Archeologia e Calcolatori
34.2, 2023, 51-68
doi 10.19282/ac.34.2.2023.03

DEFINING SOUTHERN ETRURIA FINAL BRONZE AGE SETTLEMENT MODELS USING AN INTEGRATED GIS AND MACHINE LEARNING APPROACH

1. Introduction

Southern Etruria has been extensively studied in recent Italian protohistory (di Gennaro 1979, 1982, 1986, 1988; Potter 1985; Negroni Catacchio 1995; Pacciarelli 2001; Belardelli *et al.* 2007; Schiappelli 2008; Barbaro 2010). In terms of historical reconstruction, this region is crucial during the transition between the Bronze and Iron Age (around 950/925 BC, Pacciarelli 2001, 67-69), when a significant change in settlement patterns occurred. The change entailed a transition from a polycentric village system, typical of the Bronze Age, to proto-urban centers (Peroni 1989, 2004; Bietti Sestieri 2010; Cardarelli 2018), sites of future Etruscan cities (Bartoloni 1989, 2012; Pacciarelli 2001, 12).The study of the last phase of the Bronze Age (Final Bronze Age, 1150-950/925 BC, Pacciarelli 2001, 67-69) in this region contributes to our understanding of the reasons for this change.

The discovery and study of these contexts (primarily settlements) have been undertaken by the Roman protohistoric school and its scholars, who have developed a comprehensive framework for historical reconstruction (Schiappelli 2008, 21-28; Barbaro 2010, 17-19). The topographical and territorial study, the position of settlements and their relationship to the surrounding area can be a valuable methodological support for the reconstruction of protohistoric society (Potter 1985, 65-106; Pacciarelli 2001, 71; di Gennaro 2010, 13-16). In this sense, digital analyses can be of valuable support: while GIS needs no introduction as its use in archaeology is almost customary by now (Scianna, Villa 2011), the application of Machine Learning (ML) is becoming more and more present in archaeology (Bickler 2021) thanks also to the availability of opensource and well-documented resources which have certainly enabled the integration of disciplines in recent years.

The aim of this paper is to propose an approach that combines GIS raster analysis with ML techniques to identify formal or quantitative characteristics of the Final Bronze Age (FBA) settlements in Southern Etruria. To achieve this goal, a specific pipeline is proposed, which includes raster morphological analyses, simulations and techniques derived from ML and data analysis. The characteristics of the FBA settlements will be defined through quantitative and reproducible raster analyses and compared with those derived from a simulation of random points within the same territory. This comparison will

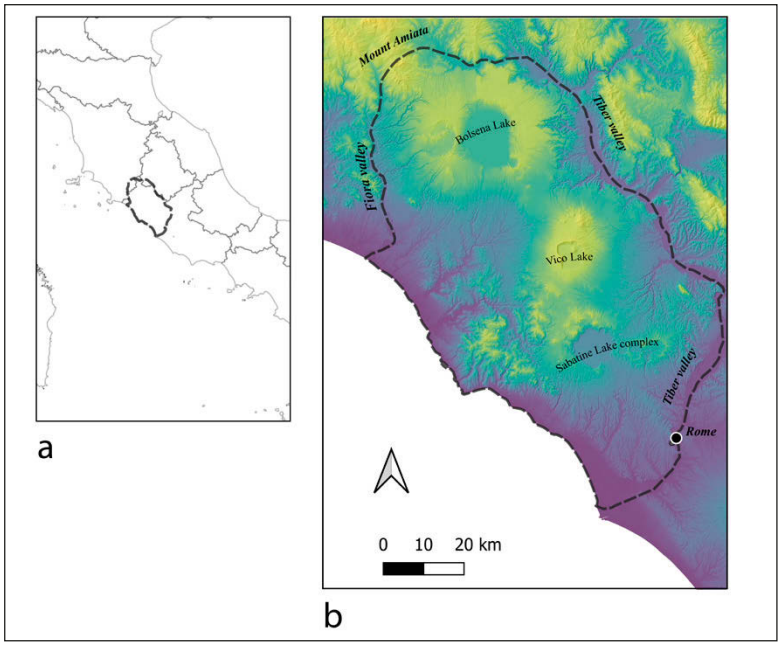

Fig. 1 – a) Territory of Southern Etruria with the current regional administrative boundaries; b) DEM of the territory of Southern Etruria. The boundaries and main geographical features mentioned in the text and the location of the city of Rome are indicated.

help us distinguish between the common characteristics in the area and those that are specific to the FBA's own settlement choices.

2. Dataset

The territory of Southern Etruria is located between the Fiora valley to the W, Amiata Mount to the N, and the Tiber River to the E and S (Fig. 1b) (di Gennaro 1986, 7-8; Barbaro 2010, 19). It mainly falls within the present-day regions of Latium, Tuscany, and Umbria (Fig. 1a). B. Barbaro (2010) comprehensively analysed evidence about the region to create a cohesive picture. She extensively collected relevant archaeological evidence, resulting in a topographical classification of settlements (Barbaro 2010, 27-35), a chronology based mainly on decorative elements (Barbaro 2010, 71-118), and an extensive catalogue of settlements, hoards, and funerary areas (Barbaro 2010, 147-330). The commonly accepted settlement pattern in the region involves sites located on plateaus with very steep flanks, which are considered an effective form of natural defence (Pacciarelli 2001, 72-74;

Barbaro 2010, 27-36). These sites are typically 5-6 km apart and cover an average of 5 to 6 hectares (Cardarelli 2018, 374).

There are exceptions to this rule, such as perilacustrian settlements or settlements located in open positions (Barbaro 2010, 33-35). However, the 'natural defenced' model is predominant: M. Pacciarelli (2001, 100) calculates that over 90% of FBA settlements (in a sample of 70 sites he analysed) are located on plateaus with very steep flanks. A. Schiappelli (2008) and B. Barbaro (2010) developed measures to quantify the 'defensive potential' of settlement contexts in their respective monographic works. In both cases, more than half of the settlements were attributed a high 'defensive potential'.

Regarding funerary practices, burials consist of cremations within biconical urns with covers. Grave goods are scarce, suggesting that the ritual followed extremely strict rules (De Angelis 2006). According to these considerations, A. Cardarelli (2018, 375) describes this territorial organization as polycentric, with villages having similar socio-political structures and governed by elites in potential competition with each other.

The sites analysed in this study were drawn from the monographic work of B. Barbaro (2010). The data used includes settlement contexts, as classified in Barbaro's work: those referred to as 'settlement' (insediamento) or 'probable settlement' (probabile insediamento) in the Barbaro 2010 classification and catalogue (Barbaro 2010, 147-330), for a total of 166 contexts (Fig. 2a). The sites were positioned as points in the UTM coordinates indicated by the author (Barbaro 2010, 150).

The analyses were performed using the Tinitaly DEM (https://tinitaly.pi.ingv.it/), a digital terrain model with a resolution of 10 m. The cells W47070, W47075, W46570, W46575 and W46075 (https://tinitaly.pi.ingv.it/Download_Area2.html) were combined into a single raster. The analysis area was generated by creating a circular buffer of 1 km around each site and using it to cut out the raster (Figs. 3, 4, 5). This resulted in a series of small, circular DEMs centred around each point. This procedure was mainly necessary to reduce the processing load without losing information by resampling the raster.

To identify the characteristics of the FBA settlements in Southern Etruria, a simulation was conducted by placing random points within the same territory. The rationale behind this approach is as follows: if there is a specific settlement pattern in this region, as proposed by numerous authors, most archaeological sites should exhibit specific attributes. However, it is important to ensure that these attributes are not merely common features shared with the surrounding territory but are indeed unique to the settlement pattern. To address this issue, a comparison with random sites was performed. These simulated sites were distributed randomly over the territory to identify common characteristics, which can be considered as a sort of background noise. This comparison allowed for the identification of unique features specific to

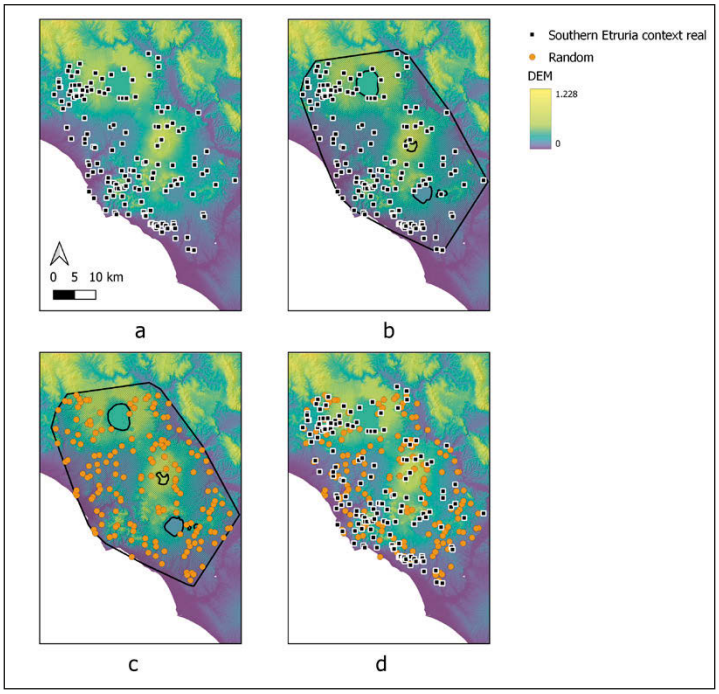

Fig. 2 – a) Position of the 166 archaeological contexts within the database; b) definition of the delimitation polygon created from the archaeological contexts; c) random points within the delimitation polygon; d) archaeological contexts and random points compared.

the settlement pattern of the FBA. To position the random points, a delimitation polygon was created relative to the real archaeological contexts and the main bodies of water, such as the Sabatine Lake complex, Lake Bolsena and Lake Vico, were excluded from the polygon to prevent random points from ending up within these areas (Fig. 2b). Thus, random points were generated amounting to the same number of archaeological contexts (166), and randomly positioned within the polygon (Fig. 2c). This method allowed for a fair comparison between archaeological and simulated sites (Fig. 2d).

3. Methods and tools

3.1 *Raster analysis*

To achieve the objectives of this study, various raster analyses were carried out to identify features that can be used as predictors in the ML algorithm (Tab. 1). These predictors, developed using the open source GIS software SAGA

(https://saga-gis.sourceforge.io/), are grouped into macro-classes, including Morphometry (Fig. 3), Lighting and visibility (Fig. 4), Channels and hydrology, and Terrain classification (Fig. 5). This division allows us to organise the results of the analyses and interpret their meaning in the context of the study. Each algorithm used returns a raster grid, and the results of the analysis can be either quantitative or continuous (Morphometry, Lighting and visibility, Channels and hydrology) or categorical (Terrain classification).

To prepare the data for analysis, each point was linked to the corresponding raster value by a spatial join. This resulted in a multivariate dataset

PREDICTOR	**TOOL**
Morphometry	
Slope	Basic Terrain Analysis[1]
Aspect	
Relative Slope Position	
Normalised Height	Relative Heights and Slope Positions[2]
Standardised Height	
Mid-slope Position	
Terrain Ruggedness Index (TRI)	Terrain Ruggedness Index[3]
DEM	*None*
Lighting and visibility	
Visible sky	Sky View Factor[4]
Sky View Factor	
Average View Distance	
Topographic Positive Openness	Topographic Openness[5]
Topographic Negative Openness	
Channels and hydrology	
Valley Depth	Basic terrain analysis[1]
Channel Network Distance	
Topographic Wetness Index (TWI)	
Terrain classification	
TPI based Landforms	TPI based Landforms[6]
Geomorphons	Geomorphons[7]

Tab. 1 – Table showing the various raster predictors divided by macroclass. The software tool used and its documentation are given in the footnotes.

[1] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_compound_0.html.

[2] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_morphometry_14.html.

[3] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_morphometry_16.html.

[4] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_lighting_3.html.

[5] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_lighting_5.html.

[6] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_morphometry_19.html: the tool returns the following values: 1: Streams; 2: Midslope Drainage; 3: Upland Drainage; 4: Valleys; 5: Plains; 6: Open Slopes; 7: Upper Slopes; 8: Local Ridges; 9: Midslope Ridges; 10: High Ridges.

[7] https://saga-gis.sourceforge.io/saga_tool_doc/8.4.1/ta_lighting_8.html: the tool returns the following values: 1: Flat; 2: Peak; 3: Ridge; 4: Shoulder; 5: Spur; 6: Slope; 7: Hollow; 8: Footslope; 9: Valley; 10: Pit.

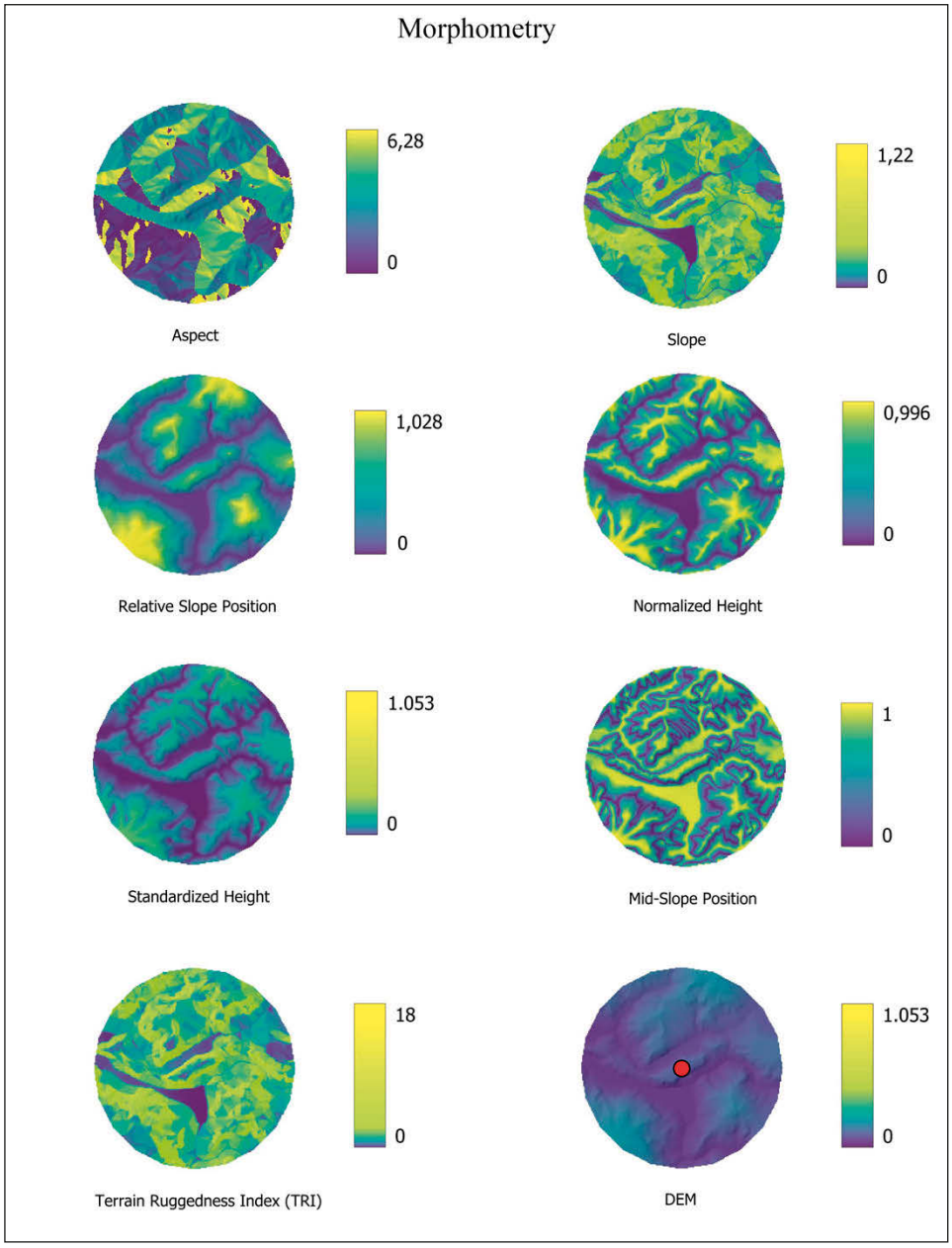

Fig. 3 – Illustration of the raster predictors related to the morphometry macro-class used for the analysis. The example context is Luni sul Mignone and the displayed area corresponds to the 1 km buffer around the site (its position is indicated by the red marker in the DEM grid).

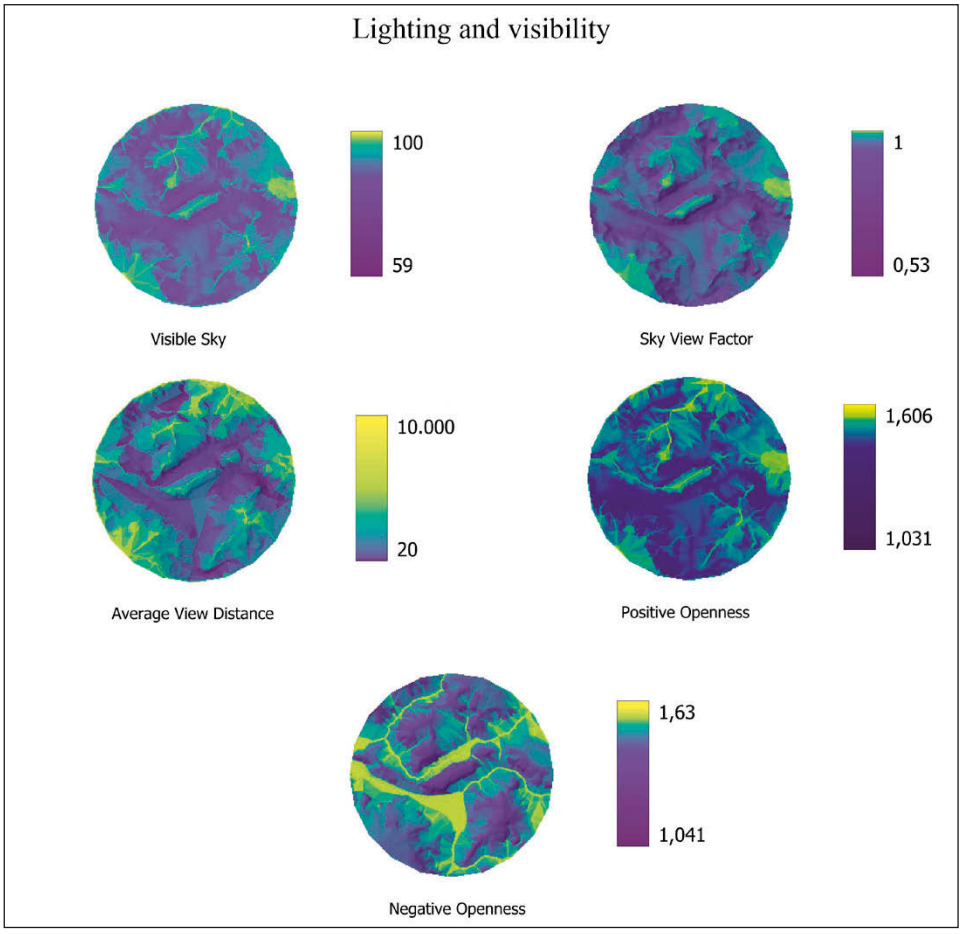

Fig. 4 – Illustration of the raster predictors related to the Lighting and visibility macro-class used for the analysis.

where each row represents a point (archaeological site or random) and each column represents a raster analysis value (Tab. 2).

3.2 *Defining most important features: a Machine Learning based pipeline*

As already mentioned in the introduction and aims of this paper, the objective of this article is to use a quantitative approach to identify morphological characteristics of settlements in Southern Etruria. To achieve this, we will use methods specific to data analysis and ML. Feature Selection (FS) is a common technique in ML that involves selecting a subset of the most relevant

index	id	name	type	Visible Sky	…	Valley Depth	TWI	TPI Landforms	Geomorphon
0	15	Sorano-Castelvecchio	real	89,30	…	54,14	2,15	9	3
1	10	Pitigliano	real	99,38	…	4,43	5,25	9	3
2	13	Monte Rosso	real	90,26	…	12,95	5,52	7	6
3	16	Sovana	real	98,91	…	3,40	6,32	9	3
4	4	Il Gaggio	real	93,63	…	68,94	6,94	4	10
5	11	Le Sparne di Poggio Buco	real	96,32	…	0,53	4,64	9	5
6	6	Meletello	real	94,13	…	18,03	4,86	6	7
7	7	Monte Tellere	real	98,94	…	0,00	3,68	10	3
…	…	…	…	…	…	…	…	…	…
324	0	random	random	98,60	…	2,04	6,18	5	4
325	0	random	random	98,04	…	3,06	6,17	5	6
326	0	random	random	95,83	…	14,17	8,90	6	6
327	0	random	random	99,36	…	2,00	9,05	5	6
328	0	random	random	98,14	…	4,39	7,54	5	6
329	0	random	random	95,49	…	12,41	7,00	6	6
330	0	random	random	96,63	…	45,67	6,65	5	6
331	0	random	random	99,22	…	42,21	13,46	5	1

Tab. 2 – Example table with some archaeological sites and random points with the values of each predictor from the spatial join operation.

features to use in the construction of a predictive model (Liu 2010; James *et al.* 2013, 204). Unlike dimensionality reduction methods, such as PCA or more recent methods like t-SNE or UMAP (Cardarelli, Lapadula 2022), FS does not create new variables, but simply selects the most important ones from the existing data. While FS is typically performed before training a model, Feature Importance (FI) is performed after training a model and it is used to evaluate the relative importance of different features in the context of a specific model (Hastie *et al.* 2009, 593-595; Saarela, Jauhiainen 2021).

To support the explanation within the text, a diagram of the proposed pipeline is provided (Fig. 6). The various steps within the procedure are indicated by a dashed box in the diagram and will be described in the following text. The application of preliminary FS (1st step, Fig. 6) involves the correlation threshold (Saeys *et al.* 2007; Van Hulse *et al.* 2012; Tang *et al.* 2014). This is a conceptually simple but powerful method for reducing redundancy and noise in the dataset: if two measures are highly correlated, it means that one of the variables can be used to predict the other. This means that, within the large number of variables in the dataset, it is possible that some of them define the same phenomenon. Therefore, within each raster macro-class, we will calculate the correlations between the variables and eliminate those that are highly correlated.

To account for the possible non-linear nature of the correlations, we will use the Spearman correlation coefficient (Spearman's ρ). This is a measure of

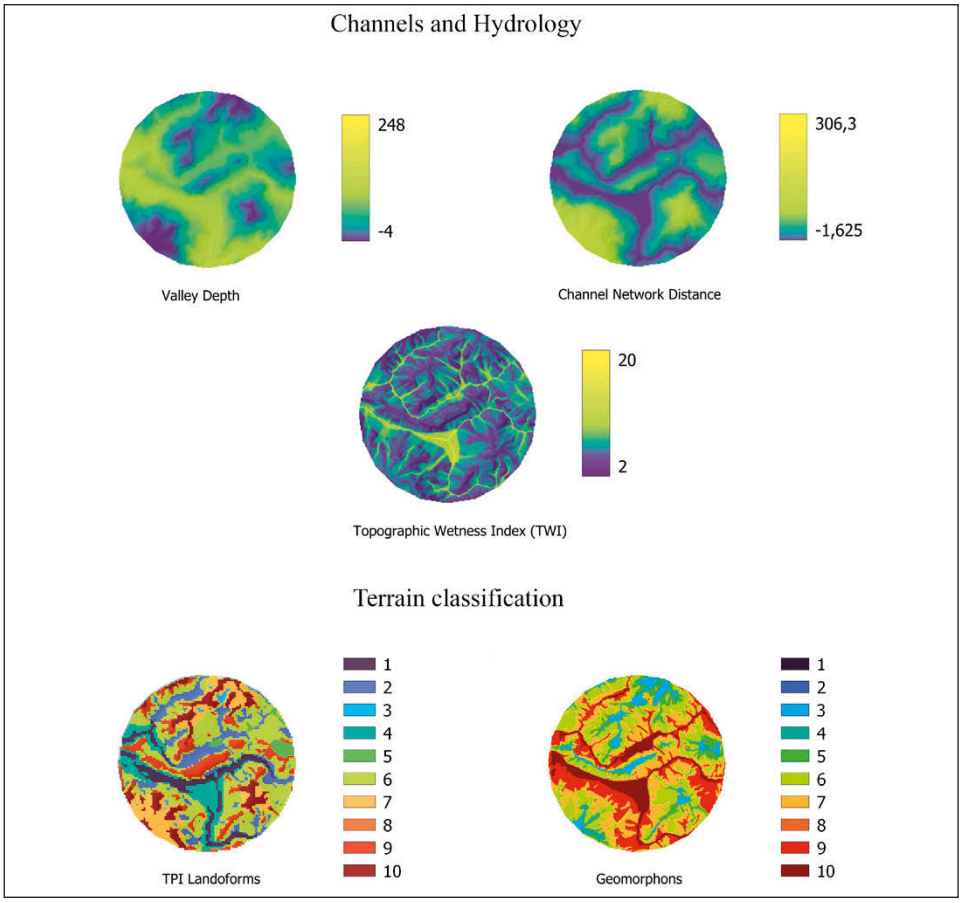

Fig. 5 – Illustration of the raster predictors related to the Channels and hydrology and Terrain classification macro-classes used for the analysis.

the statistical dependence between two variables, which considers the ranks of the data rather than the raw values (Spearman 1904; Corder, Foreman 2014, 140-145). This allows us to capture also non-linear relationships between the variables and to identify the most relevant features more accurately. The absolute value of 0.8 (i.e., ± 0.8) was chosen as the limit value, which is generally recognized as an extremely high correlation limit (Schober *et al.* 2018). As this method is based on correlation, it is not possible to apply it to categorical predictors. The second step (2nd step, Fig. 6) involves training the ML model that will subsequently be used to calculate the FI. Prior to training the model, the dataset needs to be divided into a Train Set and a

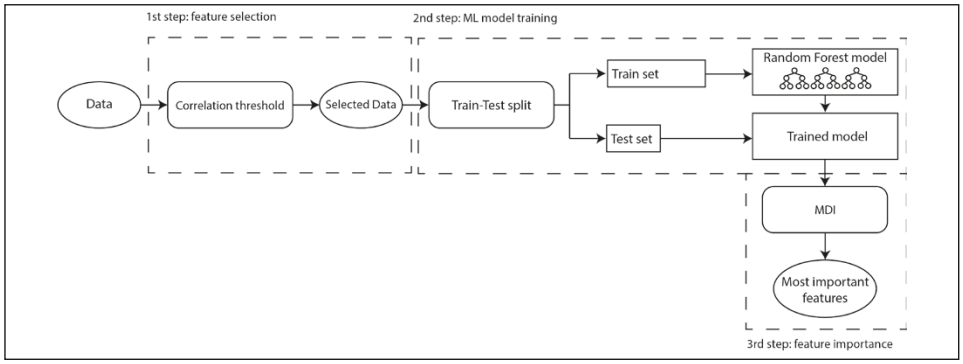

Fig. 6 – Schematic representation of the pipeline used for data analysis. The dashed boxes indicate the various steps described in the text.

Test Set (James *et al.* 2013, 176). This is a common procedure for many ML algorithms and allows the model to be trained on a significant portion of the data (Train Set - 80% of the total) before testing its performance on unknown data (Test Set - 20% of the total).

In this specific case, real and random sites are equally distributed in both the Train and Test sets, ensuring that our results are fair and unbiased. For this analysis, we have chosen to use the Random Forest model (Ho 1998). Random Forest is a ML algorithm that utilizes multiple decision trees to make predictions (Hastie *et al.* 2009, 587-603). The decision tree algorithm creates a tree-like model of decisions and their consequences by splitting the data into smaller subsets based on input variable values until a stopping criterion is met. Each node in the tree represents a decision or a test on an input variable and each branch represents the outcome of that decision or test. The leaves of the tree represent the final output or prediction (Harrington 2012, 37-60).

Although decision trees are intuitive ways to classify or label objects, they tend to overfit, i.e., memorize the training data and fail to predict new data (Harrington 2012, 39), defeating their substantial purpose. Random Forest combats this issue by combining multiple decision trees, each trained on a random subset of the data, to make the final prediction (Vanderplas 2016, 426-433). Both Random Forest and decision trees use criterions called Gini impurity or entropy (Harrington 2012, 40-43; James *et al.* 2013, 311-314) to determine how to split the data at each node in the decision tree. The goal of the algorithm is to minimize the criterion of the subsets created by each split. In this sense, a split that results in subsets with lower criterion is considered better than a split that results in subsets with higher criterion. Regarding the parameters chosen, 1000 decision trees were used in the proposed Random Forest model and the Gini impurity was chosen as criterion.

After training the model, we move to the last step (3rd step, Fig. 6), determining the FI by using data from the Test Set to identify the most important features that distinguish real sites from simulated sites. In this analysis, we use the mean decrease in impurity (MDI) method. The idea behind this method is that features that are more frequently selected for splitting are more important in predicting the target variable. For each decision tree in the forest, we can calculate the total amount of impurity (e.g., entropy or Gini, *supra*) that is decreased by splitting on a particular feature. We can then average these values across all trees in the forest to obtain an estimate of the feature importance (Hastie *et al.* 2009, 593-595). In this case, for the calculation of the FI, it was chosen to treat quantitative and categorical measures separately.

To summarize the proposed model, redundant variables will be eliminated during the initial FS step. Then, the ML model will be trained on the remaining data. The trained model will be used to calculate the most important features using the FI process, which will help to identify the most significant characteristics of the settlement pattern of Southern Etruria in the FBA.

3.3 *Hardware and software*

The supplementary material includes details on the hardware and software used, as well as the raw data in XLSX format (the table containing values resulting from spatial join between real/random site and raster analysis), along with the analysis procedure with commented code, tables, and graphs (http://www.archcalc.cnr.it/indice/suppl-material/34.2/3/Cardarelli_2023_supplementary.zip).

4. Results

The results of correlation threshold process for FS are exemplified using the visibility macro-class as an example. In this case, we can visualize the relationship between all variables using a scatterplot matrix (Fig. 7). From the scatterplot matrix, we can point out that some measures are highly correlated, such as Positive Openness and Visible Sky, which have a linear relationship. Other measures, such as Sky View Factor and Visible Sky, exhibit a non-linear relationship, such as an exponential or logarithmic relationship. Then, the correlations between the variables are quantified using the Spearman's coefficient and a matrix can be used to visualise them.

In the case of the visibility measures (Fig. 8a), the highly correlated measures (with an absolute correlation value greater than 0.8) are Positive Openness/Visible Sky, Positive Openness/Sky View Factor, and Visible Sky/Sky View Factor. Positive Openness can be eliminated because it appears twice as a highly correlated feature, and in the case of Visible Sky/Sky View Factor, it is preferable to eliminate Sky View Factor because its average correlation

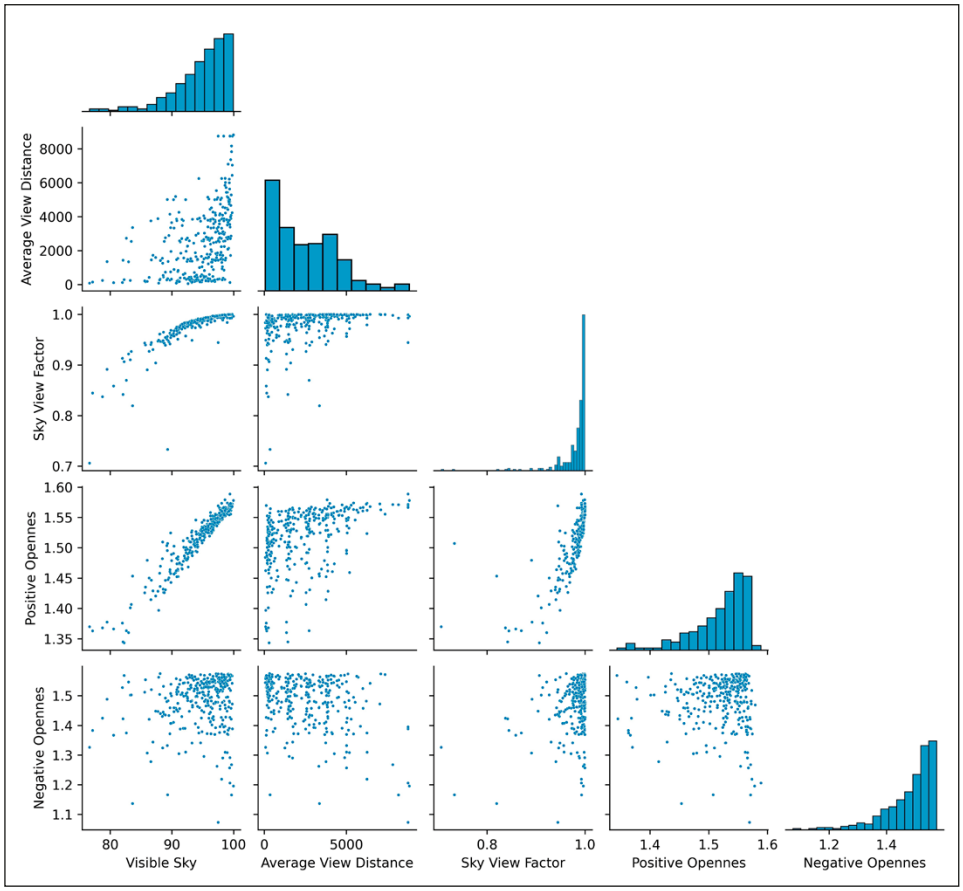

Fig. 7 – Scatterplot matrix related to the Lighting and visibility macro-class. Related measures are clearly visible. On the diagonal are histograms relating to each univariate variable.

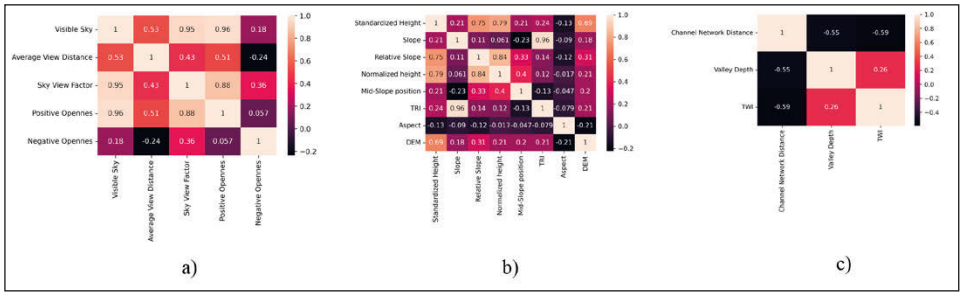

Fig. 8 – a) Correlation matrix for the Lighting and visibility macro-class; b) correlation matrix relating to Morphometry macro-class; c) correlation matrix relating to Channels and hydrology macro-class.

with the other features is higher (0.722 vs 0.724). By applying the same process to the morphometric measurements, we can identify and eliminate the features Normalized Height and Terrain Ruggedness Index (TRI) (Fig. 8b). In the case of the channels and hydrology measures, lastly, there are no highly correlated measures (Fig. 8c). As previously mentioned, this process is not applied to qualitative measures, such as those related to Terrain classification. After training the Random Forest model, we can measure model performance through accuracy, i.e., the percentage of predictions (real or simulated site) that are corrected by the model.

The model trained on quantitative data had an accuracy score of 0.65, meaning that it was able to correctly classify 65% of the data, while the model trained on categorical data obtains an accuracy of 0.71 (71%) suggesting that categorical data seems to be able to discriminate more accurately between real and simulated data. After this verification, we can finally move on to feature importance (3rd step, Fig. 6). For a better interpretation of the results, it was decided to create a random variable that serves as a threshold for importance. Features that exceed this threshold are considered important, while those that fall below it are considered random and basically insignificant within the model. In terms of quantitative measures, five features exceed and overcome the random variable threshold: Negative Openness, Mid-Slope Position, Channel Network Distance, TWI, Visible Sky, Valley Depth, Slope, Average View Distance and Relative Slope (Fig. 9a).

However, only the first four predictors were found to be effective in correctly discriminating real sites from random ones. This conclusion was drawn from boxplots that visually represented the values of the predictors (Drennan 2010, 37-41). These plots showed the difference in values between the real and simulated sites. The variables that differed markedly from the importance value obtained from the random variable (Negative Openness, Mid-Slope Position, Channel Network Distance, TWI, Visible Sky) and one below the random limit (DEM) were entered within the grid of boxplots. The predictors Negative Openness (Fig. 10, 1), Mid-Slope Position (Fig. 10, 2), TWI (Fig. 10, 3) and Channel Network Distance (Fig. 10, 4) exhibited relevant differences at the interquartile range level. On the other hand, the difference using the variable Visible Sky mainly manifested itself at the level of outliers, which were more present in the random sites, particularly with respect to low value levels (Fig. 10, 5).

The last boxplot (Fig. 10, 6) represented a variable considered unimportant (that of the DEM). The boxplots were extremely similar and overlapped, making it impossible to discriminate real sites from random ones. Based on these results, the characteristics of the real sites were identified by having lower Negative Openness values (indicating they are more topographically closed), higher Mid-Slope Position values (indicating they are further from the slopes), lower TWI values (indicating that they are located in areas sheltered

from waterlogging), and higher Channel Network Distance values (indicating that they are positioned above the underlying hydrological network) than the random sites. When analysing qualitative categorical data (Fig. 9b), we need to consider each value individually within the model, rather than comparing them as we would with quantitative data. This allows us to determine which individual categories within the data are the most important, rather than trying to compare the overall importance of different data sets.

In this case, the most informative variables are TPI Landforms (Middle Ridge) value 9, TPI Landforms (Plains) value 5, TPI Landforms (High Ridge) value 10, and Geomorphon (Ridge) value 3. These categories characterise the archaeological sites, except for TPI Landforms (Plains) value 5, which appears to describe simulated sites instead (Fig. 11).

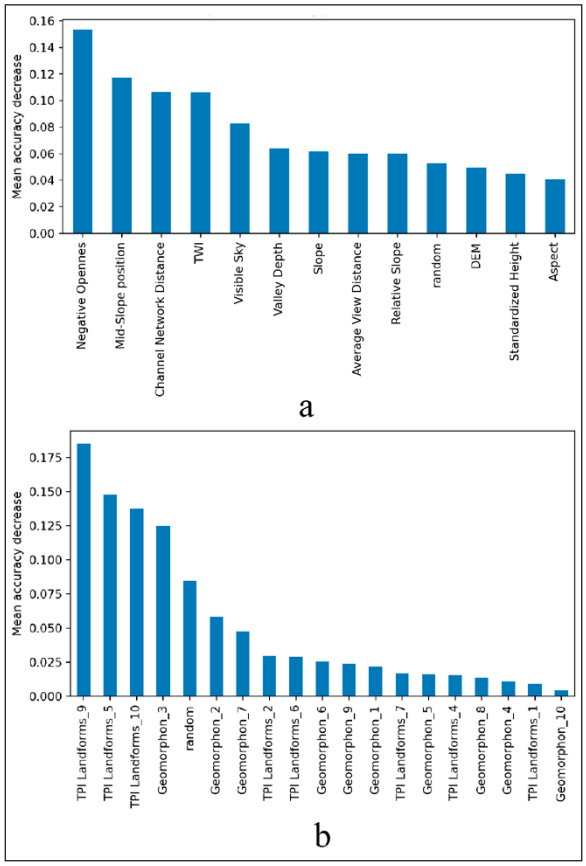

Fig. 9 – Feature importance for continuous (a) and categorical (b) values.

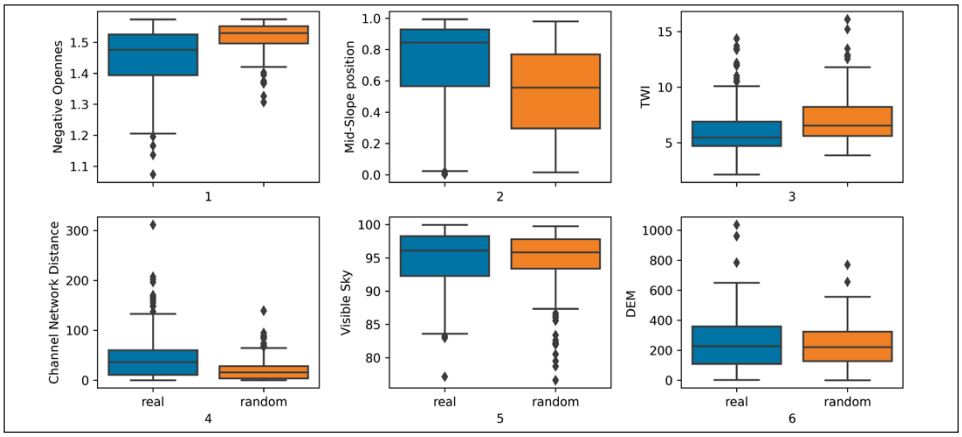

Fig. 10 – Series of boxplots showing the difference between the various 'important' continuous measurements (1-5) and 'unimportant' measure (6).

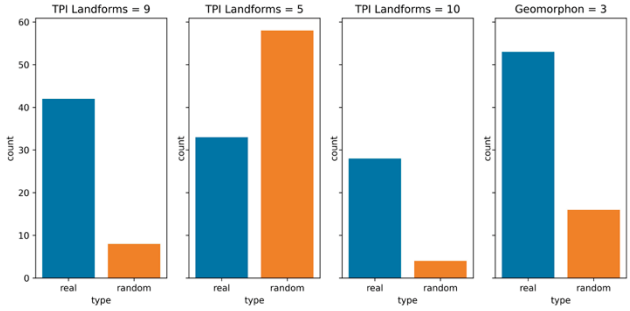

Fig. 11 – Series of bar graphs showing the number of sites per category for each 'important' categorical variable.

5. Discussion and conclusion

The creation of numerous raster predictors allows us to build a comprehensive database for the creation of objective morphologic measurements. SAGA is an excellent solution for this purpose as it is open source and offers a wide range of powerful algorithms. By integrating this raster-database with the dataset of archaeological and random contexts and using the Random Forest model and the FI process, we are able to identify the attributes that characterise the actual FBA settlements in Southern Etruria: by comparing these attributes with random points within the same landscape, we can discern the characteristics that are specific to the settlement pattern and not influenced

by the overall landscape morphology. As far as the characterization of the macro-classes is concerned, the average impact of the various features is highest for the classes referring to visibility (0.097) followed by hydrology (0.090) and finally the general morphology of the territory (0.06), suggesting a strong impact of the visibility concept for the characterization of the analysed sites.

The features selected in this work can be linked to the settlement pattern model described empirically for the FBA in Southern Etruria: settlements are located in a prominent position with respect to the underlying hydrographic network (Channel network distance, TWI), are not on slope (Mid-slope position) and occupy mainly summit positions (informative variables for TPI Landforms and Geomorphons). In contrast, more intuitive measures such as altitude above sea level (DEM) do not seem to be significant factors in characterizing the settlement pattern. Specifically, the Negative Openness measure is the variable that best discriminates real from simulate site (Fig. 9a) and seems to describe extremely effectively the concept of 'defensive potential' (i.e., a closed location in landscape) described by numerous authors and considered as one of the fundamental settlement characteristics in the territory. In fact, the analyses carried out confirm that a settlement model exists, and it is generally well characterized, as the accuracy of the Random Forest model is around 70% for both quantitative and categorical data.

The quantification of such phenomenon, as well as confirming the settlement model proposed by numerous authors, allows for a broader field of investigation. For example, we can generate predictive maps of the territory using the features considered most important or take into account different targets. By using sites from a different chronological phase (e.g., those from the Iron Age) instead of random points, it is possible to identify new characteristics and features that discriminate contexts and can be used as the basis for new historical-archaeological interpretations. In this sense, combining FE and FI is a useful tool that allows us to identify and select the most important features from a dataset regarding a specific objective. In our case, we were able to reduce the number of continuous predictor variables from 14 to 4 and categorical variables from 20 to 3. This not only simplifies the interpretation of the results, but also allows us to conduct further analyses without the need for dimensionality reduction algorithms.

Obviously, the use of this pipeline is not limited to the GIS and geographical ambit but can be used as an excellent alternative to dimensionality reduction methods in any multivariate archaeological dataset.

Lorenzo Cardarelli
Dipartimento di Ricerca e Innovazione Umanistica
Università degli Studi di Bari Aldo Moro
Istituto di Scienze del Patrimonio Culturale - CNR
lorenzo.cardarelli@uniba.it

REFERENCES

Barbaro B. 2010, *Insediamenti, aree funerarie ed entità territoriali in Etruria meridionale nel Bronzo finale*, Firenze, All'Insegna del Giglio.

Bartoloni G. 1989, *La cultura villanoviana. All'inizio della storia etrusca*, Roma, La Nuova Italia Scientifica.

Bartoloni G. (ed.) 2012, *Introduzione all'Etruscologia*, Firenze, Hoepli.

Belardelli C., Angle M., di Gennaro F., Petitti P., Trucco F. (eds.) 2007, *Repertorio dei siti protostorici del Lazio: province di Roma, Viterbo e Frosinone*, Firenze, All'Insegna del Giglio.

Bickler S.H. 2021, *Machine Learning arrives in archaeology*, «Advances in Archaeological Practice», 9, 186-191 (https://doi.org/10.1017/aap.2021.6).

Bietti Sestieri A.M. 2010, *L'Italia nell'età del Bronzo e del Ferro: dalle palafitte a Romolo (2200-700 a. C.)*, Roma, Carocci.

Cardarelli A. 2018, *Before the city. The last villages and proto-urban centres between the Po and Tiber rivers*, «Origini», 42, 359-382.

Cardarelli L., Lapadula A. 2022, *Dimensionality reduction for data visualization and exploratory analysis of ceramic assemblages*, «Archeologia e Calcolatori», 33, 33-52 (https://doi.org/10.19282/ac.33.2.2022.03).

Corder G.W., Foreman D.I. 2014, *Nonparametric Statistics: A Step-by-Step Approach*, Hoboken NJ, Wiley.

De Angelis S. 2006, *Il passaggio tra Bronzo Finale 2 e Bronzo Finale 3 in Etruria meridionale sotto il profilo delle sepolture*, in R. Peroni (ed.), *Studi in onore di Renato Peroni*, Firenze, All'Insegna del Giglio, 581-589.

di Gennaro F. 1979, *Contributo alla conoscenza del territorio etrusco meridionale alla fine dell'età del Bronzo*, in *Atti XXI Riunione Scientifica Istituto Italiano Preistoria e Protostoria*, 267-274.

di Gennaro F. 1982, *Organizzazione del territorio nell'Etruria meridionale protostorica: applicazione di un modello grafico*, «Dialoghi di Archeologia», 2, 102-112.

di Gennaro F. 1986, *Forme di insediamento tra Tevere e Fiora dal Bronzo finale al principio dell'età del Ferro*, Firenze, L.S. Olschki.

di Gennaro F. 1988, *Il popolamento dell'Etruria meridionale e le caratteristiche degli insediamenti tra l'età del Bronzo e l'età del Ferro*, in C. Bettini, G. Colonna, R. Staccioli (eds.), *Etruria meridionale. Conoscenza, conservazione. Atti del convegno (Viterbo 1985)*, Roma, Edizioni Quasar, 59-82.

di Gennaro F. 2010, *Introduzione*, in Barbaro 2010, 13-16.

Drennan R.D. 2010, *Statistics for Archaeologists: A Commonsense Approach*, New York, Springer.

Harrington P. 2012, *Machine Learning in Action*, Shelter Island, Manning Publications Co.

Hastie T., Tibshirani R., Friedman J.H. 2009, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, New York, Springer.

Ho T.K. 1998, *The random subspace method for constructing decision forests*, «IEEE Transactions on Pattern Analysis and Machine Intelligence», 20, 832-844 (https://doi.org/10.1109/34.709601).

James G., Witten D., Hastie T., Tibshirani R. 2013, *An Introduction to Statistical Learning*, New York, Springer (https://doi.org/10.1007/978-1-4614-7138-7).

Liu H. 2010, *Feature selection*, in C. Sammut, G.I. Webb (eds.), *Encyclopedia of Machine Learning*, Boston, Springer US, 402-406 (https://doi.org/10.1007/978-0-387-30164-8_306).

Negroni Catacchio N. (ed.) 1995, *Sorgenti della Nova: l'abitato del Bronzo Finale*, Firenze, Istituto Italiano di Preistoria.

Pacciarelli M. 2001, *Dal villaggio alla città: la svolta protourbana del 1000 a.C. nell'Italia tirrenica*, Firenze, All'Insegna del Giglio.

Peroni R. 1989, *Protostoria dell'Italia continentale: la penisola italiana nelle età del Bronzo e del Ferro*, Roma, Biblioteca di Storia Patria.

Peroni R. 2004, *L'Italia alle soglie della storia*, Roma, Laterza.

Potter T.W. 1985, *Storia del paesaggio dell'Etruria meridionale. Archeologia e trasformazioni del territorio*, Roma, La Nuova Italia Scientifica.

Saarela M., Jauhiainen S. 2021, *Comparison of feature importance measures as explanations for classification models*, «SN Applied Sciences», 3, 2, 272 (https://doi.org/10.1007/s42452-021-04148-9).

Saeys Y., Inza I., Larrañaga P. 2007, *A review of feature selection techniques in bioinformatics*, «Bioinformatics», 23, 2507-2517 (https://doi.org/10.1093/bioinformatics/btm344).

Schiappelli A. 2008, *Sviluppo storico della Teverina nell'età del Bronzo e nella prima età del Ferro*, Firenze, All'Insegna del Giglio.

Schober P., Boer C., Schwarte L.A. 2018, *Correlation coefficients: Appropriate use and interpretation*, «Anesthesia & Analgesia», 126, 1763-1768 (https://doi.org/10.1213/ANE.0000000000002864).

Scianna A., Villa B. 2011, *GIS applications in archaeology*, «Archeologia e Calcolatori», 22, 337-363 (http://www.archcalc.cnr.it/indice/PDF22/AC_22_Scianna_Villa.pdf).

Spearman C. 1904, *The proof and measurement of association between two things*, «The American Journal of Psychology», 15, 1, 72-101 (https://doi.org/10.2307/1412159).

Tang J., Alelyani S., Liu H. 2014, *Feature selection for classification: A review*, in C.C. Aggarwal (ed.), *Data Classification*, New York, CRC Press, 37-64 (https://doi.org/10.1201/b17320).

Vanderplas J.T. 2016, *Python Data Science Handbook: Essential Tools for Working with Data*, Sebastopol, Ed. O'Reilly Media.

Van Hulse J., Khoshgoftaar T.M., Napolitano A., Wald R. 2012, *Threshold-based feature selection techniques for high-dimensional bioinformatics data*, «Network Modeling Analysis in Health Informatics and Bioinformatics», 1, 47-61 (https://doi.org/10.1007/s13721-012-0006-6).

Ward Perkins J.B. 1955, *Notes on Southern Etruria and the Ager Veientanus*, «Papers of the British School at Rome», 23, 44-72 (https://doi.org/10.1017/S0068246200006620).

ABSTRACT

This research aims to use quantitative and repeatable GIS techniques, as well as Machine Learning algorithms, to study the settlement patterns in Southern Etruria during the final phase of the Bronze Age (1150-950/925 BC). The region of Southern Etruria is located in present-day Latium, Tuscany, and Umbria. The study, which includes 166 settlements, focuses on identifying the morphological characteristics of these settlements by means of raster analysis. Using a Machine Learning approach, the research will compare real settlements with random points within the region to understand the specific characteristics of the settlement pattern in the landscape. The study will also examine the use of feature selection and features importance methods to select the most significant features of a multivariate dataset.