Functional data analysis as a tool to correlate textural and geochemical data

C. Ordóñez a,*, C. Sierra a, T. Albuquerque b, J.R. Gallego a

a Department of Mining Exploitation and Prospecting, Polytechnic School of Mieres, University of Oviedo, Campus de Mieres, Gonzalo Gutiérrez, s/n, 33600 Mieres, Asturias, Spain
b IPCB-CVRM, Polytechnic Institute of Castelo Branco, IPCB, Geo-Systems Centre, ISTUTL, Lisbon, Portugal

A R T I C L E   I N F O
Keywords:
Functional data analysis
Granulometric curve
Element concentration

A B S T R A C T
This paper discusses the use of functional data analysis to determine the interactions between the chemical composition and grain size. The proposed method is compared for a case study with others that involve the determination of the coefficient of correlation between the concentration of each chemical element in a bulk sample and representative points in the granulometric curve. The results obtained by the two approaches are consistent and indicate an increasing concentration of certain metals in the fine particle-size fractions. Moreover, functional data analysis provides useful additional information on the relationship between the concentration of the chemical elements in bulk samples and particle size distribution.

1. Introduction

Grain-size distribution can offer information about several features of sediment, thus providing evidence of sedimentary origin (allochthonous, autochthonous, anthropogenic), transport conditions (rolling, creeping, saltation, flotation, dissolution, suspension etc.), or even chemical composition [1,2].

Sediment texture can be correlated with its chemical composition and genesis since the physical and chemical weathering of rocks combined with selective processes of segregation during transport and deposition cause the accumulation of particular materials in certain grain-size fractions [3,4]. Thus, during grain-size fractionation coarser fractions commonly have a higher concentration of SiO$_2$ or calcium oxide, while the finest fraction may be enriched in alkalies such as Na and K, as well as Al$_2$O$_3$, MgO and TiO.

Similarly, there is a close relationship between decreasing grain size and pollutant enrichment [5]. This occurs because the surface per unit area increases with smaller sizes, thus sediments tend to retain metals and other contaminants with surface-type chemical reactions, and smaller grain sizes are more chemically active [6,7]. As a result, some recovery and remediation techniques exploit heavy metal enrichment within the fine fractions [8].

Various parameters, such as particle size and the concentration of certain chemical elements like Al, Fe, and Ti (which are used as proxies for grain size because they tend to co-vary with it), are commonly used in geochemistry as normalization elements in order to compensate for the variability generated by mineralogical composition and to facilitate comparisons between areas with different grain sizes [9,10]. These elements are characterized by presenting a notable linear correlation...
between grain size and concentration. If normalization is not carried out, erroneous conclusions can be reached, attributing high concentrations of pollutants to an anthropogenic source, when they are in fact due to grain size.

In the absence of grain-size variations, geochemical data of siliciclastic sediments may be used to characterize size-independent processes as sediment provenance, weathering, mixing or diagenesis [11].

For all the reasons above, it is of interest to determine the chemical composition of distinct grain-size fractions. In this regard, the standard procedure consists of a sieving (size fractionation of the soil) and subsequent chemical analysis of each fraction, but this is a laborious procedure and it favors contamination during manipulation and exponentially increases the analytical costs. For this reason, it is pertinent to establish correlations between particle size and chemical composition of the bulk sample. This procedure is based on the use of descriptive parameters of the particle size distribution, such as the mean or median size of grain [12], or parameters such as the quartile diameters (d75, d50, and d25), others like d10, d50, d90 (grain size below which 10%, 50%, 90% of the grains are smaller than these sizes) and even the sand, clay and silt content. Such studies are almost always performed using Pearson’s correlation coefficient [13,14], and can be supported by other techniques such as principal component analysis and other methods of factorial analysis [15].

However, the disadvantage of using granulometric parameters in the form of discrete variables is common to all these procedures, and a great deal of information held in the granulometric curve is wasted, particularly when these data have been obtained by modern methods of laser diffraction and the function is defined by dozens of points.

Here we discuss and compare these methods of correlation between the grain-size and chemical composition of the bulk sample with other approaches based on functional data analysis. Functional data analysis has recently been applied to address other kinds of problems, such as environmental pollution [16], GPS measurement in forest environments [17], and the analysis of multispectral data [18], among others.

The usefulness of the proposed method is tested for alluvial detrital sediment sampled in an area notably affected by heavy metal pollution. Here we consider its applicability for selecting normalizer elements, as well as its efficacy for identifying and describing the behavior of contaminants and natural allochthonous inputs.

2. Materials and methods

2.1. Data collection and chemical analyses

Samples were collected in Avilés Estuary (North of Spain), where heavy industrial activity has contaminated the sediments during several decades [15,19].

A total of 54 bed-deposited sediment samples from the upper 0–20 cm beneath a water depth of about 20 cm were collected in the Llodero Cove area in January 2012. Samples were collected using a stratified systematic sampling method at short, medium and long distances in order to obtain a representative picture of the total variability of the variables. Fig. 1 shows the spatial distribution of the sample points.

The samples were packed and sealed in pre-washed polyethylene bags and dried at room temperature. They were then fractionated into two sub-samples in order to determine their composition and grain-size distribution.

Chemical analyses were performed by Inductively Coupled Plasma – Optical Emission Spectroscopy after sample leaching by means of an ‘Aqua regia’ digestion (HCl + HNO3).

Grain-size distribution was obtained using a Laser Dispersion Particle Analyser (Beckman Inc. Coulter) after subjecting samples to a disaggregation treatment with two dispersants, namely sodium hexametaphosphate and sodium carbonate, and eliminating the organic matter with hydrogen peroxide and acetic acid.
2.2. Mathematical model

The aim of this study is to determine the relationship between the concentration of pollutants in a soil and the grain size of each sample. This relationship can be approached as a linear regression problem:

\[ E(Y|X) = \alpha + \sum_{i=1}^{p} \beta_i x_i = \alpha + \langle X, \beta \rangle \]  \hspace{1cm} (1)

where \( Y \) is the response variable, \( \alpha \) and \( \beta = (\beta_1, \ldots, \beta_p)^T \) are the regression coefficients and \( X = (x_1, \ldots, x_p)^T \) the p-dimensional input variable.

In the case analyzed, the dependent variable is a vector of real numbers, representing the concentration of a particular pollutant in the different samples, while the independent variable for each sample is given by a discrete set of values for an increasing sequence of grain sizes (so-called grain-size curve of the sample).

A functional linear regression model [20] for the particular case of scalar response and functional predictors is an extension of the multivariate linear regression model to the case of infinite-dimensional data. Then Eq. (1) is written as follows:

\[ E(Y(s)|X) = \alpha + \int_t X(s)\beta(s)ds = \alpha + \langle X, \beta \rangle \]  \hspace{1cm} (2)

where coefficient \( \alpha \) is a scalar and the input variable \( X : T \rightarrow \mathbb{R} \) and the regression coefficients \( \beta : T \rightarrow \mathbb{R} \) are represented as functions.

Several approaches can be used to estimate the functions \( X(s) \) and \( \beta(s) \) from discrete points, smoothing by means of the decomposition in basis functions being one of the most frequent in functional data analysis [21]:

\[ X(s) = \sum_{k=1}^{n} c_k \phi_k = c^T \phi(s) \]

\[ \beta(s) = \sum_{p=1}^{m} b_p \psi_p = b^T \psi(s) \]  \hspace{1cm} (3)

where \( c \) and \( b \) are vectors of coefficients, and \( \phi(s) = (\phi_1(s), \ldots, \phi_n(s))^T \) and \( \psi(s) = (\psi_1(s), \ldots, \psi_m(s))^T \) are the basis functions. These functions can be of a different type, such as polynomial, exponential, B-splines, Fourier functions, etc.

By substituting expressions in (3) in Eq. (2), an estimation of \( Y \) can be expressed as follows:

\[ \hat{Y} = \hat{\alpha} + c^T J_{\psi\phi} b \]  \hspace{1cm} (4)

where the \( n \times m \) matrix \( J_{\psi\phi} \) is given by:

\[ J_{\psi\phi} = \int_T \psi(s)\phi(s)ds = \langle \psi, \phi \rangle \]  \hspace{1cm} (5)

When the basis functions are orthonormal, such as with Fourier functions, the matrix \( J_{\psi\phi} \) results in identity.

Introducing the notation \( \xi = (\alpha, b_1, \ldots, b_m) \), \( Z = [1, c^T J_{\psi\phi}] \), the prediction \( \hat{Y} \) becomes simply

\[ \hat{Y} = Z\xi \]  \hspace{1cm} (6)

The estimation of vector \( \xi \) is obtained by minimizing the residual sum of squares

\[ \text{LMMSE}(\alpha, \beta) = \| Y - \alpha - \langle X, \beta \rangle \|^2_{L^2(T)} \]  \hspace{1cm} (7)

using the least squares method, as follows:

\[ \hat{\xi} = (Z^T Z)^{-1} Z^T Y \]  \hspace{1cm} (8)

However, minimizing the square of the residuals is not the only aim of the estimation. It is also often necessary to avoid excessive local fluctuation in the estimated function. Regularization with roughness penalty can be used for this purpose. To this end, given any twice differentiable function \( \beta \), it is possible to define the penalized residual sum of squares:

\[ \text{PENSSE}(\alpha, \beta) = \left( Y - \alpha - \int_T X(s)\beta(s)ds \right)^2 + \lambda \int_T |D^2 \beta(s)|ds \]  \hspace{1cm} (9)

where the operator \( D^2 \) represents the second derivative, and the smoothing parameter \( \lambda > 0 \) controls the trade-off between roughness and infidelity to the observed data [20].

The solution to the regularization problem with roughness penalty is analogous to that of Eq. (8):

\[ \hat{\xi} = (Z^T Z + \lambda R)^{-1} Z^T Y \]  \hspace{1cm} (10)

being \( R \) a \( m \times m \) matrix with elements \( R_{jk} = \int_T D^2 \psi_j(s)D^2 \psi_k(s)ds = \langle D^2 \psi_j, D^2 \psi_k \rangle_{L^2(T)} \)

The smoothing parameter \( \lambda \) can be chosen either subjectively or by means of cross-validation. When the latter approach is used, a cross-validation score can be defined as:
CV(\(\lambda\)) = \sum_{j=1}^{N} \left( y_j - x_j^{(j)} - \int_{\Omega} X_j(s)b_j^{(j)} ds \left/ C_0 \right. \right)^2 \tag{11}

\(N\) being the sample size and \(x_j^{(j)}\) and \(b_j^{(j)}\) the estimates of \(a\) and \(b\) obtained by minimizing the penalized residual sum of squares based on all the data except \((x_j,y_j)\). The smoothing parameter \(\lambda\) is obtained by minimizing \(CV(\lambda)\).

3. Results and discussion

3.1. Sample descriptive statistics and Pearson's correlation

The functional regression model discussed in the previous section was applied to the data set taken in the field. Each sample has a concentration value for each chemical element and the corresponding grain curve (differential curve) in a set of

<table>
<thead>
<tr>
<th>Minimum (mg/kg)</th>
<th>Maximum (mg/kg)</th>
<th>Mean (mg/kg)</th>
<th>Stand. deviation (mg/kg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>20.00</td>
<td>75.00</td>
<td>32.06</td>
</tr>
<tr>
<td>Ti</td>
<td>2.00</td>
<td>16.00</td>
<td>4.09</td>
</tr>
<tr>
<td>Fe</td>
<td>65.00</td>
<td>775.00</td>
<td>131.63</td>
</tr>
<tr>
<td>Cd</td>
<td>5.00</td>
<td>137.00</td>
<td>35.91</td>
</tr>
<tr>
<td>Ca</td>
<td>20.00</td>
<td>320.00</td>
<td>81.13</td>
</tr>
<tr>
<td>Na</td>
<td>1.00</td>
<td>103.00</td>
<td>29.02</td>
</tr>
</tbody>
</table>

| d_{10} | 40.25 | 220.87 | 141.69 | 46.67 |
| d_{50} | 142.53 | 311.23 | 230.82 | 27.35 |
| d_{90} | 221.30 | 424.06 | 319.89 | 32.10 |

Table 2

| Pearson's correlation coefficient between pollutants concentration and representative parameters of the granulometric curve. |
|-----------------|-----------------|--------------|-----------------|-----------------|-----------------|
| d_{10}          | -0.796**        | -0.380       | -0.289          | -0.781**        | -0.405**        |
| d_{50}          | -0.725**        | -0.268       | -0.145          | -0.558**        | -0.416**        |
| d_{90}          | 0.562*          | -0.188       | -0.090          | -0.400**        | -0.278**        |

* The correlation is significant at 0.05 (bilateral).

** The correlation is significant at 0.01 (bilateral).
discrete grain sizes. The whole grain size curves were obtained by smoothing using 15 cubic B-splines, as shown in Fig. 2. The Y axis represents the percentage volume of the soil sample for each grain size.

The parameters studied were the concentrations of the elements Al, Ti, Fe, Cd, Ca, and Na, the grain size parameters d_{10}, d_{50}, d_{90}, and the mean grain size. The main statistical descriptors related to the chemical characterization and particle size are shown in Table 1. Textural characterization obtained by means of GRADISTAT v 10.0 classified the sediments as sands and muddy sands [22].

**Fig. 3.** Regression coefficient functions for the 6 chemical elements studied.
The relationships between element concentration and grain sizes were studied by means of Pearson’s correlation coefficient. Table 2 shows the correlation coefficients and their significance levels for parameters $d_{10}$, $d_{50}$ and $d_{90}$. It can be observed that the correlation coefficients are negative, which is consistent with the idea that the smaller the grain size, the greater the concentration of certain elements, and particularly that of contaminants.

Parameter $d_{10}$ (10% of the sediment is smaller than this size), which represents the smaller grain sizes, was often better correlated with the chemical composition than $d_{50}$ (median size grain) and $d_{90}$ respectively, except for Ca and Na, for which the correlation coefficient with the elements increased from $d_{10}$ to $d_{50}$ and then decreased again for $d_{90}$.

### 3.2. Functional data analysis

Conversely, functional analysis offers a continuous representation of the regression coefficients for the different particle sizes, together with the 95% confidence lines (Fig. 3). For all the cases, the regression coefficients ($\beta(s)$) were negative, and the higher, in absolute value, the grain size, the lower the coefficient (except for the Ca). This observation indicates that the influence of the granulometric curve, or what is the same, the effect of the volume percentage of a given grain size in the concentrations of the elements, is higher in the finer fractions. Table 3 lists the coefficients of determination obtained for each chemical element.

The optimum penalty factors for the 6 models were obtained by minimizing the cross-validation score, following Eq. (11). Fig. 4 shows the variation of this score for different values of the penalty coefficient corresponding to Cd. The minimum value corresponds to $\lambda = 10^{10}$. For the 5 remaining elements, similar optimum penalty factors were obtained. This result is coherent with the notion that a higher coefficient of determinations requires lower penalization factors, thus producing wiggle shapes of the regression coefficient with no physical interpretation.

For the penalty factors determined, the highest coefficient of determination (RSQ) between grain size and composition was observed for Al (RSQ = 0.72), which is consistent with the common use of this metal as a normalizer [23,24]. Cd also showed high correlations between composition and grain size (RSQ = 0.59), thus corroborating a clear relationship between increasing Cd concentrations with smaller grain sizes. The presence of Cd in the Avilés area is very low in the geologic substrate; however, there is a notable anthropogenic input of this metal promoted by the emissions of small grain-size particles from the metallurgical industry [15].

Ti, which represents silicate minerals, showed a less clear correlation with the grain size (RSQ = 0.36) than the former metals, thus suggesting that its use as a normalizer is less desirable than Al. In the same way, Fe, which is also frequently used as a normalizer [24,25], correlated poorly with the particle size (RSQ = 0.19). This finding is coherent with the fact that the use of this element as a normalizer can be compromised in anoxic environments, such as that of these sediments [26], and with the non-significant correlation obtained using Pearson’s correlation coefficient.

<table>
<thead>
<tr>
<th>Element</th>
<th>Cd</th>
<th>Al</th>
<th>Ca</th>
<th>Fe</th>
<th>Ti</th>
<th>Na</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSQ</td>
<td>0.59</td>
<td>0.72</td>
<td>0.48</td>
<td>0.19</td>
<td>0.36</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Fig. 4. Variation of the cross-validation score with the penalty factor.
Elements such Ca and Na showed a regression coefficient ($r(s)$) with a different shape. Ca presented a non-significant correlation coefficient for the extreme grain-size values and negative shape for the intermediate ones. This irregularity can be attributed to the dismantling of the limestone bedrock in the vicinity, as well as to the multiple inputs produced by the remains of bivalves in the area (biogenic origin). The different behavior of Na – shown in the plot as a straight line – may be the consequence of sea salt contribution to sediment samples. This anomalous behavior for Ca and Na was also detected using Pearson’s correlation coefficient, which did not increase with the decreasing grain size.

4. Conclusions

We propose functional data analysis as an alternative approach to study the relationship between the concentrations of chemical elements and grain-size since it offers a continuous representation of the correlation between these two variables, unlike conventional procedures that focus only on discrete points of the granulometric curve.

The procedure tested for a marine detrital sediment showed similar behavior for the 6 elements studied in terms of the correlation between concentration and grain-size distribution, except for Ca, which showed a relationship between concentration and particle size distribution curve that was not significant in the extreme grain sizes but in the intermediate sizes. Al and Cd concentration showed higher correspondence with the granulometric curve, while Fe showed the lowest association. These results are consistent with those obtained by discrete parameters representing the particle size.

Acknowledgement

Carlos Sierra obtained a grant from the “Severo Ochoa” programme (Ficyt, Asturias, Spain).

References


