Archeologia e Calcolatori
33.2, 2022, 33-52
doi 10.19282/ac.33.2.2022.03

DIMENSIONALITY REDUCTION FOR DATA VISUALIZATION AND EXPLORATORY ANALYSIS OF CERAMIC ASSEMBLAGES

1. Introduction

One of the most common problems in archaeological studies is organising and interpreting multivariate datasets. Within a dataset, as in most common spreadsheets, each archaeological record is connected to a wide range of variables, both quantitative and qualitative, such as provenance, size, typological classification, petrographic composition – if we are talking about ceramics – but also altitude above sea level, presence of certain structures, chronology and other characteristics related to architecture or to the different classes of materials – if we are dealing with sites.

In order to address this problem, it is necessary to organize these variables to facilitate, if not obtain, a possible interpretation of the archaeological data under consideration. Reducing the size of the dataset is particularly useful for the following reasons (Harrington 2012, 270-271):

- Reduce data storage space.
- Fewer dimensions require a shorter computation time.
- Some algorithms do not perform well with large dimensions.
- Remove redundant features.
- Data visualization.

Dataset dimension reduction can be performed mainly in two ways (James *et al.* 2013, 204): i) by keeping the most relevant variables from the original dataset (Subset or Attribute Selection) (Sammut, Webb 2010, 332), and ii) by creating new variables reworking input variables (Dimensionality Reduction) (Sammut, Webb 2010, 326)

To handle quantitative data, Principal Component Analysis (PCA) is often used, especially for the analysis of archaeometric or petrographic pottery composition (Baxter 1994; Marengo *et al.* 2005; Erdem *et al.* 2008). In addition to ceramics, PCA is also useful when dealing with other archaeological problems, such as lithic analysis (Prentiss 1998; Scerri *et al.* 2016), spatial analysis and landscape archaeology (Šmejda 2007; Klinger *et al.* 2011; Janovský, Horák 2018; Piña-Torres *et al.* 2018). This method is also the most detailed in several handbooks, thus building a bridge between archaeology and statistics (Shennan 1997; Drennan 2009, 299-303; VanPool, Leonard 2011, 285-303).

PCA is not the only one of its kind. Other methods with different characteristics and purposes can also be employed. A selection of *unsupervised* dimensionality reduction algorithms (§ 3) (Vanderplas 2016, 333-334) will be applied to a quantitative multivariate archaeological dataset (§ 2). The idea is to assess the strength of these algorithms for exploratory analysis (Tukey 1997) and data visualization. As mentioned earlier, one of the purposes of these techniques is to aid data visualization by reducing the data size. The relationships between variables, such as the height and width of a group of vessels, can be visualized and more easily understood through a scatter plot (Drennan 2009, 200-201). In this way, three variables at most can be displayed simultaneously on the same graph. Consequently, tools for dimensional reduction and for the complex visualization of multivariate datasets must be introduced, one of which will be described below. This dataset includes about 1500 vessels from the protohistoric necropolis of Osteria dell'Osa, 20 km East of Rome. The site was chosen as it was systematically investigated and the materials are comprehensively and exhaustively published (Bietti Sestieri 1992).

The application of dimensionality reduction methods to a multivariate dataset allows us to formulate new interpretations through a visualisation of the overall data structure. Reduction algorithms can be a tool as valuable in data interpretation as more traditional approaches to the study of ceramic, e.g. the creation of typologies or classifications. Displaying a large amount of data on a single graph allows us to identify structures and distributions based on ceramic characteristics. In the case of Osteria dell'Osa, the application of this methodology confirms that the Functional Classes (§ 2), archaeologically defined, correspond to effective clusters characterised by morphological similarities.

Therefore, the aims of this work are to:

1) Apply and evaluate a range of dimensionality reduction methods on a dataset of vessel profiles.
2) Identify the most efficient algorithm based on the ability to identify correspondences between archaeological interpretations and profile morphology.
3) Proceed with some exploratory analyses and evaluate the results according to the specificity of the applied algorithm.
4) Relating the method to 'more traditional' approaches.

2. The dataset

The necropolis of Osteria dell'Osa, in chronological terms, covers the Latial Periods II, III and IV (9th-6th century BC) (Bietti Sestieri 1992, 276). The extensive excavation of this cemetery makes it one of the best-documented contexts of the period. For this reason, it is one of the best-known sites that offers the possibility of carrying out analysis on numerous materials, especially for the Early Iron Age.

Every Early Iron Age 1 vessel (Fig. 1) that is complete or can be entirely reconstructed from the excavation catalogue (Bietti Sestieri 1992, 536-537,

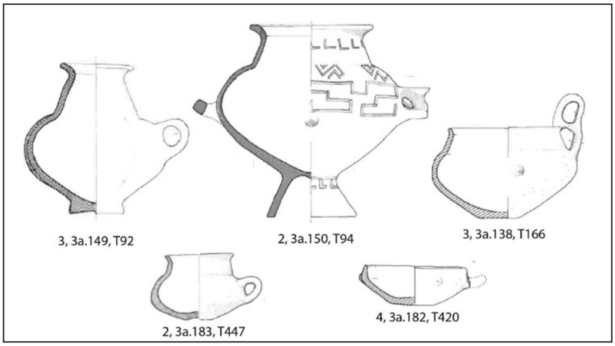


Fig. 1 – A selection of the drawings of the original vessels from the catalogue. Bibliographic references (figure, table and tomb number) are indicated for each vessel.

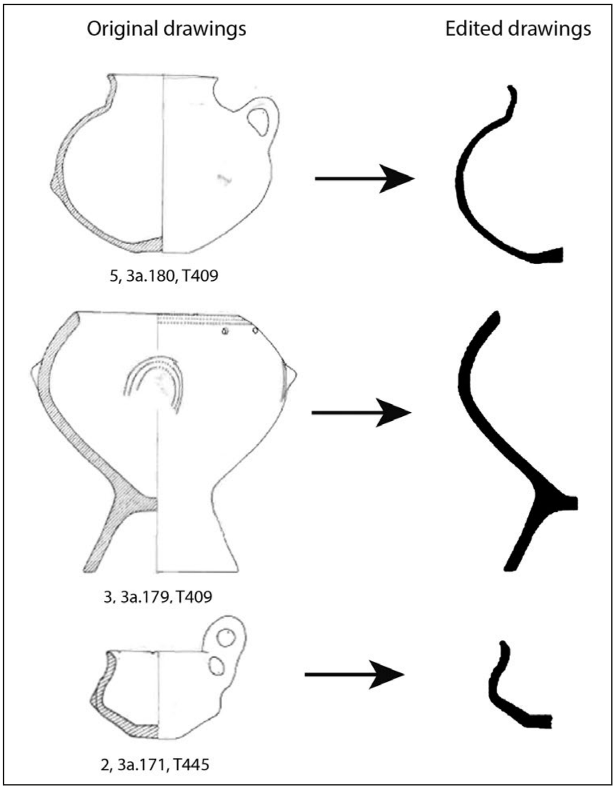


Fig. 2 – Editing process representation. For bibliographic references see Fig. 1.

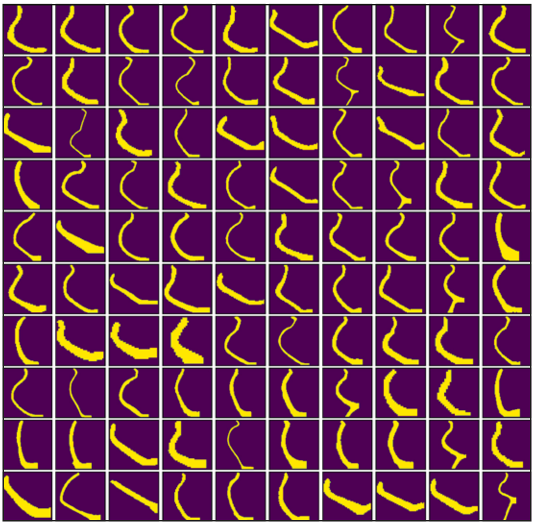


Fig. 3 – A selection of the vessels in the dataset, represented by binary matrices.

Osteria dell'Osa phases IIA-IIB; 900-770 BC) was included in the dataset (for a total of 1576 vessels). All the vessels' drawings are edited using a photo-editing program (Fig. 2), and the drawing sections are extrapolated, rescaled, and standardized in resolution (0,125 px/cm). A Python code[1] is used to create binary images on which the analysis is carried out. Subsequently, the vessel profiles are transformed into a two-dimensional array where the presence of the profile (yellow) is identified with the value number 1, while the absence of the latter (purple) corresponds to the value number 0 (Fig. 3).

Finally, all binary images are scaled within a 256×256 pixel frame, to obtain a coherent and standardized database.

The following information, of an archaeological-interpretative nature, is associated with each of the images:

– Functional Class: each vessel has been distributed into functional categories, according to the following criteria:

Class 1: open vessel with a horizontal handle and non-articulate profile (*bowls*).

Class 2: open vessel with an articulated profile or one or two vertical handles (*cups*, *mugs*, *goblets*).

[1] This Python script is part of Lorenzo Cardarelli's PhD project and it will be made available at the end of the PhD period.

Class 3: closed vessel with horizontal handle (*jars* and *necked jars*).

Class 4: closed vessel with one or two vertical handles (*jugs* and *amphoras*).

– Morphology: Class 1 and 2 are combined in Open forms, Class 3 and 4 in Closed forms.

Summing up, the multivariate dataset consists of archaeological information and the profile image of the vessel itself, described by a binary array of 256×256 values. This profile is translated into a very high number of dimensions: if a common spreadsheet is used, 65,536 columns are connected to each record (row). The best approach to explore this type of dataset is using dimensional reduction tools. Taken individually, variables defining the image are meaningless, therefore it is not possible to select more relevant ones (feature selection), but they must be created to represent the characteristics of the dataset using dimensionality reduction algorithms.

3. The algorithms

From the highest level, dimensionality reduction algorithms can be divided into *supervised*, which consider one or more attributes when reducing the dimensions, and *unsupervised*, which consider only the data structure. An example of a supervised size reduction algorithm is the Linear Discriminant Analysis (LDA) (Fisher 1936; Sammut, Webb 2010, 745-747). All the algorithms used in this paper, as mentioned above, are unsupervised. Dimensionality reduction algorithms can further be roughly divided into linear/non-linear and parametric/non-parametric methods. In linear methods, the projection in the lower dimension is a linear operation, a condition not met in non-linear methods (Fig. 4). Parametric methods require the construction of an explicit function for dimensionality reduction, unlike non-parametric methods. Both non-linear and non-parametric techniques include *Manifold learning* methods (Vanderplas 2016, 445). The lower dimensions in linear dimension reduction are easily interpretable because of their linear combinations of the input variable. Thus, they do not perform well when the data present non-linear relationships (Vanderplas 2016, 445-446). Non-linear algorithms and especially Manifold learning methods can highlight complex non-linear relationships between data, but only some properties of the original data are preserved, making the meaning of these features difficult to understand. Consequently, they are mainly used for data visualization (Das, Pal 2022, 1).

VanPool and Leonard (2012, 287) stated that archaeologists are generally not interested in algebra or mathematical structures defining and characterising these methods, therefore only a brief empirical introduction to the algorithm is proposed here. For those who are interested in more mathematical or technical aspects, details are however provided in the bibliography/web bibliography section of this paper.

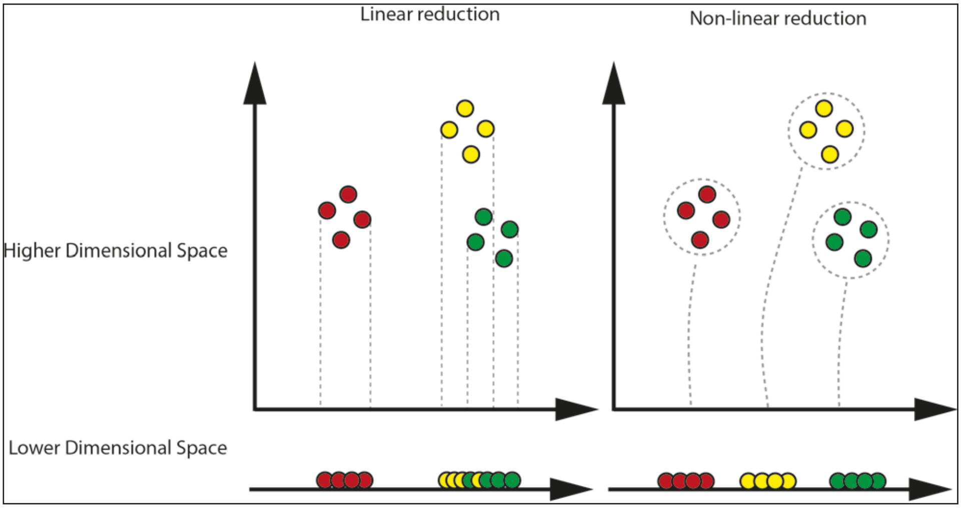


Fig. 4 – A simple representation of how linear and non-linear techniques for dimension reduction works.

This paper compares results from ten dimensionality reduction algorithms. The linear and parametric algorithms used here are: Principal Component Analysis and some of its implementations (Sparse PCA, Incremental PCA) and Truncated singular value decomposition (TruncSVD). The non-linear and non-parametric algorithms used are: Multidimensional scaling, Locally Linear embedding, ISOMAP, t-SNE, UMAP. Lastly, Kernel PCA is a non-linear but parametric method. Except the UMAP, all the other algorithms are part of the Scikit-learn library (https://scikit-learn.org/stable/), an open-source Python library for data analysis. Since UMAP is not included in this library, the official repository was used.

3.1 *Principal Components Analysis (PCA)*[2]

PCA is one of the oldest and best-known multivariate technique (Pearson 1901; Hotelling 1933). Its goal is to reduce the dataset dimension preserving as much variability as possible. Subsequently PCA finds new variables, called Components, that are linear combinations of variables from the original dataset. These linear variables maximize the variance and are uncorrelated with each other (Shlens 2014; Nanga *et al.* 2021, 192). The biggest variance is always represented by the first component, the second one by the second component and so on. The second axis is also orthogonal to the first one and to the

[2] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.

direction of the largest variance. The total number of components is equal to the original size of the dataset (Harrington 2012, 270). A significant amount of information is extracted using this analysis: PCA can give insight into the data structure using the correlation between component scores and variables. These correlations are known as Component Loadings, and their purpose is to improve the interpretation of each component (Guerra-Urzola *et al.* 2021).

This method has some limitations: since the new components are linear combinations, the PCA fails if the analysed dataset is characterized by non-linear relationships (Nanga *et al.* 2021, 192) and it is sensitive to outliers. This method is widely used and known in archaeology. For further information on this technique, see Jolliffe's monographic work (2002).

3.2 *Sparse PCA*[3]

It is an implementation of PCA. While PCA components are usually linear combinations of all input variables, Sparse PCA finds linear combination that contains a small subset of original variables (Zou *et al.* 2006; Chen, Rohe 2021, 2; Guerra-Urzola *et al.* 2021). It often has applications in anthropology (Zhao *et al.* 2017).

3.3 *Incremental PCA*[4]

This PCA implementation finds similar projections while processing only a few samples per time. This process has the advantage of using less memory with similar result (Ross *et al.* 2008).

3.4 *Truncated singular value decomposition*[5]

This method (TruncSVD) is closely related with PCA but works better on sparse data (sparse data refers to the data with many zero values – https://en.wikipedia.org/wiki/Sparse_matrix). The differences with PCA are found from a computational point of view: it shares all the pros and cons (Shlens 2014, 7). This method is applied in electromagnetic prospection, in connection with landscape archaeology (Catapano *et al.* 2014).

3.5 *Kernel PCA*[6]

This method is a non-linear implementation of PCA. The non-linearity is achieved using the so-called *kernel trick* to divide the input variables (Vanderplas 2016, 413; Nanga *et al.* 2021, 200-201). It has applications in

[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.SparsePCA.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.KernelPCA.html

ceramics studies, with the comparative use of Kernel PCA and PCA methods in archaeometric analysis (He *et al.* 2019).

3.6 *Multidimensional scaling (MDS)*[7]

This non-linear method performs a visual representation of distance or dissimilarities between pairs of data points (Kruskal, Wish 1978). Records that have shorter distances are more similar and are close in the graph; conversely, records that are less similar have longer distances in the graph preserving the global structure of the data (Saaed *et al.* 2018; Nanga *et al.* 2021, 201-202; Wang *et al.* 2021, 5). For instance, this method is used in zooarchaeology to identify regional and functional variability in the exploitation of some species in different sites (Orchard, Clark 2005). Another application of the method concerns the use of non-Euclidean distances in point pattern analysis (Pérez 2015).

3.7 *Locally Linear Embedding (LLE)*[8]

LLE is a non-linear dimension reduction that can preserve only the local properties of data (Roweis, Saul 2000). This method learns the global structure to recreate it in a local linear reconstruction (Nanga *et al.* 2021, 202-204).

3.8 *ISOMAP*[9]

This algorithm combines some characteristics of PCA and MDS. Isomap seeks a lower-dimensional embedding, which maintains geodesic distances (https://en.wikipedia.org/wiki/Geodesic) between all points (Tenenbaum *et al.* 200; Nanga *et al.* 2021, 202). Isomap can be seen as an extension of Multi-dimensional Scaling (MDS) or Kernel PCA.

3.9 *t-distributed Stochastic Neighbor Embedding (t-SNE)*[10]

t-SNE Embedding (van der Maaten, Hinton 2008) converts affinities of data points to probabilities (Nanga *et al.* 2021, 206-205). If two points are close to each other in the high dimensional space, they have a high probability of being close to each other in the low dimensional embedding space. Unlike the other non-linear methods mentioned above, t-SNE reveals structure at many different scales, translating into better data visualization. This local structure is well preserved, but the algorithm fails to preserve the

[7] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.locally_linear_embedding.html.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.Isomap.html.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

global structure of the data (Wang *et al.* 2021, 7). The method is used in flint study (Elliot *et al.* 2021)

3.10 *Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)*[11]

UMAP is a dimensionality reduction technique introduced in 2018 (McInnes *et al.* 2018). This algorithm projects high-dimensional data in a lower space. UMAP is similar to t-SNE, but it is considered by the authors more performing, specifically in relation to computational time (McInnes *et al.* 2018, 28-30) and with a better ability to preserve the global structure of the data (McInnes *et al.* 2018, 36-38). This method is applied for the study of archaeological pottery (Navarro *et al.* 2021).

4. Clustering metrics

Some clustering metrics will be calculated to quantify the results of the dimensionality reduction: it is assumed that the Functional Class, previously defined, corresponds to a homogeneous set of vessels identifiable through the binary matrices. The following metrics are used.

4.1 *Silhouette score*

This score relates to a reduction with better defined clusters. The result is defined between -1 (incorrect clustering) and +1 (correct clustering). Scores around 0 indicate overlapping clusters (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html).

4.2 *Calinski-Harabasz Index*

This score is defined as the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters (where dispersion is defined as the sum of squared distances): a higher Calinski-Harabasz score relates to a model with better-defined clusters (https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index).

4.3 *Davies-Bouldin Index*

This index indicates the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves. 0 is the lowest possible score and a lower Davies-Bouldin index relates to a model with better separation between the clusters (https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index).

[11] https://umap-learn.readthedocs.io/en/latest/.

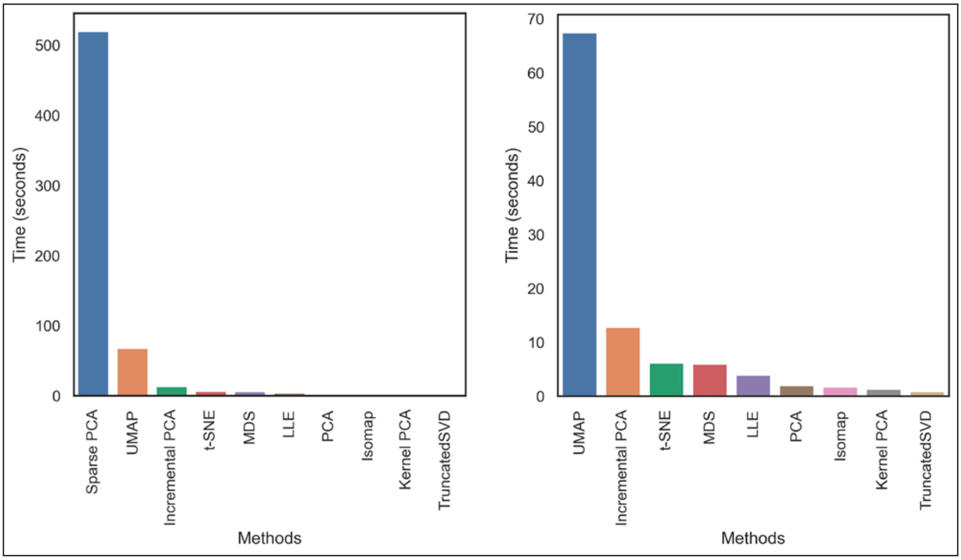


Fig. 5 – Barplots showing the execution time of each algorithm on the dataset.

5. Results

The dataset was pre-processed by using Scikit-learn *MinMaxScaler* (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html), to scale all values in a range between 0 and 1, and algorithms are tested on the dataset. Two dimensions are used, so the results could be seen on a simple scatterplot (*embeddings*): every point, therefore, represents a vessel. This feature is very interesting, as it goes beyond simple visualization: it is possible to globally observe the characteristics of ceramic assemblages on a single graph such as macro or micro differences between the vessels, quantity of objects of a certain class, presence of outliers, etc.

Before discussing metrics and reductions' shapes, let us have a brief look at the algorithms' executions time[12].

The barplots show, in order, the highest to the lowest execution time (Fig. 5): the Sparse PCA algorithm was extremely slow when compared to other methods (nearly 10 minutes to run). Eliminating this outlier from the graph, the second most expensive algorithm is UMAP (67 seconds), followed by Incremental PCA (12 secs); t-SNE (6,1 secs); MDS (5,9 secs); LLE (3,8

[12] The script is run on Acer notebook, Windows 10, Intel Core i7-10750H, 2.60GHz; 16 GB DDR4 RAM; NVIDIA GeForce RTX 3060 6GB.

Method	**Silhouette score**	**Davies Bouldin index**	**Calinski Harabasz index**	**Davies Bouldin index (*inversed*)**
PCA	0,252	1,510	798,325	0,662
Sparse PCA	0,256	1,473	802,966	0,679
Kernel PCA	0,258	1,478	817,184	0,677
Incremental PCA	0,252	1,510	798,325	0,662
TruncatedSVD	0,180	1,600	654,367	0,625
MDS	0,112	3,614	229,309	0,277
LLE	0,034	1,811	422,399	0,552
Isomap	0,262	1,290	871,737	0,775
t-SNE	0,298	1,187	759,926	0,842
UMAP	0,330	1,072	903,297	0,933

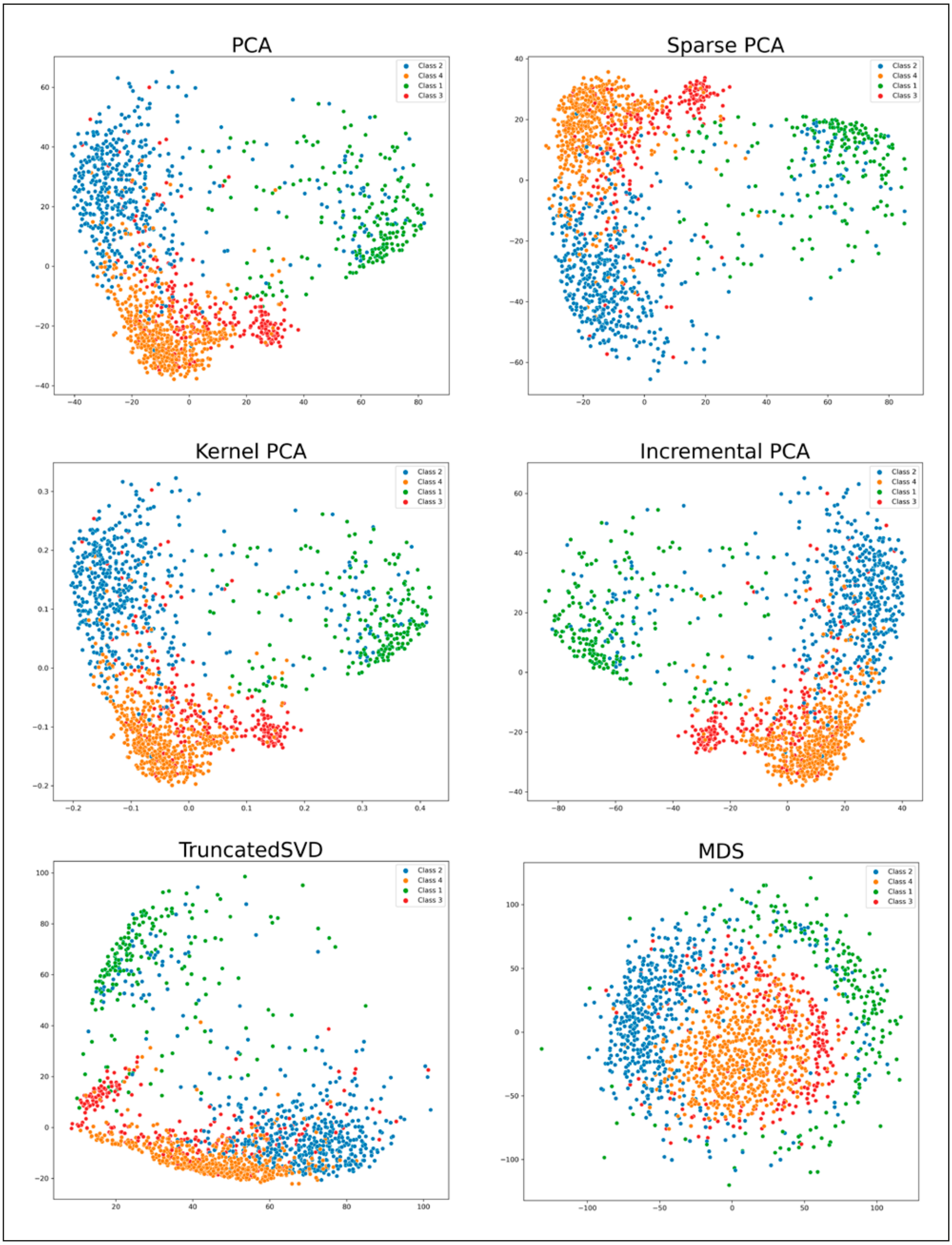
Tab. 1 – Clustering metrics.

secs); PCA (1,9 secs); Isomap (1,6 secs); Kernel PCA (1,2 secs); TruncSVD (0,8 secs). The result is surprising: UMAP is believed to be extremely faster than t-SNE (https://umap-learn.readthedocs.io/en/latest/performance.html). However, this can be explained by the performance of linear size reduction algorithms such as PCA and specifically TruncSVG, which is built for sparse matrices.

Besides execution time, these algorithms produced different results and they can be useful to visualize or explore this type of data (Fig. 6). LLE algorithm did not seem to perform correctly (Vanderplas 2016, 456), providing a result that is not easily interpretable by archaeologists. The other methods, unlike LLE, seem to place the vessels correctly: these are closer to each other within the Functional Class. Implementations of PCA, on a global level, seem to provide a reduction with a different rotation from classical PCA. The algorithms offering more visually interesting results are t-SNE and UMAP. The structure of the data, and therefore of a generic ceramic assemblage, is characterized by non-linearity relationships.

Regarding the metrics, a higher value of the Silhouette score and Calinski Harabasz index also corresponds to a better result in clustering similar objects. As for the Davies Bouldin index, it is the opposite: when it tends to zero, it corresponds to a better result. To make the three indices comparable, the inverse of the Davies Bouldin index is calculated. In this case, higher values also correspond to better performances. The three indices are then standardized[13] on the same scale (Tab. 1).

[13] The standardized measure (or Z-score) (Drennan 2009, 49; VanPool, Leonard 2011, 139) is obtained by subtracting, for each value (X), the mean (μ) and dividing the result by the standard deviation (σ).



PCA

Sparse PCA

Kernel PCA

Incremental PCA

TruncatedSVD

MDS

Fig. 6.1 – Dimensionality reduction algorithm embeddings output.

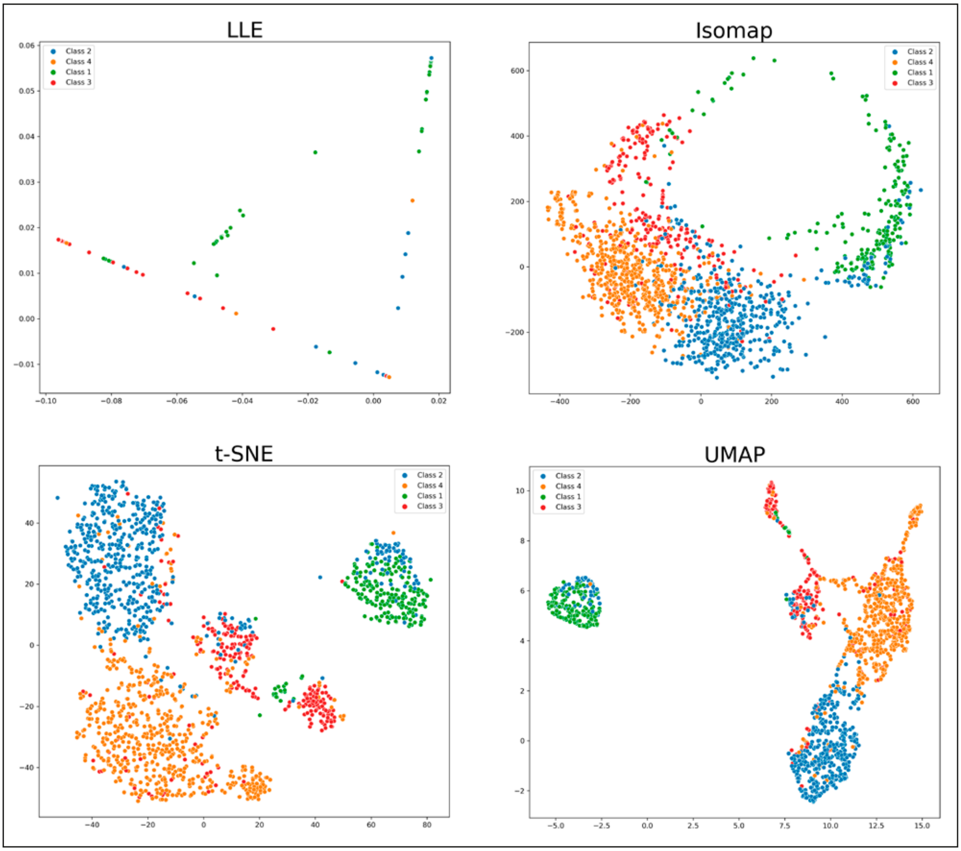


Fig. 6.2 – Dimensionality reduction algorithm embeddings output.

Method	Mean
UMAP	1,265
t-SNE	0,727
Isomap	0,634
Kernel PCA	0,333
Sparse PCA	0,307
PCA	0,250
Incremental PCA	0,250
TruncatedSVD	-0,340
LLE	-1,435
MDS	-1,992

Tab. 2 – Average metrics scores.

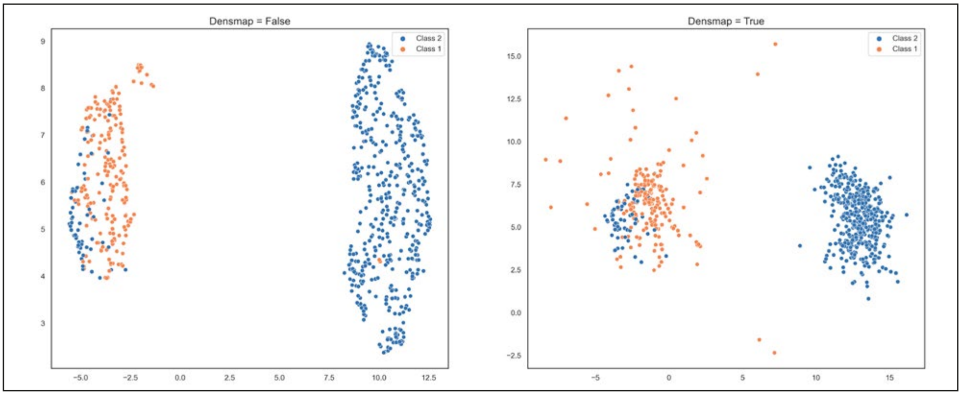


Fig. 7 – Scatterplot showing UMAP reduction on Open shapes, comparison between activated and deactivated *densmap* is proposed.

Once the measures have been standardized, the average is calculated for each algorithm (Tab. 2). Relying on this process, the methods that can best represent the data are UMAP, followed by t-SNE and ISOMAP.

Based on the score obtained, UMAP is the preferred algorithm for displaying and exploring this type of data. It is a very powerful tool, which preserves the local structure and the global one to a greater extent than the t-SNE (McInnes *et al.* 2018, 28-30). The uses of UMAP are therefore not limited to data visualization: it can be applied to other applications, such as clustering tasks (McInnes *et al.* 2018, 39; Allaoui *et al.* 2020; https://umap-learn.readthedocs.io/en/latest/clustering.html). These features make UMAP a very powerful tool, but like all algorithms for dimensionality reduction, it has some weaknesses: such as the interpretability of methods like PCA. The PCA components, as previously mentioned, represent the directions of the greatest variance. In the case of UMAP, these dimensions are meaningless. There are some key points to remember when reading and interpreting a UMAP result: although it is true that the global structure is better preserved, the distances between the clusters could be not significant (https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17).

Therefore, it is dangerous to quantify the distance between clusters because UMAP and t-SNE use local notions of space to proceed with dimensionality reduction. Furthermore, using the default settings of UMAP, the size of the clusters is meaningless, because the density is not preserved (Narayan *et al.* 2021). For better preservation of the density in the reduction, the *densmap* function from the UMAP library can be used (Narayan *et al.* 2021, https://umap-learn.readthedocs.io/en/latest/densmap_demo.html#better-preserving-local-density-with-densmap).

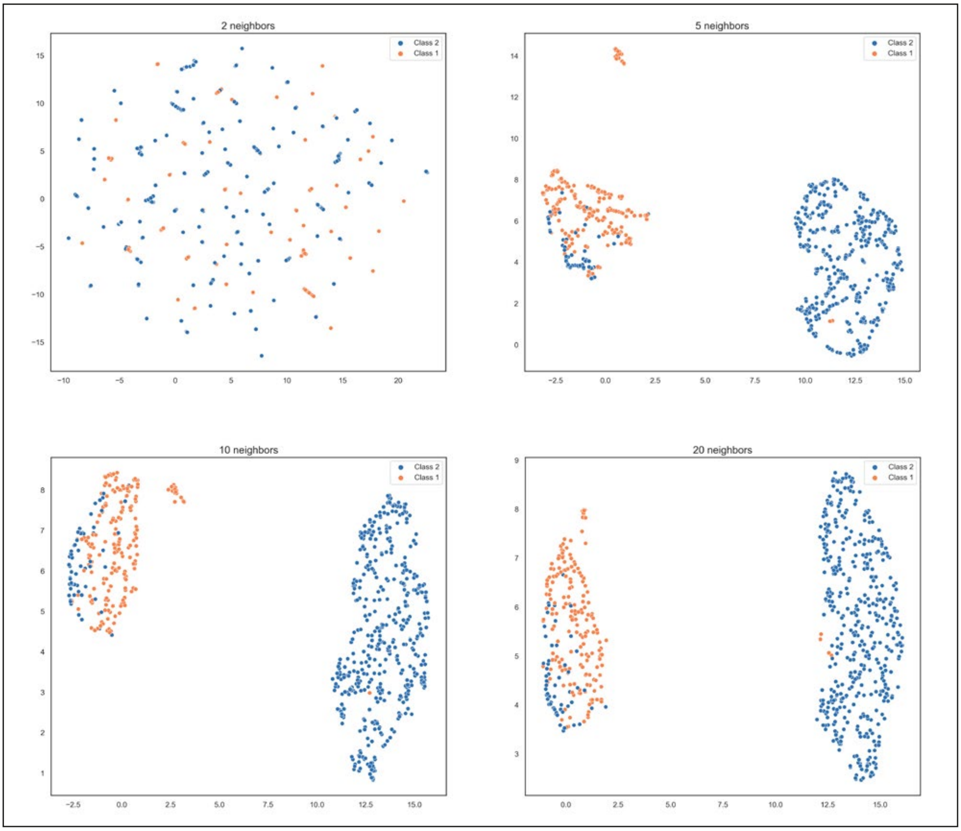


Fig. 8 – Comparison between different number of *neighbors* during UMAP dimensionality reduction.

We propose here another example using Open shapes from the dataset, presented in two scatterplots: one with UMAP without density preservation and one with density preservation active (Fig. 7).

By these embeddings and their comparison, some interesting information on vessels is obtained: first, Open shaped vessels can be divided into two groups, which would seem to roughly correspond to Functional classes. Specifically, in the cluster on the left, there are some cups (and therefore with a vertical handle), but morphologically they are very similar to bowls (Class 1). In the second scatterplot, the Class 1 (*bowls*) shows a greater dispersion, or at most a greater number of outliers, if compared to the cluster on the right. This class is clearly characterized by greater morphological heterogeneity. This type of investigation offers a lot of possibilities for the morphometric study of ceramics, such as variability and standardization. If variability is defined as the relative degree of heterogeneity in a class of manufacts (Kotsonas 2014, 8), and the standardization is a reduction in variability within the same class (Rice 1991, 268), by calculating metrics on these clusters, their variability is easily defined.

Particularly useful for archaeologists is the possibility, in UMAP (but also in other Manifold learning algorithms), to control the balance between local and global structures. In UMAP, this can be controlled through a parameter called *n_neighbors*: if this parameter has low values, UMAP will highlight the local data structure. Higher values instead will emphasize the global data structure. A low value of this parameter leads to a combination of small clusters, containing extremely similar vessels but without memory of the global structure, as it results in the macro-division between Class 1 and Class 2 highlighted by using a higher value of the parameter (Fig. 8). This parameter can be very useful in the study of ceramics. In its classification, especially in protohistoric contexts, a hierarchical approach is often used Peroni 1994, 26.

For other details on UMAP and on the parameters that can be used, please refer to the algorithm's website: https://umap-learn.readthedocs.io/en/latest/index.html#umap-uniform-manifold-approximation-and-projection-for-dimension-reduction.

Summing up, there are several algorithms for size reduction with different characteristics and purposes: to visualize the data, t-SNE and UMAP are the best choices. To explore the data structure, focusing on the local and global structure, UMAP is the best algorithm. To create clusters, UMAP is also the best choice. However, this type of consideration does not mean that methods such as PCA should always be discarded. The main purpose of PCA is dimension reduction and using it as a data visualization tool is allowed, at the cost of performance, as shown above. PCA is a great and useful tool in the hands of archaeologists. If non-linear reduction methods (specifically UMAP) performed better in exploring and displaying this type of data, it is also necessary to highlight some criticalities. With Manifold learning methods, there is no possibility to manage missing data, unlike PCA. The presence of noise in the data can lead the Manifold learning algorithms to be mistaken, leading to an incorrect dimension reduction. The PCA filters out the most important components by choosing the greatest variance. The result of a reduction through Manifold learning strictly depends on the number of neighbours chosen, and there are no criteria for defining this choice. The number of optimal dimensions for the Manifold methods is difficult to determine, while in the case of PCA this is defined by the variance. If attributing a meaning to dimensions in Manifold methods is cryptic and often unclear, the main components of PCA have a very clear meaning (Vanderplas 2016, 455-456).

6. Conclusions

Using dimension reductions algorithms on a multivariate dataset has undoubted benefits: in this example, their use simplifies the understanding of a high-dimensionality dataset, ranging from 65,536 dimensions to 2. The quantification of the pot profile is also a rather interesting feature, and it can be insightful for the study of pottery. Using these methods, it is possible to explore large quantities of material in a short time (1500 vessels processed in just over a minute in the case of UMAP, e.g.). Manifold learning methods (especially t-SNE and UMAP) are successful in identifying the structure of the data, creating clusters of vessels that are similar in shape and function. By performing a quick analysis on the data, some cups have a shape more similar to the bowls and the latter group is characterized by a greater morphological heterogeneity. The analysis also fits a hierarchical approach widely used in the study of ceramics.

In conclusion, the approach proposed in this work is particularly useful because it allows the use of statistical-mathematical tools for the morphological study of ceramic profiles. This method can be used independently or in integration with a traditional approach. In this perspective, the correspondence between archaeological and statistical data, defined by the correspondence between functional forms and vessel profiles, is particularly relevant.

Lorenzo Cardarelli
Dipartimento di Ricerca e Innovazione Umanistica
Università di Bari Aldo Moro
Istituto di Scienze del Patrimonio Culturale - CNR
lorenzocardarelli2@gmail.com

Annalisa Lapadula
Dipartimento di Scienze dell'Antichità
Sapienza Università di Roma
annalisalapadula@outlook.it

REFERENCES

Allaoui M., Kherfi M.L., Cheriet A. 2020, *Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study*, in A. El Moataz, D. Mammass, A. Mansouri, F. Nouboud (eds.), *Image and Signal Processing, Lecture Notes in Computer Science*, Cham, Springer International Publishing, 317-325 (https://doi.org/10.1007/978-3-030-51935-3_34).

Baxter M.J. 1994, *Principal Component Analysis in archaeometry*, «Archeologia e Calcolatori», 5, 23-38.

Bietti Sestieri A. (ed.) 1992, *La necropoli laziale di Osteria dell'Osa*, Roma, Quasar.

Catapano I., Affinito A., Gennarelli G., di Maio F., Loperte A., Soldovier. F. 2014, *Full three-dimensional imaging via ground penetrating radar: Assessment in controlled conditions and on field for archaeological prospecting*, «Applied Physics A», 115, 1415-1422 (https://doi.org/10.1007/s00339-013-8053-0).

Chen F., Rohe K. 2021, *A new basis for sparse Principal Component Analysis*, arXiv:2007.00596, 1-46 (https://doi.org/10.48550/arXiv.2007.00596).

Das S., Pal N.R. 2022, *Nonlinear dimensionality reduction for data visualization: An unsupervised fuzzy rule-cased approach*, «IEEE Transactions on Fuzzy Systems», 1-13 (https://doi.org/10.1109/TFUZZ.2021.3076583).

Drennan R.D. 2009, *Statistics for Archaeologists: A Common Sense Approach*, New York Springer, 2nd ed.

Elliot T., Morse R., Smythe D., Norris A. 2021, *Evaluating machine learning techniques for archaeological lithic sourcing: A case study of flint in Britain*, «Scientific Reports», 11, 10197 (https://doi.org/10.1038/s41598-021-87834-3).

Erdem A., Çilingiroğlu A., Giakoumaki A., Castanys M., Kartsonaki E., Fotakis C., Anglos D. 2008, *Characterization of Iron Age pottery from eastern Turkey by laser-induced breakdown spectroscopy (LIBS)*, «Journal of Archaeological Science», 35, 9, 2486-2494.

Fisher R.A. 1936, *The use of multiple measurements in taxonomic problems*, «Annals of Eugenics», 7, 179-188 (https://doi.org/10.1111/j.1469-1809.1936.tb02137.x).

Guerra-Urzola R., Van Deun K., Vera J.C., Sijtsma K. 2021, *A guide for sparse PCA: Model comparison and applications*, «Psychometrika», 86, 893-919 (https://doi.org/10.1007/s11336-021-09773-2).

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

He J., Liu Y., Pan C., Du X. 2019, *Identifying ancient ceramics using laser-induced breakdown spectroscopy combined with a back propagation neural network*, «Applied Spectroscopy», 73, 10, 1201-1207 (https://doi.org/10.1177/0003702819861576).

Hotelling H. 1933, *Analysis of a complex of statistical variables into principal components*, «Journal of Educational Psychology», 24, 498-520 (https://doi.org/10.1037/h0070888).

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

Janovský M., Horák J. 2018, *Large scale geochemical signatures enable to determine landscape use in the deserted medieval villages*, «Interdisciplinaria archaeologica», 9, 71-80 (http://dx.doi.org/10.24916/iansa.2018.1.5).

Jolliffe I.T. 2002, *Principal Component Analysis*, New York, Springer, 2nd ed.

Klinger R., Schwanghart W., Schütt B. 2011, *Landscape classification using Principal Component Analysis and fuzzy classification: Archaeological sites and their natural surroundings in Central Mongolia*, «DIE ERDE – Journal of the Geographical Society of Berlin», 142, 3, 213-233.

Kotsonas A. 2014, *Understanding standardization and variation in Mediterranean ceramics*, in A. Kotsonas (ed.), *Understanding Standardization and Variation in Mediterranean Ceramics: Mid 2nd to Late 1st Millennium BC*, BABESCH Suppl. 25, Leuven, Peeters, 7-23.

Kruskal J., Wish M. 1978, *Multidimensional Scaling*, Newbury Park, Inc., California, SAGE Publications (https://doi.org/10.4135/9781412985130).

Marengo E., Aceto M., Robotti E., Liparota M.C., Bobba M., Pantò G. 2005, *Archaeometric characterisation of ancient pottery belonging to the archaeological site of Novalesa Abbey (Piedmont, Italy) by ICP-MS and spectroscopic techniques coupled to multivariate statistical tools*, «Analytica Chimica Acta», 537, 1-2, 359-375.

McInnes L., Healy J., Melville J. 2018, *UMAP: Uniform Manifold approximation and Projection for dimension reduction*, arXiv:1802.03426, 1-63 (https://doi.org/10.48550/arXiv.1802.03426).

Nanga S., Bawah A.T., Acquaye B.A., Billa M.-I., Baeta F.D., Odai N.A., Obeng S.K., Nsiah A.D. 2021, *Review of dimension reduction methods*, «Journal of Data Analysis and Information Processing», 9, 189-231.

Narayan A., Berger B., Cho H. 2021, *Assessing single-cell transcriptomic variability through density-preserving data visualization*, «Nature Biotechnology», 39, 765-774 (https://doi.org/10.1038/s41587-020-00801-7).

Navarro P., Cintas C., Lucena M., Fuertes J.M., Delrieux C., Molinos M. 2021, *Learning feature representation of Iberian ceramics with automatic classification models*, «Journal of Cultural Heritage», 48, 65-73 (https://doi.org/10.1016/j.culher.2021.01.003).

Orchard T.J., Clark T. 2005, *Multidimensional scaling of northwest coast faunal assemblages: A case study from Southern Haida Gwaii, British Columbia*, «Canadian Journal of Archaeology/Journal Canadien d'Archéologie», 29, 1, 88-112 (http://www.jstor.org/stable/41103518).

Pearson K. 1901, *LIII. On lines and planes of closest fit to systems of points in space*, «The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science», 2, 559-572 (https://doi.org/10.1080/14786440109462720).

Perez J.N. 2015, *Non-Euclidean distance in Point Pattern Analysis: Anisotropic measures for the study of settlements networks in heterogeneous regions*, in J.A. Barceló, I. Bogdanovic (eds.), *Mathematics and Archaeology*, Boca Raton, CRC Press, 369-382.

Peroni R. 1994, *Introduzione alla protostoria italiana*, Roma, Laterza.

Piña-Torres C., Lucero-Gómez P., Nieto S., Vázquez A., Bucio L., Belio I., Vega R., Mathe C., Vieillescazes C. 2018, *An analytical strategy based on Fourier transform infrared spectroscopy, Principal Component Analysis and linear Discriminant Analysis to suggest the botanical origin of resins from Bursera. Application to archaeological Aztec Samples*, «Journal of Cultural Heritage», 33, 48-59.

Prentiss W. 1998, *The reliability and validity of a lithic debitage typology: Implications for archaeological interpretation*, «American Antiquity», 63, 4, 635-650.

Rice P.M. 1991, *Specialization, standardization and diversity: A retrospective*, in R.L. Bishop, F.W. Lange (eds.), *The Legacy of Anna O. Shepard*, Boulder, The University Press of Colorado, 257-279.

Ross D.A., Lim J., Lin R.-S., Yang M.-H. 2008, *Incremental learning for robust visual tracking*, «International Journal of Computer Vision», 77, 125-141 (https://doi.org/10.1007/s11263-007-0075-7).

Roweis S.T., Saul L.K. 2000, *Nonlinear dimensionality reduction by locally linear embedding*, «Science», 290, 2323-2326 (https://doi.org/10.1126/science.290.5500.2323).

Saeed N., Nam H., Haq M.I.U., Muhammad Saqib D.B. 2018, *A survey on multidimensional scaling*, «ACM Computing Survey», 51, 1-25 (https://doi.org/10.1145/3178155).

Sammut C., Webb G.I. (eds.) 2010, *Encyclopedia of Machine Learning*, Boston MA, Springer US (https://doi.org/10.1007/978-0-387-30164-8).

Scerri E.M.L., Gravina B., Blinkhorn J., Delagnes A. 2016, *Can lithic attribute analyses identify discrete reduction trajectories? A quantitative study using refitted lithic sets*, «Journal of Archaeological Method and Theory», 23, 669-691 (https://doi.org/10.1007/s10816-015-9255-x).

Shennan S. 1997, *Quantifying Archaeology*, Saint Louis, Elsevier Science.

Shlens J. 2014, *A tutorial on Principal Component Analysis*, arXiv:1404.1100, 1-12 (https://doi.org/10.48550/arXiv.1404.1100).

Šmejda L. 2007, *Time as a hidden dimension in archaeological information systems: Spatial analysis within and without the geographic framework*, «Archaeology», 26, 517-540 (http://archive.caaconference.org/2009/articles/Smejda_Contribution316_c.pdf).

Tenenbaum J.B., Silva V. de, Langford J.C. 2000, *A global geometric framework for nonlinear dimensionality reduction*, «Science», 290, 2319-232 (https://doi.org/10.1126/science.290.5500.2319).

Tukey J.W. 1977, *Exploratory Data Analysis*, Addison-Wesley Series in Behavioural Science, Reading, Mass., Addison-Wesley Pub. Co.

van der Maaten L., Hinton G. 2008, *Visualizing data using t-SNE*, «Journal of Machine Learning Research», 9, 2579-2605.

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

VanPool T.L., Leonard R.D. 2011, *Quantitative Analysis in Archaeology*, Malden, MA., Chichester, West Sussex, U.K., Wiley-Blackwell.

Wang Y., Huang H., Rudin C., Shaposhnik Y., *Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization*, arXiv:2012.04456 [cs, stat] (http://arxiv.org/abs/2012.04456).

Zhao J., Duan F., Pan Z., Wu Z., Li J., Deng Q., Li X., Zhou M. 2017, *Craniofacial similarity analysis through sparse Principal Component Analysis*, «PLoS ONE», 12, 6, e0179671 (https://doi.org/10.1371/journal.pone.0179671).

Zou H., Hastie T., Tibshirani R. 2006, *Sparse Principal Component Analysis*, «Journal of Computational and Graphical Statistics», 15, 265-286.

ABSTRACT

Size reduction algorithms are essential in the study of multivariate datasets. Many variables make it difficult to visualize data. In Archaeology, this problem often concerns the study of some variables, which can be quantitative or qualitative. In this article, several methods for dimension reduction are applied to a pottery dataset from the protohistoric necropolis Osteria dell'Osa, located 20 km East of Rome. These methods offer the possibility of visualising and analysing large amount of data in a very short time. Our results show that non-linear and non-parametric algorithms such as t-SNE and UMAP are the best choice for visualising and exploring this type of data.