This item is the archived peer-reviewed author-version of:

Chemometric analysis of Mesoamerican obsidian sources

Reference:

Lopez-Garcia Pedro, Argote Denisse L., Beirnaert Charlie.- Chemometric analysis of Mesoamerican obsidian sources
Quaternary international - ISSN 1040-6182 - 510(2019), p. 100-118
Full text (Publisher's DOI): <https://doi.org/10.1016/J.QUAINT.2018.12.032>
To cite this reference: <https://hdl.handle.net/10067/1604310151162165141>

Accepted Manuscript

Chemometric analysis of Mesoamerican obsidian sources

Pedro Lopez-García, Denisse L. Argote, Charlie Beirnaert

PII: S1040-6182(18)31002-4

DOI: <https://doi.org/10.1016/j.quaint.2018.12.032>

Reference: JQI 7696

To appear in: *Quaternary International*

Received Date: 28 August 2018

Revised Date: 27 December 2018

Accepted Date: 31 December 2018

Please cite this article as: Lopez-García, P., Argote, D.L., Beirnaert, C., Chemometric analysis of Mesoamerican obsidian sources, *Quaternary International* (2019), doi: <https://doi.org/10.1016/j.quaint.2018.12.032>.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

CHEMOMETRIC ANALYSIS OF MESOAMERICAN OBSIDIAN SOURCES

Pedro Lopez-García^a, Denisse L. Argote^b and Charlie Beirnaert^c

- (a) Posgrado de Arqueología, Escuela Nacional de Antropología e Historia. Periférico Sur esq. Calle Zapote, Col. Isidro Fabela, Deleg. Tlapan, CDMX, Mexico. E-mail: dplopez14@gmail.com
- (b) Dirección de Estudios Arqueológicos, Instituto Nacional de Antropología e Historia. Lic. Primo Verdad no. 2, Col. Centro, Deleg. Cuauhtémoc, CDMX, Mexico. E-mail: efenfi@gmail.com *Corresponding author*
- (c) Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium. E-mail: charlie.beirnaert@uantwerpen.be

ABSTRACT

We propose a combination of portable X-ray fluorescence (pXRF) and chemometrics to discriminate between Mesoamerican obsidian sources and to assign archaeological artifacts of unknown origin to their respective deposits using a procedure that does not require any type of calibration or reference standards. A set of 109 samples of known origin and a total of 257 samples of unknown origin were analyzed with a portable XRF spectrometer. The resultant spectra were used as spectral signatures for the chemometric data analysis. First, we applied spectral pre-treatment techniques, such as CluPA algorithm for peak alignment and the Savitzky-Golay and Extended Multiplicative Signal Correction for data smoothing and noise removal, combined with methods for the selection of a spectral range containing the variables with the most relevant information (iPLS and). The full spectrum of the obsidian samples was divided into 20 subintervals, fitting a local regression model (PLS) to each subinterval. The performance was evaluated by the Root Mean Square Error of Cross-Validation, the Root Mean Square Error of Prediction and the correlation coefficient. The selected spectral regions were then analyzed with ROBPCA algorithm for the discrimination of outliers and the projection of the observations in the PCA space. For the classification, we propose a robust procedure (RSIMCA) which is based on a ROBPCA method for high-dimensional data. The classification rules were obtained by using the orthogonal and the score distances, from which it is possible to distinguish samples that belong to a given group. Using the proposed methodology, we were able to provide evidence about which variables were meaningful for the classification and provided information about group membership or provenance. This approach proves to be a valid technique for the quantitative analysis of XRF spectra.

47

48 **KEYWORDS:** Data analysis; Chemometrics; pXRF; Provenance; Obsidian sources;
49 Mesoamerica.

1. INTRODUCTION

Nowadays, a wide range of applications involve archaeometric techniques for the compositional analysis of diverse archaeological materials, aiming to distinguish chemical data groups by applying quantitative approaches like multivariate methods. For the case of provenance analysis of archaeological artifacts, the basic goal is to relate analyzed samples with their sources or place of origin; in other words, to classify artifacts into compositional types that reflect their geological origin (Frahm, 2014). The strategies employed for this kind of study are different depending on the type of artifacts under examination, implementing several approximations that consider several aspects that are closely related. One of these aspects is the selection of the instrument to be used for the data acquisition; some of the most widely employed are INAA, ICP-MS, SEM-EDS, PIXE, XRF, XRD, among others. Once the samples have been analyzed, researchers have to deal with choosing the right components that allow the better discrimination of the different chemical groups.

Another important aspect is the selection of the statistical method to process the data. The most common multivariate methods applied to the data analysis include bivariate plots of specific components or component ratios, Principal Component Analysis (Baxter, 1999; 2001; Doran and Hodson, 1975), cluster analysis (Hall et al., 1999), Factor analysis (Golitzko et al., 2012), discriminant analysis (Hall, et al., 1999; Kovarovic et al., 2011), canonical discriminant analysis (Neff et al., 2000) and Partial Least Squares regression (Lemberge et al., 2000). When using the conversion of the measurements into elemental concentrations, expressed as wt % or ppm, the calibration is an important step. For

example, in XRF, the photon counts or peak intensities of the analytes have to be measured along with certified reference materials or CRM (Rousseau, 2001).

A central consideration when selecting CRM is that they should exhibit the entire range of variation expected from the materials to be analyzed (Shackley, 2011). However, the instrument factory calibrations are generally not appropriate for several archaeological materials analysis. The calibrations are not matrix matched for archaeological materials and rarely contain all the elements of interest or an adequate elemental dynamic range suitable for archaeological materials characterization, and it is often not possible for the user to create a matrix that matches an empirical calibration (Hunt and Speakman, 2015). In essence, the chemical concentrations of a sample are obtained from the conversion of the intensities measured from the characteristic radiation of the elements (analytes) present which are fitted into a linear regression model called empirical calibration (Shackley, 2011). Linearity is critical given that non-linear regions of a calibration will adversely affect the accuracy of samples falling within those regions (Speakman, 2012). Only then, the matrix effects in all samples are similar and linear relationship between radiation intensity and analyte concentration can be obtained (Sitko and Zawisza, 2012).

It is common practice to check the linearity of a calibration curve by inspecting the correlation coefficient (r^2). Nevertheless, a correlation coefficient close to 1 is not a reliable indicator of linearity. In several cases, the calibration line adjustment is good, the slope is positive, the ordinate is close to zero, and $r^2 = 1$ and the data points are close to the straight line. But in many cases, although the adjustment is almost perfect ($r^2 = 1$), the straight line was forced to linearity by a single point that was located in an extreme of the dynamic

range of the analysis. Even if they are not outliers, these points will have a high degree of leverage and will have a significant effect on the position of the regression line (Barwick, 2003). Therefore, some of the reference materials could show concentrations grouped near the ordinate due to an inadequate dynamic range of certain elements within the calibration. On the other hand, since there is no linearity test as such, a hypothesis test by means of an *F*-test has to be applied as a reliable approach to check the linearity of any calibration function. Furthermore, to evaluate the analytical method, it is necessary to consider other validation parameters such as the selectivity/specificity, accuracy, precision, range, limit of detection, limit of quantitation, ruggedness and robustness (Araujo, 2009). Hence, calibration experiments with only one standard per concentration level are a poor calibration strategy and must be avoided unless the standard solutions are effectively error-free (Araujo, 2009). It is desirable to use several reference materials so that the error due to calibration is kept as small as possible.

For obsidian, which is quite a homogenous material, MURR2 Obsidian reference set (developed by Jeff Fergusson and used for Bruker's Tracer factory calibrations) seem to be reliable (Speakman, 2012). To confirm this, Glascock and Ferguson (2012) analyzed a set of 40 calibration samples with XRF, INAA, LA-ICP-MS and ICP-MS at the Missouri University Research Reactor (MURR). By a procedure of 17-hour continuous measurements and recording the spectra every 200 seconds, they obtained 307 analyzes with a relatively low variation (low relative standard deviation or %RSD) for most of the components according to the values obtained with the lab-based EDXRF instruments. The only exception was Sr that presented a higher %RSD. The plots of observed values versus

expected values showed that the calibration was highly linear and the values of the determination coefficient R^2 gave a good fit of the model (Speakman, 2012).

For other more complex (heterogeneous) materials, like ceramics or sediments, things are quite different. Hunt and Speakman (2015), based on Speakman et al. (2011), warn about the use of portable X-ray fluorescence (pXRF) factory calibrations, since they frequently do not include all the elements of interest for archaeological pottery and/or the dynamic range for their adequate quantification. For example, the 'mudrock' calibration computes inaccurate concentrations for most of the elements and other calibrations report values below the certified values. When managing pottery samples, it is recommended to homogenize the material prior to the analysis to avoid matrix effects. Rowe et. al. (2012) compared the efficiency of five internationally accepted standards and a suit of mudrock and other 85 reference materials developed by them. They determined that the standards are not diverse enough to properly calibrate the entire range of elemental concentrations in most mudrock studies.

Undoubtedly, the accuracy and reliability of the results are strongly dependent on the calibration of the device. pXRF analysis using empirical calibrations requires a large number of reference standards with similar physical and chemical properties as the unknown samples; but the preparation of these standards is expensive and time-consuming, which prevents the application of pXRF as a tool in many quality control processes (Gullayanon, 2011). But, even if the calibration of the archaeological materials is performed successfully, the researcher faces another problem. Regularly, in the analysis of compositional data, the original data is rarely processed, while applying some kind of

transformation is more common. Baxter (2008) mentions 5 types of transformations, like the standardization of the original data to $z(0,1)$, $\log_{10}$ and log-ratios, and Jajuga and Walesiak (2000) compare the performance of more than 13 different functions to standardize data. Further, it is frequent to use bivariate plots or multivariate methods for the discrimination of chemical groups, but the results given by these statistical techniques are scale dependent. The problem of data transformation has given place to a great debate on which is the suitable geometrical space for the analysis of data (Aitchinson, 1986; Baxter and Freestone, 2006).

When analyzing archaeological materials with different instruments, a large number of components can be reported; for example, INAA can report 33 components and pXRF around 20 (Bishop et al., 1990). Although it is possible to process high dimensional data, it is common to discard several of these components for different reasons, such as: missing data, contamination, observations below the limit of detection, values above a determined %RSD, poor analytical precision, or interference between some components. This is equivalent to making a selection of variables where the criteria for retaining a component in the model are established by a subjective criterion determined by the investigator. It should be more adequate that the criterion for the selection of variables were established by a procedure based on a mathematical model.

On the other hand, in archaeometric studies an inefficient use of statistical methods can be observed for the formation of chemical groups, with a clear absence of diagnostic tools like the detection of outliers. Furthermore, the employment of classical multivariate statistics can present problems when there are deviations from the theoretical assumptions of the

models or these are not fulfilled. In Archaeology, it is common to analyze data with unsupervised statistical methods like Principal Component Analysis or PCA (Santos et al., 2006; Tanasi et al., 2017), Cluster analysis or Correspondence analysis (Baxter and Cool, 2010). However, it is important to note that these are exploratory techniques and the groups so formed cannot be taken as conclusive. PCA is based on the proportion of variability explained, but the first PCs obtained frequently register relatively low percentages of the variability explained (i.e. 80% or less), which evidence a lack of adjustment of the model. On other occasions, the existence of outliers in the dataset inflates the variance which results in variability explained of more than 90%, giving the wrong idea of a good adjustment of the model. Regarding this, Varmuza and Filzmoser (2009) asserts that PCA should not only be based on the components, but should be based on the loadings.

PCA has other disadvantages, like the tricky interpretation of the PCs when most of the loadings are neither very small nor very large in absolute values (Croux et al., 2007; Hubert et al., 2016). Another important issue comes when the computation of the optimal number of components is not taken into account, which must be obtained from the cross-validation. After the PCA or cluster analysis, some applications label the groups detected in the exploratory analysis and use a supervised technique (like the Mahalanobis distance, or Discriminant Analysis) to check if the formation of groups has a greater statistical significance and to assign the unknown samples to the detected groups. Due to the large number of problems associated to compositional analysis, from calibration of the data to transformations and statistical methods employed, in this work we decided to analyze the data with the chemometrical approach by using the pXRF spectra. These spectra register all the fluorescent energy emitted by a sample and can be numerically expressed as matrices of

the distribution of the frequencies of photon counts *versus* energy levels or channel counts.

Therefore, it is possible to classify the samples in their respective chemical groups.

The application of chemometrics in archaeometry is an interesting subject. Nevertheless, few references have applied spectral chemometric analysis for the same aim. For example, Lemberge et al. (2000) employed Electron Probe X-ray Microanalysis (EPXMA) intensities and micro X-ray Fluorescence (μ -XRF) spectra for the analysis of 16-17th century archaeological glasses (180 samples), applying a Partial Least Squares regression (PLS) to quantify the major constituents of glass (SiO_2 , K_2O and CaO) and to build a calibration model with a relatively small error ($< 3\%$ for the EPXMA data and 5% in the case of μ -XRF). The PLS model summarized the information of the whole spectrum in some latent variables (LV's), orthogonal among themselves, avoiding possible multicollinearity. For this case, the spectra were used as explanatory variables and the measured SiO_2 as the response variable. This procedure was able to perform a diagnosis of the data from outlier maps. Later on, Hubert et al. (2005) applied a Robust Principal Components Regression (RPCR) for outlier detection to Lemberge et al. (2000) data. From the calculation of the robust score distance (SDi) and the orthogonal distance (ODi) in the PCA sub-space, they were able to discriminate between regular observations (non-contaminated) and two large groups of contaminated spectra (bad leverage points).

Another example is the one presented by Janssens et al. (1998), who used 180 samples archaeological glass vessels excavated in Antwerp, Belgium. From each vessel, they measured a spectrum at 1920 different energy levels corresponding to different chemical elements. In order to distinguish differences between imported or locally produced glasses,

the samples were chemically analyzed by employing EPXMA and μ -SRXRF (Synchrotron Radiation Induced X-Ray Fluorescence). Their data have been the basis of several analysis performed by Filtzmozer and Todorov (2011), Hubert et al. (2005), and Ortner et al. (2017). On the other hand, Seetha and Velraj (2016) obtained the elemental composition of archaeological ceramics from Energy Dispersive X-ray spectroscopy (EDS) and classify the samples with a Factor analysis (FA) and a Cluster analysis (CA).

This study proposes a procedure for the classification of obsidian archaeological artifacts and the discrimination of their geological sources by combining X-ray Fluorescence (XRF) and chemometrics. This procedure allows employing data expressed as spectra, although it can be used for elemental concentrations as well. With this in mind, a set of 109 obsidian samples of known origin was analyzed using a portable XRF instrument in order to obtain their characteristic spectra. The samples are from some of the most important Mesoamerican sources: Sierra de Pachuca, Otumba volcanic complex, Zinapécuaro, Ahuisculco, El Paredón, Zacualtipán, Tulancingo, Oyameles, San Martín Jilotepeque, El Chayal, Ixtepeque and La Esperanza. The number of samples analyzed for each obsidian source is detailed in Table 1. As a pilot test, 257 obsidian artifacts from the archaeological site of Xalasco were analyzed by the same procedure. The results confirmed the efficiency of the proposed methodology, which proves to be a valid quantitative approach to handle semi-quantitative data in a way that does not require the calibration of the samples by using certified reference materials.

2. REGIONAL SETTING

Mesoamerica contains several important obsidian sources which were intensively exploited by pre-Hispanic lithic industry since early times. This natural resource was deemed of great value and was distributed via large exchange networks (Clark, 1989; Cobean, 2002). According to Gaxiola (1989) and Cobean (2002), some of the more intensively and extensively exploited natural deposits in Mesoamerica are: Sierra de Pachuca, Otumba complex, Tulancingo, Zacualtipan, El Paredón, Oyameles-Zaragoza, Pico de Orizaba (in Central Mexico), Zinapécuaro, Ucareo, Tequila volcanic region (in western Mexico), El Chayal, Ixtepeque, San Martín Jilotepeque and La Esperanza (in Guatemala and Honduras). Due to their physical qualities, some sources were more extensively and intensively exploited than others, like Sierra de Pachuca, Zinapécuaro, Zacualtipán and Otumba. In this research, we used samples coming from the twelve sources (Figure 1) described below.

Sierra de Pachuca, also known as Sierra de las Navajas, is located between the modern cities of Pachuca and Tulancingo in the state of Hidalgo, on the northern edge of the Trans-Mexican Volcanic Belt (TMVB). It is the largest obsidian deposit in central Mexico, covering an area of approximately 250 km². Las Navajas is a per-alkaline volcanic center already extinct. The eastern portion of this area consists of volcanic cones and domes of the Late Pliocene with a composition ranging from andesite to rhyolite; the western portion contains Pleistocene volcanic structures with a composition ranging from andesite to dacite (López et al., 1989). The green-translucent obsidian comes from Las Minas flow, the second of four main eruptive events, with an age of approximately 2 My (Ponomarenko, 2004). Obsidian from this deposit can be found as part of a lava flow or in blocks carried down the slopes as part of an avalanche. Its exploitation started thousands of years before the Spanish conquest and is still in process by modern artisans. Due to its almost unique

color, its clean conchoidal fracture and its lack of crystals, this green obsidian is ideal for the manufacture of a diverse variety of artifacts.

The Otumba complex is located in the limits of the states of Mexico, Hidalgo and Tlaxcala, in the central portion of the TMVB. It covers an area of approximately 40 km² and is the closest deposit to Teotihuacan, and was exploited by this ancient city during its early phases (Clark, 1989). This volcanic region was affected by two main fault systems: the east-west Chapala-Acambay fault system and the north-south Soltepec fracture zone (García-Palomo et al., 2002; Hernández, 2007; Mooser, 1968). Several lava flows and obsidian domes were originated in these two systems. Soltepec range is a rhyolite calc-alkaline volcanic center consisting in a sequence of obsidian flows intercalated with microcrystalline rocks; it is related to the emplacement of rhyolitic domes from the early Pleistocene around 1.5 My, according to K-Ar datings of two of our samples.

The northern part of the Otumba volcanic field corresponds to the Chapala-Acambay system and contains several rhyolitic domes and scoria cones aligned with the Cerro Gordo complex. This system also originated obsidian flows from the Pleistocene (2-1 My, according to K-Ar dating of six of our samples), found in the range formed by Malpais, Pacheco, Ixtepec and Tepayo hills (Argote-Espino et al., 2012). Although Soltepec mines were more intensively exploited in pre-Hispanic times, Malpais sub-source has also been reported in archaeological literature (Mora, 1981; Cobean, 2002).

The Tulancingo source system is located in the state of Hidalgo, approximately 10 km east of Sierra de Pachuca, and covers an area of around 80 km². Considerable concentrations of

Zr and Zn provide a slight greenish color to this obsidian (Argote Espino et al., 2010), showing a certain degree of devitrification. The quality of the glass makes it suitable only for the manufacture of monofacial and bifacial artifacts (Gaxiola and Guevara, 1989). This source has been exploited since the Archaic period (7000 BC), although its most extensive quarrying appears in the Late Classic (750-900 AD) by Huapalcalco regional center (Cobean, 2002; Gaxiola and Guevara, 1989). The largest pre-Hispanic mine complex is found at El Pizarrín, near the southern edge of Huapalcalco and on the slopes of Cerro Tecolote, in the northeastern portion of the modern city of Tulancingo.

Zacualtipan is located in the eastern portion of the state of Hidalgo, within the province of the Sierra Madre Oriental; this massive obsidian flow covers a minimum area of 60 km². This volcanic glass exhibits a very dark lustrous aspect with varying degrees of translucency. Previous geochemical studies have determined a wide geographic distribution of artifacts coming from this source in the Late Classic and Postclassic periods, documented in near sites within Hidalgo, San Luis Potosi, Veracruz, Puebla and Morelos states and reaching areas as far as Oaxaca, Chiapas and Guatemala (Cobean, 2002; Mandujano et al., 2002).

The El Paredón source is located at the north-eastern end of Tecocomulco sub-basin in the state of Puebla, near the border with the state of Hidalgo, covering an area of approximately 30 km² (Cobean, 2002; Millhauser et al., 2017). The region presents a volcanic sequence of dacite flows and gaps with rhyolitic clasts and glasses dated between 3 and 0.24 My (García-Palomo et al., 2002). Coyaco, Tres Cabezas and El Paredón are three localities associated with this source that present evidence of pre-Hispanic quarries (Charlton and

Spence, 1982). Tres Cabezas occurs as a discontinuous flow exposed by a gully, while El Paredón consists of two hilltop extrusions near the town of El Paredón (Millhauser et al., 2017). This deposit has a long history of exploitation, being a primary source of obsidian for sites of the Puebla-Tlaxcala Valley and Central Morelos during the Formative (2000 BC – 100 AD) and the Postclassic (900-1521 AD) periods (Carballo et al., 2007; Charlton et al., 1978; Millhauser et al., 2015; Millhauser et al., 2017).

The Oyameles deposit is located few kilometers north of the pre-hispanic center of Cantona, in the eastern portion of the state of Puebla and near the border of Veracruz state. This source extends for over 30 km from the town of Oyameles to the south and west of Zaragoza city and consists of obsidian outcrops and rhyolitic domes that form part of the Xaltipan Ignimbrite, with an age of 0.46 My (Cobean, 2002). Its exploitation began in the Formative period, reaching its maximum peak during the Late Classic and until the Late Postclassic periods; several workshops associated to this source produced prismatic blades and bifaces (Evans and Webster, 2001)

The Zinapécuaro obsidian source is located within and south of the city with the same name, in the northwest portion of the Mexican state of Michoacán. The geology of the region includes three obsidian flows (Zinapécuaro, Cruz Negra and Ucareo), dated between 1 and 1.6 My (Nelson and Healan, 1995), covering an area of about 80 km² (Cobean, 2002). Its obsidian has a translucent dark-gray color that presents bands of micro-crystals following the direction of the flow. According to some researchers, this outcrop alongside with Ucareo were probably the second in importance for the production of prismatic blades

in the Formative period, formed between 1100-100 BC, having a widespread distribution (Boksenbaum et al., 1987; Cobean, 2002; Darras, 2012 ; Healan, 1997).

Ahuiscalco is in western Mexico, 32 km southwest of Guadalajara city, Jalisco state, within the Selva Negra natural sanctuary. Several individual flows compose this source, spread over an area of 3.6 km². Black and red mecca obsidian with excellent fracturing properties can be found around the area, as well as large spherulites formed by the crystallization of the obsidian. Over 300 obsidian pre-Hispanic mines and workshops have been located, from which the earliest dated materials are from the Late Formative period (300 BC to 200 AD), although some Classic and Postclassic materials are also present (Glascok et al., 2010).

The Guatemala highlands contain three major obsidian sources, El Chayal, San Martín Jilotepeque and Ixtepeque. El Chayal is located in the Canchon Plateau, about 20 km northeast from the capital city of Guatemala and 25 km from the ceremonial center of Kaminaljuyu; it covers an area of approximately 525 km². San Martín Jilotepeque is located 50 km northwest from the capital, in the municipality of Chimaltenango, and covers an area of about 280 km² (Carr, 2015). Ixtepeque is a stratovolcano with several rhyolitic domes, located north of Lake Guija in the southwestern border of Guatemala, near Honduras and El Salvador; it covers an approximate area of 300 km² (Sorensen and Hirth, 1984).

The beginning of the exploitation of Guatemalan sources started in the middle Formative period (1000 to 400 BC), with San Martín Jilotepeque being more representative of this period (Brown et al., 2004); from the Late Formative to the Terminal Classic periods (400

BC to 900 AD), El Chayal became the dominant source, being overshadowed by Ixtepeque in the Postclassic period (900-1500 AD; Nazaroff et al., 2010). Their major production consisted in prismatic blades, which were distributed through large commercial networks to the Pacific coast, Belize and the Mayan Lowlands, to the north and west (Brown et al., 2004; Carr, 2015; Nazaroff et al., 2010), and El Salvador, Honduras and Nicaragua, to the south (Sheets et al., 1990).

La Esperanza is located near Lake Yajoa, municipality of Intibuca, in the southwestern highlands of Honduras. Evidence of occupation and obsidian use from this source has been reported at least since the Formative, through the Classic and until the Postclassic periods. The production of prismatic cores and blades was a major component in lithic manufacture with obsidian coming from this source, with minor assemblages of bifacial and monofacial artifacts (Sheets et al., 1990). A total chronology of its exploitation is still uncertain, though assemblages from the Late Classic and Postclassic periods are documented (Sorensen and Hirth, 1984).

3. MATERIALS AND METHODS

Volcanic glasses are formed by the rapid cooling of lavas in contact with the atmosphere. The tendency of lava to produce obsidian is based on its cooling range and its composition, with felsic lavas (with higher contents of SiO₂, Na and K) been more suitable for obsidian formation (Argote-Espino et al., 2012). One advantage of obsidian over other archaeological materials, such as ceramics and soils, is its homogeneous composition within the same lava flow and its heterogeneous composition with respect to different

volcanic regions (Pollard and Heron, 2008). Multi-elemental analysis of obsidian artifacts can characterize distinct chemical signatures that are traceable to their natural source. Therefore, the employment of analytic techniques in archaeological artifacts has allowed precise provenance identifications (Carballo et al., 2007; Carter et al., 2006; Glascock et al., 2007; Smith et al., 2007; Tykot, 2010; among others).

Spectroscopic methods have been increasingly applied in archaeology and art history since they are quick, simply to apply and capable of measuring the chemical characteristics of diverse materials. Currently, numerous papers have focused on the employment of portable X-ray Fluorescence (pXRF) instruments for discerning the provenance of obsidian artifacts due to their positive characteristics: non-destructive analysis, minimal sample preparation, fast elemental identification, simultaneous multi-elemental determination (from Na to U), low costs, transportability, short analysis time and ease of use (Hunt and Speakman, 2015; Nazaroff et al., 2010; Scharlotta and Quach, 2015; Shackley, 2010). Although pXRF has several disadvantages such as the low detection limit of some light elements (like Na and Mg) and the need for empirical standards (Speakman and Shackley, 2013), the reliability and validity of this technique for obsidian sourcing has been tested with confident results (Drake et al., 2009; Frahm, 2014; Millhauser et al., 2011; Millhauser et al., 2017; Nazaroff et al., 2010).

XRF uses the emission of secondary radiation generated by exciting a sample with a primary X-ray radiation source. As each element has a different number of protons, the resulting secondary radiation emission is characteristic of the different atoms that compose the sample. This emission can be displayed as a spectrum composed by the photon counts

at different regions of energy or channel counts p . Therefore, X-ray spectroscopy can provide $n \times p$ data matrices, where the rows refer to the obsidian samples and each column is associated with a channel or energy region of the spectrum. In this study, the matrix was formed by $n = 109$ and $p = 1939$ (Table S.1), so $n \ll p$.

For this research, we used a Bruker TRACER III-SD portable analyzer (S/N: T3S2615), with a Rh anode angled 52 degrees from the sample and a glancing angle of 62 degrees toward a Silicon Drift Detector with a 7.5 μm Be window. The acquisition parameters were set with an energy of 40 kV and a current of 30 μA , using a filter composed of 150 μm Cu, 25 μm Ti, 300 μm Al; no vacuum pump was used. The measuring time was 200 live seconds. In order to be able to compare the results, these parameters were set in the same way for the analyses of all samples (Kaiser and Shugar, 2012). A table with the chemical concentrations of the sources (Table S.2) obtained with Bruker's Tracer factory glass calibration (based on the MURR2 calibration obsidian set) is added for others in the field to evaluate the results.

3.1. Spectrum pre-processing

Experience has shown to us that portable XRF instruments are susceptible to shifts in the energy calibration over time (X-axis), this being a reality for instruments using a wide variety of detectors. Differences due to unstable experimental and/or instrumental conditions are some of the causes of these shifts. One small variation in the position of a peak is an acceptable error, but a large sum of these displacements can propagate the error. If the horizontal drift goes unnoticed by the researcher, the analysis will be wrong. On the

other hand, if calibrations are performed over misaligned peaks, the reported concentrations will also be incorrect. Therefore, an alignment of the intensity peaks of the spectra is necessary to eliminate the effects of such variations on the data sets and to ensure that the same components are precisely identified in each of the analyzed samples. Peak alignment procedures aim to compensate for the small variations in the position of the corresponding peaks between the spectra. Figure 2a shows an example of spectra with shifted peaks in an interval of 590 to 830 channel counts (11.8 to 16.6 keV). Although the position of the peaks is not that variable, it can be considered that they are misaligned.

To address the problem of misaligned spectra, we employed a robust and highly reliable alignment algorithm known as Hierarchical Cluster-based Peak Alignment or CluPA, originally developed by Vu et al. (2011) to work with Nuclear Magnetic Resonance spectroscopy and modified by Beirnaert et al. (2018). Although the CluPA algorithm works for raw spectral alignment, it can also be used to quantify spectral regions that differ between case and control classes (Beirnaert et al., 2018). The algorithm aligns a target spectrum with the reference spectrum in a top-down fashion by building a hierarchical cluster tree from peak lists of reference and target spectra and then dividing the spectra into smaller segments based on the most distant clusters of the tree.

For the detection of peaks, the algorithm uses the continuous wavelet transform (CWT) to the spectrum, dividing it into predefined equal window-sized segments. CluPA uses the peak lists to find the shifts needed to align corresponding to the segments, in a recursive scheme that moves from global (the entire spectrum) to local (a small subsection of the spectrum) in order to go from a crude to a more refined alignment (Vu and Laukens, 2013).

In Figure 2b can be observed the result of the application of this algorithm. We consider that is imperative to align the spectra as the first step in data treatment; afterwards further processing can be performed.

When reaching the material, some of the x-rays will be absorbed and some scattered; therefore, overlapping bands may occur. Moreover, besides the useful information, the collected spectra may contain other irrelevant information and noise. Differences in the relief along the surface of obsidian samples can hold undesired effects on the measurements, due to the physical dispersion of the light over the sample. These effects on light translate directly to complexities when interpreting the fluorescence radiation that is detected in the XRF spectrometer (La Tour, 1989). Spectra are often pre-processed in order to remove systematic noise and other spurious data such as baseline variation and multiplicative scatter effects. Depending on the subject under research, for the pre-processing there are several computer applications that make use of specific filtering methods or a combination of filters in order to enhance the quality of the spectra. Some pre-processing methods commonly employed are: normalization, weighting, smoothing and baseline corrections, multiplicative scatter correction and derivatives for baseline correction of spectroscopic data (Lazim et al., 2016; Martelo and Vázquez, 2014).

The Savitsky-Golay (SG) filter employs a method for data smoothing and noise removal. It is based on a local least-squares polynomial approximation, which reduces random noise of the instrumental signal while maintaining the shape and height of waveform peaks (Schafer, 2011). Another pre-processing technique that has shown to be good at spectra filtering is the Extended Multiplicative Signal Correction method (EMSC), which has

proved to be a reliable tool for the correction of additive baseline, multiplicative scaling and interference effects (Afseth and Kohler, 2012). The EMSC performs model-based background correction and normalization of the spectra, removing identified but undesired physical and chemical interference effects, while retaining identified but desired effects as well as unidentified effects in the data (dos Santos Panero et al., 2013). While the use of SG and EMSC filters have been common in FTIR analysis, this is the first case where they have been applied to XRF data.

The unknown parameters of the EMSC are estimated using an ordinary or weighted least squares estimation. In addition to correcting the spectra with the desired model, the model parameters are returned (Liland et al., 2016). The EMSC can extend the fitting to any desired order (linear, quadratic, polynomial), which is used to correct more complex fluctuating baseline features. More details about preprocessing techniques can be found in (Liland et al., 2016; Varmuza and Filzmoser, 2009; Wehrens, 2011). In order to verify the robustness and prediction ability of several pre-processing methods, dos Santos Panero et al. (2013) compared the performance of the EMSC and the PLS models with Short-Wavelength Near-Infrared (SW-NIR) spectra. The authors found that the EMSC models always produced low minimum values of the root mean square error of cross validation (RMSEC). A low RMSEC indicates that predictions are close to the reference values of the samples on average. The authors conclude that EMSC methods provide the best overall performance in terms of prediction ability and calibration.

3.2. Variable selection

Commonly, chemometric applications work with the whole data spectrum, paying little attention to methods for the optimal selection of variables and reduction of dimensionality. Spectral data contain a high degree of covariance and large amount of redundant information (Nørgaard et al., 2000). Variable selection techniques can be used to improve chemometric models and to enhance spectral features. These techniques are, in fact, able to retain and preserve only the variables that contain significant information for a specific task (Ballabio, 2006). Recently, considerable efforts have been made to develop algorithms that can distinguish informative energy regions or remove non-informative ones from hundreds of spectroscopic data points (Xu et al., 2014).

Among the high-performance energy region selection methods are: Interval Partial Least Squares or iPLS (Xu et al., 2015), Synergy Interval Partial Least Squares or siPLS (Nørgaard et al., 2000), Moving Window Partial Least Squares PLS or MW-PLS (Xu et al., 2014), Competitive Adaptive Reweighted Sampling or CARS (Hongdong et al., 2009), Genetic Algorithm PLS or *GA-PLS* (Leardi, 2000) and Interval Random Frog or iRF (Yun et al., 2013). The selection of specific spectral regions, or variables, is key to decreasing the complexity of the calibration model and increasing its prediction precision (Leardi, 2000; Nørgaard et al., 2000). These spectral regions correspond to certain elements that emit energy in those intervals. In this work, the iPLS and the siPLS were applied to select the spectral range that provided the lowest prediction error in comparison to the full-spectrum model. These methods can search for the optimal spectral interval or intervals respectively, that is, those that contain the most important information to model the parameters of interest (Gao et al., 2015).

The objective of the iPLS is to split the spectra into smaller equidistant subintervals, developing PLS models on each subinterval using the same number of latent variables (Xu et al., 2015), and the predictive performances of all the local models as well as the global full spectrum model are compared. The model comparison is based on the validation parameter RMSECV/RMSEP (root mean square error of cross-validation/prediction) for a given validation scheme (Leardi and Nørgaard, 2004). The best subintervals are determined on the basis of the (RMSECV) values (Suhandy et al., 2013), and the root mean square error of prediction (RMSEP) but other parameters such like the squared correlation coefficient (r^2), slope and offset (Nørgaard et al., 2000). The RMSECV can be calculated with the following equation:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (1)$$

where n is the number of samples in the calibration sample set, y_i is the actual value for sample i and $\hat{y}_i$ is the predicted value for sample i , when the model is constructed with sample i removed (Romia and Bernardez, 2009; cited by Suhandy et al., 2013).

RMSEP is also an external validation method which evaluates whether the model constructed from calibration data sets correctly predicts the validation data. The RMSEP can be calculated with the following equation:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

where $\hat{y}_i$ denotes the predicted value of the i th sample of the calibration set, y_i denotes the observed value of the i th sample of the validation set, and n is the sample size of observed values in the validation set (Lin et al., 2015).

The best treatment method and spectral range are decided by the model's evaluation parameters. One of the main advantages of this method is the possibility to represent a local regression model in a graphical display, focusing on a choice of the best intervals and allowing a comparison between interval models and the full-spectrum model (Xu et al., 2015; Nørgaard et al., 2000)). In our case, the whole spectral data range was used as the independent variable X and Sr concentration values were used as the dependent variable y . The selection of the interval that held most information and less redundancy was based on the model whose value had lower RMSECV than the average. The prediction accuracy of the established *iPLS* model was evaluated by an external test validation and a leave-one-out cross validation.

Zhiming et al. (2014) point out that the selected features are usually scattered throughout the spectrum; therefore, the underlying principle of the siPLS tell us that the full spectrum must be split into a number of smaller intervals. Then, PLS regression models are employed in all possible combinations of two, three or four intervals, calculating the RMSECV for every combination of intervals. The optimal model was verified by the independent samples of the prediction set. Then, the combination of intervals with the lowest RMSECV is chosen. Generally, a good model should have lower RMSECV and RMSEP, and higher r^2 (Pan et al., 2015; Zhiming et al., 2014). The *iPLS* and the *siPLS* routine generate

graphical information indicating the optimal number of latent variables used in each interval model and the RMSECV values. The selected intervals can be subsequently used in any other modeling method, like robust principal component analysis or ROBPCA, PLS and PLS-DA.

3.3. ROBPCA in high dimensionality

Due to the large number of variables involved in a spectral analysis, the existence of outliers should be taken into account. Hawkins (1980) defines an outlier as an observation that deviates much from other observations and arouses a suspicion that it was generated by a different mechanism. An outlier is also defined as an unusual observation outside the majority of the data set, behaving differently from the rest of the observations (Van den Boogaart and Tolosana-Delgado, 2013). Many multivariate methods base their computations in the vector of means and the covariance matrix of the sample, which are very sensitive to the presence of outliers (Filzmoser and Todorov, 2011; Hubert et al., 2005), thus being important to identify them prior to modeling and analysis (Ben-Gal, 2005; Hron et al., 2010; Liu et al., 2004; Williams et al., 2002).

A large number of methods have been proposed to detect outliers in different kinds of data sets. With the aim of avoiding erroneous interpretations, it is advisable to employ robust methods, which are less sensitive to the presence of outliers. A robust estimation finds a 'robust' fit, similar to the fit we would have found without outliers (Hubert and Rousseeuw, 2011). If $n > p$, then the minimum covariance determinant (MCD) method can be used (Rousseeuw and Van Driessen, 1999). This is a robust and affine equivariant estimator;

being asymptotically normal, it is statistically more efficient than other estimators. Nevertheless, in chemometrics there is the curse of dimensionality; therefore, it is necessary to use estimators that work when $n \ll p$.

To overcome this difficulty, Hubert et al. (2005) proposed the ROBPCA method. The reduction of dimensionality of ROBPCA consists in performing a decomposition of the singular data values, in order to express the information in the n -dimensional space (Filzmoser and Todorov, 2011). ROBPCA method combines ideas from both projection pursuit (PP) and robust covariance estimation. The PP is used to pre-process the data, so that the transformed data can lie in a subspace whose dimension is at most $n-1$. Afterwards, the fast-MCD estimator is used to get the robust center and covariance matrix (Nurunnabi et al., 2012). This technique maximizes a robust measure of spread and obtains consecutive directions in which the data points are projected (Hubert et al., 2005).

The advantages of ROBPCA are twofold: (1) it finds those linear combinations of the original variables that contain most of the information, even if there are outliers, and (2) it flags outliers and determines their type. A diagnostic plot can be used to detect and classify possible outliers; this plot allows discerning four types of observations: regular observations, good leverage points, orthogonal outliers and bad leverage points (Hubert et al., 2005). In the outlier map, each observation is displayed: the score distance (SDi) within the PCA subspace is represented on the x -axis, while the orthogonal distance to the PCA subspace is in the y -axis (Verboven and Hubert, 2005). To classify the observations, two cutoff lines are drawn. The cutoff value for the horizontal axis is determined by $\sqrt{X_{k,0.975}^2}$.

The cutoff value for the vertical axis is obtained through the Wilson-Hilferty approximation, which is a scaled version of the χ^2 distribution (Hubert et al., 2005; Verboven and Hubert, 2005).

3.4. Classification

The goal of supervised chemometric methods is to create classification/discrimination rules using a set of training samples of known origin and then to use the created rules to assign new samples of an unknown origin to one of the available classes (Stanimirova et al., 2010). Several supervised classification algorithms are based upon calculating the distance of each one of the observations to each of the centroids of the reference group (Atkinson, 2009). Some provenance studies of archaeological artifacts have established certain rules of classification, like the Mahalanobis Distance (Glascok, 1992; Little, et al., 2004; Stahl, et al, 2008), Discriminant Functions (Smith et al., 2007) and the Squared Euclidean distance (Kilikoglou et al., 2007). When dealing with obsidian provenance, the sources are divided in G groups ($y = \{ 1, 2, \dots, G \}$) where each group is characterized by a matrix of dimension $n \times p$. Subsequently, a decision rule is established to assign unlabeled data to one of the pre-defined groups.

One of the best known approximations for the calculation of membership probabilities of each specimen to his group centroid is the Mahalanobis Distance (MD). This distance takes into account the complete covariance matrix to construct a dispersion matrix for each of the groups. Assuming a multivariate Gaussian distribution in terms of standard units, the MD is converted into group membership probabilities for each specimen. If the MD of an

observation with respect to the centroid is within the dispersion matrix defined by the covariance of the reference group, then the probability that this observation belongs to the reference group is greater than 0. The smaller the distance from an observation to the centroid of a group, the greater the probability of the observation belonging to that group. However, the dispersion between groups can cause overlap, with the result that some observations may belong to more than one reference group. Similarly, if an observation has an MD outside of the dispersion matrix, then it has probability 0 of belonging to any of the reference groups (Grunski, 2009). As a consequence, many observations are not assigned to any group and are sometimes labeled as outliers because of the masking problem, which frequently occurs when the data distribution is not normal. Another problem with MD is the sample size, which must fulfill $n \gg p$; when this condition is not fulfilled, the MD can be calculated from the first components of a PCA.

A classification by means of a Linear Discriminant Analysis (LDA) is another common archaeometrical approach to predict the group membership of observations (Smith et al., 2007). The procedure for the classification of unknown samples is based on discriminant functions that assign an object of unknown origin to a reference group. One of its theoretical assumptions says that the covariance matrices must be equal, which is rarely fulfilled. Another problem is that LDA is very sensitive to the presence of outliers and multicollinearity between the predictors, which can reduce its discriminant power. Also, when groups are of unequal size or there is overlap between them, the computation of the error rate is highly impacted. In other words, the more the subgroups overlap with one another, the larger the disparity in classification accuracy between the equal and unequal

subgroup size conditions (Finch et al., 2014). Due to these problems, LDA has the highest rates of misclassification in situations when the groups were not homogeneous.

To avoid the disadvantages of the above mentioned methods, a novel approach is introduced to archaeometric analyses. Soft Independent Modelling of Class Analogies (SIMCA) is a supervised classification technique whose goal is to obtain a classification rule for a set of g known groups in such a way that additional information about the group structure is exposed (Hubert et al., 2005; Vanden Branden and Hubert, 2005). The algorithm can work with both low-dimensional and high-dimensional data. With SIMCA, a PCA is performed on each of the groups of samples by using the optimal number of PCs, which allows each group to be summarized in a different dimension (Vanden Branden and Hubert, 2005). To find the optimal number of components, it uses a cross-validation procedure. Nonetheless, SIMCA also has some inconvenience: its sensitivity to the presence of outliers, its inability to detect these outliers, and the simultaneous classification of an observation in several groups.

To attend these deficiencies, Vanden Branden and Hubert (2005) modified SIMCA into a more robust classifier called RSIMCA, based on the ROBCA algorithm described. This method assigns each observation to only one group and the outliers are labeled as such. When the sample size is large enough, the performance of the classification is evaluated by dividing the dataset into a training set and a validation set. Otherwise, if the sample is too small to divide, it is best to estimate the PCA models from the full data set and evaluate the classification rules by means of leave-one-out cross-validations. The first step is to apply ROBPCA method to all the training sets to develop a decision rule for the construction of a

classification model that is not affected by outliers. For this purpose, the orthogonal distance (OD) and the score distance (SD) are employed. So, for a new observation y , the SD with respect to the l th group is taken into account, along with largest robust eigenvalues in the l th and the diagonal matrix of the eigenvalues. The distances are standardized by means of two reference or cut-off values; for the SD, the cut-off value for the score distance in group l (c_{SD}^l) is based on the X^2 -distribution with k_l degrees of freedom at a determined α significance level

$$c_{SD}^l = \sqrt{X_{k_l;0.975}^2} \quad (3)$$

The cut-off value for OD is given by

$$c_{OD}^l = (\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2} \quad (4)$$

with $z_{0.975} = \Phi^{-1}(0.975)$ the 97.5% quantile of the Gaussian distribution. With these cut-off values, a classification rule is constructed by creating a confidence region for each one of the groups: if $k = 1$, the boundary is a straight line; if $k = 2$, the region is a square; if $k = 3$, the region is a box. Multidimensional boxes are obtained for higher dimensions (Daszykowski et al., 2007a; Vanden Branden and Hubert, 2005). The classification of the new observations is performed with a linear combination of these two standardized distances. The addition of the tuning parameter γ , which can vary from 0 to 1, emphasizes in the role of a certain distance in the construction of the final rules of classification. To

obtain satisfactory classification results, the parameter γ must be optimized with the test set or a cross-validation (Daszykowski et al., 2007a).

3.5. Software

Most of the statistical computations were done using free access software working in R environment version 3.4.3 (R development Core Team, 2011), which includes a wide variety of functions for chemometric analyses. For data pre-processing, CluPA was implemented with the package called “speaq” version 2.0 (Beirnaert et al., 2018); considering it as an important step in data pre-processing and of a more general use, the modified version of this script can be found as an appendix at the end of the article. The SG filter was applied with the “Prospectr” package version 0.1.3, which can be found in Stevens and Ramirez-Lopez (2015). For filtering with the Extended Multiplicative Signal Correction algorithm, we used the “EMSC” package version 0.9.0 (see Liland, 2017).

The iPLS and siPLS, for the selection of the most informative variables, were conducted using iToolbox for Matlab version R2016a (Nørgaard, 2013). The iPLS and SyPLS independent variable matrix (X matrix) was generated by assembling the spectral data of each sample as a row vector; the dependent variable (Y matrix) of each model was composed by the concentration values of Sr for each sample. For the ROBPCA computation, we used the toolboxes TOMCAT (Daszykowski et al., 2007b) and LIBRA (Verboven and Hubert, 2005) for MATLAB. ROBPCA can also be performed in R with the function PCAHubert from library “rrcov”. The RSIMCA algorithm was computed using

the LIBRA Toolbox for Matlab; for the prediction and calculation of classification error, we used the “nnet” package version 7.3-12 for R (Verboven and Hubert, 2005).

4. RESULTS AND DISCUSSION

For this analysis, the entire spectral range (1 to 1939 channel counts, equivalent to the energy range of 0.02 to 38.78 keV of the detector resolution) was employed. Figure 3 presents the raw spectra of all obsidian samples with the peaks previously aligned using the CluPA algorithm. As the Figure shows, the shape of the spectra is similar, but the intensities are different. The variations in the global intensities (pulse counts) of the spectra are evident in the Figure; larger fluctuations are present at 5.6, 15.4-15.6 and 20.4-20.6 keV. On the other hand, in addition to the sample information, the raw spectra contain background information and noise. To correct these undesired effects, the spectra were smoothed using the SG filter and the EMSC algorithm. For the SG filter, we used the following parameters: p = polynomial order (3), w = window size (11) data points, and m = m -th derivative (0 = smoothing). For the EMSC, we increased the polynomial degree to six to provide a major improvement to the spectral data. In Figure 4 we show the reduction of noise and the smoothed spectra, which decrease the effects of multiple scattering (caused by sample surface roughness) and enhancing small spectra differences in the samples.

Once the data was pre-treated, the spectra were analyzed by two methods for the detection of highly informative bands: iPLS and siPLS. Firstly, we applied the iPLS to eliminate

collinearity and redundancy of information in spectral data. The algorithm iPLS focuses on important spectral intervals and removes interference in the full-spectral region (Nørgaard et al., 2000; Ferrao et al., 2011). The full spectral region (1– 1939; 0.02 to 38.78 keV) was divided into 20 equidistant non-overlapping intervals; each interval underwent a separate PLS modelling to determine the most useful variable range. The optimal model performance was evaluated using the RMSECV and the correlation coefficient r . Bar plots of the RMSECV are shown in Figure 5 for each range of variables and for the models developed with the optimal amount. The iPLS model produced the best results of RMSECV with 4 PLS components (0 to 3).

In Figure 6, the bar plots of the RMSECV are shown for each range of variables (20 intervals) and for the models developed with one to four principal components (PCs). The dotted line represents the RMSECV of the full-spectrum iPLS model. The best modelling parameters in each iPLS interval model were selected. The iPLS calibration model of the 7th interval, correspondent to the peak of Sr located in the 12-14 keV range of the PLS model, achieves the minimum RMSECV (with 3 optimal LVs components). The local iPLS calibration model of the 7th interval (Figure 7), used for predicting the Sr value, is reasonably good with 3 PLS components, a lower RMSECV (6.9991), a value of $r = 0.9882$ and a bias of 0.0216. Although a satisfactory RMSECV value was obtained, the presence of outliers can be detected, for example sample no. 17 which is located far from adjustment of the predicted values. These observations usually cause an undesired effect in the calibration of the model and tend to distort the results (Nørgaard, 2013).

To detect outliers, we applied ROBPCA algorithm to our data set. In Figure 8 are displayed the ODs versus the SDs for the obsidian data, where observations 17 and 109 are classified as bad PCA-leverage points and observation 53 is labeled as an orthogonal outlier. Sample 17 proceeds from Ixtepeque source and sample 109 from Zinapécuaro; these two samples were prepared different (pressed pellets) than the rest of the samples (bulk fragments), which explains why the half back of their spectra diverge (see Figures 3 and 4). In some occasions, the researcher does not take into account the presence of divergent spectra and analyzes the whole data set as outlier free. Hence part of the need of applying outlier detection methods. Although the method was effective in discriminate them as outliers, it is also advisable to analyze and process samples prepared the same way. There are also two small orthogonal outliers and some good leverage points correspondent to the samples proceeding from Zacualtipán source (sample numbers 92 to 101). However, the good leverage points do not impact negatively the estimations, but rather stabilize the estimation of the PCA space. On the other hand, the bad leverage points are due to their distance from the center in the projection of the PCA space (Todorov and Filzmoser, 2009).

According to the previous results, the two bad leverage points and the orthogonal outlier were eliminated and the model was re-calibrated, observing an improvement in the model performance (Figure 9) represented by a lower RMSECV value (3.4327) and a higher r value (0.9972). The RMSEP of the prediction was 3.03 for the new model. For this last iPLS model, the results showed a good correlation between reference and predicted values, indicated by the correlation coefficient as shown in Figure 7. By comparing the figures 7 and 9, we can see that the measured and predicted values are similar and that the model's

predicted effect was good. Therefore, it was possible to fit a narrower region for Sr determination with small prediction errors and a reduced variable number.

The siPLS algorithm develops PLS regression models for all possible combinations of two, three or four intervals; siPLS combine several informative spectral regions in building calibration models (Zhiming et al., 2014). The combination of intervals with the lowest RMSECV was chosen, using the model fit without the presence of outliers; the selected spectral regions or intervals were 4 and 6 (Figure 10). Figure 11 is a scatter plot showing the correlation between the reference values, the measured and predicted spectral data. The PLS method obtained a predicted correlation coefficient of 0.9973 and a RMSECV value of 3.3540.

Thus, the presence of outliers and the interference of uninformative variables not only decrease the precision of the prediction, but also complicate the multivariate model. Hence the importance of obtaining a reduced model by applying variable selection methods to the processed data. In our case, the local model (interval 7) of the iPLS model was reduced to a number of only 92 variables (compared to the initial 1939 used in the full-spectrum model) and 4 LVs (compared to the 10 LVs used in the full-spectrum model), resulting in a more robust model with better predictive power. On the other hand, the siPLS model holds a total of 184 variables with 3 LV's.

Afterwards, the ROBPCA method was applied to the obsidian samples data; the input data had now only $p = 92$ variables, including only the informative variables selected by the iPLS model. When applying ROBPCA to a data set, it is important to decide how many

principal components to keep. “Libra” toolbox (Verboven and Hubert, 2005) offers two options. In the first option, the number of principal components to retain is set to zero ($k = 0$); in the second option, used when the optimal number of components are known, this value is specified in the function (e.g. $k = 5$). Then, a screen plot and a robust predicted residual error sum of squares (PRESS) curve are displayed. For the case of TOMCAT toolbox (Daszykowski et al., 2007b), the program generates its own screen plot based on the data matrix and the user can select directly over it the number of components to retain. For our case, we decided to retain $k = 3$ components.

In the previous step it could be perceived that the ROBPCA algorithm was able to operate efficiently with the presence of outliers. Although ROBPCA is not a classification technique, it can reveal latent structures within the data by projecting points on a two-dimensional plot, separating the obsidian sources into groups by using the spectra without any type of calibration. Through the projection of the obsidian samples data in the PCA space, the formation of several compact groups can be perceived (Figure 12). The group labels correspond to the next obsidian sources: Ahuis = Ahuisculco; Espz = La Esperanza; Chay = El Chayal and La Joya; Ixtep = Ixtepeque; SMJ = San Martin Jilotepeque; Ot-Solt = Otumba-Soltepec; Ot-Ixt-Pach-Malp = Otumba-Ixtepec, Otumba-Pacheco, Otumba-Malpaís; Ot-Rinc = Otumba-Rinconada; Oyam = Oyameles; Par = El Paredón; Piz = El Pizarrín, Tulancingo; SP = Sierra de Pachuca; Zac = Zacualtipán; Zin = Zinapécuaro.

The effective discrimination between groups performed by the ROBPCA algorithm, with respect to their spectral information, confirms that the range of the variables selected by the iPLS model contains reliable information about the group membership of each sample. It

also supports the statement that the obsidian flows have distinctive chemical compositions as a result of their specific volcanic properties, which can be measured and detected by spectroscopic techniques such as pXRF. In the case of samples pertaining to the groups labelled as Ot-Solt, Ot-Ixt-Pach-Malp and Ot-Rinc, although they proceed from the same volcanic region (Otumba Complex), certain chemical variability was detected corresponding to three different volcanic flows or sub-sources, as exposed by Argote-Espino et al. (2012). On the other hand, the green obsidian samples from Sierra de Pachuca contain a very homogeneous chemical composition, forming a very compact group.

Concerning the identified outliers, sample 109 had a different preparation than the rest of the analyzed samples, corresponding to a pressed-powder pellet instead of a raw obsidian bulk. This suggests that the presentation of the sample does affect the measurement; therefore the need to analyze all the samples prepared the same way. When the outliers were excluded from the analysis and the model was re-adjusted, the estimations turned out to have a better fit with lower RMSECV values, resulting in a model with a better performance. A fundamental step in archaeometric studies is to relate archaeological artifacts to their geological sources. Therefore, in the next sub-section we performed a pilot study to test the efficiency of this methodology to data of archaeological interest by analyzing set of 257 obsidian artifacts recovered through systematic excavations in the archaeological site of Xalasco, Tlaxcala, Mexico. These data were then compared with our sample population of known origin (obsidian sources).

4.1 Archaeological pilot study

The archaeological site of Xalasco is located within the municipality of Atlitzayanca towards the northeastern portion of the state of Tlaxcala, Mexico. Xalasco was inhabited during the Classic period (AD 100 - 650) by local populations and non-locals proceeding from Teotihuacan, the Mexican Gulf Coast and southern Puebla. It was the seat of multiethnic groups of elite, artisans or emissaries of Teotihuacan or people who shared Teotihuacan cultural heritage. The archaeological materials found in the area showed a constant and dynamic interaction between different areas to the east of the Mexican plateau, so we can surmise that this settlement, along with other sites in the Tlaxcala valley, had an important role in the sustainable development of the great city of Teotihuacan by allowing the passage of resources that came from the Gulf as well as from southeastern Mexico and the south of Puebla. Therefore, Xalasco was a locality for the exchange and transport of raw materials and products between these regions (Manzanilla et al., 2011).

During archaeological excavations at Xalasco in 2015, a set of 257 obsidian artifacts was recovered. Identification of the artifact sources will allow hypotheses to be proposed regarding the use of resources from Teotihuacan and its control over Xalasco. These artifacts were analyzed and processed the same way as the samples from the obsidian sources. Only the data of 6 of the natural deposits were considered in the classification of these samples with RSIMCA algorithm, excluding the data from the sources in the Mayan area and Jalisco, since the materials come from an archaeological site in Central Mexico. The six sources included were Otumba, Oyameles, Paredón, Sierra de Pachuca, Zacualtipán and Zinapécuaro. Figure 13 shows the diagnostic diagrams of the data related to each one of the six sources' groups, constructed with RSIMCA. In table 2, we present the optimal

number of k components and the percentage of explained variance for each one of the groups.

First, we can see that the PCA model seems to be adequate for all the groups. In each one of the groups, some observations are above the cut-off values. For example, points 8 and 6 of group 1, points 3 and 5 for group 2, and so on for the next groups. The good leverage points lie close to the PCA space, but far from the regular observations. This outlier-type can actually stabilize the PCA fit (Varmuza and Filzmoser, 2009). The orthogonal outliers measure the distance between an observation x_i and its projection $\hat{x}_i$ in the k -dimensional PCA subspace (Rousseeuw et al., 2006); this might indicate a higher concentration of one element in that sample. Bad leverage points are not observed; only this kind of observations has an undue effect on the parameter estimates of the PCA. Therefore, it can be concluded that the boundaries of the groups obtained by the RSIMCA algorithm successfully enclose the regular data points in their respective groups. For the prediction of a new data set, the robust distance to the center of its group is computed. If this distance exceeds the cut-off value, the observation can be considered as an atypical value. In our case, none of the observations have a distance greater than the cut-off values, which means that there are no outliers among the samples classified in the groups.

As part of the classification procedure, the percentages of misclassified data points must be calculated. The validation of the classification rule was done processing the complete dataset with PCA; afterwards, a logistic regression model was used to predict the probabilities of the different possible outcomes of a categorical distribution with more than two levels as the dependent variable. The data were randomly divided into a training set

(80%) and a testing set (20%). The confusion matrices are in Table 3, showing that the training set of all the observations are correctly classified. The misclassification error of the training data was 0 and the misclassification error of the testing data was 0.033. In the validation set are some single misclassified samples in group 2, 3 and 5, and 3 misclassified samples in group 6. The results evidence that the RSIMCA approach offers satisfactory efficiency due to its robustness, which guarantees an adequate classification of new objects even when the set of models (used to derive the classification rules) is greatly influenced by distant objects.

RSIMCA assigned the artifacts to only three known sources: Oyameles ($n = 183$), Sierra de Pachuca ($n = 64$) and Zinapécuaro ($n = 10$). While the second source is strongly associated with Teotihuacan control, the first one had a more local use. Of the total of artifacts recovered from Xalasco, 71.2% correspond to raw materials from obsidian sources that were nearer to the site and were commonly distributed along the Puebla-Tlaxcala region (Oyameles), while only a 24.9% come from a source controlled by Teotihuacan, the Sierra de Pachuca). 3.9% of the artifacts from the Zinapécuaro source might also be part of the raw materials distributed by Teotihuacan, since this metropolis had close cultural and commercial ties with this region (Gómez and Gazzola, 2007). The predominance of regional deposits over sources dominated by Teotihuacan lead us to think that Xalasco was not necessarily an enclave of this metropolis and that a more balanced interaction between Teotihuacan state and the local inhabitants occurred. Therefore, the hypothesis of the absolute control of Teotihuacan over Xalasco can be discarded, at least as far as the lithic materials are concerned.

5. CONCLUSIONS

In this study we were able to demonstrate that it is possible to determine the provenance of obsidian samples by using the raw spectra obtained with a portable XRF spectrometer and a chemometric approach without calibrating reference materials. A careful selection of the pre-processing method allows eliminating undesired variations in the spectra and improving the quality of the data. This was achieved through the use of the CluPA algorithm for raw spectral alignment and the EMSC algorithm in combination with the SG smoothing filter, which are ideal for eliminating scattering effects. Using variable selection methods, we were able to choose a spectral sub-set of 92 independent variables from a total of 1950 original variables which compose the full set of spectra. With this procedure, only the relevant information was retained and the redundant information was eliminated.

The two methods applied for the selection of the most informative variables were the iPLS and the siPLS, which are based on the adjustment of PLS local models. The optimal model was selected through a cross-validation, choosing the one having the lowest RMSECV value and the highest correlation coefficient of calibration. The prediction accuracy of the different PLS methods and their different interval partition number were compared, with similar results in both cases. For the case of the iPLS model, only one spectral region was obtained, while the siPLS model found two spectral regions. These algorithms, used for the selection of the intervals, allowed to construct regression models with the lowest errors compared with the PLS model of the full-range spectra. With the outcomes of both models, a ROBPCA algorithm for high dimensional data was then applied. The associated

diagnostic plot was a useful graphical tool that allowed us to visualize and classify the outliers.

Once the outliers were recognized and removed, the variable selection method with the iPLS model was run again in order to determine a more suitable range of informative variables. This resulted in an improvement in the calibration of the models, although the spectral intervals remained the same. Afterwards, using these spectral regions, the ROBPCA algorithm was applied to project the observations into a low-dimensional PCA subspace. By this means, we were able to attain a precise classification of the obsidian samples and to correlate them with their source. This corroborates that the spectral regions or intervals selected by the iPLS and siPLS models contain relevant information concerning the membership of the samples, which makes it possible to identify the differences between the chemical patterns of the obsidian sources for classification purposes.

The classification of unknown samples was carried out using a robust method known as RSIMCA, based on a robust Principal Component Analysis method for high-dimensional data (ROBPCA). To reduce the dimensionality of each group, a PCA analysis was performed. This algorithm employs outlier maps, which are helpful graphical tools to gain insight to the data and help to recognize outliers. Then a robust classification rule was developed to determine the assignment of new observations. The classification rule is based on robust covariance matrices and a linear combination of scaled orthogonal and scaled score distances. In order to compare the classification accuracies, we also divided the dataset into the training and the testing sets. Then, we calculated the classification rule based on the training set and evaluated the estimated class membership of the observations

in the testing set by projecting them onto the directions calculated from the training set and classifying them using the rule obtained from the training set.

The results obtained in this work demonstrate that the approach proposed here is a valid technique for the quantitative analysis of obsidian samples, as well as other types of archaeological and geological materials.

ACKNOWLEDGMENTS

Special thanks to RAISA (Radiación Aplicada a la Industria S.A. de C.V.), Armando Macías and Enrique Franco, for providing us with the XRF spectrometer. We thank the reviewers for their assertive comments on how to clarify and improve the article. We would also like to thank deeply to Gustavo Tolson for his valuable comments and the language revision of this article.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Afseth, N.K., Kohler, A., 2012. Extended multiplicative signal correction in vibrational spectroscopy, a tutorial. *Chemometrics and Intelligent Laboratory Systems* 117, 92-99.

Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Chapman & Hall, London.

Atkinson, P.M., 2009. Spatial data handling and GIS, in: Atkinson, P.M. (Ed.), *Geoinformatics*, vol. 2. EOLSS publishers, Oxford, UK, pp. 65-81.

Araujo, P., 2009. Key aspects of analytical method validation and linearity evaluation. *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences* 877(23), 2224-2234.

Argote Espino, D., Solé, J., López García, P., Sterpone, O., 2010. Análisis composicional de seis yacimientos de obsidiana del centro de México y su clasificación con DBSCAN. *Arqueología* 43, 197-215.

Argote-Espino, D., Solé, J., López-García, P., Sterpone, O., 2012. Obsidian subsource identification in the Sierra de Pachuca and Otumba volcanic regions, Central Mexico, by ICP-MS and DBSCAN Statistical Analysis. *Geoarchaeology* 27, 48-62.

Ballabio, D., 2006. Chemometric characterisation of physical-chemical fingerprints of food products. PhD unpublished dissertation, University of Milano.

Barwick, V., 2003. *Preparation of calibration curves: a guide to best practice*. LGC, National Measurement System, Valid Analytical Measurement Program, UK.

- 1044 Baxter, M.J., 1999. Detecting multivariate outliers in artefact compositional data.
1045 Archaeometry 41 (2), 321-338.
1046
- 1047 Baxter, M. J., 2008. Statistical modelling of artefact compositional data. Archaeometry 43
1048 (1), 131-147.
1049
- 1050 Baxter, M.J., Cool, H.E.M., 2010. Correspondence analysis in R for archaeologist: an
1051 educational account. Archeologia e Calcolatori 21, 211-228.
1052
- 1053 Baxter, M.J., Freestone, I. 2006. Log-ratio compositional data analysis in
1054 archaeometry. Archaeometry 48, 511-531.
1055
- 1056 Beirnaert, C., Meysman, P., Vu, T.N., Hermans, N., Apers, S., Pieters, L., Covaci, A.,
1057 Laukens, K. 2018. speaq 2.0: A complete workflow for high-throughput 1D NMR spectra
1058 processing and quantification. PLOS Computational Biology 14 (3), e1006018.
1059
- 1060 Ben-Gal, I., 2005. Outlier detection, in: Maimon, O., Rockach, L. (Eds.), Data mining and
1061 knowledge discovery handbook: A complete guide for practitioners and researchers.
1062 Springer, New York, pp. 131-146.
1063
- 1064 Bishop, R.L., Canouts, V., Crown, P.L., De Atley, S.P. 1990. Sensitivity, precision, and
1065 accuracy: their roles in ceramic compositional data bases. American Antiquity 55, 537-546.
1066

- 1067 Boksenbaum, M.W., Tolstoy, P., Harbottle, G., Kimberlin, J., Neivens, M., 1987. Obsidian
1068 Industries and Cultural Evolution in the Basin of Mexico before 500 B.C. *Journal of Field*
1069 *Archaeology* 14, 65-75.
- 1070
- 1071 Brown, D.O., Dreiss, M.L., Hughes, R.E., 2004. Preclassic obsidian procurement and
1072 utilization at the maya site of Colha, Belize. *Latin American Antiquity* 15 (2), 222-240.
- 1073
- 1074 Carballo, D.M., Carballo, J., Neff, H., 2007. Formative and classic period obsidian
1075 procurement in Central Mexico: A compositional study using laser ablation-inductively
1076 coupled plasma-mass spectrometry. *Latin American Antiquity* 18, 27-43.
- 1077
- 1078 Carr, S.P., 2015. Geochemical characterization of obsidian subsources in Highland
1079 Guatemala. Unpublished dissertation, Pennsylvania State University, USA.
- 1080
- 1081 Carter, T., Poupeau, G., Bressy, C., Pearce, N.J.G., 2006. A new program of obsidian
1082 characterization at Catalhöyük, Turkey. *Journal of Archaeological Science* 33, 893-909.
- 1083
- 1084 Charlton, T.H., Spence, M.W., 1982. Obsidian exploitation and civilization in the Basin of
1085 Mexico. *Anthropology* 6, 7-86.
- 1086
- 1087 Charlton, T.H., Grove, D.C., Hopke, P.K., 1978. The Paredón, Mexico, obsidian source and
1088 Early Formative exchange. *Science* 201, 807-9.
- 1089

Clark, J., 1989. The distribution of obsidian, in: Voorheis, V. (Ed.), Ancient trade and tribute: Economies of the Soconusco region of Mesoamerica. University of Utah Press, Salt Lake City, pp. 268–284

Cobean, R.H., 2002. A world of obsidian: The mining and trade of a volcanic glass in ancient Mexico. INAH and Pittsburgh University, Mexico.

Croux, C., Filzmoser, P., Oliveira, M., 2007. Algorithms for Projection-Pursuit Robust Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems 87, 218–225.

Darras, V., 2012. Development of pressure blade technology in North-Central and Western Mexico, in: Desrosiers, P.M. (Ed.), The emergence of pressure blade making: from origin to modern experimentation. Springer, New York, pp. 417–462.

Daszykowski, M., Kaczmarek, K., Stanimirova, I., Vander Heyden, Y., Walczak, B., 2007a. Robust SIMCA-bounding influence of outliers. Chemometrics and Intelligent Laboratory Systems 87(1), 95–103.

Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., Walczak, B., 2007b. TOMCAT: A MATLAB toolbox for multivariate calibration techniques. Chemometrics and Intelligent Laboratory Systems 85 (2), 269–277.

- 1113 Doran, J., Hodson, F.R., 1975. Mathematics and computers in Archaeology. Harvard
1114 University Press, Cambridge.
- 1115
- 1116 dos Santos Panero, P., dos Santos Panero, F., dos Santos Panero, J., Bezerra da Silva, H.E.,
1117 2013. Application of Extended Multiplicative Signal Correction to short-wavelength near
1118 infrared spectra of moisture in marzipan. Journal of Data Analysis and Information
1119 Processing 1, 30-34.
- 1120
- 1121 Drake, B.L., Nazaroff, A.J., Prufer, K.M., 2009. Error assessment of portable X-ray
1122 fluorescence spectrometry in geochemical sourcing. SAS Bulletin 32 (3), 14-17.
- 1123
- 1124 Evans, S.T., Webster, D.L., 2001. Archaeology of ancient Mexico and Central America: an
1125 encyclopedia. Garland Publishing, New York.
- 1126
- 1127 Ferrao, M.F., Viera, M.D.S., Panta Pazos, R.E., Fachini, D., Gerbase, A.E., Marder, L.,
1128 2011. Simultaneous determination of quality parameters of biodiesel/diesel blends using
1129 HATR-FTIR spectra and PLS, iPLS or siPLS regressions. Fuel 90, 701–706.
- 1130
- 1131 Filzmoser, P. and Todorov, V., 2011. Review of robust multivariate statistical methods in
1132 high dimension. Analytica Chimica Acta 705 (1-2), 2-14
- 1133
- 1134 Finch, W.H., Bolin, J.H., Kelley, K., 2014. Group membership prediction when known
1135 groups consist of unknown subgroups: a Monte Carlo comparison of methods. Frontiers in
1136 Psychology 5, 337.

1137

1138 Frahm, E., 2014. Characterizing obsidian sources with portable XRF: accuracy
1139 reproducibility, and field relationships in a case study from Armenia. *Journal of*
1140 *Archaeological Science* 49, 105-125.

1141

1142 Gao, H., Mao, H., Zhang, X., 2015. Determination of lettuce nitrogen content using
1143 spectroscopy with efficient wavelength selection and extreme learning machine.
1144 *Zemdirbyste-Agriculture* 102 (1), 51–58.

1145

1146 García-Palomo, A., Macías, J. L., Tolson, G., Valdez, G., Mora, J. C., 2002. Volcanic
1147 stratigraphy and geological evolution of the Apan region, east-central sector of the Trans-
1148 Mexican Volcanic Belt. *Geofísica Internacional* 41, 133–150.

1149

1150 Gaxiola, M., 1989. Prólogo, in: Gaxiola, M., Clark, J.E. (Eds.), *La obsidiana en*
1151 *Mesoamérica*. INAH, México, pp. 7-12.

1152

1153 Gaxiola, M., Guevara, J., 1989. Un conjunto habitacional en Huapalcalco, Hgo,
1154 especializado en la talla de obsidiana, in: Gaxiola, M., Clark, J.E. (Eds.), *La obsidiana en*
1155 *Mesoamérica*. INAH, México, pp. 227-242.

1156

1157 Glascock, M.D., 1992. Characterization of archaeological ceramics at MURR by neutron
1158 activation analysis and multivariate statistics, in: Neff, H. (Ed.), *Chemical characterization*
1159 *of ceramic pastes in Archaeology*. Prehistory Press, Madison, Wisconsin, pp. 11-26.

1160

- 1161 Glascock, M. D., Ferguson, J. R., 2012. Report on the analysis of obsidian source samples
1162 by multiple analytical methods. Archaeometry Lab, Research Reactor Center, University of
1163 Missouri, Columbia. [https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/X-](https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/X-rayDiffraction_ElementalAnalysis/HH-XRF/LabReports/Obsidian_Source_Samples.pdf)
1164 [rayDiffraction_ElementalAnalysis/HH-XRF/LabReports/Obsidian_Source_Samples.pdf](https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/X-rayDiffraction_ElementalAnalysis/HH-XRF/LabReports/Obsidian_Source_Samples.pdf)
1165 (accessed 15 September 2017).
1166
- 1167 Glascock, M.D., Speakman, R.J., Neff, H., 2007. Archaeometry at the University of
1168 Missouri research reactor and the provenance of obsidian artifacts in North America.
1169 Archaeometry 49, 343–357.
1170
- 1171 Glascock, M.D., Weigand, P.C., Esparza, R., Ohnerson, M.A., Garduño, M., Mountjoy,
1172 J.B., Darling, J.A., 2010. Geochemical characterisation of obsidian in Western Mexico: the
1173 sources in Jalisco, Nayarit, and Zacatecas, in: Kuzmin, Y.V., Glascock, M.D. (Eds.),
1174 Crossing the straits: prehistoric obsidian source exploitation in the North Pacific Rim.
1175 Archaeopress, Oxford, pp. 201-217.
1176
- 1177 Golitko, M., Dudgeon, J., Neff, H., Terrell, J., 2012. Identification of post-depositional
1178 chemical alteration of ceramics from the north coast of Papua New Guinea (Sanduan
1179 province) by time-of-flight-laser ablation–inductively coupled plasma–mass spectrometry
1180 (TOF–LA–ICP–MS). Archaeometry 54 (1), 80-100.
1181
- 1182 Gómez, S., Gazzola, J., 2007. Análisis de las relaciones entre Teotihuacan y el occidente de
1183 México, in: Faugère-Kalfon, B. (Coord.), Dinámicas culturales entre el occidente, el centro-

norte y la cuenca de México, del preclásico al epiclásico. COLMICH and CEMCA, México, pp. 113-135.

Grunsky, E.C., 2009. Statistical analysis in the Geosciences, in: Atkinson, P.M. (Ed.), Geoinformatics, vol. 2. EOLSS publishers, Oxford, UK, pp. 1-64.

Gullayanon, R. (2011) A calibration methodology for Energy Dispersive X-Ray Fluorescence measurements based upon synthetically generated reference spectra. PhD Thesis, Georgia Institute of Technology, Atlanta.

Hall, M.E., Moreva, O.L., Popov, A.N., Kusmin, Y.V., 1999. Chemical analyses of pottery from the Boisman 2 site, Primorye (Russian Far East). Northeast Asian Studies 4, 247-262.

Hawkins, D., 1980. Identification of outliers. Chapman and Hall, London.

Healan, D.M., 1997. Pre-Hispanic quarrying in the Ucareo-Zinapécuaro obsidian source area. Ancient Mesoamerica 8, 77-100.

Hernández, I., 2007. Geología y geomorfología volcánica de la región de los yacimientos de obsidiana de Otumba en el sector Norte de la Sierra Nevada de México. Unpublished dissertation, Facultad de Filosofía y Letras, UNAM, Mexico.

- 1206 Hong-Dong, L. Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using
1207 competitive adaptive reweighted sampling method for multivariate calibration. *Analytica*
1208 *Chimica Acta* 648 (1), 77-84.
- 1209
- 1210 Hron, K., Templ, M., Filzmoser, P., 2010. Imputation of missing values for compositional
1211 data using classical and robust methods. *Computational Statistics and Data Analysis*
1212 54(12), 3095–3107.
- 1213
- 1214 Hubert, M., Rousseeuw, P., 2011. Multivariate techniques: Robustness, in: Lovric, M.
1215 (Ed.), *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg, pp.
1216 932-935.
- 1217
- 1218 Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: A new approach to
1219 robust Principal Component Analysis. *Technometrics* 47, 64–79.
- 1220
- 1221 Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T., 2016. Sparse PCA for high-
1222 dimensional data with outliers. *Technometrics* 58 (4), 424-434.
- 1223
- 1224 Hunt, A.M.W., Speakman, R.J., 2015. Portable XRF analysis of archaeological sediments
1225 and ceramics. *Journal of Archaeological Science* 53, 1-13.
- 1226
- 1227 Jajuga, K., Walesiak, M., 2000. Standardisation of data set under different measurement
1228 scales, in: Decker, R., Gaul, W. (Eds.), *Classification and information processing at the turn*

of the millennium: studies in classification, data analysis, and knowledge organization.

Springer, Berlin, Heidelberg.

Janssens, K., Deraedt, I., Freddy, A., Veeckman, J., 1998. Composition of 15-17th Century Archaeological Glass Vessels Excavated in Antwerp, Belgium, in: Love, G., Nicholson, W.A.P., Armigliato, A. (Eds.), Modern developments and applications in microbeam analysis. *Mikrochimica Acta Supplement*, vol 15. Springer, Vienna, pp. 253–267.

Kaiser, B., Shugar, A., 2012. Glass analysis utilizing handheld X-ray fluorescence, in: Shugar, A., Mass, J. (Eds.), *Studies in archaeological sciences*. Leuven University Press, Belgium, pp. 449-470.

Kilikoglou, V., Grimanis, A.P., Tsolakidou, A., Hein, A., Malamidou, D., Tsirtsoni, Z., 2007. Neutron activation patterning of archaeological materials at the National Center for Scientific Research ‘Demokritos’: the case of black-on-red neolithic pottery from Macedonia, Greece. *Archaeometry* 49 (2), 301-319.

Kovarovic, K., Aiello, L.C., Cardini, A., Lockwood, C.A., 2011. Discriminant function analyses in archaeology: are classification rates too good to be true? *Journal of Archaeological Science* 38 (11), 3006-3018.

La Tour, T., 1989. Analysis of rocks using x-ray fluorescence spectrometry. *The Rigaku Journal* 6 (1), 3-9.

- Lazim, S.S.R.M., Nawi, N.M., Chen, G., Jensen, T., Rasli, A.M.M., 2016. Influence of different pre-processing methods in predicting sugarcane quality from near-infrared (NIR) spectral data. *International Food Research Journal* 23 (Suppl), S231-S236.
- Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics* 14, 643-655.
- Leardi, R., Nørgaard, L., 2004. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *Journal of Chemometrics* 18 (11), 486-497.
- Lemberge, P., De Raedt, I., Janssens, K.H., Wei, F., Van Espen, P.J., 2000. Quantitative Z-analysis of the 16th–17th century archaeological glass vessels using PLS regression of EPXMA and μ -XRF data. *Journal of Chemometrics* 14, 751–763.
- Lin, W-S., Cheng, S-C., Lee, C-P., Hong, M-C., Kuo, B-J., 2015. Using orthogonal signal correction to correct the Near Infrared Reflectance spectroscopy for determination of protein content in brown rice. *Crop, Environment & Bioinformatics* 12, 42-56.
- Liland, K.H., 2017. Extended Multiplicative Signal Correction. <https://cran.r-project.org/web/packages/EMSC/EMSC.pdf> (accessed 22 August 2016).
- Liland, K.H., Kohler, A., Afseth, N.K., 2016. Model-based pre-processing in Raman spectroscopy of biological samples. *Journal of Raman Spectroscopy* 47 (6), 643-650.

1278

1279 Little, N.C., Kosakowsky, L.J., Speakman, R.J., Glascock, M.D., Lohse, J.C., 2004.
1280 Characterization of Maya pottery by INAA and ICP-MS. *Journal of Radioanalytical and*
1281 *Nuclear Chemistry* 262 (1), 103-110.

1282

1283 Liu, H., Shah, S., Jiang, W., 2004. On-line outlier detection and data cleaning. *Computers*
1284 *and Chemical Engineering* 28, 1635–1647.

1285

1286 López, F., Nieto, R., Cobean, R., 1989. La producción de obsidiana en la Sierra Las
1287 Navajas, Hgo., in: Gaxiola, M., Clark, J.E. (Eds.), *La obsidiana en Mesoamérica*. INAH,
1288 Mexico, pp. 193–197.

1289

1290 Mandujano, C., Elizalde, S., Cassiano, G., Soto, D., Tenorio, D., Jiménez-Reyes, M., 2002.
1291 Provenance and use wears of pre-Hispanic obsidian scrapers from Metztitlan, Hidalgo,
1292 Mexico. *Journal of Radioanalytical and Nuclear Chemistry* 252 (1), 81-88.

1293

1294 Manzanilla, L.R., Aguayo, E., Hernández, M., Bautista, A., 2011. Informe técnico final del
1295 salvamento de un conjunto con materiales teotihuacanos en Xalasco, Atltzayanca, Tlaxcala
1296 2008. Unpublished report, Archivo Técnico-INAH, Mexico.

1297

1298 Martelo, M.J., Vázquez, V., 2014. Classification of red wines from controlled designation of
1299 origin by ultraviolet-visible and near-infrared spectral analysis. *Ciência e Técnica*
1300 *Vitivinícola* 29(1), 35-43.

1301

1302 Millhauser, J.K., Rodríguez-Alegría, E., Glascock, M.D., 2011. Testing the accuracy of
1303 portable X-ray fluorescence to study Aztec and Colonial obsidian supply at Xaltocan,
1304 Mexico. *Journal of Archaeological Science* 38, 3141-3152.

1305
1306 Millhauser, J.K., Fargher, L.F., Espinoza, V.Y.H., Blanton, R.E., 2015. The geopolitics of
1307 obsidian supply in Postclassic Tlaxcallan: a portable X-ray fluorescence study. *Journal of*
1308 *Archaeological Science* 58, 133-46.

1309
1310 Millhauser, J.K., Bloch, L., Golitko, M., Fargher, L.F., Xiuhtecutli, N., Heredia, V.Y.,
1311 Glascock, M.D., 2017. Geochemical variability in the Paredón obsidian source, Puebla and
1312 Hidalgo, Mexico: a preliminary assessment and inter-laboratory comparison.
1313 *Archaeometry*, online. DOI: 10.1111/arc.12330

1314
1315 Mooser, F., 1968. Geología, naturaleza y desarrollo del Valle de Teotihuacan, in: Lorenzo,
1316 J.L. (Ed.), *Materiales para la arqueología de Teotihuacan*. INAH, Mexico, pp. 29–37

1317
1318 Mora, R., 1981. Yacimientos y talleres de obsidiana en el sureste del estado de Hidalgo, in:
1319 Gaxiola, M. (Ed.), *Yacimientos y talleres prehispánicos de obsidiana del estado de Hidalgo*.
1320 INAH, Mexico, pp. 21–38.

1321
1322 Nazaroff, A.J., Prufer, K.M., Drake, B.L., 2010. Assessing the applicability of portable X-
1323 ray fluorescence spectrometry for obsidian provenance research in the Maya lowlands.
1324 *Journal of Archaeological Science* 37, 885-895.

- 1326 Neff, H. Glascock, M.D., Charlton, T.H., Charlton, C.O., Nichols, D.L., 2000. Provenience
1327 investigation of ceramics and obsidian from Otumba. *Ancient Mesoamerica* 11 (2), 307-
1328 321.
- 1329
- 1330 Nelson, S.A., Healan, D.M., 1995. Obsidians from the Ucareo and Zinapécuaro areas,
1331 Michoacan, Mexico. *Geological Society of America: Abstract with Programs* 27, 358.
- 1332
- 1333 Nørgaard, L., 2013. *iToolbox manual*. KVL, Denmark.
1334 http://www.models.life.ku.dk/sites/default/files/iToolbox_Manual.pdf (accessed 02 April
1335 2017).
- 1336
- 1337 Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck L. and Engelsen S.B., 2000.
1338 Interval Partial Least Squares Regression (iPLS): A comparative chemometric study with
1339 an example from Near-Infrared Spectroscopy. *Applied Spectroscopy* 54, 413-419.
- 1340
- 1341 Nurunnabi, A., Belton, D., West, G., 2012. Diagnostic-robust statistical analysis for local
1342 surface fitting in 3D point cloud data. Conference paper: *ISPRS Annals of the*
1343 *Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume I-3, XXII
1344 *ISPRS Congress*, Melbourne, Australia.
- 1345
- 1346 Ortner, T., Filzmoser, P., Zaharieva, M., Breiteneder, C., Brodinova, S., 2017. Guided
1347 projections for analysing the structure of high-dimensional data.
1348 <https://arxiv.org/pdf/1702.06790.pdf> (accessed 02 April 2017).
- 1349

Pan, X., Li, Y., Wu, Z., Zhang, Q., Zheng, Z., Shi, X., Qiao, Y., 2015. A online NIR sensor for the pilot-scale extraction process in fructus aurantii coupled with single and ensemble methods. *Sensors* 15, 8749-8763.

Pollard, M.A., Heron, C., 2008. *Archaeological chemistry*. Royal Society of Chemistry, UK.

Ponomarenko, A.L., 2004. The Pachuca obsidian source, Hidalgo, Mexico: A geoarchaeological perspective. *Geoarchaeology* 19, 71–91.

R Development Core Team, 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org> (accessed 24 November 2016)

Romia, M.B., Bernardez, M.A., 2009. Multivariate calibration for quantitative analysis, in: Sun, D.W. (Ed.), *Infrared spectroscopy for food quality analysis and control*. Academic Press, US, pp. 51-82.

Rousseau, R.M., 2001. Concept of the influence coefficient. *The Rigaku Journal* 18 (1), 8-21.

Rousseeuw, P.J., Debruyne, M., Engelen, S., Hubert, M., 2006. Robustness and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* 36 (3-4), 221–242.

- 1374 Rousseeuw, P.J., Van Driessen, K., 1999. A Fast Algorithm for the Minimum Covariance
1375 Determinant Estimator. *Technometrics* 41 (3), 212-223.
1376
- 1377 Rowe, H., Hughes, N., Robinson, K., 2012. The quantification and application of handheld
1378 energy-dispersive X-ray fluorescence (ED-XRF) in mudrock chemostratigraphy and
1379 geochemistry. *Chemical Geology* 324- 325, 122-131.
1380
- 1381 Santos, J.O., Munita, C.S., Valério, M.E.G., Vergne, C., Oliveira, P.M.S., 2006.
1382 Determination of trace elements in archaeological ceramics and application of Kernel
1383 Density Estimates: Implications for the definition of production locations. *Journal of*
1384 *Radioanalytical and Nuclear Chemistry* 269 (2), 441–445.
1385
- 1386 Schafer, R.W., 2011. What *Is* a Savitzky-Golay filter? *IEEE Signal Processing Magazine*
1387 28 (4), 111-117.
1388
- 1389 Scharlotta, I., Quach, T.T., 2015. Provenance analysis of porphyritic volcanic materials in
1390 San Diego using portable X-ray Fluorescence. *Journal of Archeological Science: Reports* 3,
1391 285-294.
1392
- 1393 Seetha, D., Velraj, G., 2016. Characterization and chemometric analysis of ancient pot
1394 shards trenched from Arpakkam, Tamil Nadu, India. *Journal of Applied Research and*
1395 *Technology* 14 (5), 345-353.
1396

Shackley, M.S., 2010. Is there reliability and validity in portable X-ray fluorescence spectrometry (PXRF)? *The SAA Archaeological Record* 10 (5), 17–20.

Shackley, M.S., 2011. An Introduction to X-Ray Fluorescence (XRF) Analysis in Archaeology, in: Shackley, M.S. (Ed.), *X-Ray Fluorescence Spectrometry (XRF) in Geoarchaeology*. Springer-Verlag, New York, pp. 7-44.

Sheets, P., Hirth, K., Lange, F., Stross, F., Asaro, F., Michel, H., 1990. Obsidian sources and elemental analysis of artifacts in southern Mesoamerica and the northern intermediate area. *American Antiquity* 55 (1), 144-158.

Sitko, R., Zawisza, B., 2012. Quantification in X-ray fluorescence spectrometry, in: Sharma, S.K. (Ed.), *X-Ray Spectroscopy*. InTech, Croatia, pp. 137-162.

Smith, M.E., Burke, A., Hare, T.S., Glascock, M.D., 2007. Sources of imported obsidian at postclassic sites in the Yauhtepec valley, Morelos: A characterization using XRF and INAA. *Latin American Antiquity* 18 (4), 429–450.

Sorensen, J., Hirth, K., 1984. Prehistoric mines and obsidian workshops at La Esperanza-Intibuca, Honduras. *Yaxkin* 7, 31-45

Speakman, R.J., 2012. Evaluation of Bruker's Tracer family factory obsidian calibration for handheld portable XRF studies of obsidian. Report prepared for Bruker AXS, Kennewick. https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/X-rayDif

fraction_ElementalAnalysis/HH-XRF/LabReports/Bruker_Obsidian_Report.pdf (accessed 15 September 2017).

Speakman, R.J., Shackley, M.S., 2013. Silo science and portable XRF in archaeology : a response to Frahm. *Journal of Archaeological Science* 40, 1435-1443.

Speakman, R.J., Little, N.C., Creel, D., Miller, M.R., Iñáñez, J.G., 2011. Sourcing ceramics with portable XRF spectrometers? a comparison with INAA using Mimbres Pottery from the American Southwest. *Journal of Archaeological Science* 38, 3483-3496.

Stahl, A.B., das Dores Cruz, M., Neff, H., Glascock, M.D., Speakman, R.J., Giles, B., Smith, L., 2008. Ceramic production, consumption and exchange in the Banda area, Ghana: insights from compositional analyses. *Journal of Anthropological Archaeology* 27 (3), 363-381.

Stanimirova, I., Üstün, B., Cajka, T., Ridelova, K., Hajslova, J., Buydens, L.M.C., Walczak, B., 2010. Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques. *Food Chemistry* 118 (1), 171-176.

Stevens, A., Ramirez-Lopez, L., 2015. Package 'prospectr': Miscellaneous functions for processing and sample selection of vis-NIR diffuse reflectance data. <https://cran.r-project.org/web/packages/prospectr/prospectr.pdf> (accessed 24 April 2017).

- 1444 Suhandy, D., Yulia, M., Ogawa, Y., Kondo, N. 2013. Prediction of L-Ascorbic Acid using
1445 FTIR-ATR Terahertz spectroscopy combined with Interval Partial Least Squares (iPLS)
1446 Regression. *Engineering in Agriculture, Environment and Food* 6 (3), 111-117.
1447
- 1448 Tanasi, D., Tykot, R., Pirone, F., McKendry, E., 2017. Provenance study of prehistoric
1449 ceramics from sicily: a comparative study between pXRF and XRF. *Open Archaeology* 3,
1450 222–234.
1451
- 1452 Todorov, V., Filzmoser, P., 2009. *Multivariate Robust Statistics: Methods and*
1453 *Computation*. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken.
1454
- 1455 Tykot, R.H., 2010. Sourcing of Sardinian obsidian collections in the Museo Preistorico-
1456 Etnografico “Luigi Pigorini” using Non-Destructive XRF. *L’ossidiana del Monte Arci nel*
1457 *Mediterraneo. Nuovi apporti sulla diffusione, sui sistemi di produzione e sulla loro*
1458 *cronologia, Atti del 5° Convegno internazionale (Pau, Italia, 27–29 Giugno 2008). NUR,*
1459 *Ales.*
1460
- 1461 Van den Boogaart, K.G., Tolosana-Delgado, R., 2013. *Analyzing compositional data with*
1462 *R*. Springer, London.
1463
- 1464 Vanden Branden, K., Hubert, M., 2005. Robust classification in high dimensions based on
1465 the SIMCA Method. *Chemometrics and Intelligent Laboratory Systems* 79 (1–2), 10-21.
1466

- 1467 Varmuza, K., Filzmoser, P., 2009. Introduction to multivariate statistical analysis in
1468 chemometrics. CRC Press, Boca Raton.
- 1469
- 1470 Verboven, S., Hubert, M., 2005. LIBRA: a MATLAB Library for Robust Analysis.
1471 Chemometrics and Intelligent Laboratory Systems 75 (2), 127-136.
- 1472
- 1473 Vu, T.N., Valkenborg, D., Smets, K., Verwaest, K.A., Dommissie, R., Lemièrre, F.,
1474 Verschoren, A., Goethals, B., Laukens, K., 2011. An integrated workflow for robust
1475 alignment and simplified quantitative analysis of NMR spectrometry data. BMC
1476 Bioinformatics 12, 405.
- 1477
- 1478 Vu, T.N., Laukens, K., 2013. Getting your peaks in line: a review of alignment methods for
1479 NMR spectral data. Metabolites 3, 259-276.
- 1480
- 1481 Wehrens, R., 2011. Chemometrics with R: Multivariate data analysis in the natural sciences
1482 and life sciences. Springer Verlag, Heidelberg.
- 1483
- 1484 Williams, G.J., Baxter, R.A., He, H.X., Hawkins, S., Gu, L., 2002. A comparative study of
1485 RNN for outlier detection in data mining. Conference paper: IEEE International Conference
1486 on Data-mining (ICDM'02), Maebashi City, Japan, CSIRO Technical Report CMIS-
1487 02/102.
- 1488

Xu, L., Lu, J., Yang, Q., Chen, J., 2014. An objective wavelength selection method based on moving window Partial Least Squares for Near-Infrared Spectroscopy. *Journal of Computers* 9 (1), 228-234.

Xu, J., Huang, F., Li, Y., Chen, Z., Wang, Y., 2015. Rapid detection of total nitrogen content in soy sauce using NIR spectroscopy. *Czech Journal of Food Sciences* 33 (6), 518–522.

Yun Y.H., Li, H.D., Wood, R.E., Fan, W., Wang, J.J., Cao, D.-S., Xu, Q.S., Liang, Y.Z., 2013. An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 111 (7), 31-36.

Zhiming, G., Wenqian, H., Liping, Ch., Yankun, P., Xiu, W., 2014. Shortwave infrared hyperspectral imaging for detection of pH value in Fuji apple. *International Journal of Agricultural and Biological Engineering* 7 (2), 130-137.

APPENDIX**Package “speaq” for R environment**

**Created by Charlie Beirnaert (Department of Mathematics and Computer Science,
University of Antwerp)**

```
# Run the whole script at one time
```

```
# download latest speaq (do this only 1 time)
```

```
devtools::install_github("Beirnaert/speaq")
```

```
library(speaq)
```

```
# change file folder
```

```
mexico = "/Users/Charlie/Downloads/"
```

```
# get the data. (spectra in matrix format)
```

```
matrix3 = read.csv2(paste(mexico, "recorte3.csv", sep = ""), header = F, sep =  
",", colClasses = "numeric", dec = ".")
```

```
spectra.matrix = as.matrix(matrix3)
```

```
index.vector = seq(1, ncol(spectra.matrix))
```

```
# plot the spectra
```

```
speaq::drawSpec(X = spectra.matrix, main = 'mexico', xlab = "index")
```

```
1537 # peak detection
1538 peaks <- speaq::getWaveletPeaks(Y.spec = spectra.matrix,
1539                                X.ppm = index.vector,
1540                                nCPU = 1,
1541                                raw_peakheight = TRUE)
1542
1543 # grouping (the filling step is not necessary in this case as the peaks are well formed and the
1544 peak detection threshold is set low.)
1545 groups<- PeakGrouper(Y.peaks = peaks)
1546
1547 peakfill <- PeakFilling(groups_rawIntensity,
1548                        spectra.matrix,
1549                        max.index.shift = 5,
1550                        window.width = "small",
1551                        nCPU = 1)
1552
1553 Features <- BuildFeatureMatrix(Y.data = peakfill,
1554                                var = "peakValue",
1555                                impute = "zero",
1556                                delete.below.threshold = FALSE,
1557                                baselineThresh = 1,
1558                                snrThres = 0)
1559
1560 # Aligning the raw spectra
```

```
1561
1562 peakList = list()
1563 for(s in 1:length(unique(peaks$Sample))){
1564     peakList[[s]] = peaks$peakIndex[peaks$Sample == s]
1565 }
1566
1567 resFindRef<- findRef(peakList);
1568 refInd <- resFindRef$refInd;
1569
1570
1571 Aligned.spectra <- dohCluster(spectra.matrix,
1572                             peakList = peakList,
1573                             refInd = refInd,
1574                             maxShift = 5 ,
1575                             acceptLostPeak = TRUE,
1576                             verbose=TRUE);
1577
1578
1579 drawSpec(Aligned.spectra)
1580 write.csv (Aligned.spectra, file =" aligned.csv")
1581
1582
1583
1584
```

FIGURE AND TABLE CAPTIONS

FIG. 1 Location of the analyzed Mesoamerican obsidian sources: (1) Ahuisculco, Jalisco; (2) Zinapécuaro, Michoacán; (3) Zacualtipán, Hidalgo; (4) Sierra de Pachuca, Hidalgo; (5) El Pizarrín-Tulancingo, Hidalgo; (6) Otumba volcanic complex, Edo. México; (7) El Paredón, Puebla; (8) Oyameles, Puebla; (9) San Martín Jilotepeque, Guatemala; (10) El Chayal, Guatemala; (11) Ixtepeque, Guatemala; (12) La Esperanza, Honduras.

FIG. 2 (a) Example of raw spectra with shifted peaks; (b) Spectra aligned using the CluPA algorithm.

FIG. 3 Raw XRF spectra of the 109 analyzed obsidian samples.

FIG. 4 Spectra after the application of the SG filter and the sixth order polynomial EMSC model. The results clearly demonstrate that the pre-processing improves the correction of scattering effect.

FIG. 5 Number of intervals optimized according to the root mean square error of cross validation (RMSECV) for the iPLS model.

FIG. 6 RMSECV plot obtained from the iPLS. The vertical bars are the RMSECV for each local model (20 interval model). The dotted line is the RMSECV for the global model (full-spectrum PLS model) versus interval model for one to 4 LV's. Italic numbers are the optimal LVs in each interval model.

FIG. 7 iPLS local calibration model of interval number 7 (correspondent to Sr) with 2 PLS components for y-variable number 1.

FIG. 8 Standardized robust distance vs standardized orthogonal distance plot obtained with the robust PCA for the obsidian samples data set.

FIG. 9 Calibration and validation results for Sr determination in the iPLS regression without the presence of outliers.

FIG. 10 Selected spectral intervals [4 6] for the siPLS model.

FIG. 11 Scatter plot showing the correlation between reference, measured and predicted spectral data in the calibration set by the siPLS model.

FIG. 12 ROBPCA projection of the obsidian samples data. Outliers (samples no. 17 and 109) are marked with boxes. The group labels correspond to the next obsidian sources: Ahuis = Ahuisculco; Espz = La Esperanza; Chay = El Chayal and La Joya; Ixtep = Ixtepeque; SMJ = San Martin Jilotepeque; Ot-Solt = Otumba-Soltepec; Ot-Ixt-Pach-Malp = Otumba-Ixtepec, Otumba-Pacheco, Otumba-Malpaís; Ot-Rinc = Otumba-Rinconada; Oyam = Oyameles; Par = El Paredón; Piz = El Pizarrín, Tulancingo; SP = Sierra de Pachuca; Zac = Zacualtipán; Zin = Zinapécuaro.

FIG. 13 Diagnostic diagrams built with RSMCA for each one of the sources incorporated to the classification of the obsidian artifacts recovered from the archaeological site of Xalasco: (1) Otumba, (2) Oyameles, (3) Paredón, (4) Sierra de Pachuca, (5) Zacualtipan, (6) Zinapécuaro.

TABLE 1 Number of samples analyzed for each obsidian source.

TABLE 2 Optimal number of k components and percentage of explained variance of each one of the groups: (1) Otumba, (2) Oyameles, (3) Paredón, (4) Sierra de Pachuca, (5) Zacualtipan, (6) Zinapécuaro.

TABLE 3 Confusion matrix and misclassification error of the training data and confusion matrix and misclassification error of the testing data.

TABLE S.1 Spectral data matrix obtained from the XRF analysis of the obsidian sources samples. The rows refer to the obsidian samples and each column is associated with a channel or energy region of the spectrum.

TABLE S.2 Chemical concentrations obtained from the analyzed samples of the geological sources.

Sample No.	Source name	Location	Geographic region
1-10	Ahuisculco		Jalisco
11-13	La Esperanza		Honduras
14-16	El Chayal		Guatemala
17-21	Ixtepeque		Guatemala
22-24	El Chayal	La Joya	Guatemala
25-27	San Martin Jilotepeque		Guatemala
28-37	Otumba	Soltepec	Edo. de México
38-46	Otumba	Ixtepec	Edo. de México
47-54	Otumba	Malpais	Edo. de México
55-60	Otumba	Cerro Pacheco	Edo. de México
61-62	Otumba	Rinconada gully	Edo. de México
63-69	Oyameles		Puebla
70-76	El Paredón	Tres Cabezas	Puebla
77-81	Tulancingo	El Pizarrín	Hidalgo
82-83	Sierra de Pachuca	Cembo	Hidalgo
84-85	Sierra de Pachuca	Cruz del Milagro	Hidalgo
86-88	Sierra de Pachuca	El Durazno	Hidalgo
89-91	Sierra de Pachuca	Rancho Oyamental	Hidalgo
92-101	Zacualtipan		Hidalgo
102-109	Zinapécuaro		Michoacán

	<i>k</i>-components	% Explained Variance
Group 1	$k_1 = 5$	93.05
Group 2	$k_2 = 3$	76.38
Group 3	$k_3 = 3$	83.27
Group 4	$k_4 = 5$	94.07
Group 5	$k_5 = 5$	95.99
Group 6	$k_6 = 3$	89.27

Training data

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Group 1	10	0	0	0	0	0	10
Group 2	0	155	0	0	0	0	155
Group 3	0	0	6	0	0	0	6
Group 4	0	0	0	54	0	0	54
Group 5	0	0	0	0	7	0	7
Group 6	0	0	0	0	0	16	16
Total	10	155	6	0	7	16	248

Testing data

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Total
Group 1	0	0	0	0	0	0	0
Group 2	0	34	1	0	0	0	35
Group 3	0	0	0	0	0	0	0
Group 4	0	0	0	0	0	0	0
Group 5	0	1	0	0	20	3	24
Group 6	0	0	0	0	0	1	1
Total	0	35	1	0	20	4	60

Selected intervals [4 6], with 6 PLS comp. for y-var. no.1

Interval no. 7, wavelengths 583-679, with 4 PLS comp. for y-var. no.1