Optimizing the removal of fluorescence and shot noise in Raman Spectra of minerals by ANFIS and moving averages filter

Optimización del proceso de eliminación de la fluorescencia y el ruido de disparo en espectros Raman de minerales mediante ANFIS y el filtro de medias móviles

CABRERA-CABAÑAS, Reinier†*, LUNA-ROSAS, Francisco Javier, MARTINEZ-ROMO, Julio César and HERNÁNDEZ-VARGAS, Marco Antonio

Instituto Tecnológico de Aguascalientes, Computing Science Department, Av. A. López Mateos 1801 Ote. Col. Bona Gens, C.P. 20256, Aguascalientes, Ags., México

ID 1st Author: *Reinier, Cabrera-Cabañas* / **CVU CONACYT ID**: 765973

ID 1st Coauthor: *Francisco Javier, Luna-Rosas* / **ORC ID**: 0000-0001-6821-4046, **arXiv Author ID**: arXivFco19, **CVU CONACYT ID**: 87098

ID 2nd Coauthor: *Julio César, Martinez-Romo* / **ORC ID**: 0000-0001-6242-5248

ID 3rd Coauthor: *Marco Antonio, Hernández-Vargas* / **ORC ID**: 0000-0002-8146-9307

DOI: 10.35429/EJT.2020.8.4.1.12 Received October 02, 2020; Accepted November 29, 2020

Abstract

Raman spectroscopy is a non-destructive and non-contact technique that requires minimal sample preparation so it can be used to identify a wide range of minerals and gemstones. This optical technique is capable of measuring vibrational modes of biomolecules, allowing their identification from the correct location of the Raman bands, one of the main challenges is the elimination of spectral noise composed of (a) fluorescence background and (b) high frequency noise. The objective of the article was to demonstrate that using ANFIS (Neuro Fuzzy Adaptive Inference System) in combination with moving averages filter on the MATLAB multicore platform we can eliminate these disturbances and optimize response time in the preprocessing of large volumes of data while maintaining the spectral meaning related to the structure and/or composition of the mineral to be validated.

Spectroscopy, Fluorescence, Optimum Design

Resumen

La espectroscopía Raman es una técnica no destructiva y de no contacto que requiere una preparación mínima de la muestra por lo que puede usarse para identificar una amplia gama de minerales y piedras preciosas. Esta técnica óptica es capaz de medir los modos vibratorios de las moléculas, permitiendo su identificación desde la correcta ubicación de las bandas Raman, uno de los principales desafíos es la eliminación del ruido espectral compuesto por (a) fondo de fluorescencia y (b) ruido de alta frecuencia. El objetivo del artículo fue demostrar que utilizando ANFIS (Sistema Adaptativo de Inferencia Neuro-difusa) en combinación con el filtro de medias móviles en la plataforma multinúcleo MATLAB podemos eliminar estas alteraciones y optimizar el tiempo de respuesta en el preprocesamiento de grandes volúmenes de datos manteniendo el significado espectral clave relacionado con la estructura y/o composición del mineral a validar.

Espectroscopía, Fluorescencia, Diseño óptimo

Citation: CABRERA-CABAÑAS, Reinier, LUNA-ROSAS, Francisco Javier, MARTINEZ-ROMO, Julio César and HERNÁNDEZ-VARGAS, Marco Antonio. Optimizing the removal of fluorescence and shot noise in Raman Spectra of minerals by ANFIS and moving averages filter. ECORFAN Journal-Taiwan. 2020. 4-8: 1-12

* Author correspondence (cabrera1988reinier@gmail.com)

† Researcher contributing as first author.

1. Introduction

Raman spectroscopy is one of the most used techniques in mining for the analysis and identification of compound. The geological characterization for example provides information about the formation of a site and its history. This process requires detailed morphological, physical, and compositional analyzes of rocks and sediments (Ishikawa & Gulick, 2013). Raman method is a high-resolution photonic technique that provides in a few seconds chemical and structural information of almost any organic and / or inorganic material or compound, allowing its identification.

The study by Raman spectroscopy is based on the analysis of the light scattered by a material by making a monochromatic light beam strike it, when this happens a small portion of the light is inelastically scattered, undergoing slight changes in frequency that are characteristic of the material analyzed and independent of the frequency of the incident light. This technique is currently being applied in multiple scientific areas, such as Physics and Chemistry, Biochemistry, virus detection (Yeh et al., 2020). However, one of the great problems of Raman spectra is that the Raman scattering (RS), which characterizes the composition of the sample, is accompanied by noise generated by the measuring instrument, external sources, and noise due to fluorescence. The latter may be orders of magnitude greater than the Raman signal, preventing obtaining information associated with its molecular composition. Therefore, it is necessary to eliminate the noise in the spectra before the analysis stage.

Noise removal is one of the most important data processing operations. Despite its wide use in various types of signals, there is no general strategy to carry out this procedure, since it largely depends on the problem treated, the signal-to-noise ratio (S / N) and the shape of the signals. Noise elimination process must be carried out with special care to avoid loss of information, and to adapt to the signal to be analyzed. To noise elimination, two different approaches have been used: the experimental and the computational. The methods that use the experimental approach are based on adjustments or improvements to the instrumentation and these include shifted excitation and time-limited systems (Gebrekidan et al., 2016).

Experimental methods like the previous ones are a little complex because they involve long acquisition times, for these reasons the use of computational methods has increased due to their speed, easy implementation, and low cost. Some computational methods that stand out include adjustment methods, highlighting the modified multipolinomial adjustment method and the Vancouver Algorithm (Lieber & Ahadevan-jansen, 2003; Zhao, Lui, Mclean, & Zeng, 2007), the methods based on wavelet transforms (Gebrekidan, Knipfer, & Braeuer, 2020) and the morphological algorithms (Javier et al., 2018).

In this work, ANFIS (Adaptive Neuro-Fuzzy Inference System), an integrating system between neural networks and fuzzy-logic that has previously been applied as an artificial intelligence tool in some areas such as Architecture, in the automotive industry, in Biochemistry and in Medicine (Übeyli, 2008) is used to characterize the contribution of fluorescence noise that is generated in Raman signals from different minerals, the procedure consists of developing an algorithm that allows to subtract the Raman peaks from the signal until a continuous signal at intervals is obtained, that signal will act as input to the diffuse neural network; the same, through an own adjustment system using the backpropagation error will adjust the background curve of the signal and filling in the empty spaces where the peaks were. This obtained signal will be assumed as the fluorescence background masking the signal and will be subtracted from the original signal. To eliminate small fluctuations that occur around the average value of the signal, moving averages filters are used which allow smoothing the signal, suppressing high-frequency noise.

This procedure ensures that the signal is clean of noise and, in a position to be correctly identified for future diagnostic and prediction procedures. Furthermore, in this article we demonstrate that it is possible to optimize the response time in the pre-processing of Raman signals of minerals, when we eliminate fluorescence and shot noise in large quantities of data, achieving an improvement of 53.60% in relation to the processing of data sequentially, which makes it a valuable tool in the field of mining for future applications of diagnosis and quick recognition of minerals but this can be used in other areas like medicine to quick prediction of diseases.

2. Materials and methods

2.1 (Adaptative Neuro-Fuzzy Inference System)

ANFIS (Adaptive Neuro Fuzzy Inference System) integrates Neural Networks with fuzzy logic inheriting the characteristics of both, allows you to tune or create the rule base of a fuzzy system using the backpropagation algorithm from the data collection of a process. It is an architecture functionally equivalent to a fuzzy rule system based on Takagi and Sugeno mode (Sugeno & Kang, 1988; Takagi & Sugeno, 1984).

The Neuro-Diffuse system is a traditional diffuse system in which each stage can be represented by a layer of neurons to which neural network learning capabilities can be provided to optimize the knowledge of the system. By having trainable parameters, the delta rule algorithms and backpropagation error are applicable (Jang, Sun, & Mizutani, 1997). Some parameters allow to establish the training set of the ANFIS system, and some common rules presented by the first order fuzzy model are necessary.

2.2 Moving averages

Moving Averages is a fairly simple prediction method that has been used in commerce and has not been altered for more than half a century (Soberón-celedón, Molina-contreras, Frausto-reyes, & Carlos, 2016). A window of size N is selected, and the mean or average of the variable for the N data is obtained, allowing the average to move as the new data of the variable in question are observed. This smoothes out possible strong oscillations or outliers.

The increase of any moving average depends exclusively on the shape of the function f and the size of the selected window. The mean movement of order N (MA_f) of a series f of values $Y1, Y2, Y3, ... Yn$ is defined by the sequence of values corresponding to the arithmetic means:

$$MA_f = (\frac{Y_1+Y_2+Y_N}{N}; \frac{Y_2+Y_3+Y_{N+1}}{N}; \frac{Y_3+Y_4+Y_{N+2}}{N}; ...) \qquad (1)$$

Where $Y_1, Y_2, Y_3, \dots, Y_N$ are the most recent observations of the closed interval; N is the size of the Interval within the function f.

2.3 Parallel Computing

Parallel computing is a form of computation in which many instructions are performed simultaneously, operating over the principle that big problems can often be divided into smaller ones, which are then solved simultaneously (in parallel).

The hardware that supports parallel computing consists of multicore computers, symmetric multiprocessors, distributed computers such as task clusters stations, and specialized parallel processors such as FPGA, GPU, and built in circuits of specific applications (AISC). With the development of hardware that supports parallel programming, especially the development of multicore computers, parallel programming architectures become more important than before (Gao, Kemao, Wang, Lin, & Seah, 2009).

The most common forms of parallelism, include: task parallelism, pipeline parallelism, and data parallelism (Gao et al., 2009). In the task parallelism, also known as functional parallelism, is a development structure in which, independent figures from parts of a method can be performed simultaneously in different processors. In the case of pipeline parallelism, the problem is separated in a series of tasks. Any of the tasks will be performed in a separate process or processor. Each parallel process is usually referred to as a pipeline state.

The exit as a pipeline state serves as the entrance of another state, therefore, in a given time each pipeline state is working over a different dataset. The data parallelism is mainly centered over the same process that will be applied simultaneously over different parts of a dataset. Let us say, similar operating sequences or functions are carried out in parallel over a large element data structure.

Moreover, if given enough parallel resources, the computing time of the data structure in parallel is usually independent from the problem size. One of the parallelism methods mentioned previously or their combination should be used in parallel applications.

3. Experiments

Raw Raman spectra were obtained from a complete set of high-quality spectral data from well characterized minerals shared in the project RRUFF. The database was prepared by collaborators from universities from different parts of the world and shared with the science. Several Raman spectrometers are used for to get RRUFF Raman samples but in this work were selected samples taken with a commercial Thermo Almega XR with 532 nm laser. 10500mineral samples were collected from different sources and universities around the world.

4. Results and Discussion

4.1 Description of the proposed algorithm for fluorescence and shot noise reduction

As previously mentioned, the Raman signal is made up of the useful signal, (characteristic of the molecular vibrations that occur inside the excited molecule) and a noise signal inherent in the measurement process, mainly high-frequency noise, and fluorescence background.

$$Y = X_{true} + b + n \qquad (2)$$

Where X_{true} is the Raman spectra free of noise and fluorescence, b is the background fluorescence, n is the noise in the signal and finally Y is the raw Raman signal acquired with the spectrometer containing fluorescence and shot noise. In such a way that to obtain a noise-free signal it is necessary to subtract the identified background fluorescence and shot noise from the raw signal.

The strategy followed is to design a mechanism to eliminate possible peaks in the raw Raman signal by detecting their start and end points. By suppressing these peaks, we will have a continuous signal at intervals such as the one shown in Figure 5, the objective of using ANFIS is to apply an interpolation to determine the possible shape of the fluorescence signal in the empty spaces of the signal, to finally subtract it from the measured spectrum. Figure1 shows a diagram highlighting each of the stages involved in the fluorescence removal process.

Stage 1. Load the Raman signal. In this step, the data vector is loaded from an Excel file, it contains two columns corresponding to the intensity values of the spectrum and the shift Raman respectively. In Figure 2 we have an example of a raw Raman spectrum corresponding to calcite.

Stage 2. Normalizing the data. This stage was performed in order to keep the signal under the same scale on the two coordinate axes, it was achieved through an interpolation algorithm, on the y-axis the maximum and minimum values of the spectrum were calculated, the minimum value is subtracted of the original spectrum and the result is divided by the maximum value, leaving the intensity values between 0 and 1, on the x-axis an arbitrary value of 574 values corresponding to the length of the data was taken. This is the best way to work with the signals because the signals would be on the same scale and the slope on each Raman signal will be easily located and very similar.

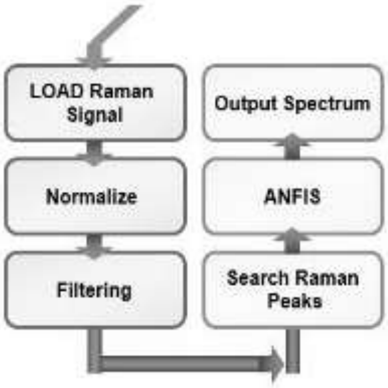

Figure 1 Program sequence designed to eliminate noise in Raman spectra

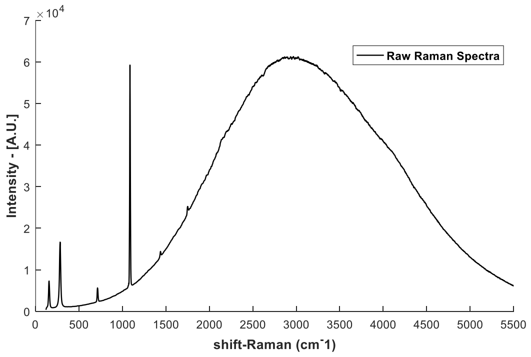

Figure 2 Raman signal of calcite ($CaCO_3$) without processing

Stage 3. Signal Filtering. A moving averages filter is applied with the purpose of smoothing the signal as a previous step to the correction of the baseline, the procedure is performed with a vector of fixed size previously defined and the central part of the result that is equal in size is rescued to the original data. With this method it is possible to overshadow the high frequency noise that affects the spectrum and dissolve the small peaks in the signal that can cause confusion in the interpretation of the data. Figure 3 shows the high frequency noise that was subtracted from the calcite Raman signal in the range of 1600 to 2400 cm^{-1}.

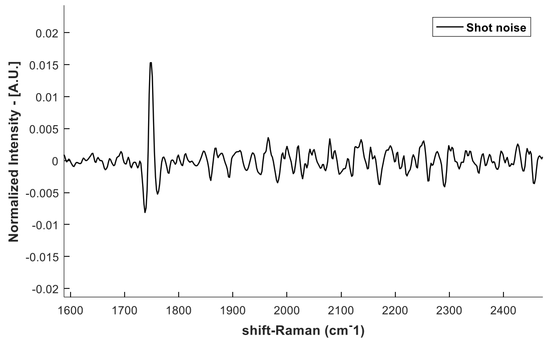

Figure 3 Shot noise removed from the Raman signal of calcite by moving averages Filter in the range of 1600 to 2400 cm^{-1}

Stage 4. Looking for Raman peaks candidates. With this objective, different algorithms were implemented that allow detecting all the signal peaks and subtracting them from the original signal, considering their starting and ending points.

The way this problem is solved is as follows:

1- Initially all the maximum points of the signal are discovered, for this it was necessary to implement a window algorithm with a window size equal to 15 samples, when we sampling the signal in this way the maximum values inside the window were found and assumed as possible Raman peaks if they were centered in the window; if they bow to one side they are discarded.

Through this method, we detect the portions of the spectrum that may have a peak shape and that could later be identified as legitimate Raman peaks.

2- Once all the possible signal peaks are obtained, the maximum values are taken for evaluation, keeping the same window size is evaluated point by point from the maximum value descending on both sides of the possible peak (to the right is increased by 1 unit and the left is decreased by one unit with respect to the x-axis), a least squares procedure is applied at each point to obtain the equation of the line that best describes the points corresponding to each window, the angle of inclination is calculated in each step respect to the abscissa and is compared taking as a criterion that the inclination of the peak is represented by an angle of more than 60º since for smaller angles they would make the appearance of the peak disappear and the structure of the signal would be affected with the introduction of morphological errors. The point where the condition is not met is taken as the starting and ending point of the peak.

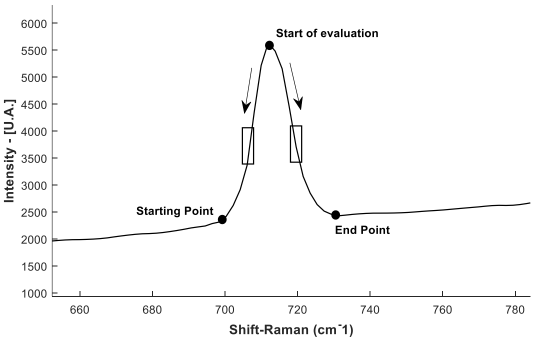

Figure 4 Evaluation to find the start and end points of the peak

In Figure 4 we can observe the procedure described above. We can see that the start and end points of the peak (points where the condition is not met) do not have the same height, and the rectangles try to represent the angle that is formed between the points with respect to the axis of the abscissa.

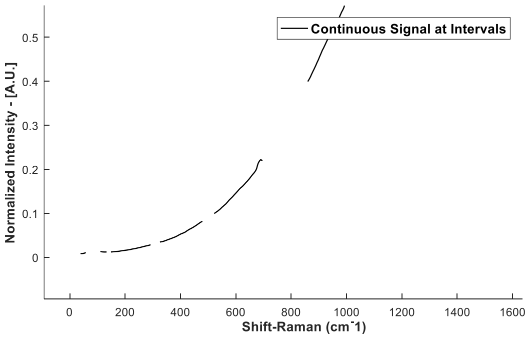

Figure 5 Raman signal continues at intervals with the peak regions removed

Stage 5. ANFIS, configuration developed. ANFIS is used at the junction of the points where each peak detected in the previous stage begins and ends. An adaptive network is constructed functionally equivalent to the fuzzy model Sugeno type whose scheme can be seen in Figure 6.

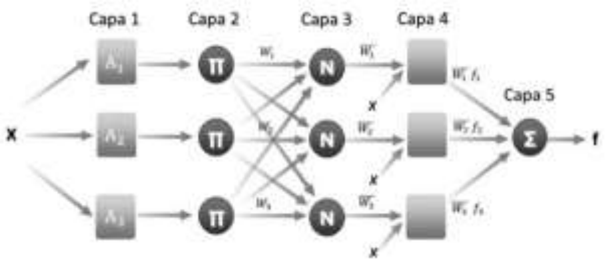

Figure 6 ANFIS network structure. One input, three rules one output

The procedure to apply ANFIS Will be explained step by step:

Preparation of the data. For explanatory purposes the vector containing the signal will be known as y or output and the vector containing the abscissa data will be known as x or input. Those definitions allow us to define the training set of the ANFIS system.

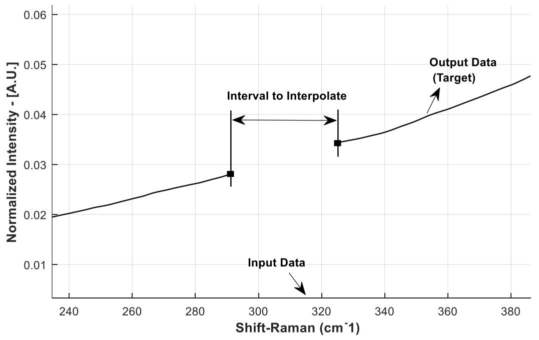

Figure 7 Portion of the spectrum selected for training

Training with ANFIS. To explain the training procedure a portion of the spectrum has been selected, the values required for training are the scale in which the Raman peaks do not appear on the x-axis as input and the amplitude y as the output variable or target that presents the Raman peak. In Figure 7 we can see a portion of the spectrum that we selected for illustrative purposes where the representation of the training vectors is framed on the x and y-axes. The scale we take on the x-axis goes from 941.7 cm^{-1} to 1051 cm^{-1} as the input variable and the amplitude y as the output or target variable.

In the mentioned interval there is a Raman peak in the interval between 962.4 cm^{-1} and 1032 cm^{-1}.

Forward propagation phase: Layer 1 (see Figure 6) receives vector X and calculates the degree of membership $\mu_{Ak}(X)$ of the fuzzy set for each of the values X_i of the vector according to the membership function A_k associated with each input; in this case a gaussian membership function with three trainable parameters

$$A_k = \text{gauss (x}, \sigma, \text{c)} = e^{-\left(\frac{x-c}{\sigma}\right)^2} \quad (3)$$

In layer 2 (layer π) the output of the nodes is the product of all the input signals, but in this case the antecedent is formed by a unique condition (if x is ...), therefore the output of this layer $w = \mu_{Ak}(X)$

In layer 3 (layer N) every element $w_{i,j}$ is normalized to the sum $w_{i,1}+w_{i,2}+w_{i,3}$

$$\overline{w_i} = \frac{w_i}{w_1+w_2} \; , i = 1, 2 \quad (4)$$

In the next stage of forward propagation layers 4 and 5 of the presented network structure are involved. The parameters (consequent parameters) p, q and r of the linear models that are weighted by the inputs are candidates for training in this phase and represent the inferred fuzzy output set. Every node in this layer is an adaptive node with a function:

$$\overline{w}_i f_i = \overline{w}_i(p_i(x) + q_i(y) + r_i) \quad (5)$$

In which f_i are those described in the rules of the fuzzy system (Sugeno & Kang, 1988). Finally, the defuzzification is carried out, the outputs are processed and integrated as a summation of all the input signals.

$$output = \sum \overline{w}_i f_i \quad (6)$$

The backpropagation algorithm uses the sum of the square error between the desired output and the output of the ANFIS system to adjust all the trainable parameters of layer 1 of the system. The error is back propagated from layer m to layer m-1; thus, from layer 5 to 4 there is an error signal of 5-4, from 4 to 3 the signal is 4-3, and so on, until the error signal 2-1 is reached.

The latter is the one used to adjust the nonlinear trainable elements of layer 1, which are the parameters σ and c of the fuzzy sets A_k. The described process of forward propagation and back propagation is performed iteratively up to a certain number of epochs (100 in this case) or until the error decreases to a specific value.

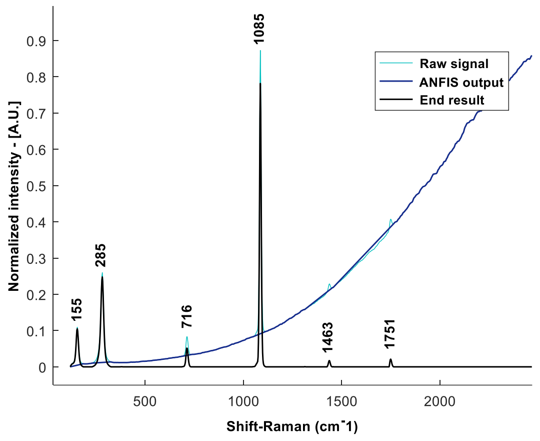

Figure 8 Raman spectrum corresponding to Calcite ($CaCO_3$). In cyan the Raman spectrum after the application of the moving averages filtering, in blue the fluorescence background rescued by the final adjustment of ANFIS and in black the noise-free Raman spectrum

Figure 8 shows the final result in three graphs; we can see the result of applying moving averages filtering and the ANFIS algorithm to eliminate the fluorescence and the shot noise signals and get a spectrum that allows us to make a prediction with the minimum error rate, the first one, of cyan color, shows the values of the Raman spectrum of $CaCO_3$ after the application of moving average filters, the blue one, shows the output values of ANFIS that translates as the fluorescence background of the analyzed spectrum and Finally, in black, the noise-free Raman spectrum after applying the subtraction of the spectra mentioned above. Raman spectra of the carbonates studied are well-known at ambient conditions.

The optical vibrations can be separated into internal vibrations of the CO_3 group (lying between 700 and 1500 cm^{-1}) and external or lattice vibrations involving translation and liberations of the CO_3 groups relative to the Ca atom (100-500 cm^{-1}). The vibrations of the "free" CO_3 group (vl: symmetric stretch; v3 asymmetric stretch; v2: out-of-plane bend; v4 in-plane bend) have been observed and assigned in Raman spectra of the carbonates of both rhombohedral and orthorhombic symmetry.

The site symmetry of the CO_3 group is determined by the cation environment and modifies the selection rules (i.e. the number of bands observed), and the frequencies to a small extent, when compared with the free group. In this figure we can observe the vibrational modes located in the low frequency range at 155, y 285 cm^{-1} that represent external or lattice modes, resulting from the interactions between Ca^{2+} and CO_3^{2-} ions. The high frequency vibrational modes at 716, 1085 and 1463 cm^{-1} represent internal modes of the CO_3^{2-} group(Gillet, Biellmann, Reynard, & McMillan, 1993).

4.2 Method Check

In this section, the efficiency of the method used to suppress noise in Raman spectra through ANFIS and moving averages filter will be verified. The raw spectra of some minerals such as Cerussite, Calomel, Hematite and Zircon will be used for this purpose. The spectra of these raw minerals are exposed, and the method described above is applied to each of them, obtaining similar results to those of the consulted literature. Each of the spectra with the most significant occurrences detected after denoising are detailed below.

Cerussite ($PbCO_3$)

The common simple pock-forming carbonates can be divided in three main groups: (1) the calcite group, (2) the dolomite group, and (3) the aragonite group. The cerussite ($PbCO_3$) is member of the aragonite group, is metastable under atmospheric conditions and therefore is less commonly found in nature than calcite. Aragonite is found in the calcareous skeletons of many organisms (e.g., shells of mollusks).

The molecular vibrations can be separated in (v$_{l:}$ symmetric stretch; v$_{3:}$ asymmetric stretch; v$_2$: out-of-plane bend; v$_{4:}$ in-plane bend) have been observed and assigned in Raman spectra, v$_1$ at 1054 cm^{-1} is the most obvious feature in this region. The FWHM of this band is much narrower that other combination bands.

This and the fact that there is a nearly constant shift between the satellite band and the main ν1 band in all aragonites suggest that combination bands are an unlikely explanation, in Figure 9 we can see v$_2$ at 837 cm^{-1} corresponding with the partial covalent bond between the lead on the covalent ions, v$_3$ at 1376 and 1476 cm^{-1} are attributed to the carbonate bending mode of cerussite, v$_4$ at 681cm-1 is attributed to the bending mode of the carbonate ion. The lattice vibrations band of cerussite are detected at 132 and 243 cm-1 corresponding to the lower frequencies(Martens, Rintoul, Kloprogge, & Frost, 2004).

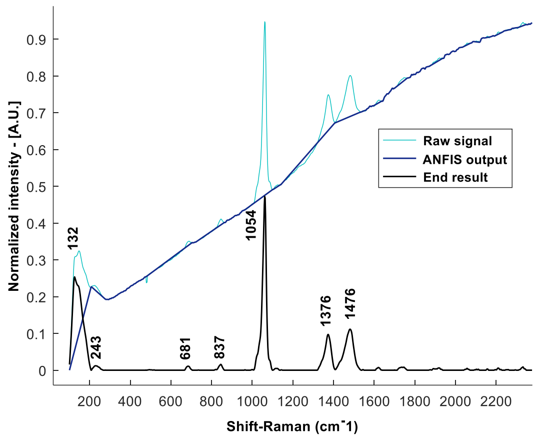

Figure 9 Raman spectrum corresponding to Cerussite (PbCO3). In cyan the Raman spectrum after the application of the moving averages filtering, in blue the fluorescence background rescued by the final adjustment of ANFIS and in black the noise-free Raman spectrum

Zircon ($ZrSiO_4$)

Zircon ($ZrSiO_4$) is an unusually common and widely distributed mineral, but the crystals are rare. The structure of zircon contains Si in tetrahedral coordination by oxygen, and Zr in 8-fold coordination by oxygen in the form of triangular-faced dodecahedra, these structural features indicate that there is a strong repulsive interaction between the neighboring Zr and Si atoms. In Figure 10 we can see the Raman spectrum of the Zircon after ANFIS and moving averages filter, the wavenumbers of all the phonon modes in zircon are shown 130 cm^{-1} represent a rigid rotation of the SiO4 , the 223 cm^{-1} is a rigid rotation of the SiO_4 around the a-axis, shearing of the Zr, 341cm^{-1} is a rigid rotation of the SiO_4 around the a-axis, shearing of the Zr, in 437cm^{-1} there are a Symmetric flattening of the SiO_4 along the c-axis, and in 1008 cm-1 have a stretching Si of Si–O that is anti-symmetric with respect to the O_{sh}–O_{sh} edge of SiO_4 .

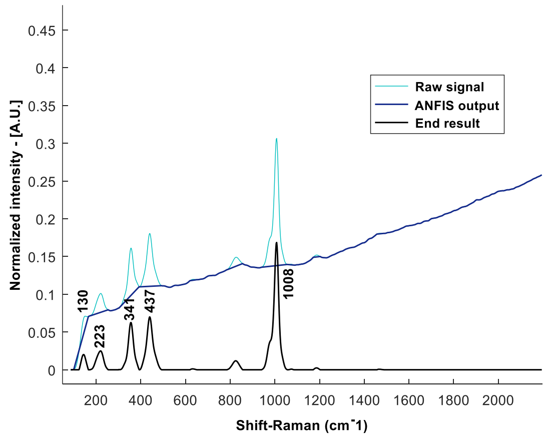

Figure 10 Raman spectrum corresponding to Zircon (ZrSiO4). In cyan the Raman spectrum after the application of the moving averages filtering, in blue the fluorescence background rescued by the final adjustment of ANFIS and in black the noise-free Raman spectrum

Calomel (Hg_2Cl_2)

Calomel is a mercury (I) chloride mineral with formula Hg_2Cl_2, mainly used nowadays as a component of reference electrodes in electrochemistry, this compound was a widespread and popular medicine until it fell out of use at the endo of 19th century due to its toxicity, and a material caller mercury white is referred to in 16th century technical literature on painting. The spectrum showed in Figure 11 Reveal two lines whose polarization correspond to fully symmetric vibrations at 167 and 275 cm^{-1}, matching those of a reference spectrum of calomel (Crippa, Legnaioli, Kimbriel, & Ricciardi, 2020).

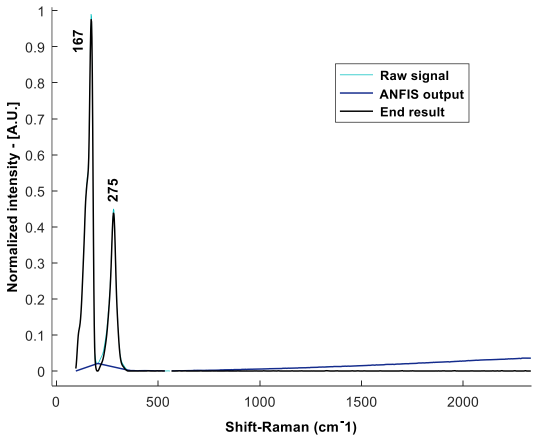

Figure 11 Raman spectrum corresponding to Calomel (Hg2Cl2). In cyan the Raman spectrum after the application of the moving averages filtering, in blue the fluorescence background rescued by the final adjustment of ANFIS and in black the noise-free Raman spectrum

Hematite (Fe_2O_3)

The hematite (Fe_2O_3) is a mineral in felsic igneous rocks, a late-stage sublimate in volcanic rocks, and in high-temperature hydrothermal veins. A product of contact metamorphism and in metamorphosed banded iron formations. A common cement in sedimentary rocks and a major constituent in oolitic iron formations. Abundant on weathered iron-bearing minerals.
In the Raman spectrum of the hematite showed in Figure 12 are expected seven phonon lines, namely two A_{1g} modes (225 and 498 cm^{-1}) and five E_g modes (246, 299, 410, 497 and 613 cm^{-1}) Hematite is an antiferromagnetic material and the collective spin movement can be excited in what is called a magnon. The intense feature at 1320 cm^{-1} is assigned to a two-magnon scattering which arises from the interaction of two magnons created on antiparallel close spin sites(De Faria, Venâncio Silva, & De Oliveira, 1997).

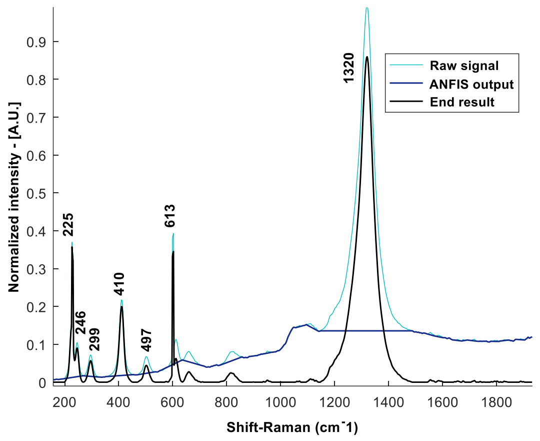

Figure 12 Raman spectrum corresponding to Hematite (Fe2O3). In cyan the Raman spectrum after the application of the moving averages filtering, in blue the fluorescence background rescued by the final adjustment of ANFIS and in black the noise-free Raman spectrum

4.3 Optimum design in the removal of fluorescence and shot noises in Raman spectrum of minerals

Today, data is in all kinds of formats; from traditional databases to hierarchical data stores created by end users, through OLAP systems, text documents, email, measurement data, signal data, video, audio, stock information and financial transactions, among many others. According to some calculations, 80% of the data of the organizations are not numerical (Prajapati, 2013). However, these should also be included in the analysis and decision-making process.

Speed designates how quickly data is generated and how fast it must be processed to satisfy demand. In this section we show how we can optimally preprocess high-volume of Raman signals from mineral examples (data parallelism) by applying MATLAB multicore technology (Cartagena, Juan, Rodríguez, & Autor, 2010). As we have seen, there are many sources of noise that attack the weak Raman signal. In order to achieve material identification, it is essential to have a good signal-to-noise ratio in the Raman spectrum.

Currently, there are different methods implemented to combat these imperfections in Raman spectra, experimental methods such as shifted excitation and computational methods such as morphological filtering (Mª José Tosina Muñoz, n.d.) and polynomial algorithms are used to suppress noise from high and low frequency highlighting the advantage of seconds in terms of low cost and ease of implementation, however, the implementation of these methods involves computation time especially when we preprocess large amounts of Raman signals, for this reason we decided to use Parallel computing with multicore technology to optimize the response time of the preprocessing of large volumes of Raman spectroscopic signals in samples of minerals.

In Table 1 we can observe the sequential and parallel processing time what takes for the suppression of fluorescence and shot noise in Raman spectra of minerals. On this occasion we use our entire database to consult the processing time in large volumes of data. It was used the moving averages filtering to smooth the signal as a previous step to operations of the developed ANFIS algorithm, which provides a close baseline in the regions where there are Raman bands which are subtracted from the raw Raman spectrum leaving the signal in the base band. To eliminate the shot noise on signal f, we use a moving average filter with a window size of N = 7, guaranteeing smoothing of the signal without damaging the identifying characteristics of the spectrum.

The arithmetic mean in this case is calculated as:

$$MA = \frac{\sum_{i=1}^{N}(f)}{N} \quad (7)$$

In equation (7), N is the base of moving averages. Although there is no specific rule on how to select the bases for moving averages (N), it is recommended that N be large when the behavior of the data is stable over time. Conversely, if the variable shows changing patterns; it is recommended to use a small value of N. In practice, values for N between 2 and 10 are normal.

Raman Spectra	**ANFIS Algorithm and Moving Averages Filter**	
Mining Examples	Sequential Process (Seg)	Parallel Process (Seg)
10	4.064	3.611
100	41.708	18.644
300	122.304	49.286
500	198.434	79.698
700	276.138	110.399
900	353.525	142.045
1100	428.817	178.942
1300	526.259	222.955
2600	996.946	425.122
5200	2016.7	911.549
10400	4106.3	1817.77

Table 1 Sequential and parallel processing times in suppression of fluorescence and shot noise in Raman spectra of minerals

As we can see in **¡Error! No se encuentra el origen de la referencia.**, we started with 10 spectra of samples of mineral, we continued with increments until we achieve the suppression of fluorescence and shot noise of 10400 Raman spectra of different minerals.

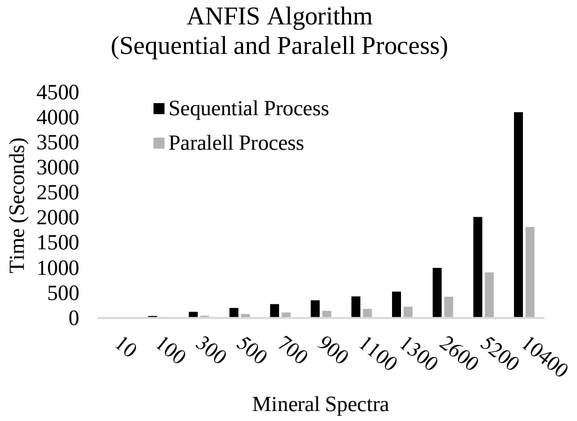

Figure 13 Sequential and Parallel Processing Time using the ANFIS algorithm to eliminate the fluorescence background and moving averages filtering for high frequency noise (shot noise)

In the graph in ***¡Error! No se encuentra el origen de la referencia.***, we clearly observe that we obtain an improvement in the processing time of the data for the elimination of fluorescence and high frequency noise when we implement the ANFIS algorithm and the filtering of moving averages with multicore technology to the set of mineral spectra (data parallelism), we can clearly observe the significant reduction in processing time with a gain of approximately 53.60%.

5. Acknowledgments

The authors thank the collaborators of the RRUFF Project for sharing their database with the scientific community for the benefit of science.

6. Conclusions

A Raman spectrum is a fingerprint of the material being analyzed since it is composed of (a) Raman scattering (RS), which characterizes the molecular composition of the sample through the position of the Raman peaks described by the wave number; there are also numerous disturbances that are added to the spectrum during the measurement process such as: (b) fluorescence noise, and (c) shot noise, sometimes these are several orders of magnitude greater than the useful signal so that could be masked making it difficult to appreciate correctly, for this reason it must be eliminated.

We used an own algorithm based on ANFIS (Adaptive Neuro Fuzzy Inference System) to reveal the fluorescence background of the spectra and the filtering of moving averages to eliminate the shooting noise; both disturbances, causing the masking of the data and the difficult appreciation of its useful content. This preprocessing takes considerable computation time when we process large amounts of Raman spectroscopic signals.

In this work we have shown that it is possible to optimize the preprocessing time of large volumes of Raman spectroscopic signals in samples of minerals through parallel processing that consists of dividing the tasks to be performed in a multicore environment. This optimized method can have specific applications in the field of medicine or industry since it guarantees to carry out diagnostic and classification applications in a considerably small time.

References

Cartagena, U. P. D. E., Juan, A., Rodríguez, F., & Autor, E. (2010). *Programación Matlab En Paralelo Sobre Clúster Computacional : Evaluación De Prestaciones*.

Crippa, M., Legnaioli, S., Kimbriel, C., & Ricciardi, P. (2020). New evidence for the intentional use of calomel as a white pigment. *Journal of Raman Spectroscopy*, (March), 1–8. https://doi.org/10.1002/jrs.5876

De Faria, D. L. A., Venâncio Silva, S., & De Oliveira, M. T. (1997). Raman microspectroscopy of some iron oxides and oxyhydroxides. *Journal of Raman Spectroscopy*, *28*(11), 873–878. https://doi.org/10.1002/(sici)1097-4555(199711)28:11<873::aid-jrs177>3.0.co;2-b

Gao, W., Kemao, Q., Wang, H., Lin, F., & Seah, H. S. (2009). Parallel computing for fringe pattern processing: A multicore CPU approach in MATLAB® environment. *Optics and Lasers in Engineering*, *47*(11), 1286–1292. https://doi.org/10.1016/j.optlaseng.2009.04.018

Gebrekidan, M. T., Knipfer, C., & Braeuer, A. S. (2020). Vector casting for noise reduction. *Journal of Raman Spectroscopy*. https://doi.org/10.1002/jrs.5835

Gebrekidan, M. T., Knipfer, C., Stelzle, F., Popp, J., Will, S., & Braeuer, A. (2016). A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference. *Journal of Raman Spectroscopy*. https://doi.org/10.1002/jrs.4775

Gillet, P., Biellmann, C., Reynard, B., & McMillan, P. (1993). Raman spectroscopic studies of carbonates part I: High-pressure and high-temperature behaviour of calcite, magnesite, dolomite and aragonite. *Physics and Chemistry of Minerals*, *20*(1), 1–18. https://doi.org/10.1007/BF00202245

Ishikawa, S. T., & Gulick, V. C. (2013). An automated mineral classifier using Raman spectra. *Computers and Geosciences*, *54*, 259–268. https://doi.org/10.1016/j.cageo.2013.01.011

Jang, J.-S. R., Sun, C.-T., & Mizutani, E. (1997). Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. In *Prentice Hall*.

Javier, F., Rosas, L., Cesar, J., Romo, M., Veloz, G. M., Contreras, J. R. M., … México, A. (2018). *Optimal Design in the Removal of Fluorescence and Shot Noise in Raman Spectra from Biological Samples*. 78–84.

Lieber, C. A., & Ahadevan-jansen, A. M. (2003). *Automated Method for Subtraction of Fluorescence from Biological Raman Spectra*. *57*(11), 1363–1367.

M^a José Tosina Muñoz, R. P. P. (n.d.). *Filtro Morfológico, eliminacion de fluorescencia*.

Martens, W. N., Rintoul, L., Kloprogge, J. T., & Frost, R. L. (2004). Single crystal raman spectroscopy of cerussite. *American Mineralogist*, *89*(2–3), 352–358. https://doi.org/10.2138/am-2004-2-314

Prajapati, V. (2013). *Big Data Analytics with R and Hadoop*. Retrieved from http://books.google.com/books?hl=en&lr=&id=8eotAgAAQBAJ&oi=fnd&pg=PT12&dq=Big+Data+Analytics+with+R+and+Hadoop&ots=vdKhha6hNe&sig=umigf1-1Rbqs-d-Bp5itHKxNEJA

Soberón-celedón, C. C., Molina-contreras, J. R., Frausto-reyes, C., & Carlos, J. (2016). *Removal of fluorescence and shot noises in Raman spectra of biological samples using morphological and moving averages filters*. *0869*(3), 14–19.

Sugeno, M., & Kang, G. T. (1988). Structure identification of fuzzy model. *Fuzzy Sets and Systems*, *28*(1), 15–33. https://doi.org/10.1016/0165-0114(88)90113-3

Takagi, T., & Sugeno, M. (1984). Derivation of Fuzzy Control Rules From Human Operator'S Control Actions. *IFAC Proceedings Series*, *16*(13), 55–60. https://doi.org/10.1016/S1474-6670(17)62005-6

Übeyli, E. D. (2008). Adaptive neuro-fuzzy inference system employing wavelet coefficients for detection of ophthalmic arterial disorders. *Expert Systems with Applications*, *34*(3), 2201–2209. https://doi.org/10.1016/j.eswa.2007.02.020

Yeh, Y. T., Gulino, K., Zhang, Y. H., Sabestien, A., Chou, T. W., Zhou, B., … Terrones, M. (2020). A rapid and label-free platform for virus capture and identification from clinical samples. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(2), 895–901. https://doi.org/10.1073/pnas.1910113117

Zhao, J., Lui, H., Mclean, D. I., & Zeng, H. (2007). Automated autofluorescence background subtraction algorithm for biomedical raman spectroscopy. *Applied Spectroscopy*, *61*(11), 1225–1232. https://doi.org/10.1366/000370207782597003