MACHINE LEARNING ALGORITHMS APPLIED TO RAMAN SPECTRA FOR THE IDENTIFICATION OF VARISCITE ORIGINATING FROM THE MINING COMPLEX OF GAVÀ

Díez-Pastor J.F.[1], Jorge-Villar S.E.[2,3], Arnaiz-González Á. [1], García-Osorio C.I. [1], Díaz-Acha Y.[4], Campeny M. [4,5], Bosch J. [6], Melgarejo J.C.[5]

[1]Departament of Civil Engineering, Universitad de Burgos, Av. Cantabria, s/n, 09006, Spain
[2]Facultad de Educación, Universidad de Burgos, C/ Villadiego, 1, 09001 Burgos, Spain,
[3]CENIEH, Paseo Sierra de Atapuerca 3, 09002, Spain,
[4] Museu de Ciències Naturals de Barcelona, Edifici Martorell, Passeig Picasso, s/n, 08003 Barcelona,
[5] Departament de Mineralogia, Petrologia i Geologia Aplicada, Universitat de Barcelona, c/Martí i Franquès s/n, 08028 Barcelona,
[6] Museu de Gavà, Plaça de Dolors Clua, 13-14, 08850 Gavà, Barcelona, Spain

E-mail: susanajorgevillar@hotmail.com; seju@ubu.es

ABSTRACT

Variscite is an aluminium phosphate mineral widely used as a gemstone in antiquity. Knowledge of the ancient trade in variscite has important implications on the historical appreciation of the commercial and migratory movements of human population. The mining complex of Gavà, which dates from the Neolithic, is one of the oldest underground mine sites in Europe, from where variscite was extracted from several mines and at different depths, providing minerals with different properties and a range of colours. In this work, Machine Learning algorithms have been used to classify variscite samples from Gavà with regard to the identification of their mine of origin and extraction depth. The final objective of the study was to see if the Raman spectroscopic signatures selected by these algorithms had a key spectral significance related to mineral structure and/or composition and validating the use of these computational procedures as a useful tool for detecting variances in the mineral Raman spectra that could facilitate the assignment of the specimens to each mine.

Keywords: Archaeometry, Mineral classification, Raman spectroscopy, High Dimensional Data, Neolithic mines of Gavà.

INTRODUCTION

Variscite is a phosphate mineral, with chemical formula $AlPO_4 \cdot 2H_2O$, which crystallises in the orthorhombic system; it is polymorphous with metavariscite and forms a solid solution with strengite ($FePO_4 \cdot 2H_2O$). Minerals from the variscite group are commonly found together in geological deposits formed by either hydrothermal water circulation through phosphate primary deposits[1] or by the weathering of aluminium minerals in phosphate-rich waters [2,3]. Variscite is often mistaken for turquoise, jade, as well as other green minerals[3] and green rocks[2]. The use of variscite as an ornamental stone for necklaces, earrings, bracelets, beads, etc. has been widely identified from the Palaeolithic age [4].

The mines of Gavà, discovered in the 1970s, is an ancient mining complex situated in Catalonia (Spain). It was exploited during the Neolithic age and is, probably, the oldest underground mining complex in Europe [5,6]. Although the entire extension of the Neolithic mining complex has not been fully excavated , more than 80 mines have already been reported which extend to more than 1000 metres in total at five different depths[6]. Pits and galleries have been found excavated at different

depths, despite the occurrence of variscite in the shallower levels. The reason for the Neolithic excavation extending to the deeper levels when the valuable green mineral could be obtained at shallower depths is still unknown.

Raman spectroscopy has been demonstrated as a non-destructive technique that has been widely applied to the study of archaeological artefacts. It is sensitive to structural and compositional mineralogical variations by the observation of changes in Raman band positions or shapes such as width, intensity and the presence of shoulders. Solid solutions, such as variscite/strengite, as well as polymorphous minerals (variscite/metavariscite) can be detected by changes in their Raman spectra[1,3,7]; these changes can be quite subtle when the variations in the mineral composition or structure are also small. Furthermore, the reported differences in colour observed in variscite specimens could be related to small compositional and structural changes which may result in only minor changes in the shape or shift of the Raman bands[8]. These small spectral changes can be difficult to identify visually and directly in the spectra and new spectral computational tools would be essential for their detection and structural interpretation. This is of particular interest when an assessment of the mineral origin is required. Geological formations and their geological environments imprint small differences on the ores which can be used as footprints for relating the origin of the mineral and its source mine , but the discrimination between those subtle differences implies the requirement of the study of a large number of samples and an associated large number of analyses. To accomplish the processing of such a large quantity of data , computational algorithms are necessary.

In the mining complex of Gavà, a wide range of variscite colours has been already reported. The colour changes from the surface deposits to the deepest levels; the shallower variscite is yellowish or brownish whereas at greater depths it shows a beautiful apple green colour[6]. However, despite the different analyses already performed hitherto on the variscite from Gavà [6,9], no analytical characteristic useful for the mine assignment and location has been found . Furthermore, there has been no correlation determined between the specimen colour, source mine and Raman spectral changes.

Tools for the automatic identification of Raman spectra can be classified into two main groups: *a*) tools based on finding the correspondence between key spectral data characteristics previously identified by spectroscopists and *b*) tools based on machine learning that use the entire spectral range for multivariate analyses.

The first methods are grouped in what is known as univariate analysis. Sobron et al.[10] have developed the automatic analysis of Raman spectra for the identification of minerals relevant to the future ESA ExoMars Rover Mission to be launched in 2020. There is also software reported for the automatic identification of Raman spectra which has been developed without the need for spectral pre-processing[11]. A disadvantage of this type of analysis is that it is not able to accurately estimate the presence of minerals in samples of complex composition[12].

The second group of computational techniques is known as multivariate analysis; these techniques try to correlate the interrelationships amongst all the variables and have become the most commonly used today. The most frequently used technique in the literature has been the SVM method (Support Vector Machines); for example, Thissen et al.[13] and Ghesti et al.[14] use SVMs and Partial Least Squares (PLS) analysis , respectively, in the context of process quality control. Muehlethaler et al.[15] use SVMs for drug detection in human urine samples and Pierna et al.[16] to determine the origin of honey specimens.

Focusing on mineral classification, it is found that there is a greater diversity of techniques available : the application of Artificial Neural Networks (ANN) to XRay diffraction spectra has been used by Gallagher & Deacon[17]. Sometimes, ANN has been used in combination with dimensional reduction techniques, such as Principal Component Analysis (PCA)[18, 19]; Artificial Neural Networks have not been the only classifier used in combination with PCA analysis and algorithms based on closest neighbours have also been used. Kelloway et al.[20] have used PCA analysis together with the Mahanalobis distance and Carey et al.[21] used a custom trajectory similarity metric along with PCA analysis.

In this work, the computational treatment of Raman spectra is described with the objective of finding Raman spectral differences for variscite specimens from the Gavà mining complex to assist in the determination of the individual mine and the depth of origin of the mineral. Furthermore, the spectroscopic interpretation of the observed key wavenumber positions selected by multivariate analyses will be studied with the purpose of relating them to chemical or structural mineral differences.

MATERIALS AND METHODS

Samples

The one hundred Raman spectra recorded for subsequent computational treatment were achieved over eleven samples identified as variscite from the Gavà mining complex. Seven specimens were collected from Mine 11N comprising specimens from the medium level (two specimens) and from the deep level (five specimens). Four samples were collected on Mine 5_7, three specimens from the medium level and one from the deep level. All of the specimens showed different colours ranging from yellowish-green, dark-green, light green, and bluish-green through to blue (Table 1).

Raman spectroscopy

Raman spectra were carried out in the Archaeometry Laboratory of CENIEH (Spain), using a DXR Thermo Fisher confocal Raman spectrometer, working with a 532 nm laser wavelength (green light). For improving the spectral signal-to-noise ratio, 60 accumulations at a 10 second exposure time were performed at each sampling position. A high laser power increased the Raman band intensities and decreased the analysis time but could also damage the sample; to ensure that no heating damage was induced, a set of Raman spectra were compiled with laser powers at the sample of 0,1; 0,5; 1 and 2 mW and, finally, an operating laser power of 1 mW was selected for each analysis.

On each sample, between six to eleven points were randomly chosen and the Raman spectra were accumulated following the previous experimental conditions. The spectrometer was calibrated every day using a silicon wafer and cross-checking with a specimen of pure calcite mineral. This calcite cross-checking procedure allows an estimation to be made of the wavenumber accuracy of the instrument calibration for either low (calcite Raman bands situated at 281 and 156 cm^{-1}) and medium-high (strongest calcite peak at 1086 and one of weaker intensity at 713 cm^{-1}) wavenumber regions, where the characteristic Raman bands of variscite also appear. Shifts observed to the silicon band wavenumber were of a maximum of 0,04 cm^{-1}, whereas those corresponding to the calcite bands were 0,08 cm^{-1} for 1086 cm^{-1}; 0,16 cm^{-1} for 713 cm^{-1}; and 0,68 and 0,46 cm^{-1} for 281 and 156 cm^{-1}, respectively. Because of such experimentally observed wavenumber differences, all spectral wavenumbers were accepted as recorded without compensation being applied.

Pre-processing

Machine Learning or data mining is the area of computer science that is dedicated to developing algorithms that devise relationships and patterns from the data. The data (in this case, the spectra derived above) can be divided into a training set (the set used by the algorithms to learn) and a test set (used to determine the quality of the predictions made by the application of the algorithms). These data sets are formed by a set of instances; each instance (i.e. each spectrum) is formed from a

set of attributes (the spectral wavenumbers). Attribute values are the intensity values recorded at each different wavenumber. These attributes will be designated the independent variables (X) from which the dependent variables (y) will be predicted: namely, the mine of origin and the depth of extraction of the specimen.

An automatic spectrum recognition system generally has two stages: a pre-processing stage and a classification stage. In the following sections, the different techniques used in this work for each of these stages will be discussed.

In order to reduce the influence of noise and fluorescence background emission and also to eliminate intra-species variations (within the same class) whilst maximizing the extra-species distances (arising from different classes) a pipeline of pre-processing operations was applied.

Smoothing

Smoothing is an operation which allows an increase in the signal-to-noise ratio without greatly distorting the signal. One of the most common methods is the so-called Savitzky-Golay filter[22] based on the calculation of a local polynomial regression (of degree k), from at least $k + 1$ equispaced points, to determine the new value of each point. The result is a function similar to the input data, but with lower noise levels. In this work, the Savitzky-Golay filter was used and the operation of the filter has been regulated by setting the value of $k = 25$.

Cropping and interpolation

Cropping and interpolation operations are necessary because in Machine Learning the model learns the relationships between attributes and classes. Therefore, to train a model, all instances need to have the same number of attributes and the attributes have to represent the same characteristics, in this case the intensity, in a certain band. Similarly, to predict the class of an instance, it must have the same attributes as the instances used to train the model.

Despite the daily calibration verification , a very small wavenumber shift was considered acceptable during spectral collection. In this work, linear interpolation and cropping was used to convert each spectrum into a vector of 1750 values between 50 and 1800 cm^{-1}.

Baseline

Baseline variation is a problem found in the recording of spectra from specimens of different sourced materials. Basically, it consists of a linear or non-linear intensity addition which results in a

distorted measurement, whereby the observed value is higher than its real value. Multiple algorithms have been proposed to deal with the estimation and correction of baseline effects[23-30] and the fundamentals of operation of these baseline methods is beyond the scope of this paper. After a series of preliminary experiments, it was decided to use the ALS algorithm[23] which offered the best results for the spectral data recorded in this work.

Intensity normalization

Intensity normalization preserves the relative order of band intensity values whilst mitigating the effect of peak intensity differences. It is performed by scaling each spectral wavenumber based on the maximum value (L_∞ *norm*), the sum of absolute values (L_1 *norm*), or the sum of squared values (L_2 *norm*). In this work, the method used was intensity normalization based on the maximum wavenumber value.

The order in which the different pre-processing operators are applied has an influence on the final result. In this work the order has been as follows: Interpolation, Smoothing, Cropping, Baseline Correction and Intensity Normalization.

Classifiers

Once all the spectra have been pre-processed, it is possible to move on to the next phase. In the classification phase, predictive models that help to determine the class (mine of origin or depth of extraction) of new variscite samples are constructed.

From a formal point of view, a classification trial to model a function $\mathbf{X} \rightarrow \mathbf{Y}$, such as $h(x) = y$, where x are the observations and y the value to be predicted, the label or output variable is attempted; during training a collection of observations and their corresponding output values are made available, and the learning task is to try to find the model parameters that better reproduce the historical data, with the aim of generalizing the relationship between observation and prediction that can be used to predict new future observations. A large number of classification algorithms have been proposed over the years. From the point of view of interpretability, they are considered as black boxes, the models can predict the class to which a sample belongs but do not provide any information about why they make the prediction. There are also models that provide “clues” about their operation. In this work, only those classifiers that are able to provide clues about their operation have been used.

All the following classification algorithms have been evaluated and their results appear later in the "Results and Discussion" section.

Logistic Regression

Logistic Regression (LR)[31] is one of the most used classification models. It allows the modelling the probability that a class takes a certain value (p_i) from a series of independent variables (x_1,x_2,...,x_k).

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i}$$

Each independent variable is assigned to a coefficient and the higher the coefficient value, the greater the influence this independent variable has in the model. Logistic Regression is a method that usually works well with high-dimensional data sets (involving many attributes), as in the case discussed here.

Ridge Regression

Ridge Regression (RR)[32] is an improvement of Logistic Regression that uses a regularization factor to facilitate the solution of ill-posed problemswhich, in general, avoids overfitting and improves the generalization of the obtained models.

Support Vector Machines

The SVM (Support Vector Machines)[33] method is an algorithm similar to Logistic Regression, but while Logistic Regression is based on the assumption that the data are produced as the result of a probabilistic model whose parameters need to be adjusted using the available observations, SVM follows a more geometric approach and tries to find the best hyperplane that separates the classes, i.e. the one that maximizes the margin (that is, the hyperplane that separates the classes is removed as far as possible from the instances).

Linear Discriminant Analysis

When working with spectra and other data sets which have a large number of attributes, there is often the need to reduce their size. There are several linear projection techniques available to

achieve this, one of the best known being PCA. PCA projects the data set into the direction which explains most of the variance in the data set, but as PCA is an unsupervised technique it does not take into account the class labels and although it can be used in this way and indeed has been used as a processing method, it is not a classification method.

Another well known technique is LDA (Linear Discriminant Analysis)[34] which finds a linear subspace that maximizes the class separability and then improve the results of a classifier with a linear decision boundary, generated by fitting class conditional densities to the data using Bayes' rule.

Decision Trees

A well known top-down classification algorithm is Decision Trees[35. This algorithm constructs tree-like graphs. Two abstractions are used in decision trees: the nodes and the branches. Branches simply connect nodes with each other. The nodes make decisions: they can send an instance to another node (termed a child node) that is connected through a branch or they can return the estimated class of an instance if it is a final node or leaf node. In the root (the first node), all instances are used to determine which is the best attribute for splitting the instances in two subsets assigned to two new child nodes. This process is recursively repeated in each new node until the class of all instances of the subset is unique or until a stopping criterion is reached. The best attribute is determined in each node by evaluating the Information Gain or Gini Index.

An ensemble is a classifier that combines the results of several base classifiers by voting or averaging improving the performance of the individual classifiers. Decision Trees are frequently used as base classifiers of ensembles because they are efficient and unstable, that is, small changes in the training set or in the construction method will produce very different classifiers (by increasing the diversity of the ensemble for its performance improvement). A Random Forest[36] is a meta classifier (or ensemble) that trains several decision trees on various sub-samples of the original dataset and uses averaging to improve the predictive accuracy and control over-fitting. A Random Forest is able to assign importance to the different attributes of a data set and it is also one of the most cited classifiers for the best results, which makes it interesting to study it in this context.

The algorithms mentioned in this section were evaluated using a Scikit-learn[37] library containing the default parameters. The experiments were performed using a 5- fold cross-validation. Figure 1 shows two spectra before and after being pre-processed using Interpolation, Smoothing, Cropping, Baseline Correction and Intensity Normalization, as explained in the previous section. It is clearly

seen here how from two very different spectra subjected to variable phenomena such as fluorescence, this variability has been practically eliminated on the final processed spectra.

Once all spectrum pre-processing was accomplished, the issues to be investigated were:

- Is there any difference between the spectra of different samples related to the mine location or to the depth at which the samples have been extracted?
- Can Machine Learning algorithms correctly classify the particular mine and depth at which a certain variscite sample was extracted?
- Do the attributes, with more weight for the applied algorithm, have any spectral meaning?

RESULTS AND DISCUSSION

Differences in the spectra depending on the specimen origin and depth

Before checking if data mining algorithms can determine the mine of origin and the depth from which variscite samples have been extracted, the identification of Raman spectral differences between specimens was implemented. For this, the cosine distance between each pair of spectra will be computed. Cosine distance is a measure of the similarity between two vectors (given by the spectral values in this work) . The higher the distance the more different the vectors are (in this case the spectra) .Figure 2 gives graphs showing the average of pairwise differences between the spectra belonging to each of the classes examined, where dark colours represent a large difference and light colours a small difference.

In this work, the number of classes in which we can classify a specific spectrum is 2, namely a) the origin of the sample which here is specifically two origins: Mine 5_7 and Mine 11N; and b) the depth of the sample: here, medium and deep. For these 2 classes, the graph is presented as a 2 x 2 grid. In this grid, each cell represents the average of the distances between each possible pair of spectra, taking one spectrum of the class that gives its name to the column and another of the class that gives its name to the row.

Note that, ideally, the average distance between spectra of the same class should be minimal (i.e. a therefore a light colour), in the same way that the average distance for spectra of different classes should be maximum (i.e. a dark colour). Therefore, ideally, the diagonal Northwest - Southeast transect should be very light colour and the diagonal Northeast - South West transect a very dark colour (Figure 2).

Figure 2a shows the average of pairwise distances according to the mine. It is observed that spectra belonging to the class Mine 5-7 are very similar to each other (light colour) and that these are very different from the spectra of samples from Mine 11N (dark colour). However, this low average intra-class distance is not observed in the samples from the Mine 11N. This result indicates that the spectra of samples from Mine 5_7 are very uniform and possess very little variability between them, whereas the spectra of samples from Mine 11N have a greater variability amongst themselves. These results are sufficient to indicate that a predictive model can be constructed, because apparently the spectra of the samples from Mine 5_7 are very different from those of Mine 11N. Although the spectra of the samples from Mine 11N are very diverse amongst themselves, a classifier may give a greater importance only to those bands that serve to discriminate the mine and ignoring the rest, which is something that does not happen when calculating distances, since all bands are then treated equally.

To observe differences between the spectra according to the depth of extraction of the mineral, spectra will be separated into two groups. In one group the 34 spectra from the Mine 5_7 specimens are taken and in the other the 66 spectra from the Mine 11N specimens are taken . Figure 2b shows the average of pairwise distances according to the depth, considering only the 34 spectra from the Mine 5_7 specimens. In this case, the ideal situation occurs and the diagonal Northwest - Southeast transect is lighter in colour. This indicates that within the spectra of the Mine 5_7 specimens there are differences according to the depth of extraction. Figure 2c shows the average of pairwise distances according to the depth, considering the 66 spectra of the specimens from Mine 11N. In this case, it is observed that the spectra of samples obtained at a medium depth are very uniform amongst themselves and easily distinguishable from the deep analogues.

Classification Results

Classification algorithms have been used in 3 experiments: *a*) Classification of the mine of origin, *b*) Classification of the depth of extraction in the samples from the Mine 5_7 *c*) Classification of the depth of extraction in the samples from the Mine 11N. The results are presented in terms of accuracy.

Table 2.a shows the results when the class to be considered is the mine of origin, it is observed that the best results are obtained by SVM (98%). All the evaluated algorithms obtained accuracy scores higher than 0.9 (90%), which indicates that there are significant differences in the spectra of the

specimens according to the mine of origin. Table 2.b shows the results when the class to be considered is the depth of the samples from the Mine 5_7, it is observed that the best results are obtained by SVM (87%). Table 2.c shows the results when the class to be considered is the depth of the samples of the Mine 11N, it is observed that the best results are obtained by Ridge Regression (90%)

These results suggest that for a classifier it is relatively easy to discriminate the mine of origin (almost a perfect classification) while the depth level of extraction is a bit more complicated to derive.

The best results have been obtained by SVM in 2 of the 3 experiments and by Ridge Regression in the third. These results make sense because SVM and Logistic Regression (along with Ridge Regression, which is an improvement on the latter) tend to be preferential to other possible algorithms when the number of attributes is very high, as in this case.

Apart from the accuracy, it is possible to visualize where the failures occur. Figure 3a shows the confusion matrix for mine classification. It is observed how only one spectrum from each mine is misclassified. Figures 3b and 3c show the confusion matrix for the classification of depths in each of the mines. In the Mine 5_7 specimens, for spectra labelled as deep there is no error, however there are 4 intermediate depth spectra that have been erroneously classified as deep. In Mine 11N specimens, of 48 spectra labelled as deep only 3 are classified incorrectly and the same is true of the intermediate depth specimens.

Visualization of best attributes according to the Machine Learning algorithm
Some of the classification algorithms can be used to compute feature importance. Classification Trees and Random Forest algorithms calculate the information gain, which is a simple statistic measure, whereas LDA does not achieve a good result for classifying the mine of origin.

So the most interesting attributes (Raman wavenumbers) and their importance are calculated by Logistic Regression and SVM. In this case SVM was chosen, because it is the algorithm with the highest accuracy. Figure 4 shows the importance of each attribute. Green colours represent bands which have a great influence on the model, whereas red-coloured bands have a very low influence on the model.

Raman Spectroscopy

From the eleven samples analysed (Table 3), eight of them (5_7-4.2; 5_7-5.1; 11N-4.1; 11N-4.3; 11N-6.1A; 11N-6.1B; 11N-6.2; 11N-6.4) give Raman spectra that match with those published by Frost et al.[38] and Onac et al.[39]. The characteristic bands are situated at wavenumbers around 1060 and 1020 cm^{-1} for the stretching vibration of the phosphate group, at around 420 cm^{-1} for the bending phosphate modes and at 221 and 110 cm^{-1} for the lattice vibrations. Onac et al.[39] observed the presence of a shoulder at 1079 cm^{-1}, which is also present as a weak or very weak shoulder in some of the spectra collected on nine samples here.

Spectra from samples 5_7-3.1 and 5_7-4.1 show a third band, sometimes a bit stronger than just a shoulder, at around 1080 cm^{-1}, forming a triplet signature with 1080 cm^{-1} (medium-weak), 1060 cm^{-1} (medium) and 1020 cm^{-1} (strong) bands. In most of these Raman spectra, the band at 401 cm^{-1} was also well differentiated and with a similar relative intensity to the adjacent band at 430 cm^{-1}.

Raman spectra collected from the sample 11N-6.5 show the strongest bands at 1059 and 1020 cm^{-1} for the variscite[38,39]. However, signatures at 799, 742 and 130 cm^{-1} that have not been described as characteristic of variscite from Frost et al.[38] or Onac et al.[39] appear here whereas the band at 1080 cm^{-1} is not present at all in our Raman spectra, as seen in samples 5_7-3.1 and 5_7-4.1. The spectra of the sample 11N-6.5 also exhibit a more intense band at 913 cm^{-1} compared with other variscite specimens. The presence of these weak bands in the lattice wavenumber region could be related to minor compositional changes or even to the presence of polymorphic phases.

In Table 4, a summary of the attributes (Raman wavenumber) which are more intense than the 0.2 threshold for either the weight-mine, weight -Depth Mine 5_7 and weight-Depth Mine 11N are shown with the Raman vibrational band assignment. For classifying the mine, ten attributes are seen above the 0,2 threshold. The strongest signatures are related to the main variscite Raman bands, such as the attribute at 1081 cm^{-1}, which appears as a shoulder on the 1060 cm^{-1} band in most of the Raman spectra, or those at 452 and 101 cm^{-1}, which are related to the characteristic Raman bands at 429 and 109 cm^{-1}, respectively. However, for classifying the depth at which a sample was collected in one specific mine, the results are not so accurate. The algorithm works better for Mine 11N than for Mine 5_7. In the first case, only 5 attributes are above the 0,2 threshold , with the largest attribute at 1006 cm^{-1}, which is related to a shoulder on the strongest Raman band at 1020 cm^{-1}; there are also significant attributes at 1082 cm^{-1} and 397 cm^{-1}, these related with the characteristic variscite Raman band at 402 cm^{-1}. For classifying depth in Mine 5_7, only three attributes are seen to lie above the 0,2 threshold.

Wavenumbers selected for classifying spectra by the algorithms appear as shoulders on some of the main variscite Raman bands. It is well known that shifts and band shape changes can be ascribed, among other parameters, to changes in composition of the material being studied. In the case of the variscite group, shifts and bandwidth changes on the main Raman bands have already been ascribed to different ratios of Al/Fe^{3+}[40]. Frost et al.[38] relate the number of bands in the symmetric (around 1030-900 cm^{-1}) and antisymmetric (1200-1030 cm^{-1}) stretching regions with multiple PO_4 species, however, Litvinenko et al.[8] suggest that the variscite band shape and its precise wavenumber position could change because of the presence of impurities of different minerals, such as strengite and phosphosiderite. For the mine identification, for example, the main attributes selected are at 1081 and 982 cm^{-1}, both related with the phosphate mineral structure of either variscite or strengite, whereas the attribute at 101 cm^{-1}, assigned to a lattice mode, could be related with cation structural changes, although it is still under study if the attributes selected by computational algorithms for classifying variscites from the complex mines of Gavà are related to impurities, cation substitutions or different PO_4 species in the crystal structure.

The results obtained here are very interesting, not only because of the high precision obtained when assigning the mine of origin to the variscite samples, but also because of the Raman bands selected, which are significant and characteristic spectral signatures. A human expert would naturally base their decision for the mineral identification on these Raman signatures, now this could be done automatically and faster by the algorithms outlined here, providing a valuable tool to researchers who may not be experts in Raman spectroscopy.

The computational algorithm could help not only in the identification of mineral varieties and, then, be very useful for variscite mine assignment, but if those relationships can be assessed, it could additionally help to inform the assignment of mineral provenance used in prehistoric times and to facilitate the discovery of commercial trade routes or population movements.

CONCLUSIONS

Classification algorithms were used to try to determine the mine of origin and the extraction depth of specific samples of variscite from the mining complex of Gavà. Among all the possible existing algorithms, those that are able to show the importance of each attribute in the classification process have been selected and described

.

These algorithms have selected attributes related to significant spectral characteristics observed in the Raman spectrum of variscite. Most of these band wavenumbers , such as those at 1081, 1006,

397 or 101 cm^{-1}, appear as shoulders on the main variscite bands which, in some cases, have been related to the ratio of Al/ Fe^{+3} or other cations in the variscite structure as well as impurities or to the presence of different PO_4 species.

Regarding the classification results, for the mine class the SVM algorithm was particularly useful and the classification was quite precise (approximately 90%). Results obtained for the depth class are almost as good as for the mine of origin, although more instances could help to improve the accuracy of the results. Further work will be carried out to extend this study with more samples from different mines and depths as well as to investigate new ways of interpreting and visualizing the decisions made within the classification.

Acknowledgements

"These analysis/experiments were performed in the Archaeometry laboratory at CENIEH facilities with the collaboration of CENIEH Staff".

This work has been partially funded by Ministry of Economy, Industry and Competitiveness through the project TIN2015-67534-P

This paper is a contribution to the projects AGAUR 2014 SGR 1661, AGAUR 2017 SGR707 and 2014/100820 of the Generalitat de Catalunya.

REFERENCES

[1] J.T. Kloprogge, B.J. Wood, *Spec. Acta Part A: Mol. Biomolec. Spect.* **2017**; *185*, 163.

[2] C.P. Odriozola, *Archaeol. Anthropol. Sci.* **2015**; *7*, 329.

[3] E. Fritsch, S. Karampelas, J.Y. Mevellec, *J. Raman Spec.* **2017**; *48*, 1554.

[4] F. Herbaut, G. Querré, *Bull. Soc. Prehist. Fr*. **2004**; *101*, 497.

[5] M. Blasco, M. Borrell, J. Bosch, *Trabajos de prehistoria* **2000**; *57*, 77.

[6] A. Camprubí, J.C. Melgarejo, J. Proenza, F. Costa, J. Bosch, A. Estrada, F. Borrell, N. Yushkin, A. Andreiche, *Episodes: J. Int. Geosci.* **2003**; *26*, 295.

[7] W.P. Griffith, *J. Chem. Soc. A: Inorg. Physic., Theoretic.* **1970**; 286.

[8] A.K. Litvinenko, E.S. Sorokina, S. Karampelas, N.N. Krivoschekov, R. Serov, *Gems & Gemology* **2016**; *52*, 60.

[9] J.C. Melgarejo, S. Lehbib, J. Bosch, L. Arqués, T. Jawhari, L. Torró, A. Camprubí in *Abstracts Colloque International Roches & Sociétés 2015,* Callais, **2015**, pp. 25-26.

[10] P. Sobron, F. Sobron, A. Sanz, F. Rull, *Appl. Spectrosc.* **2008**; *62*, 364.

[11] I.H. Rodriguez, G. Lopez-Reyes, D.R. Llanos, F.R. Perez, *Automatic Raman spectra processing for Exomars*, Springer, Berlin-Heidelberg, **2014**, pp. 127-130.

[12] P. Vandenabeele, *Spectrochim. Acta A* **2011**; *80*, 27.

[13] U. Thissen, M. Pepers, B. Üstün, W.J. Melssen, L.M.C. Buydens, *Chemom. Intell. Lab. Syst*. **2004**; *73*, 169.

[14] G.F. Ghesti, J.L. de Macedo, I.S. Resck, J.A. Dias, S.C.L. Dias, *Energy Fuels* **2007**; *21*, 2475.

[15] C. Muehlethaler, M. Leona, J. R. Lombardi, *Anal. Chem.* **2015**; *88*, 152.

[16] J.A.F. Pierna, O. Abbas, P. Dardenne, V. Baeten, *Biotechnologie, Agronomie, Société et Environnement* **2011**; *15*, 75.

[17] M. Gallagher, P. Deacon, *Proceedings ICONIP '02* **2002**; *5*, 2683.

[18] G. Lopez-Reyes, P. Sobron, C. Lefebvre, F. Rull, *Am. Mineral.* **2014**; *99*, 1570.

[19] S.T. Ishikawa, V.C. Gulick, *Comput. Geosci.* **2013**; *54*, 259.

[20] S.J. Kelloway, N. Kononenko, R. Torrence, E.A. Carter, *Vib. Spectrosc.* **2010**; *53*, 88.

[21] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, M.D. Dyar, *J. Raman Spec.* **2015**; *46*, 894.

[22] A. Savitzky, M.J. Golay, *Anal. Chem.* **1964**; *36*, 1627.

[23] P.H. Eilers, H.F. Boelens, Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, **2005**; *1*, 5.

[24] C. A. Lieber, A. Mahadevan-Jansen, *Appl Spectrosc*. **2003**; *57*, 1363.

[25] R. Perez-Pueyo, M.J. Soneira, S. Ruiz-Moreno, *Appl Spectrosc*. **2010**; *64*, 595.

[26] Z.M. Zhang, S. Chen, Y.Z. Liang, *Analyst* **2010**; *135*, 1138.

[27] L. Shao, P.R. Griffiths, *Environ. Sci. Technol.* **2007**; *41*, 7054.

[28] W. Dietrich, C.H. Rüdel, M. Neumann, *J. Magn. Reson. (1969),* **1991**; *91*, 1.

[29] J.C. Cobas, M.A. Bernstein, M. Martín-Pastor, P.G. Tahoces, *J. Magn. Reson.* **2006**; *183*, 145.

[30] J. Kajfosz, W.M. Kwiatek, *Nucl. Instrum. Methods Phys. Res. Sect. B* **1987**; *22*, 78.

[31] D.W. Hosmer Jr., S. Lemeshow, R.X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, New Jersey, **2013**.

[32] A.E. Hoerl, R.W. Kennard, *Technometrics* **1970**; *12*, 55.

[33] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, *IEEE Intell. Syst. Their Appl.* **1998**; *13*, 18.

[34] S. Balakrishnama, A, Ganapathiraju, *Institute for Signal and information Processing* **1998**; *18*, 1.

[35] J.R. Quinlan, *C4.5: programs for machine learning*, Elsevier, **2014**.

[36] L. Breiman, *Machine learning* **2001**; *45*, 5.

[37] F. Pedregosa, G. Varoquaux,A. Gramfort,V. Michel, B. Thirion, O. Grisel, J.Vanderplas, J. Mach. Learn. Res. **2011**; *12*, 2825.

[38] R.L. Frost, M.L. Weier, K.L. Erickson, O. Carmody, S.J. Mills, *J. Raman Spec.* **2004**; *35*, 1047.

[39] B.P. Onac, J. Kearns, R. Breban, S. Cîntă Pânzaru, *Studia UBB Geol.* **2004**; *49*, 3.

[40] N. Acevedo, M. Weber, A. García-Casco, J.A. Proenza, J. Sáenz, A. Cardona, *Latin Amer. Antiq.* **2016**; *27*, 549.

FIGURE CAPTIONS

Figure 1: Previous and final Raman spectra after pre-processing following the pipeline: Interpolation, Smoothing, Cropping, Baseline Correction and Intensity Normalization. All of these processes result in the uniformity of the Raman spectra despite differences experienced in spectral noise or fluorescence.

Figure 2: Average of pairwise cosine distance between classes: a) Class Mine; b) Class Depth-Mine 5_7; c) Class Depth Mine 11N

Figure 3: Confusion matrix for the classification of matMina, matProf-Mine5_7 and matProf-Mine11N.

Figure 4: Attributes (Raman bands) and weights for each classification. A green colour represents a band which has a great influence upon the model, a red colour shows a band with a very low influence upon the model . The weights of the attributes have been obtained from the model that classifies between the following: Mine of origin (top); Depth in mine 5_7 (centre); and depth in Mine 11N (down).

Table 1: This table shows sample characteristics with regard to mine of provenance, colour and the depth at which each sample was collected, as well as the number of spectra that were used for the classification phase by computational algorithm.

Sample code	Mine	Depth	Colour	Number of Spectra
5_7-3.1	5-7	Medium	Dark green	10
5_7-4.1	5-7	Medium	Blue	8
5_7-4.2	5-7	Medium	Blue	6
5_7-5.1	5-7	Deep	Bluish-green	10
11N-4.1	11N	Medium	Bluish	7
11N-4.3	11N	Medium	Dark green	11
11N-6.1A	11N	Deep	Blue	11
11N-6.1B	11N	Deep	Blue	8
11N-6.2	11N	Deep	Dark green	10
11N-6.4	11N	Deep	Dark green	9
11N-6.5	11N	Deep	Light green	10

Table 2: Classification results (success rate). Names of the classifiers are abbreviated, apart from common use abbreviations there are others as follows: LR means Logistic Regression; RR means Ridge Regression; RF means Random Forest; and Tree is a single classification Tree. The best results are obtained with SVM (in the case of mine of origin classification and in the case of the determination of the depth of the samples extracted in Mine 5_7) and Ridge Regression (in the case of the depth of the samples extracted in Mine 11N).

a) Classification of the mine of origin			b) Depth classification (Mine 5_7)			c) Depth classification (Mine 11N)	
Clasificador	Accuracy		Clasificador	Accuracy		Clasificador	Accuracy
LDA	0.939474		LDA	0.695238		LDA	0.797619
LR	0.928947		LR	0.752381		LR	0.828571
Ridge	0.970000		Ridge	0.838095		Ridge	0.902381
RF	0.929424		RF	0.752381		RF	0.835714
Tree	0.901429		Tree	0.785714		Tree	0.866667
SVM	0.980000		SVM	0.871429		SVM	0.892857

Table 3: Characteristic Raman bands for variscite. The most significant spectrum in each sample spectrum collection was used for the construction of this table. (*) The band appearance is not consistently regular for a specific sample, it was sometimes detected whereas other times it was absent and, when it appeared, it always showed a very low relative intensity.

5_7-3.1	5_7-4.1	5_7-4.2	5_7-5.1	11N-4.1	11N-4.3	11N-6.1A	11N-6.1B	11N-6.2	11N-6.4	11N-6.5
1640w-br	1640w-br		1636w-br	1632w-br	1633w-br	1640w-br				1632w-br
		1618w-br					1620w-br	1629w-br	1630w-br	
		1355w-br								
1079m	1080w									
1061m	1062m	1059m	1057m-s	1058m	1060m	1060m	1060m-s	1060m	1058m	1059m
1020s	1022s	1019s	1019s	1020s	1020s	1021s	1021s	1019s	1021s	1020s
915vw	*	902vw	905vw	*	910vw	*	*	*	913w	908vw
									742w	
	605vw	598w	603w	603w	601w	600w	607w	607w	598w	602w
580w	575w	572w	571w	573w	569w	573w	572w	573w	568w	572w
549w	545w	542w	543w	546w	549w	546w	545w	546w	542w	544w
									451m	
429m	428m	422m	424m	426m	426m	425m	424m	426m	431m-s	45m-s
401m	402m			402m	404m	403w	403m-w	403m		
363w	364w	359w	360w	361w	362w	360w	361w	360w	360w	360m-w
									338w	
									278w	
221w	223w	220m-w	221m-w	222m-w	221m-w	223w	220w	223m-w	220w	222m-w
	172w	170w	170w	170m-w	171w	170w	169w	171m-w	170w	171w
									130m	131w
109w	110w	108m-w	108m-w	109w	110m-w	110w	110w	110w	109w	109m-w
70w	72w	70w	71w	70m-w	70m-w	71w	70w	71m-w		70m-w572

Table 4: Attributes which showed a weight more intense than the 0,2 threshold for Mine, Depth-Mine5_7 and Depth-Mine11N.

	Attribute (Wavenumber)	Weight_mine	Weight_depth11	Weight_depth5_7
Lattice mode	66		0,224330534	
Lattice mode	70			0,416584033
Lattice mode	78	0,273611408		
Lattice mode	101	0,327766812		
Lattice mode	132	0,224349938		
Lattice mode	168			0,230152572
	223	0,211938279		
Phosphate bending[39]	392	0,233489443		
	397		0,273253664	
PO_4 out of plane bends[38] Phosphate bending[39]	452	0,258777425		
	465		0,215060576	
ν_1 symmetric stretching mode[1,38,39]	982	0,301730927		
	992			0,261528511
	1001	0,224774125		
	1006		0,353563363	
	1062	0,205872033		
ν_3 antisymmetric stretching mode of the PO_4[1,38]; Broad P=O stretch[39]	1081	0,276824145		
ν_3 antisymmetric stretching mode of the PO_4[1,38]; Broad P=O stretch[39]	1082		0,260974593	