Projection-Based Classification of Chemical Groups for Provenance Analysis of Archaeological Materials

PEDRO A. LÓPEZ-GARCÍA¹, DENISSE L. ARGOTE², AND MICHAEL C. THRUN³,⁴
¹Posgrado de Arqueología, Escuela Nacional de Antropología e Historia, Mexico 14030, Mexico
²Dirección de Estudios Arqueológicos, Instituto Nacional de Antropología e Historia, Mexico 06000, Mexico
³Databionics AG, Mathematics, and Computer Science, Philipps University of Marburg, 35032 Marburg, Germany
⁴Department of Hematology, Oncology and Immunology, Philipps University of Marburg, 35032 Marburg, Germany

Corresponding author: Michael C. Thrun (mthrun@mathematik.uni-marburg.de)

ABSTRACT In provenance analysis, identifying the origin of the archaeological artifacts plays a significant role. Usually, this problem is addressed by discovering natural groups in data measured with spectroscopic techniques. Then, principal component and classical partitioning cluster analysis are employed to reveal the groups that supposedly define the origin of the investigated artefacts. However, this work shows that maximizing the variance and searching for specific cluster structures can be misleading because it fails to discriminate clearly the different archeological sources. In contrast, the new methodology reveals several acknowledged geological sources present in the materials through the exploitation of emergence and swarm intelligence without prior assumptions about the data structures. A combination of unsupervised and semi-supervised machine learning and chemometric is applied on samples of Mesoamerican geological sources and obsidian artefacts collected from the archaeological site of Xalasco in Mexico. The analysis of the artifacts showed a preference of Xalasco inhabitants to local obsidian deposits. The results show that this approach, in terms of robustness, is suitable for handling unbiased quantitative spectral analysis of archaeological materials revealing the natural groups of archeological data.

INDEX TERMS Swarm intelligence, emergence, chemometrics, chemical data grouping, provenance analysis, archaeological obsidian.

I. INTRODUCTION

In recent years, the number of archaeometric investigations that make use of analytical techniques in the analysis of compositional data of various archaeological materials continues to grow considerably. The specific approach is to identify groups of similar objects in terms of their chemical composition. One example of this type of studies is provenance analysis, which tries to relate archaeological materials to their original natural sources by discriminating their characteristic chemical fingerprint through the application of analytical and quantitative methods for compositional data. Nonetheless, when using different grouping algorithms in the same data set, one of the most obvious concerns is that the output data is not consistent, producing different results for each situation [1], [2]. Moreover, it is always crucial to apply a procedure of evaluation to the results of clustering algorithm, something that is usually set aside in archaeometric studies [3]–[8].

Although several publications on diverse clustering algorithms exists in archaeology, the knowledge of the theoretical foundations of clustering methods is limited. For example, there is no consensus definition of what a ‘cluster’ is [9]. This is because the cluster depends largely on the context. One of the main difficulties in formalizing its definition is that clustering is ambiguous [10]; that is, it can be understood or interpreted in various ways, and can vary from application to application. One objective of grouping can be to discover the classification structure of meaningful (“natural”) groups in the data. However, when identifying homogeneous groups of objects [11], many clustering algorithms implicitly assume different structures of clusters [12]–[17]. It should be noted that some clustering algorithms could detect groups even if they do not have a clustering structure [16], [18].
In conditions for which no prior knowledge is accessible (for example, the possible natural sources of collected archaeological materials), it is preferable to choose a clustering algorithm with no implicit assumptions about the data structures. In contrast, conventional clustering methods use a global cluster criterion that imposes a fixed geometric nature model to detect the groups that will be formed, without taking into account the underlying distribution of the data in the n-dimensional space [14], [16], [17]. For example, the k-means [19]–[22] clustering criterion implies specific cluster structures [12], [14], [15], [23], [24]. This causes a considerable tendency of the established clustering algorithms to produce erroneous results by creating incorrect cluster associations of samples or forcing the construction of non-existent cluster structures in the data.

In provenance analysis, linear dimensionality reduction techniques, in particular principal component analysis (PCA) [25], have been widely used in the analysis of archaeological data for searching homogeneous groups [26]–[29]. However, it has many shortcomings as a visualization technique [17], [30]. For example, the first two or three principal components seldom define the sub-space that is most informative about the cluster structure in data [31]–[34], if the task of clustering is defined as the grouping of similar objects. Reference [35] show that the PCA visualization does not demonstrate a clear separation of the data and that it is common to observe a significant mixture between the different populations, which makes this inclusion ineffective for the classification of sub-types.

PCA has other disadvantages: it is scale-dependent, the scaling of variables can cause different PCA results, and the interpretation of the principal components is tricky when most of the loadings are either very low or very high in absolute values [36], [37]. On the other hand, the calculation of the components is based on the estimation of the vector of means and the matrices of variances and covariances, which are very sensitive to anomalous observations. The presence of outliers in the data makes very difficult to accurately evaluate the covariance matrix, strongly influencing the definition of the direction of the first principal components. It is important to remark that PCA does not define groups explicitly and the leading utility of PCA is exploratory rather than inferential.

Therefore, we propose a novel methodology for chemical grouping and provenance analysis of archaeological material using Databionic swarm (DBS) [38]. Its coexistence of projection and clustering allows exploring cluster structures through a topographic map without any implicit assumptions about the data [38]. To assess the effectiveness of the method, we analyzed with a portable x-ray fluorescence (pXRF) spectrometer a set of 256 obsidian artifacts re-covered from recent archaeological excavations at the site of Xalasco [39]–[41], located northeast of Tlaxcala state in Central Mexico. The objective was to associate the obsidian artifacts with their natural sources in order to identify commercial routes and cultural ties with other contemporary urban centers.

The archaeological site of Xalasco was occupied in the Classic period (200–700 A.D.), and was the seat of multi-ethnic groups of elite, artisans or emissaries of Teotihuacan or people who shared similar Teotihuacan cultural ideals. Former archaeological excavation projects have uncovered several architectural elements, like roadways, floors, walls, staircases and other elements that corresponded to high-rank habitational or civic units [39], [42], [43]. According to the obtained characteristics of the archaeological materials, it has been proposed that Xalasco was a locality for the exchange and transport of raw materials and products between the regions of the Gulf of Mexico, South Puebla and Central Mexico.

Therefore, for the determination of the provenance of the raw material, we introduced two datasets of obsidian samples collected from Mesoamerican natural sources. For the first experiment we used 64 samples from El Paredón (Puebla), Otumba (state of Mexico), Sierra de Pachuca (Hidalgo), Oyameles (Puebla) and Zinapécuaro (Michoacan); for the second experiment, we extracted the samples from Zinapécuaro and added 29 samples collected from Zacualtipan (Hidalgo) and El Chayal (Guatemala). A full description of the obsidian geological sources can be found in references [41], [44]–[46].

II. METHODS

For the XRF analysis, we employed a TRACER III-SD portable analyzer manufactured by Bruker Corporation, with an Rh tube at an angle of 52°, a drift silicon detector and a 7.5 µm Be detector window. The instrument was set with a voltage of 40 kV, a current of 30 µA, and a measurement time of 200 live seconds. A factory filter composed of 150 µm Cu, 25 µm Ti, 300 µm Al was used. With this configuration, we obtained the fluorescent spectrum of each sample. With the numerical values obtained from the photon counts, a matrix of n x p (where n is the number of samples and p are the channel counts) was constructed and processed with the methodology proposed below. The methods section is organized in three stages, as illustrated in Figure 1. All data processing was carried out in R environment version 3.4.3 [47] on a 64-bit Windows 7 system.

A. PRE-PROCESSING THE XRF SPECTRA: PEAK ALIGNMENT, FILTERING AND VARIABLE SELECTION

Pre-processing of XRF spectra is an essential stage in the analysis that is scarcely brought into consideration in archaeometric studies. Spectra are pre-processed in order to remove systematic noise and other spurious data, such as baseline variation and multiplicative scatter effects. Due to the nature of the data, peak alignment and filtering must be the first steps in data processing. To address the problem of misaligned spectra, we employed a robust and highly reliable algorithm known as Hierarchical Cluster-based Peak Alignment or CluPA [48] using the ‘Speaq’ package for R [49]. The code for this algorithm can be found in [40]. To filter the data, the model-based background correction and normalization of
spectra was performed with the ‘EMSC’ package [50] and the Savitsky-Golay (SG) filter with the ‘prospectr’ package [51].

Afterwards, the interval partial least squares model (iPLS) was applied to select the spectral range that provided the lowest prediction error compared with the full-spectrum model. The selection of specific regions of the full spectra in a small number of variables without significant loss of relevant information is key to decrease the complexity of the calibration model and to increase its predictive accuracy in the classification, separating the useful information from which it is not [52], [53]. The best sub-intervals are determined based on several parameters: the (RMSECV) values, the root mean square error of prediction (RMSEP), as well as other parameters such as the squared correlation coefficient (r²), the slope and the offset [53]. For the Interval Partial Least Squares Regression (iPLS) model, we employed the iToolbox for MATLAB version R2016a [54]. The methods employed in this step are fully described in [41].

**B. IDENTIFICATION OF DATA CLUSTER STRUCTURES: DATABIONIC SWARM**

The following section explains an approach to cluster analysis for which no prior knowledge or implicit assumptions are required about the data, and which resolves the challenge of finding the appropriate number of clusters and investigating the cluster tendency. In the case of the estimation of an appropriate number of clusters, a wide variety of indices have been proposed [55]; for the case of the cluster tendency, one of many statistical approaches has to be selected to test for it, often also called clusterability [56].

Here, a non-linear dimensionality reduction (DR) approach was used, which exploits the concepts of swarm intelligence, self-organization, and emergence. The swarm projects high-dimensional data onto a two-dimensional plane by using intelligent agents operating on a toroidal and polar grid [38]. Hence, it is called polar swarm (Pswarm). Each agent is identified by a high-dimensional data points and called DataBot. The DataBots self-organize in the polar grid by searching for the DataBots with the most potent scent. In other words, they search for other agents that carry data with most similar features (defined through a distance matrix) to itself, by moving across the grid or staying in their current position and with a decreasing search radius. Through self-organization, novel and irreducible data structures can emerge (see emergence, c.f. [57], [58]).

Typical non-linear projection methods require several parameters to define the annealing scheme [58]–[60]. The construction of this type of projection, called the learning phase, is sensitive to the choice of parameters. Using a concept of game theory, Pswarm remains parameter-free.
After the learning phase has finished, DataBots are grouped and can be visualized as projected points in two-dimensions. The identification of emergent structures is made visible by a specific visualization called topographic map [61]. For this purpose, the concept of the U-matrix [62] is used, providing emergent structures corresponding to clusters and performing classical clustering methods.

The U-matrix, or Unified Distance matrix, is a method used for the self-organizing map (SOM) in which the Euclidian distance between the codebook vectors of neighboring neurons define the height of each position of a codebook vector [62]. There exists numerous approaches for SOMs to visualize the U-matrix, which summarizes these U-heights [61]. Recently, it was shown how to generalize the U-Matrix technique for planar projections [16]. The generalized U-matrix generates the visualization of a topographic map with hypsometric tints, which can be vividly described as a virtual 3D landscape with a specific color scale chosen with an algorithm defining the contour lines [61]. The topographic map addresses the central problem in clustering; i.e., the correct estimation of the number of clusters. It allows the assessment of the number of clusters [61] by inspecting the 3D landscape. The color scale and contour lines imitate valleys, ridges, and basins: blue colors indicate small distances (sea level), green and brown colors indicate middle distances (low hills), and shades of gray and white indicate vast distances (high mountains covered with snow and ice). Valleys and basins represent clusters, and the watersheds of hills and mountains represent the borders between clusters [61].

In the topographic map, the borders of the visualization are cyclically connected with a periodicity (L,C) because it is preferable to perform the learning phase on a toroidal grid. A central problem in clustering is the correct estimation of the number of clusters. The topographic map addresses this problem by assessing the number of clusters [61]. The clustering is valid if mountain ranges do not partition the clusters, indicated by colored points of the same color and colored regions of points. If cluster structures are visible in the topographic map, the dataset does not possess any cluster tendency [63]. The semi-automated clustering is performed by calculating the shortest paths [64] of the Delaunay graph between all projected points weighted with high-dimensional distances. This is possible because it was shown that the U-matrix approximates the abstract U-matrix [65], which is based on Voronoi cells. Voronoi cells define a Delaunay graph where the edges between every projected point are weighted by the high-dimensional distances of the corresponding data points. The clustering approach itself involves one of two choices [63].

Outliers or small clusters are represented as volcanoes, mountain peaks or tiny valleys and can be interactively marked in the visualization after the automated clustering process [63]. Together, the three modules of projection, topographic map and clustering are defined as the Databionic swarm or DBS [38] and are available as an R package on CRAN [66].

C. VALIDATION
A variety of measures of validation has been proposed in the literature. References [67] and [68] present the Silhouette plot to evaluate the quality of a clustering. Another graphical representation for visualizing high-dimensional data is the Heat map [69]. A Heat map visualizes the distances ordered by the clustering through variations in color, by placing variables in the rows and columns, and coloring the cells within the table. Heat maps are good for showing how similar the objects are within a cluster and dissimilar outside a cluster.

A supervised index can be used to assess the efficiency of the classifications:

$$\text{Accuracy} \% = \frac{\text{No. True Positive}}{\text{No. True Positive} + \text{No. False Positive}} \times 100$$

In Equation (1), the number of true positives is the number of labeled data points for which the label defined by a prior classification is identical to the label defined after the clustering process [17], (otherwise see procedure in [38], [63]).

III. EXPERIMENTAL RESULTS
We analyzed an initial data set that consisted of the addition of two matrices. The first matrix (n = 64) contained our control samples and corresponded to the geological materials collected from several Central Mexico obsidian sources: Otumba from Soltepec range (n = 10), Otumba from Ixtpec, Malpais and Pacheco hills (n = 23), Oyameles (n = 7), El Paredón (n = 7), Sierra de la Navajas (n = 10) and Zinapécuaro (n = 7). The second matrix contained a total of n = 256 artefacts of unknown origin collected through systematic excavations in the archaeological site of Xalasco. These two matrices were combined in a single matrix with a total of n = 320 data points and p = 1752, where p corresponds to the spectral range of 38 to 1790 channel counts equivalent to the energy range of 0.76 to 35.8 keV of the detector resolution.

Since the origin of the control samples (geological sources: n1, n2, . . . n64) was fully known, when combining this matrix with the matrix of the artefacts of unknown origin, we can determine whether the groups that we obtained reflect the actual structure of the groups present in the data. Next, spectral alignment, filtering and variable selection procedures were applied to the raw spectra. In Table 1 are shown the selected sub-interval, the range of selected variables, the number of latent variables (LVs), Correlation coefficient ($r^2$), the RMSECV and the RMSEP. A good calibration model should have a high correlation coefficient ($r^2$) and a low RMSEP. The results in Table 1 showed that an iPLS model based on pre-processed spectra with the EMSC filter resulted in superior performance compared to that of the Savitsky-Golay and combinations of these two. The iPLS calibration model reduced the number of variables (channels) from 1752 to only 88, corresponding to the spectral range of 617 to 704 channel counts and equivalent to the energy range of 12.34 to 14.08 keV of the detector resolution.
TABLE 1. Calibration and cross-validation result using *IPLS* regression.

| Filter   | Selected subinterval | Range (channel counts) | LVs | Correlation coefficient (r) | RMSECV   | RMSEP   |
|----------|----------------------|------------------------|-----|-----------------------------|----------|---------|
| EMSC     | 8                    | 619-706                | 5   | 0.9709                      | 0.3713   | 0.3853  |
| SG       | 8                    | 617-704                | 9   | 0.9441                      | 0.5114   | 0.5328  |
| EMSC+SG  | 8                    | 617-704                | 5   | 0.9473                      | 0.4965   | 0.5205  |
| SG+EMSC  | 8                    | 617-704                | 7   | 0.9506                      | 0.4812   | 0.5156  |

TABLE 2. Number of samples measured of each of the geological sources (*n*) and number of artifacts assigned to that source by the projection method DBS

| Sources                             | n  | Assigned artifacts | Total |
|-------------------------------------|----|--------------------|-------|
| Otumba (Soltepec range)             | 10 | 13                 | 23    |
| Otumba (Ixtepec, Pacheco and Malpais hills) | 23 | 4                  | 27    |
| Oyameles, Puebla                    | 7  | 146                | 153   |
| El Paredon, Puebla                  | 7  | 16                 | 23    |
| Sierra de Pachuca (Hidalgo)         | 10 | 66                 | 76    |
| Zinapécuar (Michoacán)              | 7  | 0                  | 7     |
| Unknown source                      | 0  | 11                 | 11    |

TABLE 3. Contingency table for the evaluation of the grouping.

| DBS  | 1  | 2  | 3  | 4  | 5  | 6  | 7  | RΣ_ | R% |
|------|----|----|----|----|----|----|----|-----|----|
| 1    | 153| 0  | 0  | 0  | 0  | 0  | 0  | 153 | 48.8|
| 2    | 0  | 76 | 0  | 0  | 0  | 0  | 0  | 76  | 23.7|
| 3    | 0  | 0  | 27 | 0  | 0  | 0  | 0  | 27  | 8.4 |
| 4    | 0  | 0  | 0  | 23 | 0  | 0  | 0  | 23  | 7.1 |
| 5    | 0  | 0  | 0  | 0  | 11 | 0  | 0  | 11  | 3.4 |
| 6    | 0  | 0  | 0  | 0  | 0  | 23 | 0  | 23  | 7.1 |
| 7    | 0  | 0  | 0  | 0  | 0  | 0  | 7  | 7   | 2.1 |
| CΣ_  | 153| 76 | 27 | 23 | 11 | 23 | 7  | 320 | 0   |
| C%   | 48.8| 23.7| 8.4| 7.1| 3.4| 7.1| 2.1| 100 |      |

The sources are accommodated in the rows and the result of the clustering in the columns. Group 1 = Oyameles, Puebla; group 2 = Sierra de Pachuca, Hidalgo; group 3 = Otumba (Ixtepec, Pacheco and Malpais hills); group 4 = Otumba (Soltepec range); group 5 = unknown source; group 6 = El Paredon, Puebla; group 7 = Zinapécuar, Michoacán. Abbreviations: RΣ_—Row sum, R%—Row percentage, CΣ_—Column sum, C%—Column percentage.

After pre-processing the data, the Databionic swarm was applied in the feature space $D = \{x_i, i = 1, \ldots, 320\} \subset \mathbb{R}^{W \times p}$. The process was unsupervised, meaning that class labels were not used in the cluster analysis. The results of this automatic grouping were used in the next stage of the analysis, in which the user can select areas and mark them manually as groups. The result of the Pswarm projection method visualized as a topographic map provided a cluster structure, detecting seven clearly separate groups according to the position of the projected points and their distances (Figure 2, upper image). Figure 2 (lower image) shows the unsupervised index of the silhouette plot.

The groups exhibit a clear (natural) structure with small intra-cluster distances and large inter-cluster distances. Each valley in the topographic map represents subsets of data points that are similar; that is to say, a group whose high dimension distances of the data points are small with high densities in the feature space. There is also a clear separation between the different clusters marked by mountain ranges with well-defined edges. On the other hand, we see an apparent absence of outliers. The data are well clustered, with no signs of observations between two groups or data-points assigned to an incorrect group. In the validation of the results, a silhouette width greater than 0.5 is considered a reasonable grouping, however, an average width below 0.2 should be interpreted as a lack of substantial clustering structure [13].

Once this non-supervised approach has been carried out, it is possible to access the polar coordinates of the projected points, as well as a vector for the group to which each of the projected samples in the multi-dimensional space belongs according to their position on the topographic map. We compared this information with the database of the geological sources and the artifacts, and we observed that each of the groups obtained through DBS corresponded to a different known source, except for one of the groups that would correspond to an unknown source. Table 2 shows the number of samples corresponding to each of the sources and the number of samples assigned by the DBS projection method.

These results suggest that the formed groups provided relevant information for the clustering of the artifacts, thereby identifying the geological source of each of the samples.
Another way to corroborate the grouping obtained by the DBS was to examine the artifact database. We found that all the artifacts assigned to the Sierra de Pachuca source group corresponded to green obsidian samples, without overlapping with any of the other groups. The Sierra de Pachuca is a well-known geological source renowned for its valuable green translucent volcanic glass [41], [44]–[46]. The grouping was also evaluated by means of a contingency table (Table 3), whose rows are the sources groups and the columns the result of the clustering. The table shows that all the elements of the main diagonal are well classified.

The previous analysis was carried out as an unsupervised task. Using the information of the group vector provided by this first projection of the DBS of high-dimensional points (see Figure 2), each of the samples was labeled according to their group membership and the algorithm was run again. Figure 3 shows the resulting topographic map with a clear solution of seven groups. In the DBS algorithm, a topographic map of the generalized U-matrix, in which the projected points are colored using the same color if the corresponding data points are assigned to the same cluster, can visualize the clustering tendency. The supervised accuracy yielded a value of 92% of 100%. The clusters are distinctly separated by both the positions of the projected points and the high-dimensional distances and densities of the generalized U*-matrix. This U*-matrix is the visualization of projected points in a two-dimensional output space and can be used to display both distance- and density-based structures.

The magenta cluster (Oyameles) has a much higher density of points than the other clusters, with 146 assigned samples of a total of 256 archaeological artifacts (57% of the artifacts); the second group with a higher density is the yellow group (25.8 %), related to Sierra de Pachuca deposit. Fewer artifacts proceed from El Paredón (dark blue cluster; 6.2%), Otumba-Soltepec (red cluster; 5.1%) and an unidentified source (green cluster; 4.3%). The clusters in black (Otumba-Ixtepec, Pacheco, Malpais) and light blue (Zinapecuaro) are of less or non-impact to the analysis, since the only samples clustered in these groups correspond to the ones collected directly from the geological deposits. It should be noted that there is no overlap between the groups.

The heat map (Figure 4, left) and the silhouette plot (Figure 4, right) were used to verify the result of grouping externally. The heat map is an easy and effective way of visually displaying multivariate data and the silhouette plot allow discriminating whether the distribution scheme of the observations within the groups is strong or weak. A heat map that displays small intra-cluster distances in every cluster and large inter-cluster distances between the seven clusters...
confirms a clear cluster structure. In this case, it shows a homogeneity of the cluster structures of DBS. This homogeneity can also be seen in the silhouette plot, indicating a good cluster structure for values above 7.0. This corroborates the results illustrated in Figure 2, where samples in the same cluster are similar to each other and different from those of other clusters. The accuracy in our classification, observed by both the supervised index as well as the unsupervised indices, showed an excellent performance, with a supervised precision index yielding 92% result.

Considering the data reported in [43], discussed in the next section, we added to our matrix data from El Chayal source \((n = 19)\) and Zacualtipan source \((n = 10)\) and deleted those from the Zinapecuaro source. The procedure was the same as the previous analysis: pre-processing, variable range selection, DBS clustering, and validation. Figure 5(A) shows that the structure of the groups is compact, with small variations in the intra-cluster distances. The heat map and the silhouette plot presented in Figure 5 (B and C) show that this dataset is defined by discontinuities because the intra-cluster distances are small and the inter-cluster distances are large. The silhouette plot indicates a good cluster structure for values above 0.5. Therefore, the obsidian sample dataset is a high-dimensional dataset with natural groups that are specified by the chemical characteristics of each of the geological sources and defined by discontinuities. The supervised accuracy decreased to value of 90% due to the presence of 3 outliers (see Fig. 5A). As was done in the previous case, the grouping was also evaluated through a contingency table (Table 4).

### A. COMPARISON TO CONVENTIONAL CLUSTERING AND PCA PROJECTION

To evaluate the performance of the DBS methodology compared to classic dimension reduction and clustering methods, the data pertaining to nine central Mexico obsidian deposits (control samples) was also analyzed using PCA method using the covariance matrix. The R package employed was ‘factoextra’ version 1.0.7 [70]. The PCA was first computed for the raw spectra matrix (Figure 6, left) and, afterwards, for the matrix of concentration values (Figure 6, right), calculating the values for Mn, Fe, Zn, Ga, Th, Rb, Sr, Y, Zr,

![Figure 4](image-url) **FIGURE 4.** Heat map of the distances with regard to the seven identified clusters of obsidian samples. Silhouette plot of the DBS clustering of obsidian samples. The colors assigned to the groups match to the ones marked in Figure 3.

![Figure 5](image-url) **TABLE 4.** Contingency table for the evaluation of the grouping in the second test.

| DBS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | \(R_\Sigma\) | R% |
|-----|---|---|---|---|---|---|---|---|--------|----|
| 1   | 153 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 153 | 44.73 |
| 2   | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 2.92 |
| 3   | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 11 | 3.21 |
| 4   | 0 | 0 | 0 | 27 | 0 | 0 | 0 | 0 | 27 | 7.89 |
| 5   | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 | 23 | 6.72 |
| 6   | 0 | 0 | 0 | 0 | 76 | 0 | 0 | 76 | 22 | 22.22 |
| 7   | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 23 | 6.72 |
| 8   | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 19 | 342 | 0 |

| C% | 44.73 | 2.92 | 3.21 | 7.89 | 6.72 | 22.22 | 6.72 | 5.55 | 0 | 100 |

The sources are accommodated in the rows and the result of the clustering in the columns. Group 1 = Oyameles, Puebla; group 2 = Zacualtipan, Hidalgo; group 3 = unknown source; group 4 = Otumba (Ixtepec, Pacheco and Malpais hills); group 5 = El Paredon, Puebla; group 6 = Sierra de Pachuca, Hidalgo; group 7 = Otumba (Soltepec range); group 8 = El Chayal, Guatemala. Abbreviations: \(R_\Sigma\)-Row sum, R%-Row percentage, C%-Column sum, C%-Column percentage.
FIGURE 5. (A) Topographic map of DBS of the obsidian samples. Eight clusters are shown: magenta (group 1) = Oyameles; yellow (group 2) = Zacualtipan; black (group 3) = unknown source group; red (group 4) = Otumba (Ixtepec, Pacheco, Malpais); green (group 5) = El Paredon; dark blue (group 6) = Sierra de Pachuca; light blue (group 7) = Otumba (Soltepec); brown (group 8) = El Chayal. (B) Heat map showing a small variance of intra-cluster distance with regard to the 8 identified clusters of obsidian samples, including the sources of El Chayal and Zacualtipan. (C) Silhouette plot of the DBS clustering of obsidian samples, including the Zacualtipan and El Chayal groups.

and Nb through the instrument factory calibration for glasses (Rousseau, 2001). The data from the second matrix were transformed to log10 [71]. In both cases, although the variance explained by the first two PCs was high (87.5% in the raw data and 87.4% in the concentration data), the models were not able to properly separate all the groups.

There is an evident overlap of the groups formed: Zinapecuaro samples are mixed with the samples from Otumba, and Oyameles is mixed with El Paredón. It is important to note that larger sample sizes generate more complex overlapping, making group discrimination more difficult. The only groups with a clear separation from the other are Sierra
de Pachuca and El Pizarrín-Tulancingo; this can be explained by the distinctive greater concentrations of zirconium and zinc of these sources compared to other regional deposits [44], [45]. Zinapecuaro and Zacualtipan also show a wide dispersion of the data points, not forming compact groups. If the provenance of these data points was unknown, it would be difficult to be able to assign them accurately to a specific source.

IV. DISCUSSION
The goal of this research was to develop a methodology for determining automatically the number of clusters of archaeological materials according to their chemical characteristics, in this case aiming for the determination of their provenance. In the analysis of multivariate archaeometric data, the identification of chemical groups is a task of central interest. The interpretation of the data may depend upon a valid cluster structure found in an archaeological dataset. In the case of provenance analysis of obsidian, it supports the study of raw material acquisition patterns, manufacturing processes and distribution networks [72]. Therefore, determining the optimal clusters in a dataset is a fundamental problem in cluster analysis. Although XRF spectroscopy has been widely applied for obsidian sourcing studies ([73]–[77]; between others), the common compositional approach holds several inconvenient [41].

This research introduces a classification procedure based on a swarm-based projection method combined with a human-understandable visualization technique, using a swarm of intelligent agents that interact with their environment and each other to produce collectively intelligent behavior. Swarm intelligence is a potentially fertile ground for new methods for classification. The key features of DBS are that neither the parameters (or an overall objective function) have to be set nor the type of grouping sought have to be explicitly defined at any point during the Pswarm process. As a swarm technique, DBS clustering is robust with respect to outliers. All parameters are automatically chosen according to and directly based on the appropriate high-dimensional definition of distance of the data [63].

DBS turns out to be an efficient automatic method that allows determining the number of clusters that can be deduced from the topographic map of the Pswarm projection. The use of DBS and the generalized U-matrix provides apparent advantages to classical clustering algorithms. The use of the generalized U-matrix allows visually validating the cluster structures directly in the data by the visualization of the topographic map, without the help of additional analyses or plots. If a high dimensional dataset contains distance and/or density-based clusters, the resulting topographic map landscape possesses a clear valley - mountain ridge - valley structure. For the case of archaeological provenance studies, the assumption is that the found structures will allow assigning new artefacts to known sources.

In the unsupervised and semi-supervised experiments, the groups found by the Databionic swarm were spatially distinctive: there was no overlap between the groups and each
group represented a different source. In the case of the obsidian artifacts from Xalasco, the highest frequency of occurrence of the materials came from Oyameles source, Puebla, located 40 km from Xalasco. From Early Classic through early Post-Classic periods (100-1100 AD), the extraction, reduction and long-distance exchange of the obsidian from this source was controlled by the local urban settlement of Cantona, located 13 km to the south [74]. Although Oyameles group has a higher density, this grouping may well be subdivided into two or three possible sub-sources. The result is consistent with the analysis performed by reference [77], who identified three sub-sources known as Z-O1, Potreros Caldera and Gomez Sur.

On the other hand, Sierra de Pachuca obsidian source is just over 100 km distant in a straight line. The presence of artifacts from this source is also significant, with a 25.8% of the artifacts coming from this site. Although the frequency of artifacts from other sources is low (Otumba, El Paredón and, probably, Pico de Orizaba), it is evident that the interaction of the population of the site of Xalasco with other commercial areas had a substantial effect on its economy. A 57% of artifacts made from obsidian controlled by Cantona (Oyameles source) versus 30.9% of artifacts made from obsidian controlled by Teotihuacan (Sierra de Pachuca and Otumba) exhibit certain degree of interaction between these two large pre-Hispanic settlements, although Xalasco shows a higher preference for nearby sources.

Archaeologist L. Manzanilla also recovered obsidian artifacts from surface sampling and systematic excavation in the site of Xalasco between 2006 and 2008. Its technical report [43] proposed the following possible origin of the obsidian artifacts by matching their visual properties (e.g., color and translucence): 22% of the samples from Sierra de Pachuca, 4% from El Chayal (Guatemala), 26% possibly from Pico de Orizaba or El Chayal, 14% possibly from Pico de Orizaba and 5% possibly from Zacualtipan. From their classification, the possible samples from El Chayal (Guatemala) caught our attention. First, we considered high the percentage attributed to this region and, second, artifacts from this source would infer long-distance interregional exchange with the Mayan people. On the other hand, maybe our 11% of artifacts assigned to an unknown source could come from either Zacualtipan or El Chayal, sources that were not considered in the first experiment.

Bearing in mind these questions, in the second semi-supervised experiment, the analysis included samples from the geological sources of El Chayal and Zacualtipan. In this experiment, none of the samples analyzed were assigned to these sources. The unsupervised validation indices agreed with this fact. Therefore, there is no chemical evidence to assume that the samples analyzed in Xalasco could come from either one these regions, leaving our unknown samples still unknown. Nevertheless, since we did not have obsidian samples from Pico de Orizaba geological deposit, it is still possible that this source is the origin for these 11 artifacts. Therefore, considering the lithic materials collected to date, the idea that the population of this pre-Hispanic settlement had some type of interrelation with the Maya area can be discarded.

V. CONCLUSION

Although XRF spectroscopy has been often applied in provenance analysis, this work showed that the common approach for data processing is inconvenient due to the challenging tasks that depend on the data. Moreover, dimensionality reduction methods, like PCA, provide scatter plots of the first components that can be difficult to read, do not necessarily visualize appropriate cluster structures [78]. Moreover, the information about the possible groups can be questionable, not allowing a direct interpretation of the data. It can also be concluded that the visual inspection of the materials by eye implies a considerable bias in the interpretations.

Successful resolution of this problem requires the use of chemometric and machine learning methods. The combination of XRF spectroscopy and machine learning allowed us to implement a procedure that automatically determines the number of groups in obsidian samples according to their chemical characteristics and, consequently, defines their origin.

XRF spectroscopy is an adequate tool for obsidian provenance studies and can successfully discriminate between several groups of archaeological artifacts by means of quantitative analysis. To test our method, we directly used obsidian artifact spectra collected from the archaeological site of Xalasco and samples from various obsidian sources, demonstrating that it is not necessary to convert spectra to composition data. Furthermore, it was not necessary to work with the full spectrum of the data.

The machine learning method, which was employed and extended to the semi-supervised approach here, is a flexible and robust approach for cluster analysis consisting of three separate modules that can be optionally combined into the Databionic Swarm (DBS). Its non-linear projection displays the structure of the high-dimensional data into a low-dimensional space preserving the cluster structure of the data. Then, the groups are easily recognizable using the generalized U-matrix approach. In the 3D visualization of the 3D topographic map, the observations with similar characteristics are represented as valleys, whereas the differences are represented as mountain ranges and the outliers are seen as volcanos. This robust alternative produce relatively lower error rates compared to those obtained by the classical methods. Additionally, the origin of the artefacts could be explained with the help of the framework in a way that is meaningful to the domain expert.

In sum, this work showed that the new unsupervised and semi-supervised framework is reliable in classifying archaeological materials as well as identifying their origin. The authors demonstrated that the methodology proposed here is suitable for handling unbiased quantitative spectral analysis and defining groups in materials of archaeological interest.
REFERENCES

[1] M. C. Thrun and A. Ultsch, “Clustering benchmark datasets exploiting the fundamental clustering problems,” Data Brief, vol. 3, Jun. 2020, Art. no. 105501, doi: 10.1016/j.dib.2020.105501.

[2] R. Wehrens, Chemometrics With R: Multivariate Data Analysis in the Natural Science and Life Sciences. Berlin, Germany: Springer-Verlag, 2011.

[3] M. J. Baxter, “Standardization and transformation in principal component analysis, with applications to archaeometry,” J. Roy. Stat. Soc., C, Appl. Statist., vol. 44, no. 4, pp. 513–527, 1995, doi: 10.2307/2986142.

[4] A. S. Cohen, D. E. Pierce, and C. T. Fisher, “Geochemical analysis and spatial trends of ceramics and clay from Angamumu, Michoacan,” J. Archaeol. Sci., Rep., vol. 23, pp. 216–219, Feb. 2019, doi: 10.1016/j.jarcsrep.2019.02.025.

[5] C. Descantes, R. J. Speakman, and M. D. Glascock, “Compositional studies of Caribbean ceramics: An introduction to instrumental neutron activation analysis,” J. Caribbean Archaeol., Special Publication, vol. 2, no. 2, pp. 1–14, 2008.

[6] M. D. Glascock, H. Neff, and K. J. Vaughn, “Instrumental neutron activation analysis and multivariate statistics for pottery provenance,” Hyperfin. Interact., vol. 154, nos. 1–4, pp. 95–105, 2004, doi: 10.1023/B:HYPE.000032025.37390.41.

[7] J. K. Millhauser, L. F. Fargher, V. Y. H. Espinoza, and R. E. Blanton, “The geopolitics of obsidian supply in postclassic tlaxcallan: A portable X-ray fluorescence study,” J. Archaeol. Sci., vol. 58, pp. 133–146, Jun. 2015, doi: 10.1016/j.jas.2015.02.037.

[8] R. H. Tykot, “Using nondestructive portable X-ray fluorescence spectrometers on stone, ceramics, metals, and other materials in museums: Advantages and limitations,” Appl. Spectrosc., vol. 70, no. 4, pp. 42–56, Jan. 2016, doi: 10.1366/15-7616745.

[9] C. Hennig, “What are the true clusters?” Pattern Recognit. Lett., vol. 64, pp. 53–62, Oct. 2015, doi: 10.1016/j.patrec.2015.04.009.

[10] M. Ackerma, “Towards theoretical foundation of clustering,” Ph.D. dissertation, David R. Cheriton School Comput. Sci., Univ. Waterloo, Waterloo, ON, Canada, 2012.

[11] P. Arabie, L. J. Hubert, and G. De Soete, Clustering and Classification. New York, NY , USA: Springer, 2007.

[12] M. J. Brownstein and A. Kohodursky, Eds. Switzerland: Humana Press, 2011.

[13] J. Roy. Stat. Soc., C, Appl. Statist., vol. 66, pp. 95–104, Feb. 2017, doi: 10.1093/opar-2017-0013.

[14] E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Eds. Berlin, Germany: Springer-Verlag, 1994, pp. 160–189.

[15] W. C. Chang, “On using principal components before separating a mixture of two multivariate normal distributions,” J. Roy. Stat. Soc., C, Appl. Statist., vol. 32, no. 3, pp. 267–275, 1983, doi: 10.2307/2347949.

[16] G. De Soete and J. D. Carroll, “K-means clustering in a low-dimensional Euclidean space,” in New Approaches in Classification and Data Analysis, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy, Eds. Berlin, Germany: Springer-Verlag, 1994, pp. 212–219.

[17] M. Vichi and H. A. Kiers, “Factorial k-means analysis for two-way data,” Comput. Statist. Data Anal., vol. 37, no. 1, pp. 49–64, Jul. 2001, doi: 10.1016/S0167-9473(00)00064-5.

[18] A. Konstorum, E. Vidal, N. Ielk, and R. Laubenbacher, “Comparative analysis of linear and nonlinear dimension reduction techniques on mass cytometry data,” bioRxiv, to be published, doi: 10.1101/273862.

[19] C. Croux, P. Filzmoser, and H. Fritz, “Robust sparse principal component analysis,” Technometrics, vol. 55, no. 2, pp. 202–214, May 2013, doi: 10.1080/00401706.2012.727746.

[20] M. Hubert, T. Reynkens, S. Schmitt, and T. Verdonck, “Sparse PCA for high-dimensional data with outliers,” Technometrics, vol. 58, no. 4, pp. 424–434, Oct. 2016, doi: 10.1080/00401706.2015.1093962.

[21] M. C. Thrun and A. Ultsch, “Swarm intelligence for self-organized clustering,” Artif. Intell., to be published, doi: 10.1016/j.artint.2020.103237.

[22] P. López, D. Argote, and A. Tejero, “Informe del Proyecto Prospección arqueológica de los sitios de Los Teteles de Ocotitla y Xalasco, Noreste de Tlaxcala, Temporada 2015,” Archivo Técnico, INAH, Mexico, Tech. Rep. 1, Oct. 2016.

[23] P. López-García, D. L. Argote-Espino, and K. Faceciovica, “Statistical processing of compositional data. The case of ceramic samples from the archaeological site of Xalasco, Tlaxclla, Mexico,” J. Archaeol. Sci., Rep., vol. 19, pp. 100–114, Jun. 2018, doi: 10.1016/j.jasrep.2018.02.023.

[24] P. López-García, D. L. Argote-Espino, and C. Beirmte, “Chemometric analysis of Mexican obsidian sources,” Quaternary Int., vol. 510, pp. 100–118, Mar. 2019, doi: 10.1016/j.quaint.2018.12.032.

[25] K. Juárez, P. López-García, D. L. Argote-Espino, A. Tejero-Andrade, R. E. Chávez, and A. García-Serrano, “Magnetic and electrical prospections in the archaeological site of Xalasco, northeast of Tlaxcla, Mexico,” Global J. Archaeol. Anthrop., vol. 2, no. 1, pp. 1–10, Oct. 2017, doi: 10.19080/GJAA.2017.02.555581.

[26] L. R. Manzanilla, E. Aguyuo, M. Hernández, and A. Bautista, “Informe técnico final del salvamento de un conjunto con materiales teotihuacanos en Xalasco, Altzayanca, Tlaxcala 2008.” Archivo Técnico, INAH, Mexico, Tech. Rep. 1, Feb. 2011.

[27] D. L. A. Espino, J. Solé, O. Sterpone, and P. L. García, “Análisis compositacional de seis yacimientos de obsidiana del centro de México y su clasificación con DBSCAN,” Arqueología Segunda Época, vol. 43, pp. 197–215, Jan./Apr. 2010.
[45] D. Argote-Espino, J. Solé, P. López-García, and O. Sterpone, “Obsidian subsite identification in the Sierra de Pachuca and Otumba volcanic regions, Central Mexico, by ICP-MS and DBSCAN statistical analysis,” Geoarchaeology, vol. 27, no. 1, pp. 48–62, Apr. 2012, doi: 10.1002/gea.21389.

[46] R. H. Cobeán, K. H. Liland. (Jun. 19, 2017). Extended Multiplicative Signal

[47] T. N. Vu, D. Valkenborg, K. Smets, K. A. Verwaest, R. Dommisse, R Development Core Team. (2011). R: A Language and Environment for

[48] Package ‘Prospectr’: A. Stevens and L. Ramirez-Lopez. (Mar. 15, 2020).

[49] VOLUME 8, 2020

[50] iToolbox Manual L. Nørgaard. (Mar. 2013).

[51] A. Adolfsson, M. Ackerman, and N. C. Brownstein, “To cluster, or not to cluster: An analysis of clusterability methods,” J. Radioanal. Nucl. Chem., vol. 27, no. 1, pp. 48–62, Apr. 2012, doi: 10.1007/s10967-011-1429-9.

[52] J. Lötsch and A. Ultsch, “Exploiting the structures of the U-matrix,” in Advances in Self-Organizing Maps and Learning Vector Quantization (Advances in Intelligent Systems and Computing Book Series), vol. 295, T. Villmann, F. M. Schleif, M. Kaden, and M. Lange, Eds. Meeting in 1998 and 2005, respectively. He has been a full-time Professor of the graduate program with the National School Anthropology and History since 1992. His several scientific articles related to numerical methods in archaeology, geophysical prospections, and archaeometry. Among his research projects carried out are Location of archaeological sites using high-resolution radar images in the Mayan area; Geophysical and geochronological prospection of the archaeological sites of Teotihuacán and Xalasico; Discrimination of textures in thin plates of ceramic with Gabor filter banks; Texture discrimination of pottery thin plates with Grammatical Evolution algorithms; between others.
DENISSE L. ARGOTE received the B.S. degree in archaeology from the National School Anthropology and History, Mexico, in 1999, the M.S. and Ph.D. degrees in earth sciences and geophysical exploration from the National Autonomous University of Mexico in 2002 and 2008, respectively. She was postdoctoral experience worked as an Assistant Professor in geochemistry with the National Autonomous University of Mexico from 2008 to 2010.

She has been a full-time Researcher with the National Institute of Anthropology and History, Mexico, since 2012. Her main research interest is focused in the incorporation of Earth Sciences methodologies and technology in archaeometric research, as a means of obtaining hard scientific data that can be interpreted from the archaeological point of view, assisting in the understanding of the behavior of past societies. Some research projects: Karst detection and characterization of the original structure of the Castle Pyramid in Chichen Itza; the detection of the NW corner of the perimeter wall of Mexico-Tenochtitlán; the discovery of a cave below the Pyramid of the Moon in Teotihuacán; pXRF analysis to mural painting and artifacts from Teotihuacan and Chichen Itzá.

MICHAEL C. THRUN received the Diplom degree in physics and the Ph.D. degree in data science from the Philipps University of Marburg in 2014 and 2017, respectively.

His research interests are centered around methods of dimensionality reduction, cluster analysis, data visualization, as well as unsupervised machine learning with a specific focus on swarm intelligence, self-organization, and emergence. He researched such methods for advanced analytics in big data and generating industrial applications in data science. He currently develops analytical procedures for explainable artificial intelligence.