y. Esin, J. Wilczek, C. Yeruul-Erdene, J.-O. Gantulga, Application de la photogrammétrie à la documentation de l'art rupestre, des chantiers de fouilles et du bâti. Bull. Musee Anthrop. Prehist. Monaco 56 (2017) 69-92.

[12] A.R. Gansella, J.-W. van de Meent, S. Zairis, C.H. Wiggins, Stylistic clusters and the Syrian/South Syrian tradition of first-millennium BCE Levantine ivory carving: a machine learning approach. J. Archaeol. Sci. 44 (2014) 194-205.

[13] C. Hörr, E. Lindinger, G. Brunnett, Machine learning based typology development in archaeology, JOCCH 7 (2014) 2.

[14] J. Wilczek, F. Monna, M. Gabillot, N. Navarro, L. Rusch, C. Chateau, Unsupervised model-based clustering for typological classification of Middle Bronze Age flanged axes, J. Archaeol. Sci. Rep. 3 (2015) 381-391.

[15] G. Barone, P. Mazzoleni, G.V. Spagnolo, C. Raneri, Artificial neural network for the provenance study of archaeological ceramics using clay sediment database. J. Cult. Herit. 38 (2019) 147-157.

[16] X. Zhang, J. Cui, W. Wang, C. Lin, A study for texture feature extraction of high-resolution satellite images based on a direction measure and gray level co-occurrence matrix fusion algorithm, Sensors 17 (2017), 1474.

[17] A. Kobler, S. Dzeroski, I. Keramitsoglou, Habitat mapping using machine learning-extended kernel-based reclassification of an Ikonos satellite image, Ecol. Model. 191 (2006) 83–95.

[18] C.J. Abolt, M.H. Young, A.L. Atchley, C.J. Wilson, Rapid machine-learning-based extraction and measurement of ice wedge polygons in high-resolution digital elevation models, Cryosphere 13 (2019) 237-245.

[19] S. Raschka, Python machine learning, Packt Publishing 2015, 454 pp.

[20] T. Turbat, J. Bayarsaikhan, D. Batsukh, N. Bayarkhuu, Deer Stones of the Jargalantyn Am, Oulan-Bator, Mongolian Tangible Heritage Association NGO 2011, 192 pp.

[21] J. Magail, J.-O. Gantulga, C. Yeruul-Eredene, M. Tsengel, Inventaire et relevés des pierres à cerfs de Tsatsiin Ereg, Bull. Musee Anthropol. Prehist. Monaco 50 (2010) 77–114.

[22] F. Monna, Y. Esin, J. Magail, L. Granjon, N. Navarro, J. Wilczek, L. Saligny, S. Couette, A. Dumontet, C. Chateau, Documenting carved stones by 3D modelling – Example of Mongolian deer stones, J. Cult. Herit. 34 (2018) 116-128.

[23] Y. Esin, J. Magail, C. Yeruul-Erdene, J.-O. Gantulga, Au sujet des traces de peintures sur les stèles ornées de Mongolie de la fin de l'âge du Bronze et du début de l'âge du Fer, Bull. Musee Anthropol. Préhist. Monaco 58 (2018) 145–156.

[24] J.C. Leachtenauer, R. Driggers, Surveillance and reconnaissance imaging systems: modeling and performance prediction, Artech House, 2001, 416 pp.

[25] G. Verhoeven, Taking computer vision aloft–archaeological three-dimensional reconstructions from aerial photographs with photoscan, Archaeol. Prospect. 18 (2011) 67–73.

[26] H. Seong, H. Son, C. Kim, A comparative study of machine learning classification for color-based safety vest detection on construction-site images, KSCE J. Civ. Eng. 22 (2018) 4254–4262.

[27] S. Sural, G. Qian, S. Pramanik, Segmentation and histogram generation using the HSV color space for image retrieval. ICIP 2002, DOI:10.1109/ICIP.2002.1040019

[28] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural Features for Image Classification, IEEE Transactions on Systems, Man and Cybernetics 3 (1973) 610-620.

[29] M. Hall-Beyer, Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales, Int J Remote Sens. 38 (2017) 1312-1338.

[30] I. Vrbik, S.J. Van Nest, P. Meksiarun, J. Loeppky, A. Brolo, J.J. Lum, A. Jirasek, Haralick texture feature analysis for quantifying radiation response heterogeneity in murine models observed using Raman spectroscopic mapping, PLoS One (2019) 1-12.

[31] L.-K. Soh, C. Tsatsoulis, Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens. 37 (1999) 780–795.

[32] J. De Reu, J. Bourgeois, M. Bats, A. Zwertvaegher, V. Gelorini, P. De Smedt, W. Chu, M. Antrop, P. De Maeyer, P. Finke, M. Van Meirvenne, J. Verniers, P. Crombé, Application of the topographic position index to heterogeneous landscapes, Geomorphology 186 (2013) 39–49.

[33] J.C. Gallant, J.P. Wilson, Primary topographic attributes. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis: Principles and Applications. Wiley, New York, 2000, pp. 51–85.

[34] S.Y. Kung, Kernel methods and machine learning. Cambridge University Press, 2014, 572 pp.

[35] C.M. Bishop, Pattern recognition and machine learning. Springer-Verlag New York Inc. 2006, 738 pp.

[36] B. Lantz, Machine learning with R. Packt Publishing. 2nd edition, 2015. 452 pp.

[37] P. Domingos, M. Pazzani, On the optimality of the simple bayesian classifier under Zero-One loss. Mach. Learn. 29 (1997) 103–130.

[38] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning. Springer-Verlag New York Inc, 2013. 426 pp.

[39] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, John Wiley & Sons, USA. 3rd ed. 2013. 528 pp.

[40] P. A. Lachenbruch, M. Goldstein, Discriminant analysis. Perspectives in biometry, Biometrics 35 (1979) 69-85.

[41] C. Cortes, V. Vapnik, Support-vector networks. Mach. Learn. 20 (1995) 273-297.

[42] M. Hofman. Support vector machines - Kernel and the kernel trick. Houptseminar report 2006.

[43] T. Hastie, R. Tibshirani, J. Friedman. The elements of statistical learning, Springer, 2nd ed. 2008. 745 pp.

[44] M. Kuhn, K. Johnson, Applied predictive modeling, Springer. 2013. 600 pp.

[45] M. Piragnolo, A. Masiero, F. Pirotti, Open source R for applying machine learning to RPAS remote sensing images. Open geospatial data, software and standards 2 (2017) 16.

[46] S. Raschka, Python machine learning. Packt Publishing. 2nd edition, 2015. 454 pp.

[47] B.D. Zarit, B.J. Super, F.K.H. Quekc Comparison of five color models in skin pixel classification. ICCV'99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece, 1999, pp. 58-63.

[48] S. Rathore, M. Hussain, M.A. Iftikhar, A. Jalil, Ensemble classification of colon biopsy images based on information rich hybrid features, Comput. Biol. Med. 47 (2014) 76–92.

[49] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119-139.

Figure caption

Figure 1: Scheme of feature engineering. Colour as RGB and HSV colour space; texture features as HOM (homogeneity), ENT (entropy), and CON (contrast). TPI for topographic position index, deduced from DEM, displayed as colour hill-shaded map.

Figure 2: The machine-learning pipeline for binary pixel classification

Figure 3: Feature selection and best machine learning algorithm for the site of Jargalant. Four feature combinations are presented: RGB, texture (CON + HOM + ENT), RGB + texture, RGB + texture + TPI. Left: training and cross-validation scores with their 95% confidence interval as a function of the size of the training dataset; right: output map with the model selected.**Figure 4**: Original orthomosaic of Jargalant overlain by a vector layer corresponding to a polygonised black and white map obtained by hard voting (A); two close-ups (B and C). Three arrows (1-3) point out specific areas.

Figure 5: Comparison between delineation of stones obtained by machine learning (a) and the map resulting from pedestrian survey by topographers equipped with a total station (b) for the funeral structure B10, Mongolia. The figure (a) displays 9 orthomosaics, noted 1-9, overlain by vector layers corresponding to polygonised black and white maps, obtained by hard voting.

Figure 6: Comparison between delineation of stones obtained by a model trained using a dataset specifically acquired for tiles n°5 and n°3, respectively (a) and (c), and delineation obtained with a common training dataset acquired from tile n°4 (b) and (d).

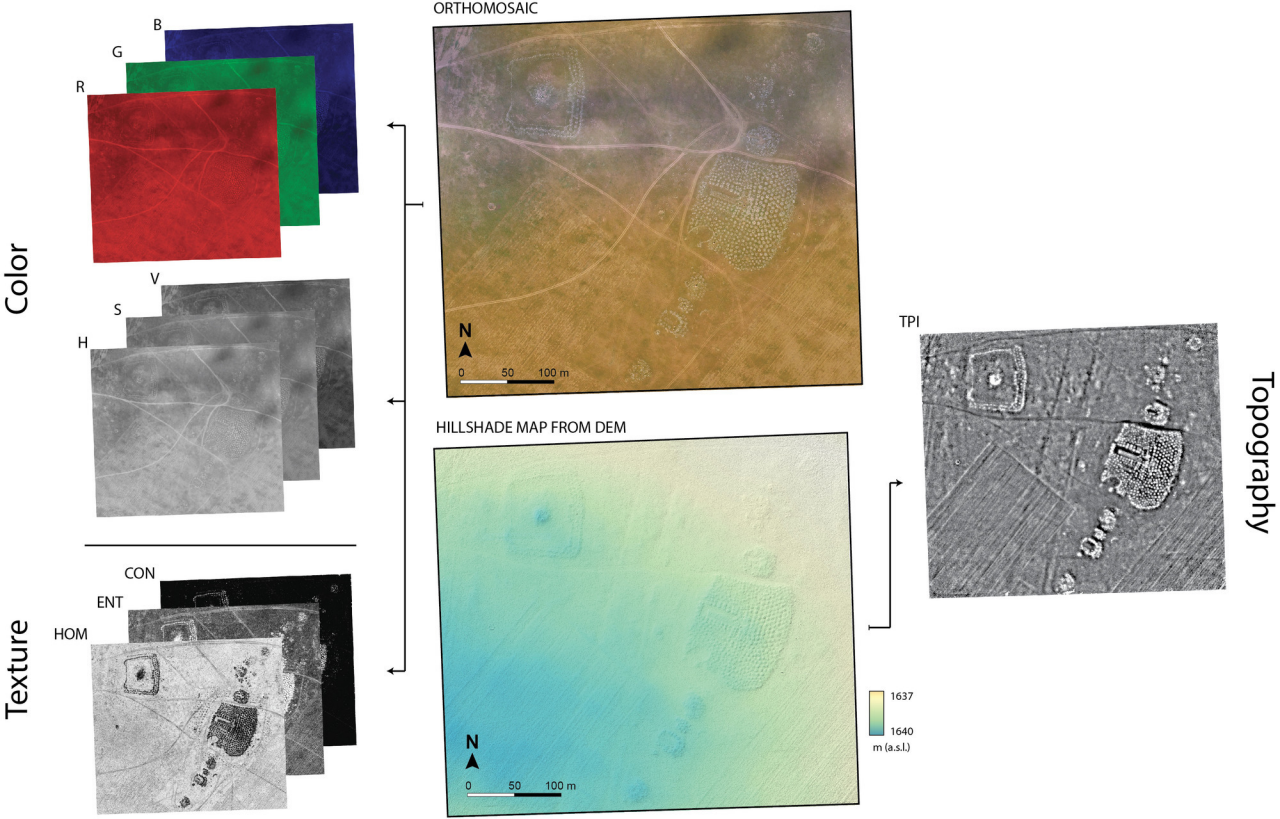
		RGB	HSV	Texture	Texture + RGB	Texture + RGB + TPI
NB	precision	90.3 ± 2.0	88.1 ± 4.1	94.8 ± 2.0	94.8 ± 2.0	94.6 ± 3.5
	recall	96.7 ± 1.1	91.7 ± 3.4	98.9 ± 1.0	99.2 ± 0.7	99.5 ± 0.7
	F1-Score	93.4 ± 1.2	89.8 ± 2.4	96.8 ± 0.8	96.9 ± 0.9	96.9 ± 2.0
	accuracy	96.4 ± 0.7	94.5 ± 1.4	98.3 ± 0.5	98.3 ± 0.5	98.3 ± 1.1
NC	precision	87.6 ± 2.5	84.9 ± 3.9	96.0 ± 1.1	96.6 ± 1.3	98.1 ± 2.3
	recall	91.1 ± 1.2	98.1 ± 1.4	97.8 ± 0.7	99.4 ± 0.7	99.4 ± 1.1
	F1-Score	92.8 ± 1.8	91.0 ± 2.6	96.9 ± 0.7	98.0 ± 0.6	98.8 ± 1.5
	accuracy	95.9 ± 1.0	94.8 ± 1.6	98.3 ± 0.4	98.9 ± 0.3	99.3 ± 0.8
KNN	precision	98.0 ± 1.5	96.4 ± 2.1	99.4 ± 1.1	99.7 ± 0.5	99.2 ± 1.1
	recall	93.4 ± 3.6	94.2 ± 2.8	97.0 ± 1.6	98.6 ± 1.5	99.2 ± 1.1
	F1-Score	95.6 ± 1.9	95.3 ± 2.3	98.2 ± 1.0	99.2 ± 0.7	99.2 ± 0.8
	accuracy	97.7 ± 1.0	97.5 ± 1.2	99.1 ± 0.5	99.6 ± 0.4	99.6 ± 0.4
LR	precision	97.2 ± 1.4	91.9 ± 2.7	**99.4 ± 1.1**	**99.7 ± 0.6**	99.4 ± 1.1
	recall	96.4 ± 1.9	93.1 ± 2.6	**98.3 ± 1.0**	**98.9 ± 0.6**	99.2 ± 1.1
	F1-Score	96.4 ± 1.3	92.5 ± 2.4	**98.9 ± 0.9**	**99.3 ± 0.4**	99.3 ± 1.1
	accuracy	98.3 ± 0.7	96.0 ± 1.3	**99.4 ± 0.5**	**99.6 ± 0.2**	99.6 ± 0.6
LDA	precision	98.0 ± 0.7	88.9 ± 3.5	98.3 ± 1.0	99.2 ± 1.1	99.2 ± 1.1
	recall	95.9 ± 2.5	96.4 ± 2.2	96.4 ± 1.9	98.6 ± 0.0	98.9 ± 1.0
	F1-Score	96.9 ± 1.5	92.5 ± 2.7	97.4 ± 1.2	98.9 ± 0.6	99.0 ± 0.8
	accuracy	98.0 ± 0.8	95.9 ± 1.6	98.6 ± 0.6	99.4 ± 0.3	99.5 ± 0.4
QDA	precision	94.7 ± 2.3	88.3 ± 3.2	95.3 ± 1.3	95.8 ± 1.9	96.1 ± 2.5
	recall	97.0 ± 2.0	92.8 ± 2.9	98.9 ± 1.0	99.4 ± 0.7	99.7 ± 0.6
	F1-Score	95.8 ± 1.8	90.5 ± 2.6	97.0 ± 0.7	97.6 ± 0.9	97.9 ± 1.4
	accuracy	97.7 ± 1.0	94.8 ± 1.4	98.4 ± 0.4	98.7 ± 0.5	98.8 ± 0.8
SVM	precision	**98.6 ± 1.5**	98.4 ± 1.6	99.5 ± 1.1	98.4 ± 1.6	98.9 ± 1.6
	recall	**96.1 ± 2.0**	96.4 ± 2.6	98.4 ± 1.0	98.6 ± 1.5	99.4 ± 1.1
	F1-Score	**97.3 ± 1.6**	97.3 ± 1.4	98.9 ± 0.3	98.5 ± 0.7	99.2 ± 1.3
	accuracy	**98.6 ± 0.8**	98.6 ± 0.7	99.4 ± 0.2	99.2 ± 0.4	99.6 ± 0.7
RF	precision	95.8 ± 1.6	96.4 ± 2.6	99.4 ± 1.1	99.5 ± 0.7	99.7 ± 0.6
	recall	94.2 ± 3.5	94.2 ± 2.9	97.8 ± 0.7	99.2 ± 1.1	99.5 ± 0.7
	F1-Score	95.0 ± 2.3	95.3 ± 2.5	98.6 ± 0.8	99.3 ± 0.6	99.6 ± 0.6
	accuracy	97.4 ± 1.1	97.5 ± 1.3	99.3 ± 0.4	99.6 ± 0.3	99.8 ± 0.3
ANN	precision	97.0 ± 1.6	96.2 ± 2.3	99.4 ± 0.7	98.9 ± 1.0	98.1 ± 1.8
	recall	96.1 ± 2.0	95.3 ± 1.9	97.5 ± 1.6	98.9 ± 0.6	98.9 ± 1.0
	F1-Score	96.5 ± 1.3	95.7 ± 0.7	98.5 ± 1.1	98.9 ± 0.5	98.5 ± 1.3
	accuracy	98.2 ± 0.7	97.7 ± 0.4	99.2 ± 0.5	99.4 ± 0.3	99.2 ± 0.7
Voting hard	precision					**99.6 ± 0.6**
	recall					**99.7 ± 0.3**
	F1-Score					**99.6 ± 0.5**
	accuracy					**99.7 ± 0.4**

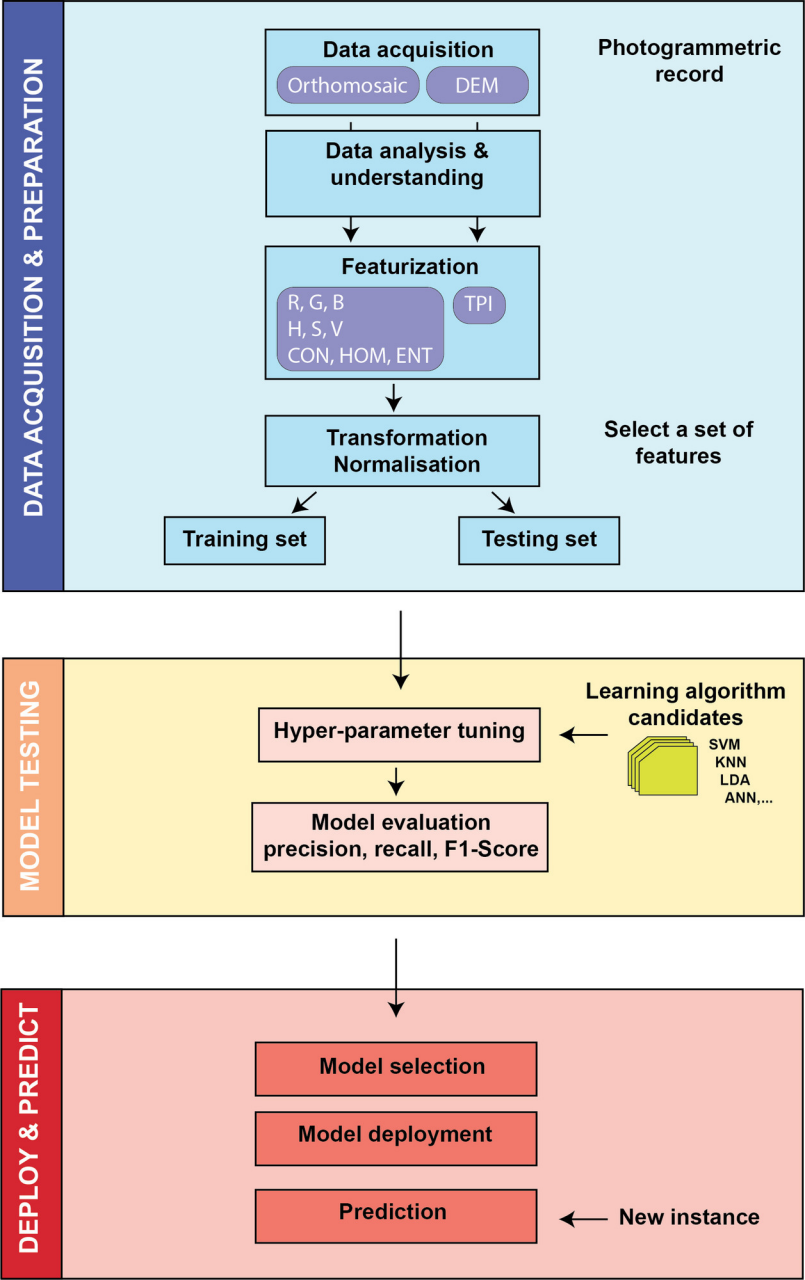
Table 1: Scores and standard deviation (obtained using an inner cross-validation on the training set) of combinations between features and machine learning algorithms for the site of Jargalant. Note

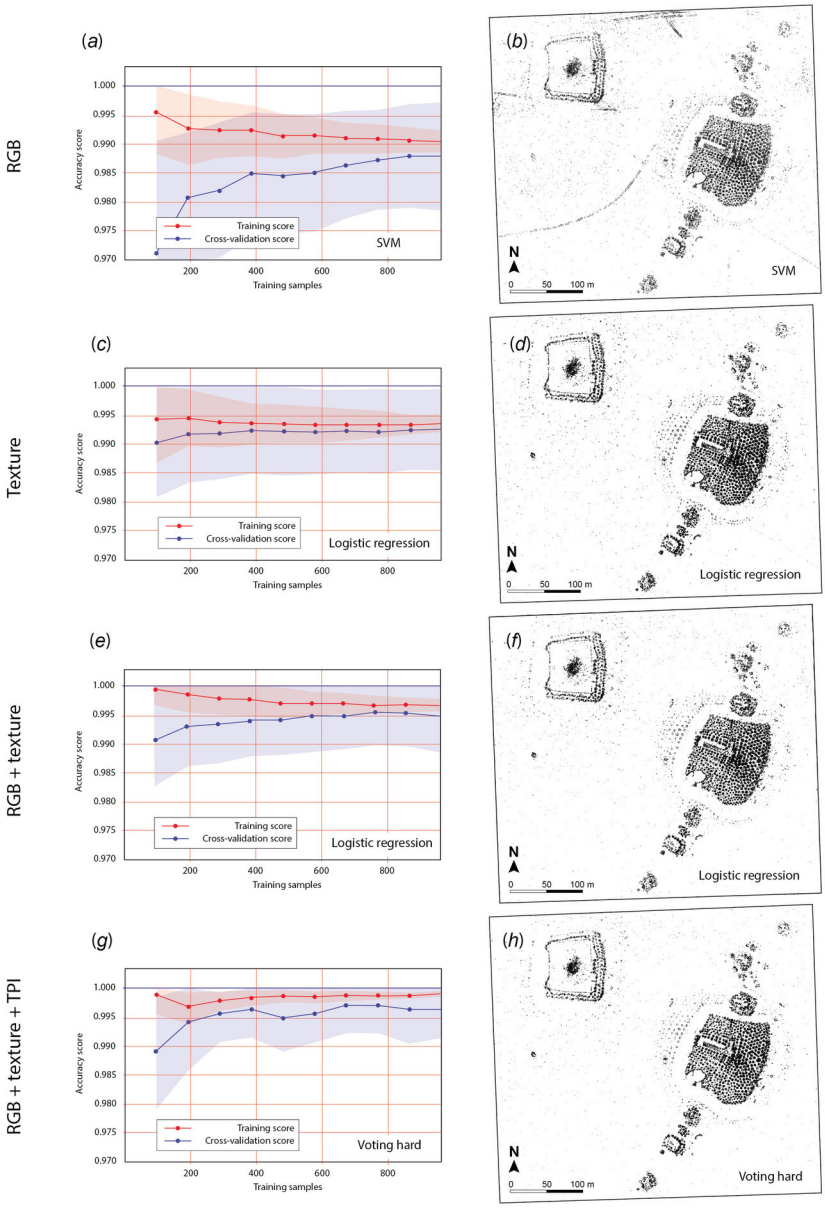
that for such high values, close to 100%, the standard deviation calculation is no longer correct. It is nonetheless provided for comparison purposes. NB for Naive Bayes, NC for nearest centroid, KNN for *k*-nearest neighbours, LR for logistic regression, LDA and QDA for linear and quadratic discriminant analyses, SVM for support vector machine, RF for random forest, ANN for artificial neural network. Texture features encompass contrast, homogeneity, and entropy. In bold, the results corresponding to the four feature combinations: RGB, texture (CON + HOM + ENT), RGB + texture, RGB + texture + TPI, presented in Fig. 3.

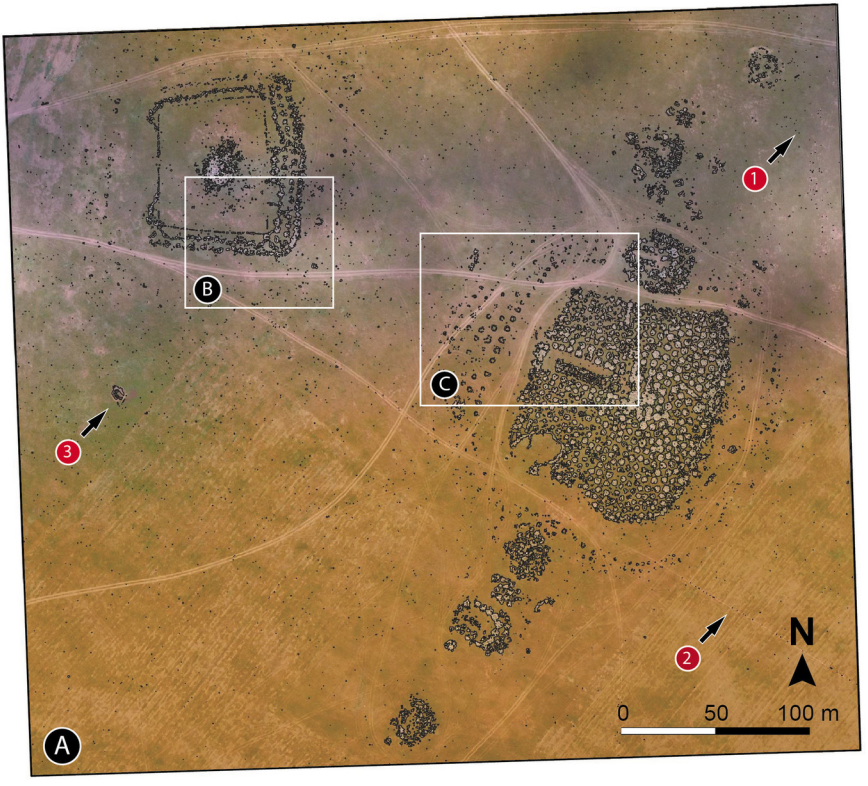
	Time consumed
Picture acquisition by drone (automatic)	15 min
Production of DEM and orthomosaic by photogrammetry (automatic)	2h30 – 3h00
Production of texture maps (automatic)	40 min
Production of TPI map (automatic)	5 min
Point selection for training dataset (manual)	10 min
Classification by hard voting and construction of vector layers (automatic)	15 min
Post-production & wrong polygon cleaning (manual)	< 30 min*

Table 2: Time necessary for processing one tile of about 300 x 300 m^2 (final images of ca. 40 Mpix). These times must be compared to the two months needed for two people in the field to record only the B10 structure (see Fig. 5). *: maximum time provided for cleaning vectorized outputs of Jergalant. Note that this value may be different, depending on the area targeted. Both DEMs and orthomosaics were produced using a computer equipped with an i7 5960X CPU, 64 Go of RAM, and two NVIDIA GeForce GTX 980 mounted in SLI.


Color
B
G
R
V
S
H
Texture
CON
ENT
HOM
ORTHOMOSAIC
N
0 50 100 m
HILLSHADE MAP FROM DEM
N
0 50 100 m
1637
1640
m (a.s.l.)
TPI
Topography

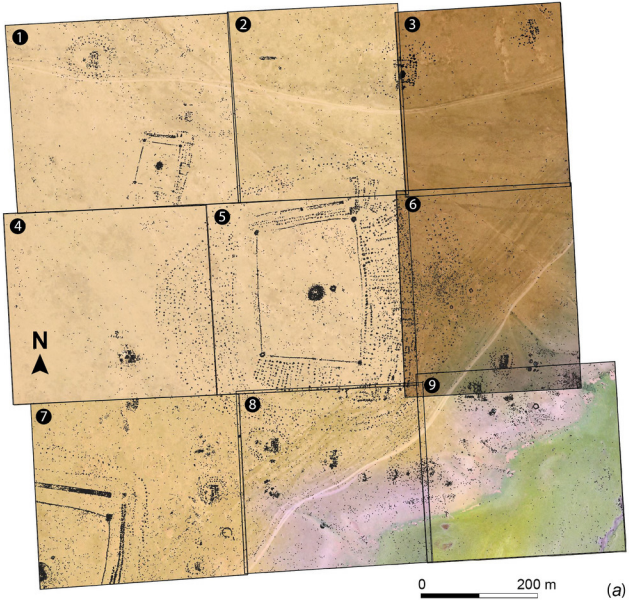

DATA ACQUISITION & PREPARATION
Data acquisition
Orthomosaic
DEM
Photogrammetric record
Data analysis & understanding
Featurization
R, G, B
H, S, V
CON, HOM, ENT
TPI
Transformation Normalisation
Select a set of features
Training set
Testing set
MODEL TESTING
Learning algorithm candidates
Hyper-parameter tuning
SVM
KNN
LDA
ANN,...
Model evaluation precision, recall, F1-Score
DEPLOY & PREDICT
Model selection
Model deployment
Prediction
New instance


RGB
Texture
RGB + texture
RGB + texture + TPI
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Accuracy score
Training samples
Training score
Cross-validation score
SVM
Logistic regression
Voting hard
N
0 50 100 m


A
B
C
1
2
3
N
0
50
100 m

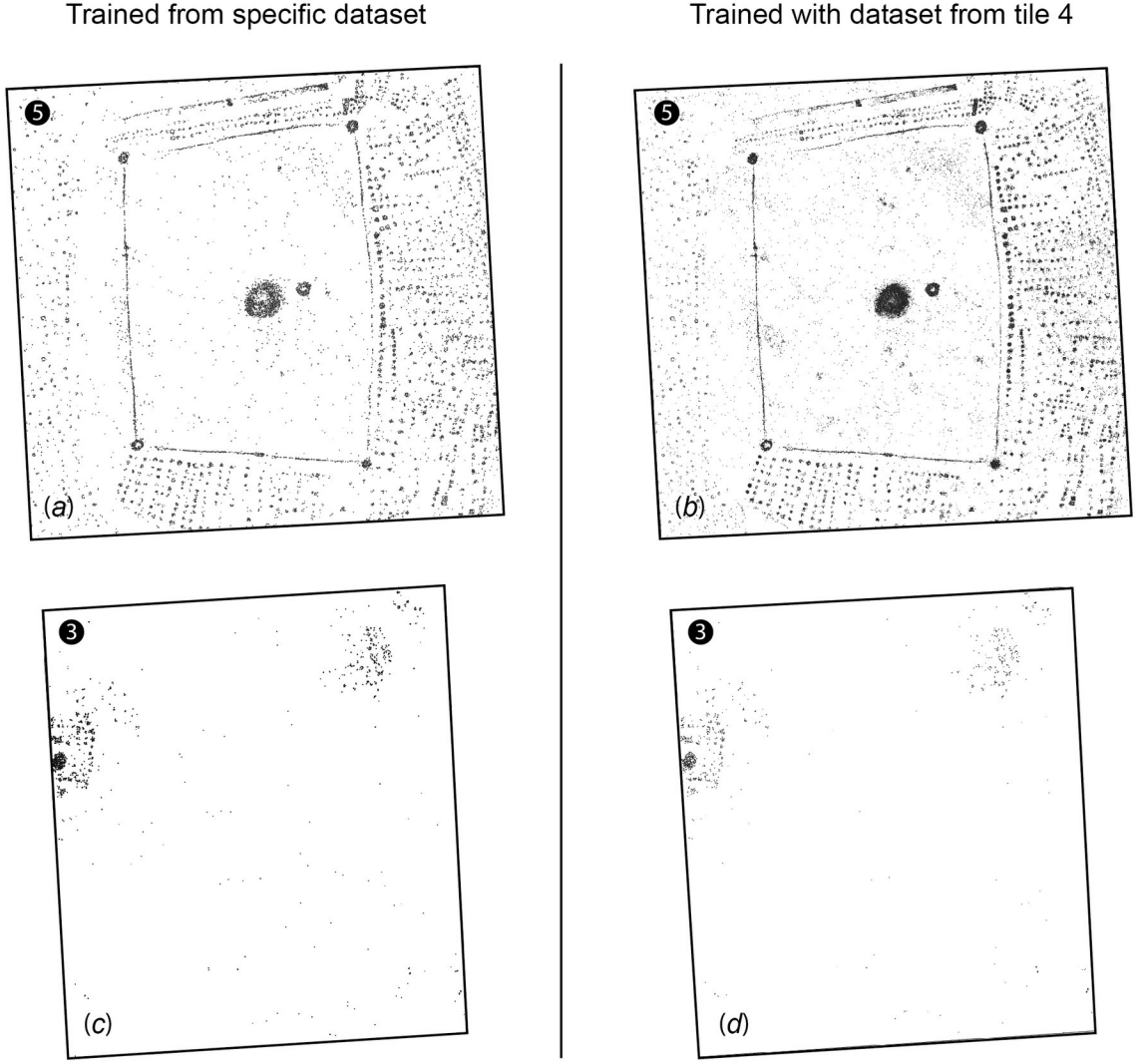

B


C



(*a*)

(*b*)


Trained from specific dataset
Trained with dataset from tile 4
5
(a)
5
(b)
3
(c)
3
(d)