

Uncovering Spatial Patterns in Archaeological Data through the Integration of Self-Organizing Maps, Principal Component Analysis and Hierarchical Clustering

Guillem Domingo-Ribas

`Guillem.Domingo-Ribas@newcastle.ac.uk`

Newcastle University

Francesco Carrer

Newcastle University

Karen Álvaro Rueda

Universitat de Barcelona

Research Article

Keywords: exploratory data analysis, multivariate techniques, principal component analysis, hierarchical clustering, self-organising maps, spatial patterns

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11096997/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

TITLE: Uncovering Spatial Patterns in Archaeological Data through the Integration of Self-Organizing Maps, Principal Component Analysis and Hierarchical Clustering

Authors: Guillem Domingo Ribas^{*1,2} (ORCID: 0000-0002-7848-1321), Francesco Carrer¹ (ORCID: 0000-0003-4397-4502), Karen Álvaro Rueda² (ORCID: 0000-0003-1321-7260)

***Corresponding author:** Guillem.Domingo-Ribas@newcastle.ac.uk

1. School of History, Classics and Archaeology, Newcastle University. Armstrong Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU, United Kingdom

2. Departament d'Història i Arqueologia, Grup de Recerca en Arqueologia Medieval i Postmedieval (GRAMPMed), Institut d'Arqueologia (IAUB), Universitat de Barcelona. Carrer de Montalegre 6-8, 08001, Barcelona, Spain

Abstract

Classification and grouping are central to archaeological interpretation, yet the multivariate techniques most commonly used for exploratory data analysis (Principal Component Analysis (PCA) and hierarchical clustering) struggle to capture non-linear relationships in complex spatial datasets. This paper introduces Self-Organising Maps (SOM), an unsupervised artificial neural network, to archaeological landscape analysis for the first time, applying it alongside PCA and hierarchical clustering to eleven environmental and topographic variables recorded at 57 early medieval sites in the Upper Arlanza Basin (Burgos, Spain). All three methods converge on a consistent four-group settlement structure, confirming the robustness of the classification. However, SOM's topology-preserving projection recovers this structure while also representing the continuous gradients between groups, identifying transitional sites whose landscape profiles resist the rigid partitions imposed by dendrograms. Its variable heatmaps further allow clusters to be interpreted directly in real-world units, without the added abstraction of principal components. The comparison demonstrates that SOM performs as reliably as established methods while offering additional interpretive depth for ambiguous, spatially structured data. This study argues that SOM complements, instead of replacing, PCA and cluster analysis, extending the multivariate toolkit of landscape archaeology to better accommodate the ambiguous, gradient-like structure of settlement data.

Keywords: exploratory data analysis, multivariate techniques, principal component analysis, hierarchical clustering, self-organising maps, spatial patterns.

Statements and Declarations

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

1. Introduction. Exploratory data analysis in Archaeology

Classification lies at the core of archaeology. Archaeologists routinely classify artifacts, features, and other archaeological information into categories based on their shared characteristics. Such priority has been fundamental in archaeology from its inception, as it allowed to put order to the vast diversity of material unearthed. This has been important long before the application of quantitative methods in archaeology, since it provided a foundation for identifying coherent patterns across both space and time.

Although some scholars stress that classification ‘involves the grouping or placement of data into groups’ (Conolly and Lake, 2006: 135), others emphasise the intellectual differences between classification and grouping (Banning 2020 *after* Dunnell 1971).

The distinction between classification and grouping is particularly relevant when it comes to working with complex datasets, especially data matrices with multiple variables (also known as multivariate data). This is a common challenge for archaeologists today, who manage large volumes of data arising from, but not exclusively, the analytical techniques applied to archaeological artefacts (Glascok and MacDonald, 2023) and the attributes from archaeological geospatial data (Conolly and Lake, 2006).

Statistical techniques for both classification and grouping have been extensively used for exploratory data analysis and spatial analysis in archaeology. Archaeologists have employed various techniques for exploring their multivariate datasets, often reducing a table (or data matrix) of n rows and p columns to a two- or three-dimensional summary of the data, which enable a better understanding of patterns and clusters. These range from basic parametrisation (e.g., mean, median, maximum, minimum) and data visualisations (e.g., box plots, density plots and scatter plots) to more complex statistical approaches. Some methods stand out for having been used more intensively in archaeology, particularly Principal Component Analysis (mentioned as PCA hereafter) and cluster analysis (Shennan, 1997; Drennan, 2009; Baxter, 2015; Carlson, 2017).

Over the past decade, machine learning (ML) methods have revolutionised classification in archaeology. As a subfield of Artificial Intelligence (AI), ML combines computational and statistical approaches to develop algorithms that learn patterns from training data and use these patterns to make classifications, predictions or discover underlying structures, without being explicitly programmed for every possible scenario. The rapid growth of ML applications and publications in archaeology during the past years has been driven by advances in computational power, the increasing availability of large and complex datasets, improvements in data-processing capabilities, and the growing accessibility of ML algorithms and software (for a comprehensive overview on AI and ML in archaeology, see Orengo, Berganzo-Besga and Lumbreras, 2026).

ML methods can broadly be divided into supervised and unsupervised approaches. Supervised learning employs categorised or labelled training data to derive mathematical classifiers that distinguish between the predefined categories, leading classification models that assign new observations to the most probable category based on quantitative features (Bickler, 2021). Unsupervised learning, on the other hand, employs uncategorised data to identify underlying structures, relationships, or groupings without predefined categories, making it particularly relevant for EDA and pattern recognition (Cacciari and Pocobelli, 2022). Among the unsupervised ML methods

available, artificial neural networks (ANN) deserve particular attention as an alternative to traditional multivariate techniques. Inspired by the structure and functioning of biological neural networks, ANN are developed for modelling complex, often non-linear relationships that are difficult to capture using conventional statistical approaches (Zou, Han and So, 2008). One type of ANN with considerable potential for organising large and complex datasets is the self-organising map (SOM), commonly employed for dimensionality reduction. Despite its potential, SOM have seen very limited use in archaeology, mostly in studies focused on specific artefact assemblages (Lopez-Molinero *et al.*, 2000; Kontogiorgos, Leontitsis and Sangole, 2007; Deufemia, Paolino and de Lumley, 2012; Korokhina, 2024), and, to our knowledge, they have not yet been applied to spatial archaeological data. In this paper, SOM are used for the first time in archaeological landscape analysis and their performance is assessed against more established exploratory methods like PCA and cluster analysis.

1.1. Principal Component Analysis (PCA) in Archaeology

Principal Component Analysis (PCA) is often considered the most widely used multivariate statistical technique, being employed in many scientific disciplines (Abdi and Williams, 2010). The objective of PCA is to minimise the number of variables in a multivariate dataset, while helping to focus on the key variables that explain most of the variability in the data. In the words of Drennan (2009: 299), archaeologists can effectively understand the data patterns shown by PCA without the need to understand all its mathematical underpinnings. This statistical method is remarkably relevant to the field of archaeology, where datasets are often complex and involve variables coming from several data sources (Carlson, 2017).

The basic premise of PCA is analysing all observations and reducing large numbers of variables (or dimensions) into fewer constituents that broadly capture the essential features of the initial dataset (Drennan 2009). As opposed to multidimensional scaling and other techniques, PCA begins with a general understanding of the variables in the dataset and their variance. After this initial step, it investigates each pair of variables, focusing on their correlation within the overall data matrix. Overall, PCA captures the variance of the original variables, and it permits the identification of the sets of variables that accounted for the maximal amount of variance in the datasets (Gantley, Whitehouse and Bogaard, 2018). Once this information has been extracted, it is provided in the form of Principal Components (hereafter, PCs), a new set of orthogonal variables that accounts for the maximum amount of variation within the dataset (Abdi and Williams, 2010). The proportion of variance explained by each PC is quantified by the eigenvalues, which, when large, represent most of the original variance and are used to rank PCs and identify a small number of PC that capture the primary axes of variability, effectively condensing the dataset while retaining its essential patterns (Greenacre *et al.*, 2022). Each of the new variables, the PCs, explains a single axis of variability in the dataset, avoiding duplication and using a smaller space that is still able to account for the initial relations between observations and variables (Verzani, 2014).

PCA imposes restriction on the use of variables based on their characteristics. Categorical variables with unranked categories, often coded as integral numbers (1, 2, 3, etc.) are statistically inappropriate for PCA due to their lack of inherent hierarchy order. Presence/absence variables can be functional, while ordinal or ranked variables perform well within PCA. Parallely, and most importantly,

continuous data is the most appropriate variable data type which is often the goal of PCA involving ordered quantitative variables that might vary along spatial, temporal, or typological dimensions (Drennan 2009: 302).

In short, the wide use of PCA in archaeology arises from its capability of reducing complex, multivariate datasets into interpretable principal components that capture the dataset's essential patterns. This dimensionality reduction and its focus on the core axes of the dataset variability enable archaeologists to interpret and visualise relationships across variables in a more effective and illustrative way, including diverse and overlapping sources of data. Moreover, although the primary aim of PCA is not to generate groups, PCs also serve as a valuable tool for exploring the formation of groups and their differences (Glascock and MacDonald, 2023). Therefore, for archaeologists, the ability to identify broad trends from complex and multivariate datasets through PCA has boosted the number of large-scale archaeological investigations, facilitating access and exploration of vast amounts of data, and enabling pattern identification for supporting inferences about past societies, environments, and material culture.

1.2. Cluster Analysis

Cluster analysis is another common quantitative approach for exploring data in archaeology. It refers to numerous techniques employed for turning observations into groups from one or various variables (Carlson, 2017). In short, it is aimed at identifying patterns in the data through grouping observations and detecting an arrangement of subsets or clusters, within which observations maximise internal similarities while minimising similarities between groups (Rencher and Christensen, 2012: 501). Conducted without supervision, as no information about the number nor the characteristics of groups is provided, clustering has equipped archaeologists with a quantitative classification method capable of defining groups of observations on artefacts, sites or other elements based on essential features, which may not be readily apparent through classical research methods, unravelling cultural or historical aspects that underlie archaeological data (Conolly and Lake, 2006; Bai, 2022).

Nevertheless, clustering is not exempt from critiques, particularly concerning the objectivity in generating groups or clusters from multivariate data and their relevance to archaeological interpretation. One of the main concerns regard the clustering outcomes and whether they correspond to real groups or are merely artificial constructs generated by the clustering algorithms (Everitt *et al.*, 2011). The other main preoccupation arises from the partition with the proper number of clusters (i.e., groups originating from the data). Numerous publications have explored this issue, and, unless the nature of the problem being researched suggests a specific number of clusters, the most common approach involves using various indexes as indicators of internally cohesive and mutually distinct clusters (Scitovski *et al.*, 2021). Thus, assessing cluster validity remains essential and, despite the concerns, the widespread use of clustering algorithms in archaeology indicates that, when properly validated, its outcome can satisfactorily contribute to the construction of sensible archaeological knowledge (Baxter, 2015)

Notwithstanding the debate, while in the past clustering was generally a subjective activity in which items were solely grouped according to the assessment of the researcher (Kaufman and Rousseeuw,

1990), cluster analysis assists archaeologists in integrating archaeological quantitative data. Statistically, it sorts the vectors into clusters based on the given variables (e.g., artefact size, geochemical analysis results, spatial data) so that members of a single cluster are closer to each other than to members of other clusters (Glascok and MacDonald 2023). With the help of such group clusters, archaeologists might be able to understand, for instance, the nature of typological variation, spatial positions, and possible time or culture separations (Carlson 2017).

Various research goals come with different types of data; hence a variety of clustering methods can be required. Among the most widely used methods in archaeology, it is worth highlighting both the hierarchical clustering and K-means partitioning methods (Carlson, 2017). Another relevant technique is spatial clustering, which involves the detection of high point density in a two-dimensional space (e.g., a distribution map) (Baxter, 2015), which some link with regionalisation, a traditional concept in geography which subsets spatially contiguous objects with similar characteristics (Wang *et al.*, 2024).

One of the main applications of clustering methods in archaeology is the analysis of attribute data associated with spatial objects. GIS-derived variables such as elevation, slope, visibility, or distance to water can be examined simultaneously to identify groups of sites sharing similar characteristics. Multivariate techniques, including PCA and clustering methods, enable the exploration and statistical assessment of spatial patterning that may not be evident through visual inspection alone. By using these, archaeologists are able to observe how each data point contributes to the formation of larger clusters through different levels, not just spatially but also contemplating the variability of each dataset (Conolly and Lake, 2006). Therefore, hierarchical clustering was considered the most suitable method for this paper. Each spatial variable contributed to calculating the similarity or "distance" between sites, hence eventually aiding in grouping the sites without predefined classes. The output of this clustering approach, normally a dendrogram, begins by linking the most similar sites at the lowest hierarchical level, gradually merging them with other clusters as it progresses upwards, resulting in a tree-like structure that illustrates nested relationships among sites. This hierarchical structure is particularly useful for identifying natural groupings within the dataset that might correspond to potentially shared socio-political, environmental, or strategic choices. By examining branches at different levels of the dendrogram, we can identify clusters that reflect varying degrees of similarity, allowing for an analysis of how site characteristics may have influenced settlement location across the landscape (Hodson, 1970).

1.3. Self-Organising Maps (SOM)

Artificial neural networks (ANN) represent a distinct class of machine learning algorithms developed for modelling complex, non-linear relationships. The basic unit of an ANN operates as a switch that receives inputs from other units or external sources, weights and sums them, and produces an output when a given threshold is surpassed (Krogh, 2008). The key advantage of ANN is their ability to model complex, pattern-oriented problems without requiring prior knowledge of the data distribution nor interaction between variables (Walczak and Cerpa, 2003), an aspect that is extremely valuable for the exploratory analysis that can help read fragmentary archaeological data.

Among the architectures and algorithms used for ANN, the self-organising map (hereafter, SOM), also known as the Kohonen map, is an unsupervised artificial neural network developed by Teuvo Kohonen (1982) and inspired by the topographic organisation of the cerebral cortex. Designed for the analysis and visualisation of high-dimensional data, SOM project multivariate datasets onto a low-dimensional grid, normally two-dimensional, while preserving the topological relationships between observations (Kiang, 2001). Through an iterative and competitive learning process, similar input vectors are mapped onto neighbouring nodes, enabling the identification of underlying structures, gradients and clusters within complex datasets. In terms of structure, SOM algorithms are made up of two layers: an input layer, which contains one neuron (i.e., model unit) per dataset variable, and an output layer (often named competition layer), comprising a geometrically ordered topological grid, normally two-dimensional (Kohonen, 2013). Each input neuron is connected to every neuron in the competition layer, and each of the latter is assigned a weight vector that represents a model of the observations (Figure 1). During training, the algorithm iteratively adjusts these prototypes so that they best describe the input layer, positioning them on the output grid so that similar models occupy neighbouring nodes and dissimilar ones lie further apart. In this way, the mapping process preserves the neighbourhood relationships of the original N-dimensional data within the low-dimensional grid, generating a topological projection of the dataset (Kohonen, 2001). A key decision left to the user, however, is the number of neurons in the low-dimensional grid, which determines its size and depends on the volume and dispersion of the input data.

Unlike more conventional linear dimensionality reduction techniques such as PCA, SOMs can model nonlinear relationships between variables while simultaneously performing dimensionality reduction, clustering, and visual exploration of the data. This makes SOMs especially suitable for exploratory analysis in datasets characterised by complex statistical interactions, multicollinearity and noisy variables (Skupin and Agarwal, 2008). The SOM algorithms establish a relationship between a high-dimensional input space and a lower-dimensional feature map through a two-layer fully connected architecture trained without external labels, allowing patterns to emerge directly from the structure of the data itself.

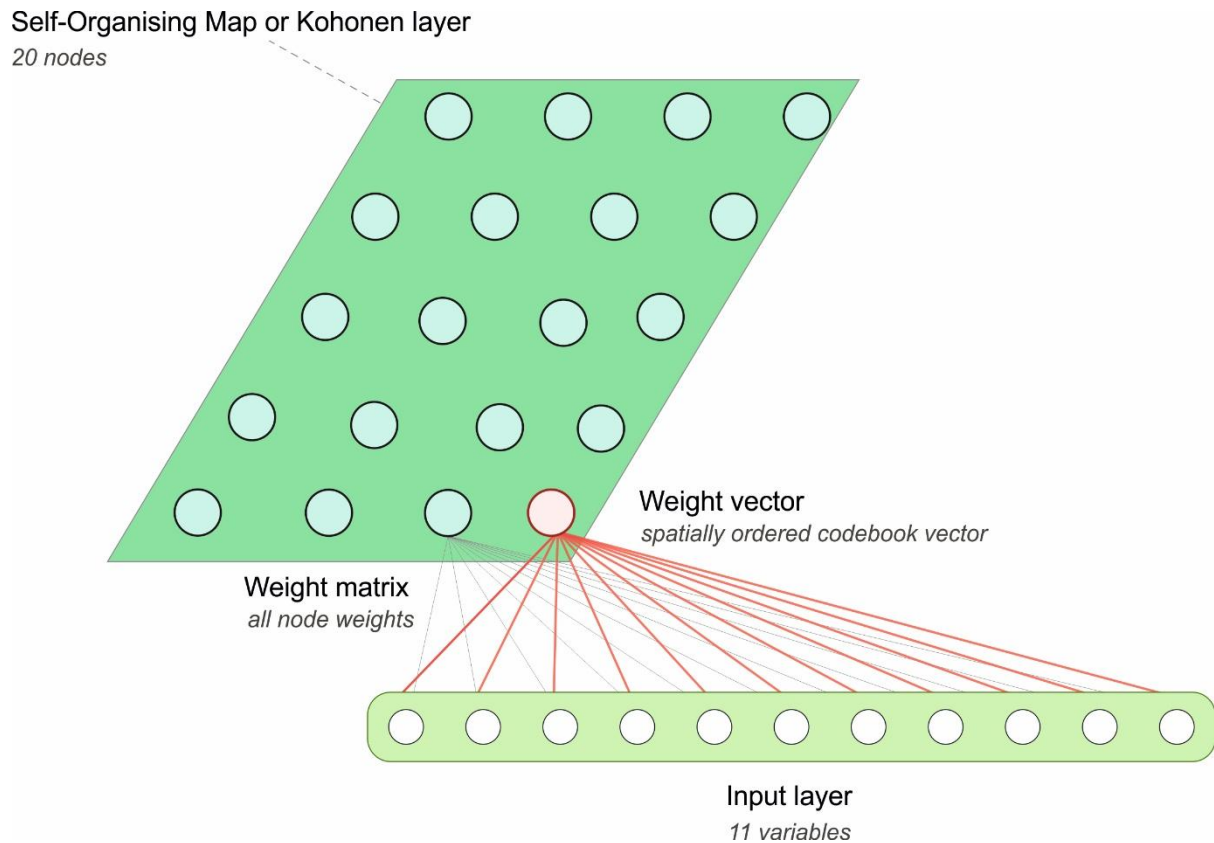


Fig. 1 Architecture of a self-organising map, configured as in the present study. The input layer corresponds to the eleven variables recorded for each site, and the Kohonen layer comprises the twenty nodes used here (a 4×5 grid). Every node is connected to all eleven input neurons, and the set of weights on these connections forms the node's weight vector, or codebook vector — a prototype in the same variable space as the sites (highlighted in red for one node). During training, each site is assigned to the node whose codebook vector most closely resembles it (its best-matching unit), and the codebook vectors are iteratively adjusted so that neighbouring nodes come to hold similar profiles, yielding a spatially ordered representation of the data. The complete set of codebook vectors across all twenty nodes constitutes the weight matrix.

Since such unsupervised training does not require predefined categories nor external guidance, SOMs have been widely applied to pattern recognition, dimensionality reduction, data visualisation and, in particular, they have become a crucial step in data mining protocols across multiple disciplines, such as speech recognition, text analysis, industrial process control, finance, image data compression, robot control, telecommunications, or biomedical analyses (Shieh and Liao, 2012; Nogales, Guadalupe and García-Tejedor, 2025). SOM approaches have been implemented in several commercial and open-source software packages, such as the R package *kohonen* employed in this work (Wehrens and Buydens, 2007), or the Python library *GEMA* (García-Tejedor and Nogales, 2022). In archaeological research, their capacity to represent complex multivariate relationships in an interpretable two-dimensional form makes SOM especially relevant for landscape and GIS-based analyses, where environmental, topographic, and spatial variables often exhibit nonlinear interactions. Despite this potential, their use in the study of the past is still limited. Existing studies have explored their use in areas such as the classification of fragmented and non-standardised ceramics (Korokhina, 2024) or settlement evolution reconstruction to identify specific morphological patterns (Lu *et al.*, 2019), but their broader applicability for archaeological spatial analysis and landscape characterisation remains entirely underexplored.

2. Case study: early Medieval landscapes. The Upper Arlanza Basin (Burgos, Spain)

The case study selected for the study is the Upper Arlanza Basin (Alto Arlanza, Burgos, Spain), an upland region characterised by the scarcity of written and archaeological data for the transitional centuries between late Roman and the early medieval times. More broadly, the northern Iberian Peninsula, where the processes of Romanisation and Christianisation developed very unevenly across the territory, constitutes a highly heterogeneous early historic landscape. Settlement patterns vary considerably between regions, a phenomenon that has been named as *micro-regionalities* (Wickham, 2005). These dynamics are particularly pronounced in mountainous environments, where topography conditioned communication and economic strategies, limiting integration into broader political structures and often leading to locally-constructed social realities (Martín Viso, 2025).

Despite these limitations, archaeological research carried out in the uplands between the headwaters of the Arlanza and Duero rivers has documented a dense network of settlements associated with rock-cut cemeteries and small churches (Figure 2), with occupation sequences from the 5th to the 13th century (Álvaro Rueda, 2012; López Perez, Álvaro Rueda and Allepuz, 2016; Álvaro Rueda, Travé Allepuz and López Pérez, 2018). Therefore, these settlements constitute an outstanding case study for analysing territorial organisation and settlement dynamics during the formation of medieval rural societies.

The Upper Arlanza Basin presents environmental characteristics typical of a mountain landscape, where altitude and rugged terrain historically constrained agricultural intensification while favouring livestock practices and forest exploitation (López Perez, Álvaro Rueda and Allepuz, 2016). These environmental constraints strongly influenced local settlement distribution, generating distinct occupation patterns across the territory.

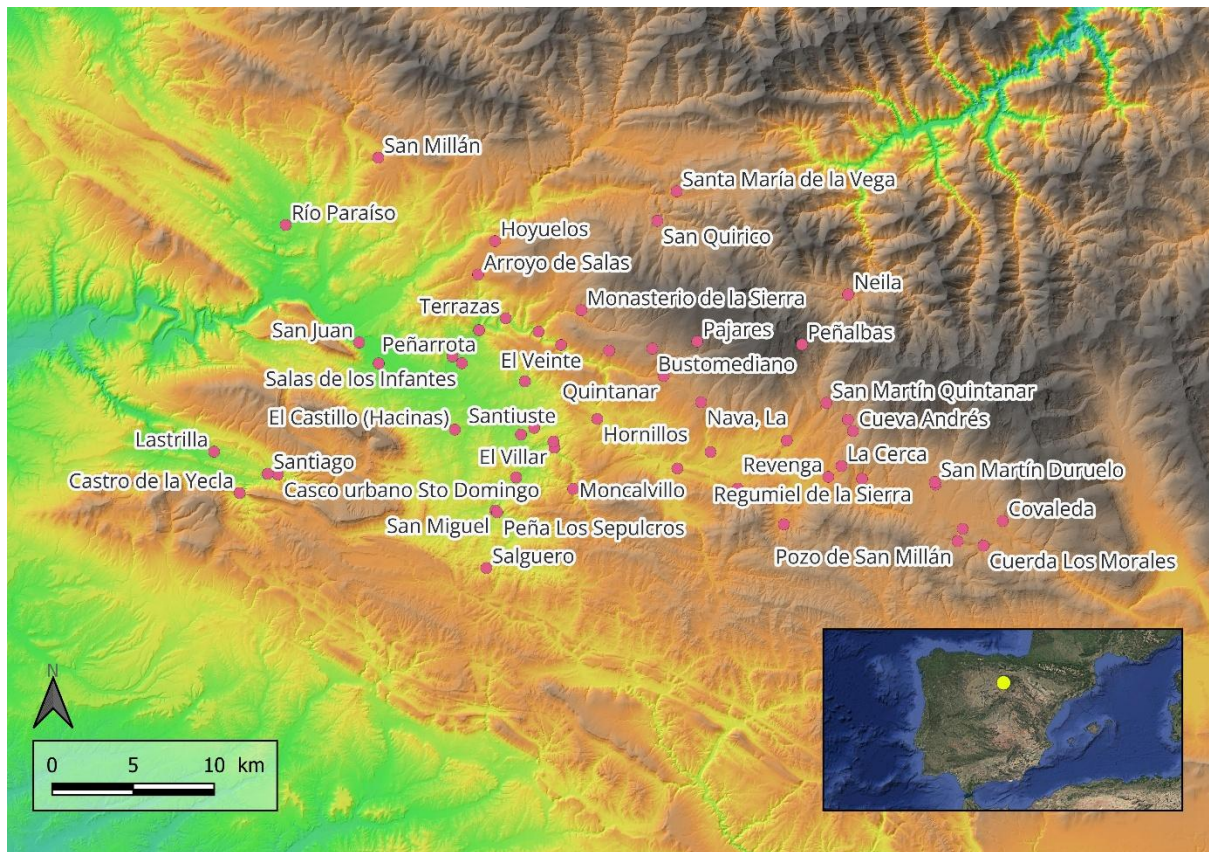


Fig. 2 Archaeological sites in Alto Arlanza with phases of occupation corresponding to the transition to the medieval world.

Archaeological evidence suggests a marked hierarchy within the settlement network. Some of the enclaves, characterised by extensive cemeteries, appear to have functioned as articulating nodes within broader local networks (López Perez, Álvaro Rueda and Allepuz, 2016: 177–179). Regardless of their size, most settlements are associated with a church and cemetery located adjacent to the habitation spaces, hence indicating the presence of stable rural communities (Álvaro Rueda, 2012). The spatial distribution of sites further reveals significant variability linked to topographic conditions: settlements above 1200 m a.s.l. are rare and likely correspond to seasonal occupations associated with pastoral activities, whereas lower-altitude sectors display denser and more complex settlement networks. Present-day inhabited villages coexist with numerous abandoned sites preserving early medieval occupation evidence, suggesting that population distribution in the past was substantially more dispersed than today.

These characteristics make the Upper Arlanza Basin particularly suitable for exploratory multivariate analysis. The combination of environmental diversity, heterogeneous settlement patterns, and complex spatial relationships provides an ideal context for evaluating the potential of SOM-based approaches for identifying landscape structures and territorial dynamics in archaeological datasets.

3. Methodology: Principal Component Analysis, Cluster Analysis and Self-Organising Maps

The selection of spatial variables is a crucial step for exploring and interpreting archaeological landscapes. In the present paper, we examine a set of spatial variables collected from early medieval sites in the Upper Arlanza Basin. These variables, which include environmental factors and visibility metrics, provide insights into how site selection, resource accessibility, and strategic positioning might have shaped settlement patterns during this period. This study applies PCA, hierarchical clustering and SOM to the same landscape dataset in order to assess whether SOM outperforms the more established techniques, revealing more meaningful or diverse patterns and supporting richer inferences about the landscape and its organisation (Drennan, 2009; Baxter, 2015; Glascock and MacDonald, 2023). The statistical analysis was carried out in R (v4.5.1; R Core Team, 2025) and the code is available on GitHub (LINK).

The variables selected for the analysis were designed to capture the environmental, topographic, and spatial characteristics that may have influenced settlement location in the Upper Arlanza Basin during the Early Middle Ages. The selected variables are proxies of local accessibility, visibility, mobility, and terrain conditions. Unless otherwise specified, raster-derived variables were sampled at site locations using the QGIS plugin *Point Sampling Tool*. All variables were standardised in R using the *scale()* command prior to the statistical analysis so as to avoid scale-dependent distortions. By default, this centres each variable to a mean of zero and scales it to unit variance using the sample standard deviation. The complete list of variables used for this analysis is provided below:

- *Altitude* (m): elevation of each settlement above sea level. Obtained from the 5 m resolution Digital Terrain Model provided by the National Geographic Institute of Spain (IGN).
- *Distance to Nearest Neighbour* (m): measures the proximity between settlements, providing an indication of spatial organisation and potential interaction between sites. Without aiming to emulate perspectives coming from point pattern analysis (Brandolini and Carrer, 2021), this metric helps to group sites by proximity. Calculated using the QGIS tool *Distance to Nearest Hub*.
- *Connectivity* (no unit): represents relative accessibility and movement potential across the landscape based on circuit theory models (Howey, 2011; Mcrae *et al.*, 2016; Anantharaman *et al.*, 2020). A cumulative current raster was generated in Julia using Circuitscape (McRae *et al.*, 2008; Anantharaman *et al.*, 2020) from a cost surface derived from slope, obtained in QGIS from the 5-m DTM and later modified with slope and hydrological barriers. Values were eventually sampled at settlement locations.
- *Aspect* (degrees): indicates the orientation of each site, reflecting differential exposure to sunlight and environmental conditions. Computed in QGIS using the *Aspect* algorithm.
- *Slope* (degrees): measures terrain steepness around each settlement, reflecting constraints and opportunities related to accessibility, agriculture, or defensibility. Computed in QGIS using the *Slope* algorithm.
- *Distance to the nearest river* (m): approaches the relationship between settlements and water resources, a key factor in site location and mobility. Computed in QGIS using the *Distance to the nearest hub* tool, employing the river shapefile downloaded from the IGN.

- *Visibility from the site*: estimates the extent of the surrounding landscape visible from each settlement, potentially reflecting strategic control over routes or territories (Brughmans and Brandes, 2017). Computed using the QGIS *Visibility Analysis* plugin (visibility index outgoing).
- *Visibility to the site*: measures how visible each settlement is from surrounding areas, approaching its visual prominence within the landscape (Eve and Crema, 2014). Computed using the QGIS *Visibility Analysis* plugin (visibility index incoming).
- *Prominence*: evaluates the overall visual dominance of settlements within the territory through cumulative viewshed analysis from 100 random points across the study area. Computed from cumulative viewsheds generated in QGIS and sampled at archaeological sites.
- *Roughness*: measures terrain irregularity surrounding each settlement, reflecting the degree of landscape ruggedness and potential mobility constraints. Computed in QGIS using the *Roughness* algorithm.
- *Global Sun Radiance*: estimates overall solar exposure, an important factor in mountain environments affecting habitability and agricultural potential (Basile and Campus, 2023). Calculated in QGIS using the GRASS GIS algorithm *r.sun.insoltime*.

PCA was performed in R using the *prcomp()* function. Results were explored and visualised using the *FactoMineR* and *factoextra* packages, complemented by *ggplot2* and *ggfortify* for the visual assessment of the results (scree plot, variable loadings, a site-variable biplot and variable contributions to the leading components).

Cluster analysis was carried out in R employing the standardised variables: pairwise Euclidean distances between sites were computed with *dist()*, and the sites were then clustered using Ward's minimum-variance method (*ward.D2*) via *hclust()*. The optimal number of clusters was assessed using an elbow plot of the within-cluster sum of squares, and the resulting dendrogram was visualised with the *dendextend* package.

4. Results

In the section the results of the three methods applied to the Alto Arlanza case study are presented and discussed. Particular emphasis is given to the key differences between each method, to establish strengths and weaknesses of their application to archaeological spatial data.

4.1. PCA

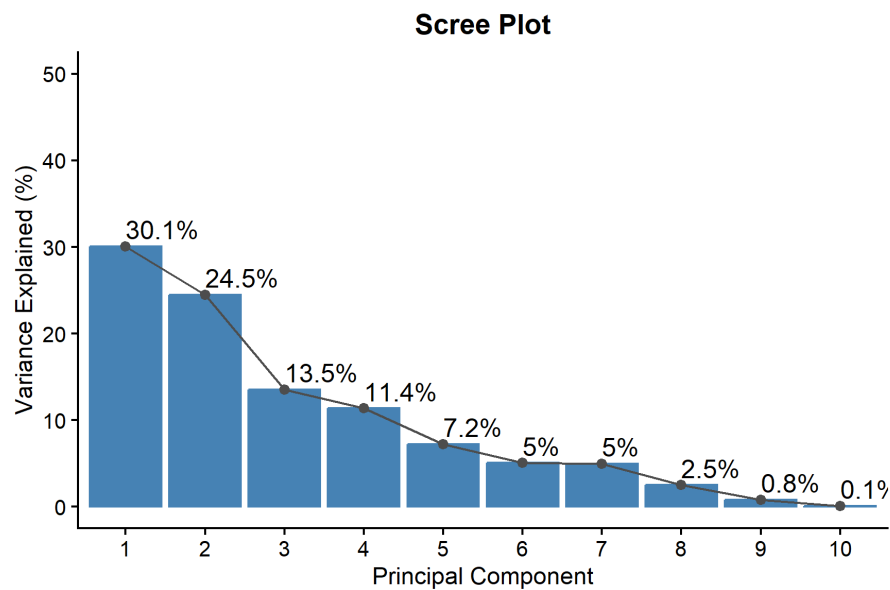


Fig. 3 Scree plot depicting the amount of variance explained by each principal component.

PCA has recombined the 11 spatial variables into 10 components (or dimensions). Figure 3 shows that the first two principal components (PC) explain nearly 55% of the total variance of the dataset. While this dimensionality reduction facilitates the exploration of the main patterns captured by the first two PCs, a substantial proportion of the variance remains distributed across the remaining components. This not only highlights the complexity of the dataset but also limits the depth of our landscape interpretations based on these PCA results, hence justifying the use of complementary approaches, such as SOMs, to explore additional structures and patterns that may not be fully uncovered in PCA.

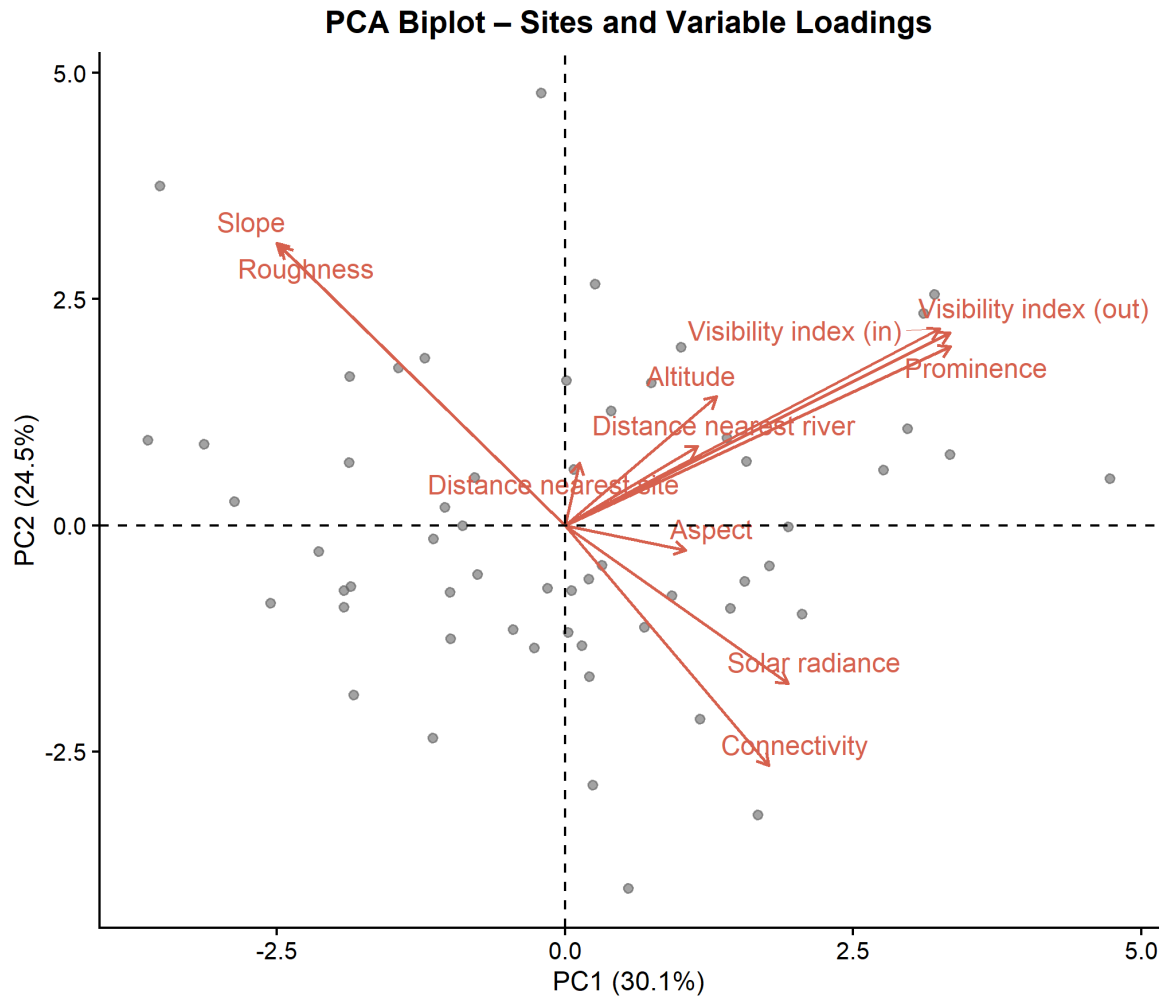


Fig. 4 PCA biplot. Points represent archaeological sites and arrows correspond to the spatial variables. Arrow's direction and length show how each landscape variable contributes to PC1 and PC2, respectively.

The PCA biplot in Figure 4 reveals that PC1 primarily reflects a contrast between visually dominant and territorially integrated settlements and sites located in steeper and more topographically constrained areas. On the other hand, PC2 distinguishes elevated and visually prominent sites from settlements characterised by higher connectivity and solar exposure, suggesting that visually dominant locations were not necessarily the most accessible within the communication network in the Upper Arlanza Basin.

In parallel, the PCA biplot also shows strong positive correlations between various landscape variables. This is evident when looking at incoming/outgoing visibility indices and prominence, which capture closely related dimensions of landscape visibility and may lead to potential multicollinearity issues. A similar pattern is observed for slope and roughness, both derived from related topographic properties and showing comparable contributions to PC2. In such cases, subsequent analysis could be simplified by maintaining only one of these variables with the highest contribution to each PC (prominence, with a 20% contribution to PC1, and slope, which accounts for over 20% of PC2 variance).

4.2. Hierarchical clustering: Dendrogram

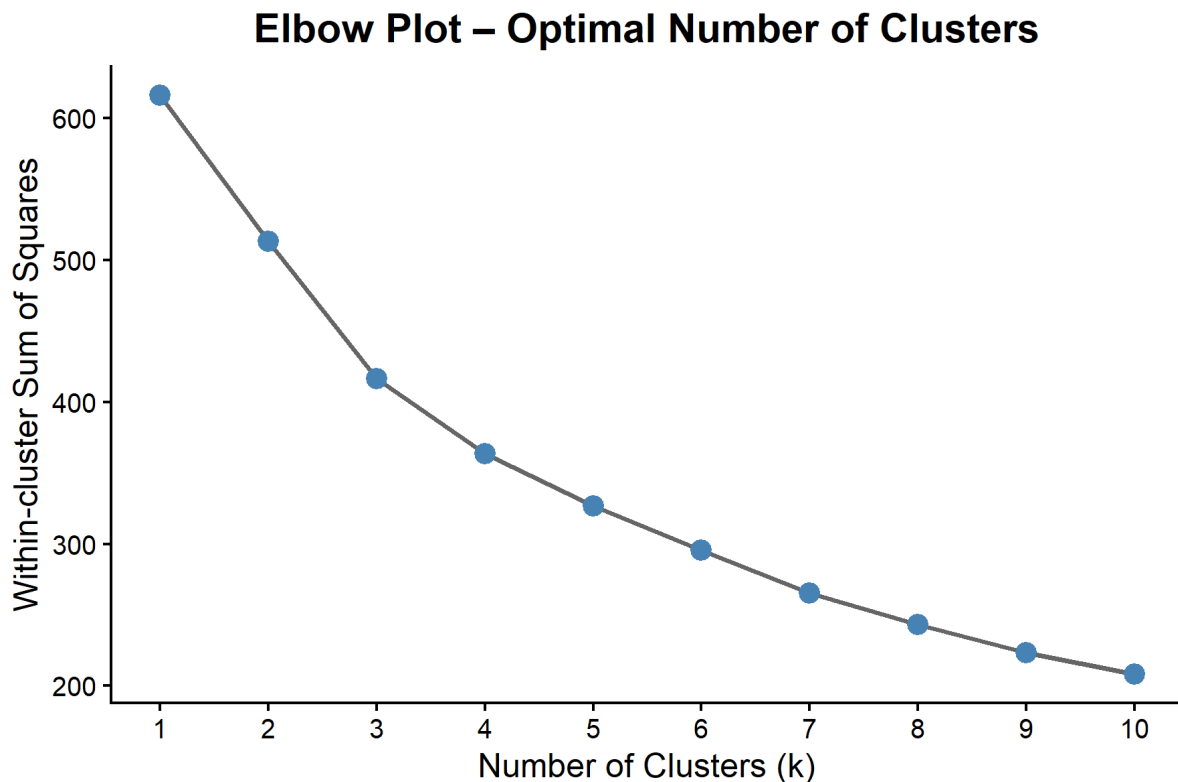


Fig. 5 Elbow plot showing the relationship between the number of clusters (k) and the within-cluster sum of squares (WCSS).

Hierarchical clustering enabled the analysis of landscape variables and the grouping of the early medieval sites according to the seven spatial dimensions employed. As a preliminary step, an elbow plot (Figure 5) was employed to determine the optimal number of clusters (k) in which the settlements can be grouped. The results suggest that the most suitable division is between 3 and 4 clusters, after which within-cluster variance becomes less relevant. Such a relatively smooth decrease of the line also indicates that archaeological sites are distributed along continuous topographical and spatial gradients, without forming sharp differences between clusters.

The hierarchical structure of the dataset is provided in a dendrogram (Figure 6), a standard visual output for hierarchical clustering approaches. The analysis was computed using Ward's method with Euclidean distance, a clustering approach that assesses the variability within a cluster and produces relatively balanced clusters, combining observations inside clusters and clusters into larger clusters (Baxter, 2015; Carlson, 2017). The dendrogram shows four broad clusters of settlements sharing similar landscape characteristics based on the selected variables. Cluster 1 contains a relatively large number of sites that merge progressively, suggesting greater internal variability. Clusters 2 and 3 depict stronger internal similarity, whereas cluster 4 is potentially the most distinctive and has a more unique combination of landscape attributes. However, the gradual merging of cluster branches across the dendrogram indicates that settlement organisation was structured along continuous multivariate gradients and complex spatial relationships. Thus, these results also reinforce the complexity already identified through PCA and denote some of the limitations of focusing only on hierarchical classifications for representing archaeological landscapes.

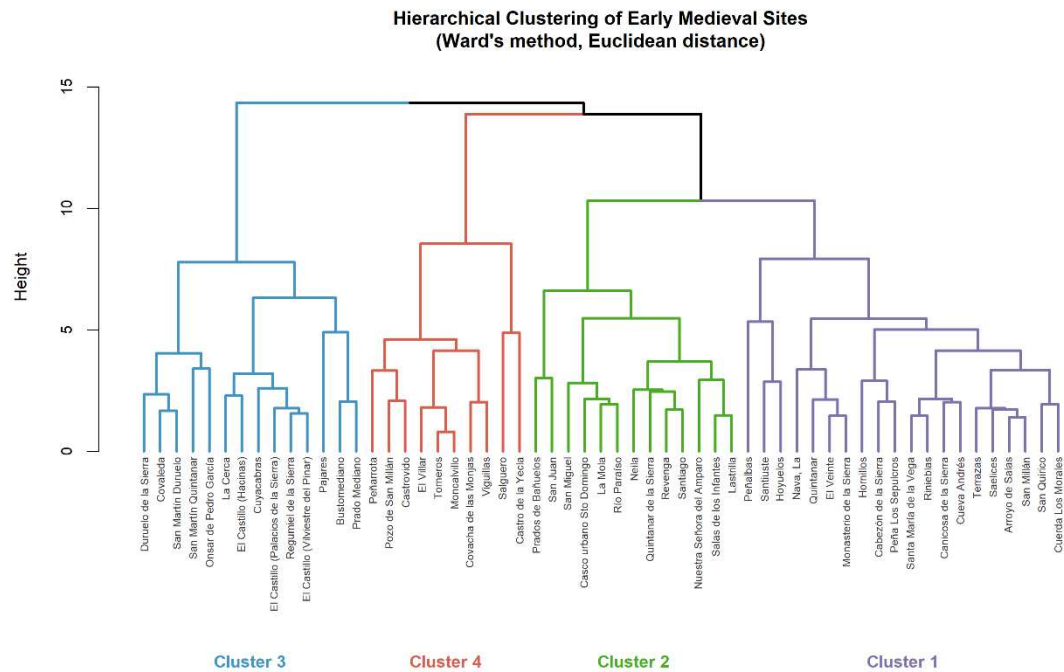


Fig. 6 Dendrogram of all sites with branches coloured by cluster. Site names are indicated.

4.3. Self-Organising Maps

To classify archaeological sites based on their 11 landscape variables, the first step was training the SOM on a 4x5 hexagonal grid using Gaussian neighbourhood function over 400 iterations with a learning rate decaying from 0.08 to 0.01. The training progress plot (Figure 7) depicts how the mean distance to the closed codebook vector (the core of SOM algorithm: feature vectors with similar characteristics are grouped) progressively decreases across all 400 iterations, with the flat ending showing that the model is stable rather than continuing learning. Such observation is further validated by the node counts map (Figure 8), since it shows that the selected grid is not too large (i.e., no empty nodes) or too small (i.e., overcrowded nodes that mask variation), with the majority of them including 2-3 sites.

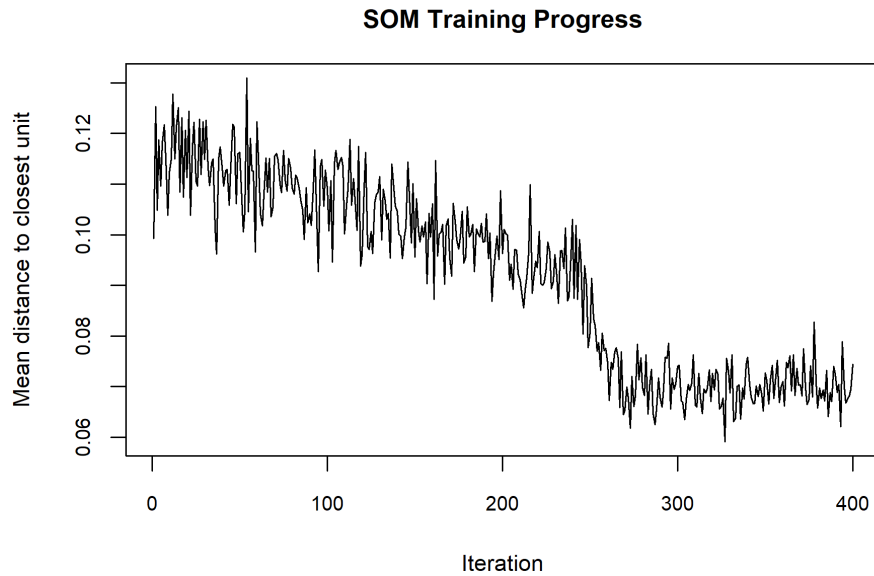


Fig. 7 SOM training progress

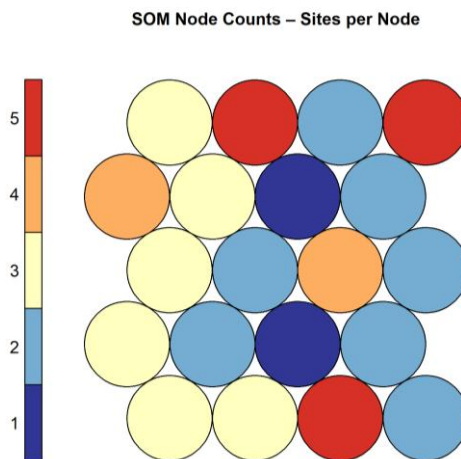


Fig. 8 SOM Node Counts. Each circle (node) is coloured according to the number of sites it represents.

In alignment with the clustering procedure described above, and to unravel which site groups underlie the dataset, the SOM codebook vectors (i.e., a prototype in the same variable space as the sites, representing the average profile of the observations assigned to it) were further analysed to determine the most suitable number of clusters. Following the same logic as the hierarchical clustering approach, an elbow plot was computed on the codebook vectors, which also supported the four-cluster solution ($k=4$, Figure 9). The convergence between two independent elbow plots (original dataset and SOM nodes) denotes that the 4-cluster structure is not an artefact produced by a single method and, instead, it represents an underlying pattern of the dataset. The SOM nodes (i.e., the codebook vectors that they contain and synthesise the different sites) were then grouped into four clusters by hierarchical clustering of the codebook vectors using Ward's method, allowing each site to receive a cluster membership through its best-matching node while enabling a direct comparison between the SOM and hierarchical clustering approaches.

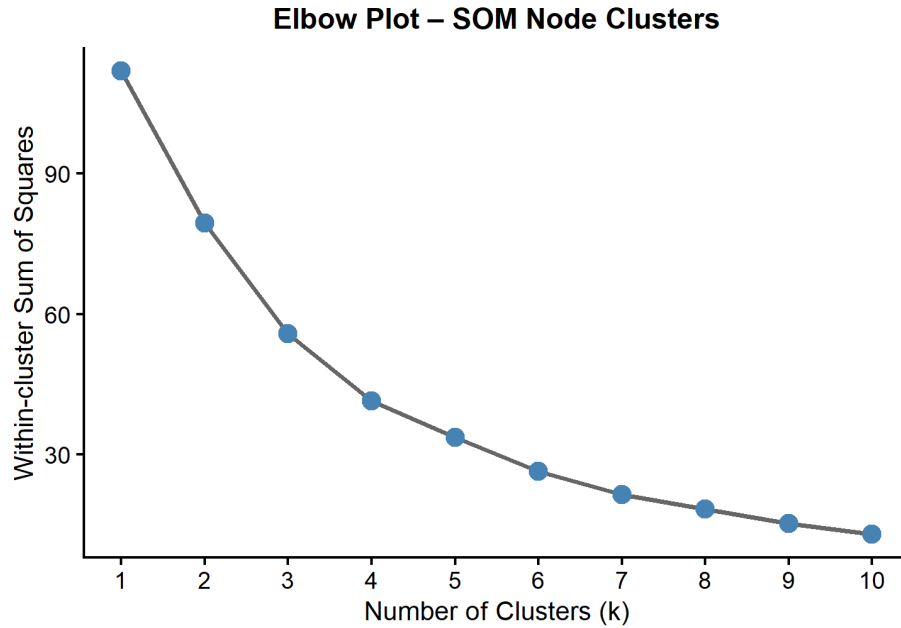


Fig. 9 Elbow plot showing the relationship between the number of clusters (k) and the within-cluster sum of squares (WCSS).

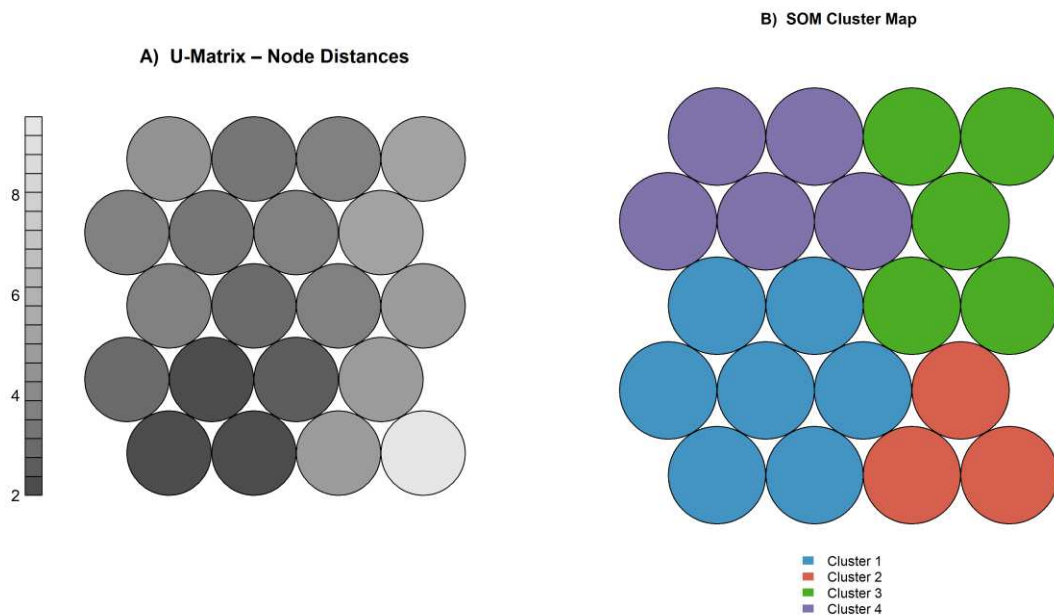


Fig. 10 A) U-Matrix Node distances. Each node is shaded by the distance to its neighbours; darker areas indicate greater dissimilarity between adjacent nodes. B) SOM cluster map with each node coloured by cluster membership, showing which areas of the node map belong to a specific cluster.

The U-Matrix Node distance map (Figure 10A) is one of the crucial outputs of SOM analysis, as it provides an internal representation of cluster structure based on the distance between neighbouring nodes, independently of clustering algorithms. At the same time, the SOM cluster map (Figure 10B) depicts the four clusters ($k = 4$) as contiguous regions within the SOM grid, revealing topological relationships that are not captured in other approaches such as hierarchical clustering. For example,

Clusters 1 and 3 share a common boundary, suggesting the existence of intermediate landscape configurations between both groups. Moreover, it is also worth noting the close correspondence between node distance patterns (Figure 10A) and the cluster distribution map (10B), reinforcing the coherence between the latter and the topographic structure of the trained SOM.

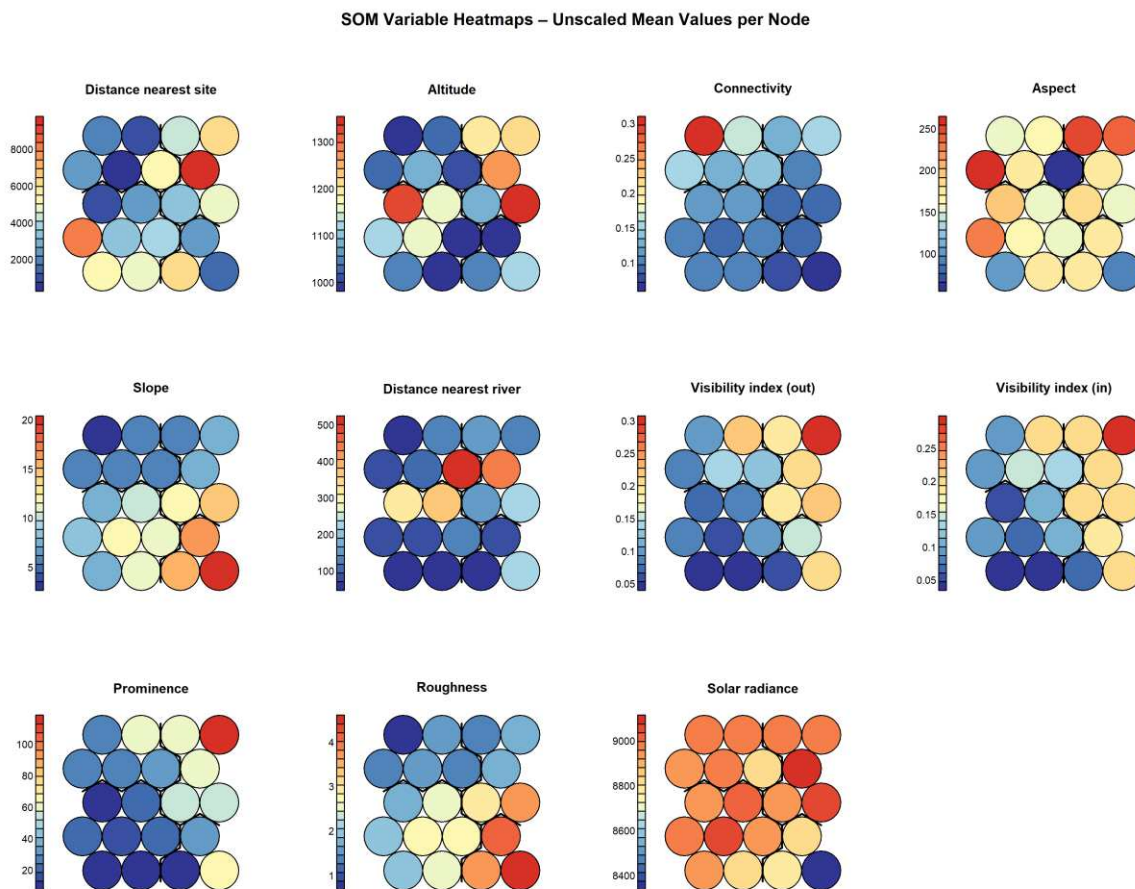


Fig. 11 Panel of heatmaps, one per variable, showing the mean value of each unscaled variable across the SOM grid nodes. Black boundaries represent the clusters.

Figure 11 presents the SOM variable heatmaps using unscaled values retaining real-world units (metres of altitude, degrees of slope, meters of distance, etc.), thus making the patterns directly interpretable in landscape terms without requiring standard deviation calculations. The panel of plots is crucial due to the several patterns that they represent. First, both visibility indices (in/out) and prominence show nearly identical behaviour on the grid, confirming what PCA loadings already suggested, with high values concentrated at the top right corner corresponding to sites in cluster 3. Roughness and slope show a contrasting but equally clear pattern, and their highest values can be found in the lower-right nodes, isolating cluster 2 as a group of sites located in strongly rugged, steep spaces. Connectivity shows a marked pattern and is concentrated in the upper-left nodes of cluster 4, mostly independent from the rest of the variables. This spatial independence on the SOM grid is analytically significant, as it confirms that connectivity captures a territorial dimension that other variables do not, and that well-connected sites form a coherent group not reducible to either the high-visibility or the rough-terrain categories. Distance to the nearest river also shows a concentration of sites in cluster 4 and populated nodes across the remaining 3 groups, which denotes a more gradual gradient and not categorical differences. A similar pattern can be observed in the altitude, with high-elevation nodes distributed across more than one cluster, suggesting that altitude alone does not

define cluster membership and interacts with other variables, as opposed to the connectivity. Solar radiance, instead, is the most uniformly distributed variable with high values across the grid, only presenting a clear outlier in cluster 2, where the combination of steep and rough spaces is also accompanied by a lower solar radiance value.

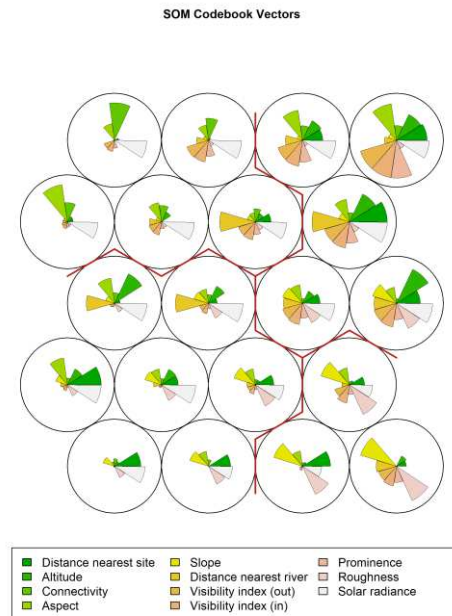


Fig. 12 SOM grid with a fan plot showing the relative values of all 11 variables for each node.

The codebook fan plot displays all landscape dimensions simultaneously (Figure 12). Looking at the grid, it is possible to see nodes in the upper-right area with large prominence and visibility values, whereas nodes in the lower-left area have high slope and roughness values. Such gradual visual presentation between node variability moving across the SOM grid illustrates one of its key advantages over hierarchical clustering, as it maintains the topology of the data and show that site types grate into one another.

5. Discussion: comparing PCA, Hierarchical Clustering and SOM

The three methods applied here are complementary rather than alternative, and their comparison reveals both the structure of the dataset and the possibilities of each approach when used in isolation.

The PCA ordination establishes that settlement organisation in the Upper Arlanza Basin is structured through continuous topographic and environmental variables rather than discrete categories. The absence of clearly separated site clusters in PC space, combined with the modest variance explained by two components (nearly 55%), indicates that the relationships between landscape variables and site location are partially non-linear. Interpretations assuming clean categorical boundaries risk imposing structure on the data instead of unravelling it.

In this line, hierarchical clustering reinforces this picture. The dendrogram confirms that meaningful groupings exist, but their boundaries are fuzzy rather than sharp, as reflected in the progressive fusion

heights of the four-cluster solution. As Baxter (2015) and Everitt *et al.* (2011) note, no single clustering method performs best in all circumstances, and standard hierarchical approaches, operating on linear distance measures and forcing binary decisions at each node, can misrepresent genuinely transitional sites by assigning them to a single branch at an arbitrary threshold.

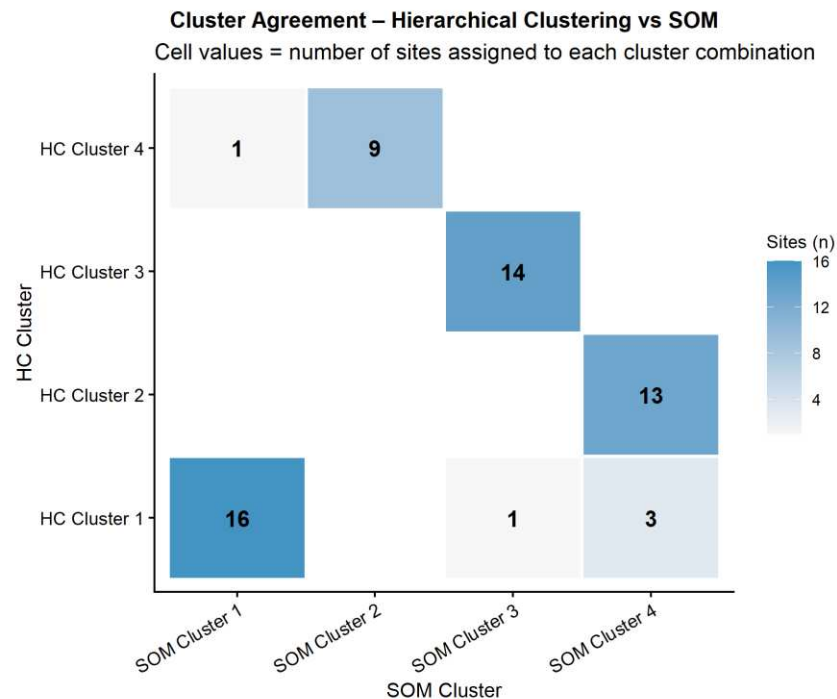


Fig. 13 Cluster agreement matrix between hierarchical clustering (HC) and the self-organising map (SOM). Cell values indicate the number of sites assigned to each cluster combination. Each HC cluster corresponds to one dominant SOM cluster (HC1–SOM1, HC2–SOM4, HC3–SOM3, HC4–SOM2). Only five sites fall outside this correspondence.

The SOM addresses such limitations directly. By preserving neighbourhood relationships in its output, the trained SOM model captures the continuous gradient structure that the dendrogram approximates but cannot fully represent. The cluster agreement matrix (Figure 13) demonstrates that both methods recover the same primary structure (16, 9, 14 and 13 sites assigned to matching cluster pairs) while the disagreements are concentrated among sites with intermediate landscape characteristics, which are more likely to be misclassified in a rigid hierarchical framework. The side-by-side PCA comparison (Figure 14) presents more evidence on that regard, and both methods produce broadly similar groupings in the PCA space, although SOM ellipses better reflect the gradient character of the data, with cluster boundaries overlapping where the underlying landscape transitions are gradual.

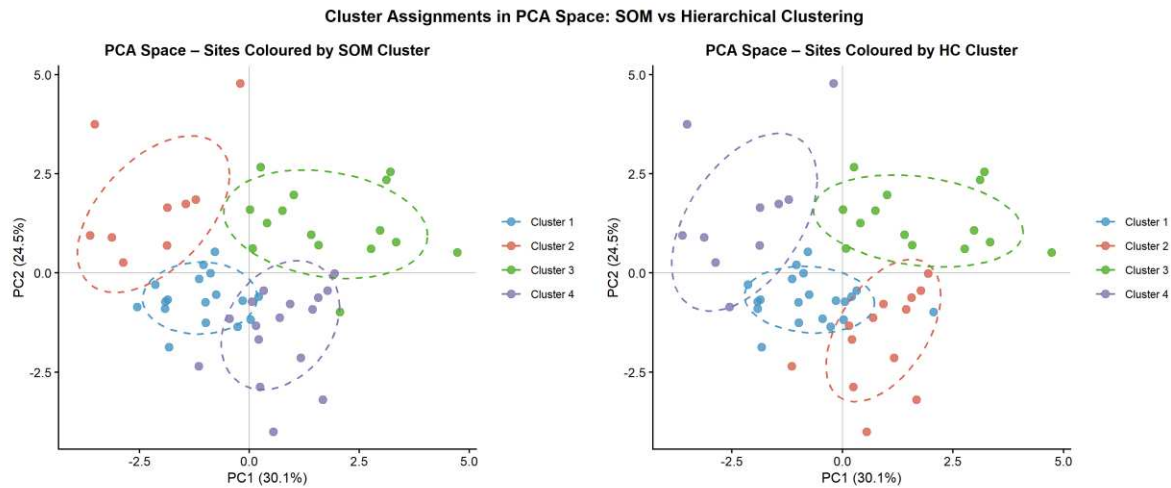


Fig. 14 Cluster Assignments: SOM vs Hierarchical Clustering in PCA space

Another crucial advantage specific to SOMs lies in the visualisation of its outputs. The ability to characterise each cluster in real-world units through the variable heatmaps, for example, eases the process of making inferences about aspects such as settlement dynamics by providing a single readable summary, which is something neither a biplot nor a dendrogram can offer in a single figure. This interpretive directness makes the SOM particularly relevant for landscape archaeology, whose goal generally goes beyond statistical classification and aims to unravel aspects such as settlement location patterns.

One thing that should be considered when using SOM for spatial data is the influence of spatial autocorrelation. This issue has been widely investigated for PCA and hierarchical clustering (Rondelli *et al.*, 2014; Ruiz-Giralt *et al.*, 2026), but, to the best of our knowledge, it has not been addressed in SOM. The identification of statistically significant hot and cold spots in the spatial distribution of codebook vectors could reveal meaningful relationships between human activities and environmental parameters and suggest the existence of missing variables.

Figure 15 projects the results of the three methods onto the Alto Arlanza area, allowing the classifications to be read against the landscape itself and offering a first, visual approximation to the spatial structure that a formal autocorrelation analysis would quantify. At a visual level, a consistent pattern emerges and distinguishes sites at higher altitudes, sites along the valley bottoms near the rivers, and sites on the more open land towards the west of the study area. In panel (a), the PC1 gradient represents this structure closely, since sites at the upper end of the scale concentrate in the highest parts of the Arlanza basin, while strongly negative scores cluster in the lower western zones, confirming that the dominant axis of variation captures an altitude-related environmental gradient. Panel (b) translates this gradient into discrete groups, with cluster 1 clearly associated with the more rugged, elevated terrain. Notably, in the easternmost part of the map, three neighbouring sites are assigned to three different clusters, providing a key reminder that geographical proximity does not imply variable similarity. Panel (c) shows the SOM classification, in which most sites retain their hierarchical clustering group, while a smaller number of settlements change group, precisely the intermediate cases where the two methods would be expected to diverge. The overall picture is therefore one of strong convergence on a spatially coherent set of site groups, with the residual disagreements concentrated at the boundaries between them, which the SOM makes explicit.

As revealed in Figure 13, only 5 of the 57 sites (around 9%) received different assignments once the cluster labels were aligned between the two methods. These disagreements are themselves informative. Three of the five (Canicosa de la Sierra, Santiuste and Cueva Andrés) occupy near-zero positions on the PC1 gradient — profiles intermediate between the low-visibility and valley-floor settlement types — and all three shift consistently. These are precisely the cases in which a dendrogram, forced to cut the tree at a fixed height, must commit a borderline site to one branch or the other, whereas the topology-preserving SOM approach assigns them to the neighbouring group (i.e., node) whose codebook vector they most closely resemble. The remaining two sites (Covacha de las Monjas and La Nava) lie at the extremes of the PC1 gradient, indicating that their reassignment is driven not by the dominant axis of variation but by secondary variables, visibility and prominence in this case. Rather than undermining the classification, these divergences exemplify one of the main advantages of SOM over other approaches, and they also identify the transitional sites within the study area: locations where the boundaries between settlement location patterns are fuzzy. In parallel, their unstable classification between methods makes them outstanding candidates for spatial autocorrelation analysis, which could shed light on whether the ambiguity arises from a genuine transition in the landscape or results from influence of variables not yet captured in the model.

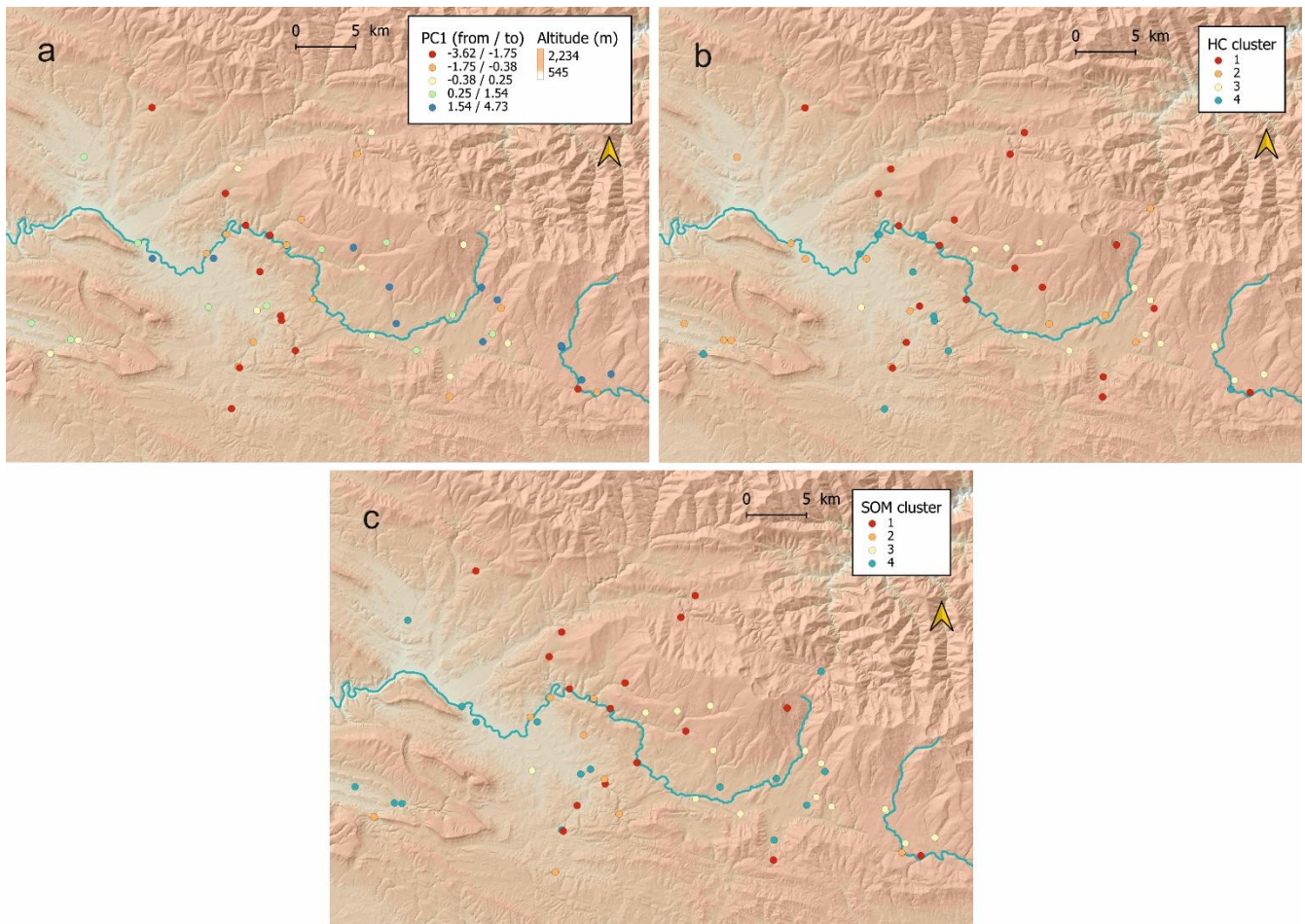


Fig. 15 The 57 sites of the Upper Arlanza Basin colour-coded by each of the three analytical approaches. To allow direct comparison of the classifications against the terrain, SOM clusters have been coloured according to the Cluster agreement matrix (Figure 13): SOM cluster 2 coincides with HC cluster 4, and SOM cluster 4 coincides with HC cluster 2. (a) Each site coloured by its score on the first principal component (PC1), the axis capturing the greatest variance across the eleven variables; the graduated scale runs from low (red) to high (blue) PC1 values, so that sites of similar colour share a comparable position along the dominant environmental–topographic gradient. (b) Sites classified into four groups by hierarchical clustering (Ward's method). (c) Sites classified into four groups by the self-organising map. In (b) and (c), cluster colours are matched between the two methods so that corresponding groups can be compared directly. The two main rivers of the area (Arlanza, left, and Duero, right) are represented in the maps for reference.

6. Conclusion

Self-Organising Maps provide a new and flexible method to classify archaeological data, which proved particularly effective to investigate locational strategies of settlements in complex archaeological landscapes. The comparison with PCA and hierarchical clustering highlighted the advantages of using SOM when relationships between variables are not linear and group assignment is not trivial, but it also showed the benefits of using these approaches in combination, to provide a deeper understanding of the landscape data being analysed.

Beyond the comparison of the results, this study highlights various aspects in which SOM outperforms PCA and hierarchical clustering for the analysis of archaeological landscapes. First, the preservation of the topological connections between sites (i.e., the neighbourhood relationships among observations in the original variable space) is ensured through SOM nodes and their codebook vectors, which implies that the relationships between sites and variables are retained in the 2-dimensional grid projection providing gradients as well as groups: similar sites are mapped onto neighbouring nodes and the SOM grid displays not only which clusters exist but also how they relate to one another, with transitional sites occupying intermediate positions rather than being forced into a single, rigid discrete category as occurs with dendrograms (Yotova *et al.*, 2021).

Second, SOM output retains the same variables as the input data, so that the model is interpreted directly in terms of the original variables, unlike PCA, which produces the principal components (Licen, Astel and Tsakovski, 2023). Third, the distribution of each variable is presented across the SOM grid, which facilitates identification of both correlations and the main variables driving each cluster. As opposed to PCA, which reduces dimensionality but does not group sites, and hierarchical clustering, which assigns groups but imposes a nested structure on a fixed distance metric, the SOM algorithm independently recovers essentially the same four site groups, meaning that the complementary use of all three methods constitutes a more robust toolset than any single method alone. Although the present analysis is based on a single regional case study, the comparison demonstrates that these features give SOM a visual and interpretive richness that facilitates establishing relationships among groups of sites more effectively than traditional exploratory multivariate techniques such as PCA or hierarchical clustering. Moreover, the two-stage design also adds an extra degree of robustness to the output, since sites are first summarised into codebook vectors and only these are eventually clustered. This extra step reduces group sensibility to individual outliers, a practical advantage given the small samples that are often used in archaeological landscape studies. These strengths should not, however, be mistaken for blind objectivity. Grid size and topology, the learning parameters, and, if clustering is employed, the choice of k all remain decisions to which the results are sensitive. Nonetheless, the core value of SOM lies in the structure it reveals and the transparency with which it can be interrogated.

Although the SOM presented test focused on spatial data, this approach is perfectly suited to contribute to the exploratory analysis of other types of archaeological data. For example, SOM would be particularly useful to investigate the assignment of different objects to specific archaeological types based on an investigation of typometric parameters.

By applying SOM, PCA and hierarchical clustering to a common dataset — eleven environmental and topographic variables recorded for 57 early medieval sites in the Alto Arlanza (Burgos, Spain) — this study has introduced SOM to archaeological landscape analysis and evaluated what it adds to the established methods. The three approaches converge on a consistent set of settlement groups, which supports the robustness of the classification while demonstrating that SOM performs as consistently as the standard techniques. Yet the comparison also reveals where SOM goes further: its topology-preserving projection and visual outputs (component planes, U-matrix and cluster maps) prove most valuable precisely where conventional clustering forces discrete partitions, identifying the transitional sites whose locational characteristics sit between distinct settlement groups and which are often the most informative for understanding locational decision-making. These capabilities extend those provided by established methods: PCA offers a gradient without groups, hierarchical clustering offers groups without gradients, and SOM integrates both within a single representation. On this evidence, the findings support the addition of SOM to the multivariate toolkit of landscape archaeology, not as

a replacement for PCA or cluster analysis, but as a particularly well-suited approach to the non-linear, spatially structured and often ambiguous data that characterise archaeological landscapes.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, 2(4), 433–459. <https://doi.org/10.1002/wics.101>

Álvaro Rueda, K. (2012). *El poblamiento altomedieval y sus manifestaciones funerarias en la cuenca del Alto Arlanza (s. IX y XI)* [Doctoral dissertation, Universitat de Barcelona]. <http://hdl.handle.net/10803/80651>

Álvaro Rueda, K., Travé Allepuz, E., & López Pérez, M. D. (2018). La acción monástica y eremítica en los procesos de poblamiento altomedievales (ss. VI–XI): Los conjuntos singulares del Alto Arlanza y sus problemas interpretativos. *Arqueología y Territorio Medieval*, 25, 145–167. <https://doi.org/10.17561/aytm.v25.5>

Anantharaman, R., Hall, K., Shah, V. B., & Edelman, A. (2020). Circuitscape in Julia: High performance connectivity modelling to support conservation decisions. *Proceedings of the JuliaCon Conferences*, 1(1), 58. <https://doi.org/10.21105/jcon.00058>

Bai, Y. (2022). Quantifying patterns in mortuary practices: An application of factor analysis and cluster analysis to data from the Taosi site, China. *Open Archaeology*, 8(1), 1231–1248. <https://doi.org/10.1515/opar-2022-0268>

Banning, E. B. (2020). *The archaeologist's laboratory: The analysis of archaeological evidence*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-47992-3>

Basile, S., & Campus, A. (2023). Integrating point pattern analysis and logistic regression approaches for exploring the settlement pattern of the Versilia and Garfagnana mountains in Roman times. *Archeologia e Calcolatori*, 34(2), 87–104. <https://doi.org/10.19282/ac.34.2.2023.05>

Baxter, M. J. (2015). *Exploratory multivariate analysis in archaeology*. Eliot Werner Publications. <https://doi.org/10.2307/j.ctv2sx9gfb>

Bickler, S. H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9(2), 186–191. <https://doi.org/10.1017/aap.2021.6>

Brandolini, F., & Carrer, F. (2021). Terra, Silva et Paludes. Assessing the role of alluvial geomorphology for Late-Holocene settlement strategies (Po Plain – N Italy) through point pattern analysis. *Environmental Archaeology*, 26(5), 511–525. <https://doi.org/10.1080/14614103.2020.1740866>

Brughmans, T., & Brandes, U. (2017). Visibility network patterns and methods for studying visual relational phenomena in archeology. *Frontiers in Digital Humanities*, 4, Article 17. <https://doi.org/10.3389/fdigh.2017.00017>

Cacciari, I., & Pocobelli, G. F. (2022). Machine learning: A novel tool for archaeology. In S. D'Amico & V. Venuti (Eds.), *Handbook of cultural heritage analysis* (pp. 961–1002). Springer International Publishing. https://doi.org/10.1007/978-3-030-60016-7_33

Carlson, D. L. (2017). *Quantitative methods in archaeology using R*. Cambridge University Press. <https://doi.org/10.1017/9781139628730>

Conolly, J., & Lake, M. (2006). *Geographical information systems in archaeology*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511807459>

Deufemia, V., Paolino, L., & de Lumley, H. (2012). Petroglyph recognition using self-organizing maps and fuzzy visual language parsing. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence* (pp. 852–859). IEEE. <https://doi.org/10.1109/ICTAI.2012.119>

Drennan, R. D. (2009). *Statistics for archaeologists: A common sense approach* (2nd ed.). Springer. <https://doi.org/10.1007/978-1-4419-0413-3>

Dunnell, R. C. (1971). *Systematics in prehistory*. The Free Press.

Eve, S. J., & Crema, E. R. (2014). A house with a view? Multi-model inference, visibility fields, and point process analysis of a Bronze Age settlement on Leskernick Hill (Cornwall, UK). *Journal of Archaeological Science*, 43, 267–277. <https://doi.org/10.1016/j.jas.2013.12.019>

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley. <https://doi.org/10.1002/9780470977811>

Gantley, M., Whitehouse, H., & Bogaard, A. (2018). Material correlates analysis (MCA): An innovative way of examining questions in archaeology using ethnographic data. *Advances in Archaeological Practice*, 6(4), 328–341. <https://doi.org/10.1017/aap.2018.9>

García-Tejedor, Á. J., & Nogales, A. (2022). An open-source Python library for self-organizing-maps. *Software Impacts*, 12, 100280. <https://doi.org/10.1016/j.simpa.2022.100280>

Glascok, M. D., & MacDonald, B. L. (2023). Multivariate analysis in archaeology. In A. M. Pollard, R. A. Armitage, & C. A. Makarewicz (Eds.), *Handbook of archaeological sciences* (pp. 1183–1192). Wiley. <https://doi.org/10.1002/9781119592112.ch58>

Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'Enza, A., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2, 100. <https://doi.org/10.1038/s43586-022-00184-w>

Hodson, F. R. (1970). Cluster analysis and archaeology: Some new developments and applications. *World Archaeology*, 1(3), 299–320. <https://doi.org/10.1080/00438243.1970.9979449>

Howey, M. C. L. (2011). Multiple pathways across past landscapes: Circuit theory as a complementary geospatial method to least cost path for modeling past movement. *Journal of Archaeological Science*, 38(10), 2523–2535. <https://doi.org/10.1016/j.jas.2011.03.024>

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley. <https://doi.org/10.1002/9780470316801>

Kiang, M. Y. (2001). Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics & Data Analysis*, 38(2), 161–180. [https://doi.org/10.1016/S0167-9473\(01\)00040-8](https://doi.org/10.1016/S0167-9473(01)00040-8)

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59–69. <https://doi.org/10.1007/BF00337288>

Kohonen, T. (2001). *Self-organizing maps*. Springer. <https://doi.org/10.1007/978-3-642-56927-2>

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, 37, 52–65. <https://doi.org/10.1016/j.neunet.2012.09.018>

Kontogiorgos, D., Leontitsis, A., & Sangole, A. P. (2007). Telling a non-linear story: The exploration of micro-artefacts' non-linear structure. *Journal of Archaeological Science*, 34(9), 1532–1536. <https://doi.org/10.1016/j.jas.2006.12.010>

Korokhina, A. V. (2024). Classification of fragmented pottery with the use of Kohonen self-organising maps (case study from the Hlyboke Ozero-2 settlement in Eastern Ukraine). *Archaeometry*, 66(6), 1205–1223. <https://doi.org/10.1111/arcm.12961>

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, 26, 195–197. <https://doi.org/10.1038/nbt1386>

Licen, S., Astel, A., & Tsakovski, S. (2023). Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review. *Science of the Total Environment*, 878, 163084. <https://doi.org/10.1016/j.scitotenv.2023.163084>

López Pérez, M. D., Álvaro Rueda, K., & Travé Allepuz, E. (2016). Necrópolis rupestres y procesos de poblamiento en el Alto Arlanza (Burgos): Un análisis del paisaje tardoantiguo y altomedieval. *Zephyrus*, 78, 173–191. <https://doi.org/10.14201/zephyrus201678173191>

Lopez-Molinero, A., Castro, A., Pino, J., Perez-Arantegui, J., & Castillo, J. R. (2000). Classification of ancient Roman glazed ceramics using the neural network of self-organizing maps. *Fresenius' Journal of Analytical Chemistry*, 367, 586–589. <https://doi.org/10.1007/s002160000433>

Lu, P., Chen, P., Tian, Y., He, Y., Mo, D., Yang, R., Lasaponara, R., & Masini, N. (2019). Reconstructing settlement evolution from Neolithic to Shang dynasty in Songshan mountain area of central China based on self-organizing feature map. *Journal of Cultural Heritage*, 36, 23–31. <https://doi.org/10.1016/j.culher.2018.08.006>

Martín Viso, I. (2025). *Las miradas y las prácticas: Perspectivas sobre los paisajes rurales en el Occidente medieval*. Editorial UCA.

McRae, B. H., Dickson, B. G., Keitt, T. H., & Shah, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, 89(10), 2712–2724. <https://doi.org/10.1890/07-1861.1>

Nogales, A., Guadalupe, D., & García-Tejedor, Á. J. (2025). Self-organizing maps to evaluate optimal strategies for balancing binary class distributions: A methodological approach. *Journal of Big Data*, 12, 141. <https://doi.org/10.1186/s40537-025-01188-5>

Orengo, H. A., Berganzo-Besga, I., & Lumbreras, F. (2026). Theory and practice of artificial intelligence in archaeology. *Journal of Archaeological Science*, 190, 106571. <https://doi.org/10.1016/j.jas.2026.106571>

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Wiley. <https://doi.org/10.1002/9781118391686>

Rondelli, B., Lancelotti, C., Madella, M., Pecci, A., Balbo, A., Ruiz Pérez, J., Inserra, F., Gadekar, C., Cau Ontiveros, M. À., & Ajithprasad, P. (2014). Anthropic activity markers and spatial variability: An ethnoarchaeological experiment in a domestic unit of Northern Gujarat (India). *Journal of Archaeological Science*, 41, 482–492. <https://doi.org/10.1016/j.jas.2013.09.008>

Ruiz-Giralt, A., Biagetti, S., Lancelotti, C., Parque, Ó., Koutroumpas, A., Rix, K. S., Ibáñez-Insa, J., & Madella, M. (2026). Anthropic activity markers 2.0: A shift towards compositional data analysis. *Journal of Archaeological Method and Theory*, 33(4), 64. <https://doi.org/10.1007/s10816-026-09799-9>

Scitovski, R., Sabo, K., Martínez-Álvarez, F., & Ungar, Š. (2021). *Cluster analysis and applications*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-74552-3>

Shennan, S. (1997). *Quantifying archaeology* (2nd ed.). Edinburgh University Press. <https://doi.org/10.1515/9781474472555>

Shieh, S.-L., & Liao, I.-E. (2012). A new approach for data clustering and visualization using self-organizing maps. *Expert Systems with Applications*, 39(15), 11924–11933. <https://doi.org/10.1016/j.eswa.2012.02.181>

Skupin, A., & Agarwal, P. (2008). Introduction: What is a self-organizing map? In P. Agarwal & A. Skupin (Eds.), *Self-organising maps* (pp. 1–20). Wiley. <https://doi.org/10.1002/9780470021699.ch1>

Verzani, J. (2014). *Using R for introductory statistics* (2nd ed.). CRC Press.

Walczak, S., & Cerpa, N. (2003). Artificial neural networks. In *Encyclopedia of physical science and technology* (pp. 631–645). Academic Press. <https://doi.org/10.1016/B0-12-227410-5/00837-1>

Wang, H., Song, C., Wang, J., & Gao, P. (2024). A raster-based spatial clustering method with robustness to spatial outliers. *Scientific Reports*, 14, 4103. <https://doi.org/10.1038/s41598-024-53066-4>

Wehrens, R., & Buydens, L. M. C. (2007). Self- and super-organizing maps in R: The kohonen package. *Journal of Statistical Software*, 21(5), 1-19. <https://doi.org/10.18637/jss.v021.i05>

Wickham, C. (2005). *Framing the early Middle Ages*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199264490.001.0001>

Yotova, G., Varbanov, M., Tcherkezova, E., & Tsakovski, S. (2021). Water quality assessment of a river catchment by the composite water quality index and self-organizing maps. *Ecological Indicators*, 120, 106872. <https://doi.org/10.1016/j.ecolind.2020.106872>

Zou, J., Han, Y., & So, S.-S. (2008). Overview of artificial neural networks. In D. J. Livingstone (Ed.), *Artificial neural networks* (pp. 14–22). Humana Press. https://doi.org/10.1007/978-1-60327-101-1_2

Funding

Financial support was provided by a NUAct PhD scholarship from Newcastle University.

Acknowledgements

This work is framed within the research lines of the GRAMPMed Research Group on Medieval and Post-medieval Archaeology and the Institute of Archaeology of the University of Barcelona (IAUB). This research is also part of the project "El origen de la red aldeana: puesta en valor y gestión del patrimonio cultural en el Alto Arlanza" (SUBV-24090-BU), funded by the Junta de Castilla y León and the Diputación Provincial de Burgos.

Author information

Authors and Affiliations

- Guillem Domingo-Ribas, Francesco Carrer
 - School of History, Classics and Archaeology, Newcastle University. Armstrong Building, Newcastle University, Newcastle Upon Tyne, NE1 7RU, United Kingdom
- Guillem Domingo-Ribas, Karen Álvaro Rueda
 - Departament d'Història i Arqueologia, Grup de Recerca en Arqueologia Medieval i Postmedieval (GRAMPMed), Institut d'Arqueologia (IAUB), Universitat de Barcelona. Carrer de Montalegre 6-8, 08001, Barcelona, Spain

Contributions

Conceptualisation: Guillem Domingo-Ribas; Data acquisition: Guillem Domingo-Ribas, Karen Álvaro. Methodology: Guillem Domingo-Ribas; Formal analysis and investigation: Guillem Domingo-Ribas, Francesco Carrer; Writing—original draft preparation: Guillem Domingo-Ribas; Writing—review and editing: Guillem Domingo-Ribas, Francesco Carrer, Karen Álvaro.

Corresponding author

Correspondence to Guillem Domingo-Ribas (Guillem.Domingo-Ribas@newcastle.ac.uk)

Data and code availability

Data and code employed in this analysis are available in Zenodo:
<https://doi.org/10.5281/zenodo.22809935>

Additional information

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.