# A coupled multivariate statistics, geostatistical and machine-learning approach to address soil pollution in a prototypical Hg-mining site in a natural reserve.

**C. Boente[a], M.T.D. Albuquerque[b], S. Gerassis[c], E. Rodríguez-Valdés[a], J.R. Gallego[a]**

[a] INDUROT and Environmental Technology, Biotechnology, and Geochemistry Group, Universidad de Oviedo, Campus de Mieres, 33600 Mieres (Asturias), Spain

[b] Instituto Politécnico de Castelo Branco, 6001-909 Castelo Branco, Portugal and CERENA/FEUP Research Center, Portugal

[c] Department of Natural Resources and Environmental Engineering, Univ. of Vigo, Lagoas Marcosende, 36310 Vigo, Spain

## Abstract

The impact of mining activities on the environment is vast. In this regard, many mines were operating well before the introduction of environmental law. This is particularly true of cinnabar mines, whose activity has declined for decades due to growing public concern regarding Hg high toxicity.

Here we present the exemplary case study of an abandoned Hg mine located in the Somiedo Natural Reserve (Spain). Until its closure in the 1970s, this mine operated under no environmental regulations, its tailings dumped in two spoil heaps, one of them located uphill and the other in the surroundings of the village of Caunedo. This study attempts to outline the degree to which soil and other environmental compartments have been affected by the two heaps. To this end, we used a novel combination of multivariate statistical, geostatistical and machine-learning methodologies. The techniques used included principal component and clustering analysis, Bayesian networks, indicator kriging, and sequential Gaussian simulations.

Our results revealed high concentrations of Hg and, secondarily, As in soil but not in water or sediments. The innovative methodology abovementioned allowed us to identify natural and anthropogenic associations between 25 elements and to conclude that soil pollution was attributable mainly to natural weathering of the uphill heap. Moreover, the probability of

surpassing the threshold limits and the local backgrounds was found to be high in a large extension of the area.

The methodology used herein demonstrated to be effective for addressing complex pollution scenarios and therefore they are applicable to similar cases.

**Keywords:** Mercury, multivariate statistics, soil pollution, machine learning, geostatistics.

# 1. Introduction

In the last two centuries, industry, mining and traffic have left a notable anthropogenic footprint, which is reflected across environmental compartments (Wang and Yang, 2016). In this era, pollution has been particularly striking not only in urban systems (Biasioli et al., 2007; Boente et al., 2017; Zacháry et al., 2015) but also in natural or rural zones, where the safety of ecosystems is seriously threatened (Vitousek et al., 1997). Although industries are not frequently located in the latter areas, mining is a case apart as the geographical distribution of ores usually determined the initiation of mining activities, thus implying considerable disturbance of the local biogeography and biodiversity (Bamberger and Oswald, 2015; Venter et al., 2016).

National or Natural Reserves are protected in many countries. The designation of new reserves often hampers urban projects, opening up a debate about development vs. sustainability (Castillo-Eguskitza et al., 2017). Nevertheless, environmental awareness is a relatively modern concept and related laws were introduced only in the last decades of the 20$^{th}$ century. However, mines have been operating since the dawn of humanity and with marked intensity since the industrial revolution (Abraham et al., 2018). Therefore studies to address the impact of mining on these protected areas worldwide have proliferated in recent years (Abraham et al., 2018; de Mahiques et al., 2013; Hilson and Nyame, 2006; Li et al., 2018; Sánchez-Chardi et al., 2009; Zapico et al., 2017).

In the abovementioned context, mining is a frequent source of potentially toxic elements (PTEs); i.e. metallic elements or metalloids which can seriously disturb the environment and pose a threat to human health when present in high concentrations (Huamain et al., 1999). Among PTEs, Hg is a highly toxic element, even at very low concentrations, and especially when methylated (Syversen and Kaur, 2012). Moreover, it is easily absorbed, biomagnified and bioaccumulated within the food chain. Given these considerations, the EU launched a strategy in 2005 to reduce Hg emissions. In addition, under the recent the UN Minamata Convention on Mercury (2017), countries are required to put into place measures to control sources of Hg pollution (Evers et al., 2016), thus renewing interest in studying this pollutant.

In the context of southern Europe, Spain has historically been one of the major Hg exporters. Special mention is given to the Almadén district (Jiménez-Moreno et al., 2016), the largest Hg deposit ever found, although there were other important Hg-mines in northern Spain, such as La Soterraña and El Terronal, which have been widely studied as a result of their dramatic environmental impact (González-Fernández et al., 2018; Matanzas et al., 2017).

However, the effects of other abandoned Hg-sites should not be overlooked. In this regard, here we addressed the exemplary site of Caunedo (Fernández et al., 2017), which is located in the region of Asturias (NW Spain) within the Somiedo Nature Reserve. This Hg-mining area operated without environmental control until its closure in the 1970s.

Thus, in the context of pollution, it is pertinent to assess the distribution of Hg and other PTEs, and their bioavailability and toxicity in this protected area. To this end, here we propose a methodology that combines classical and contrasted statistical methodologies for PTE identification and risk estimation (McIlwaine et al., 2014; Mohmand et al., 2015) with machine-learning algorithms (Barzegar et al., 2018; Betrie et al., 2013) and geostatistical simulations (Benndorf, 2013). In this context, geostatistics allows establishment of the value of a variable using probabilistic models. For instance, the content of an element (e.g., Hg) in soil can be estimated by means of indicator kriging or by the index of geoaccumulation (Chakraborty et al., 2017; Škrbić et al., 2018). Results may be more robust when powerful interpolation methods are

applied, such as sequential Gaussian simulation (SGS), which involves dozens of simulations using optimizing mathematical algorithms (Boluwade and Madramootoo, 2015; Nunes and Almeida, 2010; Qu et al., 2014). However, to get a full picture of the nature of a given pollution problem, it is also necessary to understand how the different variables (element contents, geochemical backgrounds, etc.) are interlinked. This is where machine-learning procedures are required, since they provide key insights regarding the relationships between pollutants and other elements, thus increasing our knowledge of the variables of interest (Ransom et al., 2017; Sui et al., 2016). In this regard, machine-learning methods based on Bayesian networks have experienced a recent boom and have started to be applied in geochemistry (Albuquerque et al., 2017), including studies on pollutant removal (Fan et al., 2018) and the monitoring of wetlands (Whyte et al., 2018).

Overall, the present study has a double objective. On the one hand, it seeks to assess and disclose the impact of a paradigmatic abandoned Hg-mine and waste heaps located in a nature reserve. On the other hand, it aims to provide effective methodological approaches based on the techniques mentioned above that can be extrapolated to other sites affected by Hg pollution.

# 2. Materials and methods

## 2.1 Site description

Caunedo is located in the Somiedo Nature Reserve (Asturias, Spain, see Figure 1). This area is the natural habitat of endangered animal and plant species, including *Ursus arctos* (brown bear), *Tetrao urogallus* (capercaillie), and also *taxaceae* (taxus) and *Ilex aquifolium* (holly), among others (Naves et al., 2001; Nores et al., 2008).

The presence of a Hg mineral deposit (cinnabar) embedded within limestone and dolomite layers attracted mining companies to this area between the 1940s and 1970s (Luque and Gutierrez-Claverol, 2006). Cinnabar was extracted from several sloping planes undermined by

the room and pillar method. The ore was transported to treatment plants close to other more relevant Hg mines, while mine waste was indiscriminately dumped in two spoil heaps. Although the area was officially included in the Spanish national inventory of polluted soils in 2001, no remediation work has been done until now (Fernández et al., 2017).

In the context of pollutant mobility, topography plays an important role in this study. Boundary altitudes range from 900 m to 1200 m for a relatively minor area of study site (0.37 $km^2$) located on the western hill of an enclosed valley, with slopes that average 30%. These features facilitate the formation of surface runoff, as well as a strong downhill wind that fosters pollutant dispersion.

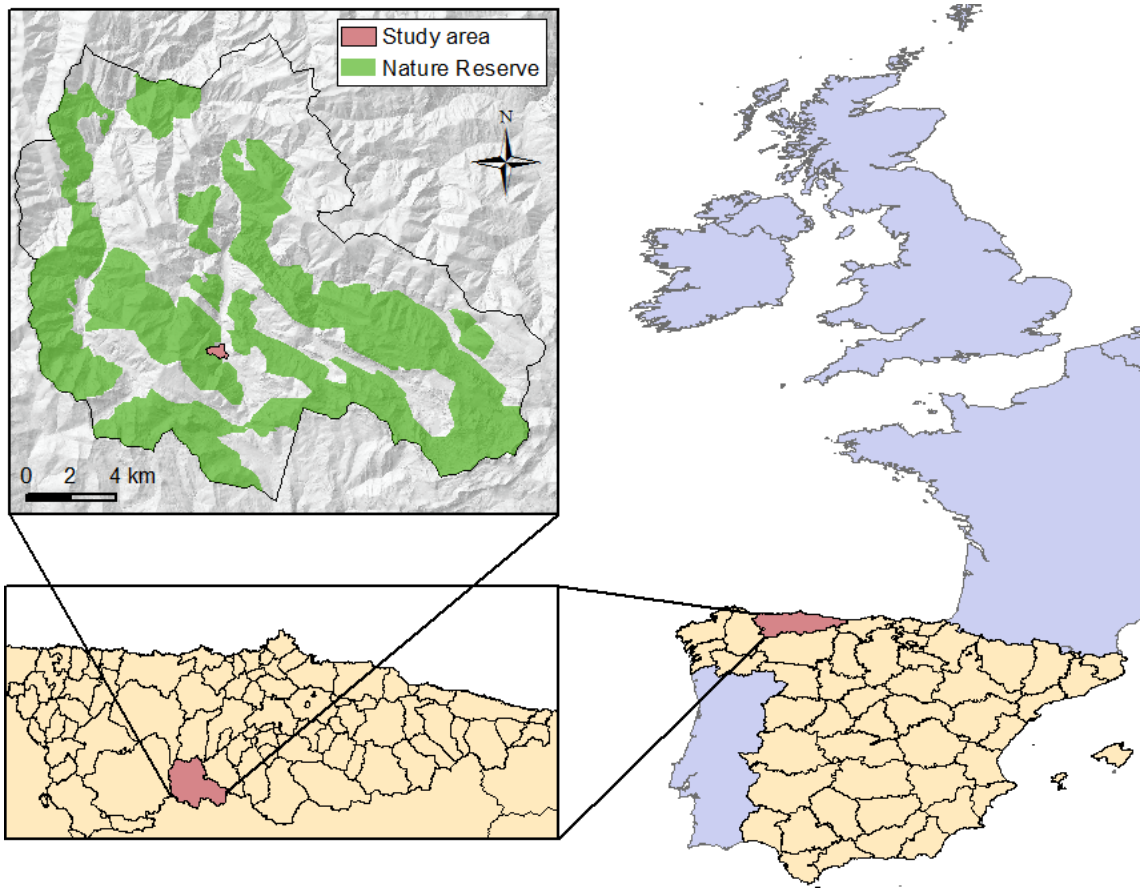



Figure 1. Location of the Caunedo study area within the Nature Reserve in the Municipality of Somiedo (Asturias, NW Spain).

## 2.2 Sampling design: Collection and preparation

Initially, topography and surficial hydrology were determined by means of a Light Detection and Ranging (LiDAR) model with an extension of 2000x2000 m and a density of 0.5 points/$m^2$. This approach allowed us to obtain a Digital Terrain Model (DTM) with a spatial resolution of 1x1 m.

The contour of the study area was defined in terms of a 30 m buffer for soil samples in the entire area, except in the northern part (Fig. SM1), where a 150 m buffer was established due to the impossibility of sampling in this area.

Regarding soils, 61 samples were taken. Most of these samples were collected in the hypothetical area of influence of the two spoil heaps, while a minor number were taken at points considered optimal for the evaluation of the geochemical background. Each sample was composed by five increases taken from each vertex of a 1 m edge square and its central point, from the top 20-25 cm of the soil, by means of an Edelman Auger. Afterwards, samples were passed through a 2 cm mesh screen in situ to remove large material such as organic matter, rocks, and gravel. The samples were transported to a laboratory, where they were dried in an oven at 30ºC to prevent the evaporation of Hg. They were then passed through a 2 mm mesh screen and the particles with a diameter of < 2 mm were quartered via a Jones riffle splitter for soil homogenization and ground in an RS100 Retsch mill at 400 RPM for 40 s until reaching < 100 μm.

In addition, five waste samples were taken: two in the western spoil heap, and three in the eastern one. These samples were subjected to the same preparation treatment as soils.

For water, the Somiedo river is the only water flow active throughout the hydrological year. Thus, another set of five samples were taken: one from an active drainage channel from the western heap and the remaining four from the course of the Somiedo river, two of them before and after the water intakes of the dumps and another two at intermediate points. Samples were collected in dry season, when the river flow is lower and consequently the concentration is expected to be higher due to a minor dilution effect. The gathering was done in polyethylene

bottles and then filtered by means of a 0.45 µm pore size mesh. The first few milliliters were used for rinsing and were then discarded. The filtrate was transferred to clean polyethylene bottles and then acidified with $HNO_3$ to pH < 2 prior to storage below 4ºC until analysis.

Finally, four composite sediment samples were taken with a sediment sampler along the Somiedo river, very close to the site where the water samples were collected. These samples were subjected to the same preparation treatment as soils.

The sampling area and different points mentioned above are shown in Figure 2. In addition, a 3D-representation of the zone by means of Lidar technology is provided in the supplementary material (Fig. SM1).

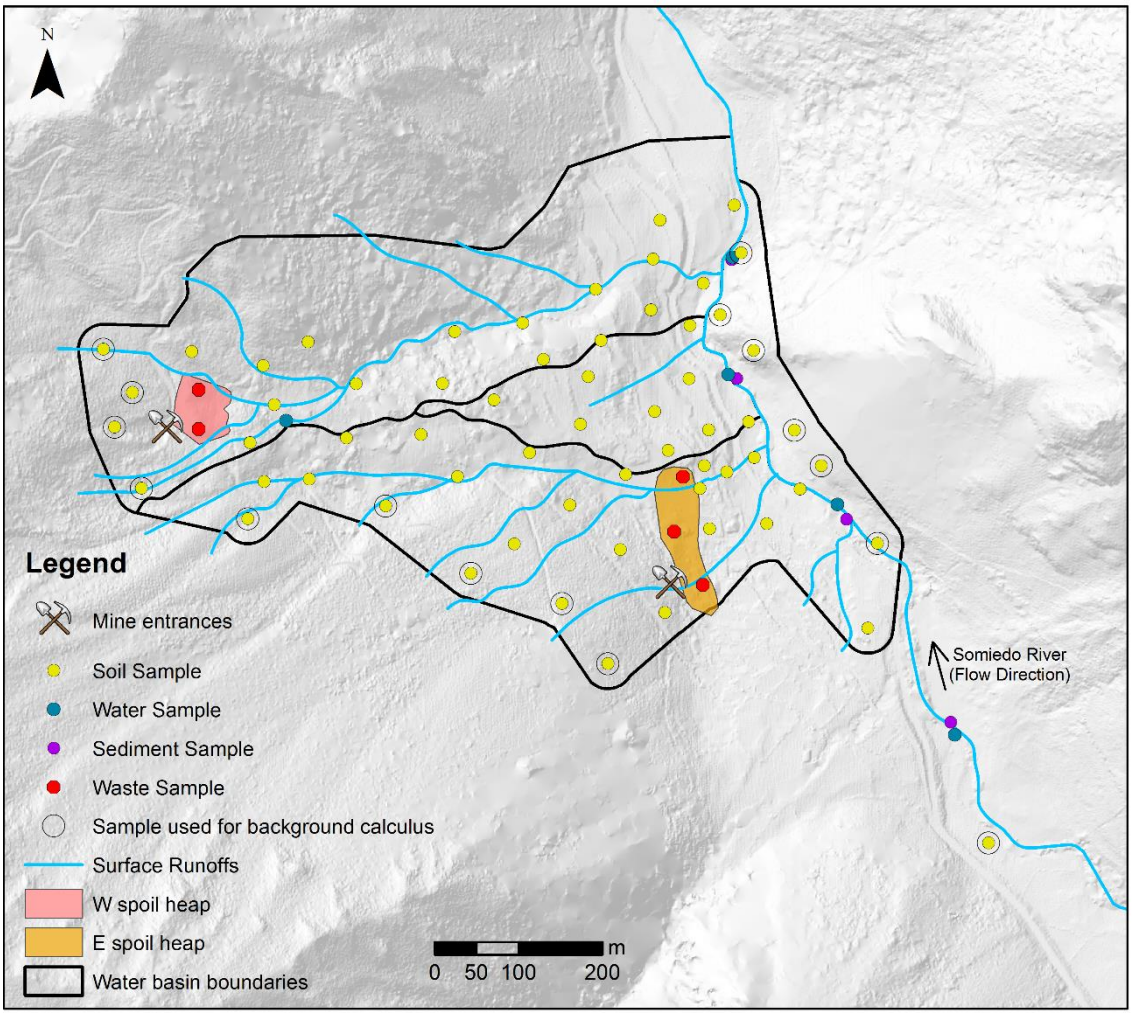



Figure 2. Bi-dimensional view of the study area: soil, water and sediment samples. Rounded soil samples were used to calculate the background. Western and eastern spoil heaps and surface run-off are also indicated.

## 2.3 Chemical analyses

1 g representative sub-samples of soils, wastes and sediments were sent to the ISO 9002-accredited Bureau Veritas Laboratories (Vancouver, Canada) and subjected to 1:1:1 "aqua regia" digestion. The total concentrations of the elements Ag, Al, As, Au, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, Ga, Hg, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Sc, Se, Sr, Te, Th, Ti, Tl, U, V, W and Zn in the digested material were determined by Inductively Coupled Plasma-Mass Spectrometry (ICP-MS) by means of the Ultratrace AQ250 analytical package of the above mentioned laboratory (Detection Limits ($mg \cdot kg^{-1}$) of As (0.1) and Hg (0.005)). Five blanks (analytical and method), five duplicates, and ten analyses of standard reference materials (internal standards and OREAS45EA) were inserted in the sequences of samples to provide a measurement of background noise, accuracy and precision.

Regarding water samples, eight PTEs (As, Cd, Cr, Cu, Hg, Ni, Pb and Zn) were quantified by means of an Inductively Coupled Plasma Mass Spectrometer (ICP-MS 7700, Agilent Technologies) using IDA (Isotopic Dilution Analysis). In addition, the main cations and anions were analyzed by ion chromatography (883 Basic IC plus, Metrohm). pH and conductivity were measured in situ using Hach Lange HQ electrodes (HQd Series Field Probe Kit).

To evaluate the mechanisms that regulate the release and mobility of As and Hg, selected soil samples were also subjected to a sequential extraction similar to that proposed in Tessier et al., (1979) (see results). In brief, extracts with reagents of increasing strengths were taken from 2.5 g samples, and exchangeable, carbonate-bound, Fe-Mn oxide bound, organic matter-bound and residual fractions were then obtained (Boente et al., 2017). Fractions were analyzed for PTE content by means of IDA-ICP-MS.

In the same samples, Hg and As speciation was also determined in order to identify the proportion of methyl- and ethyl-Hg and As (III), which are more toxic than inorganic Hg and As (V) respectively. The species were separated and subsequently quantified in a 1260 Infinity HPLC coupled to a 7700 ICPMS (Agilent Technologies), as detailed in Gallego et al., (2015).

188

189 **2.4 Descriptive statistics**

190 For all the elements analyzed in the soil samples, the mean, median, standard deviation (SD)
191 and relative standard deviation (RSD) were calculated.

192 The same database was subjected to a Factor Analysis by means of Principal Component
193 Analysis (PCA). As proposed for geochemical data (Reimann and De Caritat, 2005), factor
194 extraction was determined by the Kaiser/Gutmann criterion, whereas the Varimax/Orthogonal
195 rotation was applied to minimize the number of variables with high loadings. To obtain groups
196 of samples with similar geochemical profile, the factor score matrix was used as an input for a
197 cluster analysis. This hierarchical procedure applied the Ward's algorithm and the Squared
198 Euclidean distance, maximizing the variance between groups and minimizing it between
199 members of the same group (Murtagh and Legendre, 2014).

200 Additionally, 16 of the 61 soil samples were used to estimate the local background for each
201 element (Fig. 2). These samples were meticulously chosen from locations outside (upstream)
202 the direct influence of the heaps or in the eastern bank of the Somiedo river (also not affected by
203 the heaps influence). Regarding local background, it is a concentration range that was
204 calculated as the mean plus-minus two times the standard deviation of the 16 samples—the
205 same expression as that used for the official threshold limits of the region of Asturias
206 (Fernández et al., 2018)-, the upper value of this range is the local Soil Screening Level (SSL);
207 i.e. the level of concentration in soil (threshold) above which there is concern enough to
208 warrant site-specific study of risks (USEPA, 2007).

209 **2.5 Bayesian simulation**

210 The main potential contaminant, Hg, was the focus of an in-depth study of its patterns and
211 distribution. A machine-learning approach based on a supervised Bayesian network was used to
212 gain insights into the relationship between Hg and the other PTEs. To this end, a direct acyclic
213 graph (DAG) was machine-learnt from data, where the probabilistic relationships based on the

probability distributions of the elements' concentrations maximized the prediction of the variable of interest (Hg). Given the possibility of Bayesian machine-learning for reasoning under uncertainty through the application of information theory, the mutual information (I) was computed between Hg and each PTE. The mutual information of two discrete random variables X and Y is defined as:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y)\ log_2\ \frac{p(x,y)}{p(x)p(y)} \quad (1),$$

where p(x,y) is the joint probability distribution of X and Y, and p(x) and p(y) are the marginal probability distributions of X and Y. This computation allows identification of the PTEs that provide maximum information regarding the presence of high concentrations of Hg.

The statistical study was completed with a sensitivity analysis. Given the aleatory uncertainty of sample collection, the risk of pollution is expressed in terms of confidence intervals. To this end, a Monte Carlo simulation was performed to generate 5,000 Bayesian networks for measuring how the joint probability distribution was modified. BayesiaLab software v. 7.0.1. was used for simulation.

### 2.6 Geostatistical modeling

The spatial characterization of Hg distribution was performed through a 3-step approach:

i) Structural analysis - experimental variograms were computed and theoretical models fitted (Journel and Huijbregts, 1978). For computation, SpaceStat V. 4.0-.18. software (https://www.biomedware.com/) was used.

The variogram is a vector function used to calculate the spatial variability of regionalized variables defined by the following equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{2N(h)}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (2),$$

Its argument is h (distance) where Z ($x_i$) and Z ($x_{i+h}$) are the numerical values of the observed variable at points $x_i$, and $x_{i+h}$. The number of forming pairs for a h distance is N (h). Thus, it is the median value of the square of the differences between all pairs of points in the geometric field spaced at a h distance. The graphic study of the variograms obtained provides an overview of the spatial structure of the variable. The nugget effect ($C_o$) shows the behavior at the origin, whereas the sill ($C_1$) and the amplitude (a) define the inertia used in the interpolation process and the influence radius of the variable, respectively.

ii) To evaluate Hg behavior, indicator variables were constructed, in terms of the background applied. The indicator kriging algorithm is a non-parametric geostatistical method for estimating the probability of exceeding a specific threshold value, $z_k$, at a given location. In indicator kriging, the stochastic variable, Z(u), is transformed into an indicator variable with a binary distribution, as follows (Goovaerts, 1997):

$$I(u; z_k) = \begin{cases} 1, & if\ Z(u) \leq z_k, \quad K = 1,2,\dots,m \\ 0, & Otherwise \end{cases} \quad (3),$$

iii) Sequential Gaussian simulation (SGS) was used as a stochastic simulation algorithm for the construction of Hg spatial distribution scenarios. SGS starts by defining the univariate distribution of values. A normal score then transforms the original values to a standard normal distribution. Simulation of normal scores at grid node locations is done sequentially with simple kriging (SK) using the normal score data and a zero mean (Goovaerts, 1997). Once all normal scores have been simulated, they are back-transformed to original grade values. The outcome of a simulation is a twisted version of an estimation process, which reproduces the statistics of the known data, making a realistic view, but supplying a low prediction behavior. However, if a multiple sequence of simulations is designed, it is possible to obtain more reliable probabilistic maps.

A hundred simulations were performed, and the average image was computed, together with the Standard Deviation for spatial uncertainty visualization (Albuquerque et al., 2017).

# 3. Results and discussion

## 3.1 Pollution sources

The sources of pollution were appropriately identified and delimited. They consisted of two Hg-mining spoil heaps, named western heap and eastern heap. The western heap is located 1200 m.a.s.l., while the eastern one is 500 m away from the former and at 1000 m.a.s.l. Moreover, the western dump is exposed to mechanical dispersion, while the eastern dump is embedded in a flat area within the village of Caunedo. Each heap is located in a different water basin (Fig. 2), and therefore there is no interaction between them from the hydrological point of view.

Chemical analysis revealed that the Hg concentration in the western heap surpassed 3000 $mg \cdot kg^{-1}$ at specific points. In the eastern heap, the maximum Hg concentration found was 114 $mg \cdot kg^{-1}$. In contrast, As presented maximum values of 194 $mg \cdot kg^{-1}$ in the eastern heap and 127 $mg \cdot kg^{-1}$ in the western one.

## 3.2 Pollution in environmental compartments

### 3.2.1 Soil pollution

The results of the multielement analysis are shown in Table 1. Descriptive statistics were calculated for the 61 samples collected and analyzed, while local Soil Screening Levels, as stated before, were assessed by taking into consideration only 16 of the 61 meticulously selected samples from unpolluted soils (Fig. 2).

First, a range column allowed us to identify those PTEs that surpassed the legal limits. These limits are defined by the Risk-Based Soil Screening Levels (RBSSL) of the Asturias region, (BOPA, 2014); the remaining elements analyzed do not have RBSSL given that they are not considered contaminants of concern, and therefore they are not shown in Table 1. Thus, As, Co, Hg, Sb and V were the only elements that exceeded the limits set for Asturias on at least one occasion. On the other hand, the mean and median values were similar and low RSD (roughly

between 40 and 55%) for almost all the elements under study were observed, with the exception of Hg and Sb whose values exceeded 100%.

Focusing on those elements that exceeded the regional RBSSL, in the first place, Co and V exceeded the limits in only one sampling point but not significantly. The limit for Sb was also narrowly surpassed but in only two samples. Given that these three elements are less toxic than than As and Hg, they were not included in the study.

As regards As and Hg, both elements tend to appear together in the Hg-mining districts of Asturias (González-Fernández et al., 2018). However, in our case the relationship is not so clear given that As exceeded the official limits in only 5 samples and showed a low RSD, whereas Hg surpassed the limits in 53 of the 61 samples, showing a high RSD and very different mean (13.1 $mg \cdot kg^{-1}$) and median (6.9 $mg \cdot kg^{-1}$) values, thus highlighting the irregular distribution of this element in the area.

Table 1. Regional RBSSL (BOPA, 2014) for Asturias. Descriptive statistics (range, mean, median, standard deviation and RSD) for the analysis of 61 soil samples from the Caunedo area. And local SSL for the Caunedo study area (mean of the 16 background samples + 2 times their standard deviation –see Fig. 2- ) are also provided. All units are expressed in $mg \cdot kg^{-1}$, except RSD, which is expressed as a %.

| Element | Regional RBSSL | Local SSL | Range | Mean | Median | SD | RSD |
|---|---|---|---|---|---|---|---|
| **As** | 40.0 | 37.7 | 2.9 - 69.4 | 22.9 | 21.5 | 12.5 | 54.4 |
| **Ba** | 1540.0 | 353.5 | 40.0 - 530.0 | 173.9 | 150.0 | 97.3 | 55.9 |
| **Cd** | 2.0 | 0.4 | 0.1 - 0.5 | 0.1 | 0.1 | 0.1 | 52.4 |
| **Co** | 25.0 | 20.3 | 0.6 - 29.1 | 12.4 | 13.0 | 5.4 | 43.5 |
| **Cr** | 10000.0 | 26.0 | 4.0 - 43.0 | 14.7 | 14.0 | 8.4 | 57.0 |
| **Cu** | 55.0 | 24.1 | 4.4 - 30.3 | 17.3 | 18.4 | 6.0 | 34.6 |
| **Hg** | 1.0 | 2.5 | 0.09 - 50.0 | 13.1 | 6.9 | 14.7 | 111.6 |
| **Mn** | 2135.0 | 1185.9 | 23.0 - 1274.0 | 608.3 | 692.0 | 268.6 | 44.2 |
| **Mo** | 6.0 | 1.2 | 0.2 - 1.8 | 0.6 | 0.5 | 0.3 | 49.6 |
| **Ni** | 65.0 | 39.9 | 2.6 - 35.6 | 20.7 | 23.3 | 8.3 | 40.2 |
| **Pb** | 70.0 | 27.1 | 7.8 - 34.7 | 19.5 | 19.3 | 5.8 | 29.9 |
| **Sb** | 5.0 | 1.1 | 0.1 - 11 | 0.9 | 0.6 | 1.5 | 162.3 |
| **V** | 50.0 | 41.2 | 7.0 - 58.0 | 21.5 | 20.0 | 9.3 | 43.0 |
| **Zn** | 455.0 | 106.4 | 11.0 - 134.0 | 71.3 | 75.0 | 31.1 | 43.6 |

All things considered, it can be concluded that Hg is the leading contaminant in Caunedo, followed by As, but to a much smaller extent. Given these observations, we focused our study on Hg and As since the other PTEs studied could not be considered contaminants due to their dispersion and low concentration.

### 3.2.2 Water pollution

As described above, water samples were taken from various parts of the Somiedo river and from an active mine drainage channel flowing from the western heap (Fig. 2). Hydrochemical analyses revealed that all the samples corresponded to calcium bicarbonate waters with values of pH between 7.6 - 7.9 and conductivity ranging between 263 - 285 µS/cm. Being all values stable and in the range of low salinity and slightly alkaline waters.

This observation is coherent with the content of PTEs, whose concentrations were lower than those established in environmental standards: very low in the case of As (0.15-0.48 µg/l) or even under detection limits for Hg (< 0.1 µg/l). The absence of PTE in waters suggests that the flow rate of surface runoff that crosses the heaps, together with the residence time that this water is in contact with polluted soils, is insufficient to cause significant dissolution of the PTEs in water. However, the mechanical dispersion of the contaminants from the heaps is clearly noticeable in soils adjoining surface run-offs (Lillebø et al., 2011). Thus, it can be concluded that neither the waters of the area nor the principal stream are affected by PTE pollution.

### 3.2.3 Sediment pollution

For the four sediment samples taken from the Somiedo river, the PTE content was within the background of soils for all 37 elements. The samples showed highly stable concentrations (e.g. 17.6-19.1 $mg \cdot kg^{-1}$ for As and 0.12-1.83 $mg \cdot kg^{-1}$ for Hg), except one (the nearest point to eastern heap, Fig. 2), which registered up to 8.65 $mg \cdot kg^{-1}$ of Hg. Since there were no indications of pollution upstream, and bearing in mind the proximity of this sample site to agricultural plots near the eastern heap, a local disposal of former mining waste is the most likely cause of the presence of Hg in this sediment.

### 3.2.4 Bioavailability and toxicity

Sequential extraction was applied to three soil samples that considerably exceeded the regional RBSSL for Hg and As. Two of the samples were taken close to the heaps and the third was from the bottom of the valley. Considering the two first fractions obtained in the Tessier method as bioavailable, the results revealed very low ranges of bioavailability, between 0.40 and 0.85 % for As and between 0.01 and 0.02% for Hg. Consequently, potential absorption by plants is considered limited.

In a different approach to potential toxicity effects, the presence of As (III) was perceivable but very low (1 $mg \cdot kg^{-1}$ average out of 20 $mg \cdot kg^{-1}$) for the three samples analyzed, As (V) being the minor toxic variant prevailing. For Hg, the same was observed: methyl mercury, $CH_3Hg^+$, presenting low concentrations (1 $mg \cdot kg^{-1}$ average out of 40 $mg \cdot kg^{-1}$) in comparison with the inorganic species of Hg.

On the basis of the results in this section, it can be concluded that there is no risk of transmitting toxicity through the food chain and that contamination is largely restricted to soils.

## 3.3 Patterns, distribution and associations of contaminants in soils

As soils were identified as the only heavily polluted matrix, we performed a more in-depth study of them. First, results of a factor analysis (PCA and varimax rotation) for the multielement data are shown in Table 2. The factors extracted accounted for more than 85% of the total variance.

Factor 1 (F1) presents high loadings for natural elements that do not represent a risk in the area and that are probably linked to soil minerals such as silicates, Fe-Mn oxides and some clays (high loads in Fe, Mn and K). Factor 2 (F2) is probably also related to clayey material (high V and Al loads). Factor 3 (F3) is related to carbonates (high Ca and Sr), and Factor 4 (F4) to cinnabar mineralization, being the group that presents most loadings for Hg, As and Sb. This

353 association is typical in natural quartz-vein deposits and it is coincident with elements labeled as
354 "enriched" in section 3.1 (Hg deposits). Despite these observations, low loading for Hg in all
355 factors revealed that a large part of Hg is highly independent of the rest of elements.

356 Table 2. Factor loadings, percentage of variance explained by the Varimax-rotated factors
357 (extracted by principal components) and communalities. Element loadings higher than 0.6 are
358 marked in bold.

| Element | F1 | F2 | F3 | F4 | Communality |
|---|---|---|---|---|---|
| **Co** | **0.95** | 0.23 | 0.00 | 0.14 | 0.97 |
| **Sc** | **0.92** | 0.28 | 0.15 | 0.10 | 0.95 |
| **Ni** | **0.91** | 0.29 | 0.15 | 0.15 | 0.96 |
| **Mn** | **0.90** | 0.24 | 0.04 | 0.23 | 0.93 |
| **K** | **0.90** | 0.10 | 0.03 | -0.05 | 0.82 |
| **Th** | **0.87** | 0.36 | 0.13 | 0.00 | 0.90 |
| **La** | **0.86** | 0.26 | 0.06 | -0.14 | 0.83 |
| **Cu** | **0.82** | 0.10 | -0.27 | 0.31 | 0.85 |
| **Fe** | **0.82** | 0.43 | -0.02 | 0.21 | 0.90 |
| **Zn** | **0.73** | 0.49 | -0.01 | 0.33 | 0.87 |
| **Ba** | **0.65** | -0.06 | -0.49 | 0.03 | 0.67 |
| **V** | 0.29 | **0.89** | 0.15 | -0.05 | 0.91 |
| **Ga** | 0.34 | **0.85** | 0.12 | 0.00 | 0.86 |
| **Mo** | -0.02 | **0.83** | 0.15 | 0.20 | 0.74 |
| **Cr** | 0.54 | **0.78** | 0.16 | 0.07 | 0.93 |
| **Al** | 0.62 | **0.73** | 0.22 | 0.02 | 0.96 |
| **Pb** | 0.46 | **0.70** | -0.19 | 0.26 | 0.81 |
| **Sr** | 0.12 | 0.05 | **0.87** | -0.13 | 0.79 |
| **Ca** | 0.31 | 0.11 | **0.86** | 0.30 | 0.94 |
| **Na** | -0.25 | 0.40 | **0.71** | 0.00 | 0.73 |
| **Mg** | 0.45 | 0.16 | 0.57 | 0.47 | 0.79 |
| **Hg** | 0.33 | 0.05 | -0.63 | 0.51 | 0.77 |
| **Sb** | -0.17 | 0.08 | 0.05 | **0.92** | 0.88 |
| **Au** | 0.22 | 0.05 | -0.02 | **0.86** | 0.80 |
| **As** | 0.43 | 0.51 | 0.01 | **0.58** | 0.78 |
| **Cumulative explained variance (%)** | *52.23* | *67.20* | *77.25* | *85.29* | |

359

A step further implied the clustering of soil samples. Five clusters of samples were identified (Figure 3). Cluster 1 (C1) and Cluster 2 (C2) are related to natural material. Taking geology into consideration, the first would be related to F1, the siliceous component, and the second to F2, the calcareous one on the eastern bank of the river. Both correspond to background samples in unpolluted soils. Cluster 3 (C3) mixes background samples with others that present anomalous levels of Hg, surely influenced by the composition of the western heap. Clusters 4 and 5 (C4 and C5) have the largest number of samples and geographical scattering. This observation might be attributable to the influence of Hg. In general, C4 holds natural samples in the boundaries of the study area but that sometimes present high concentrations of Hg. Finally, sample core of C5 is located in farms or pasturelands alongside the river, comprising areas with human influence in the downstream of the western heap and also with a relatively high content of Hg.

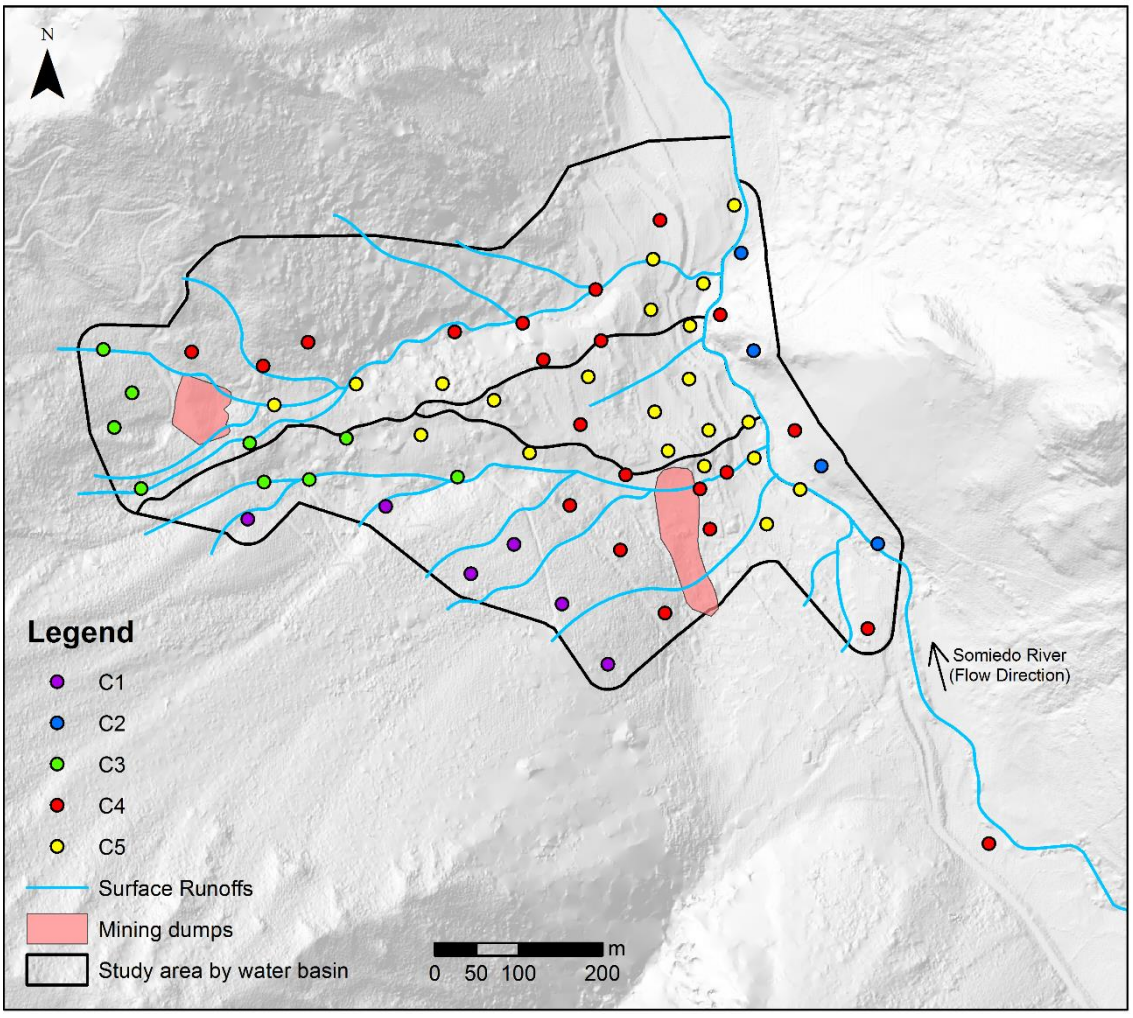


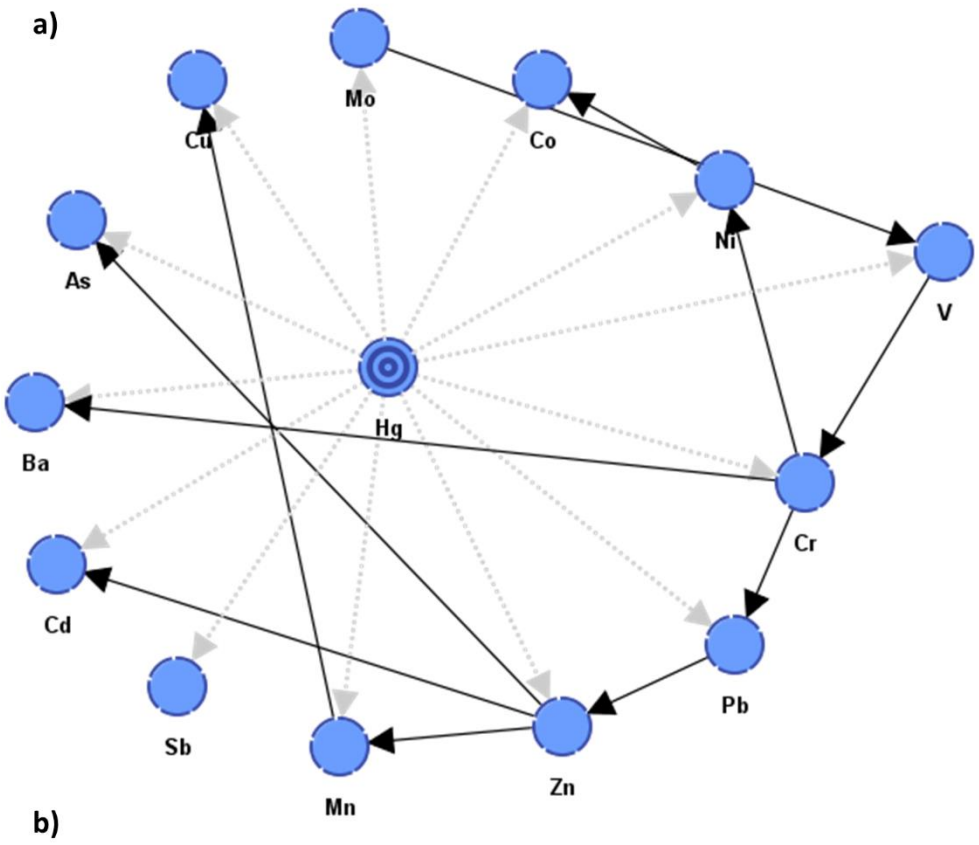

Figure 3. Graphical clustering according to Ward's algorithm.

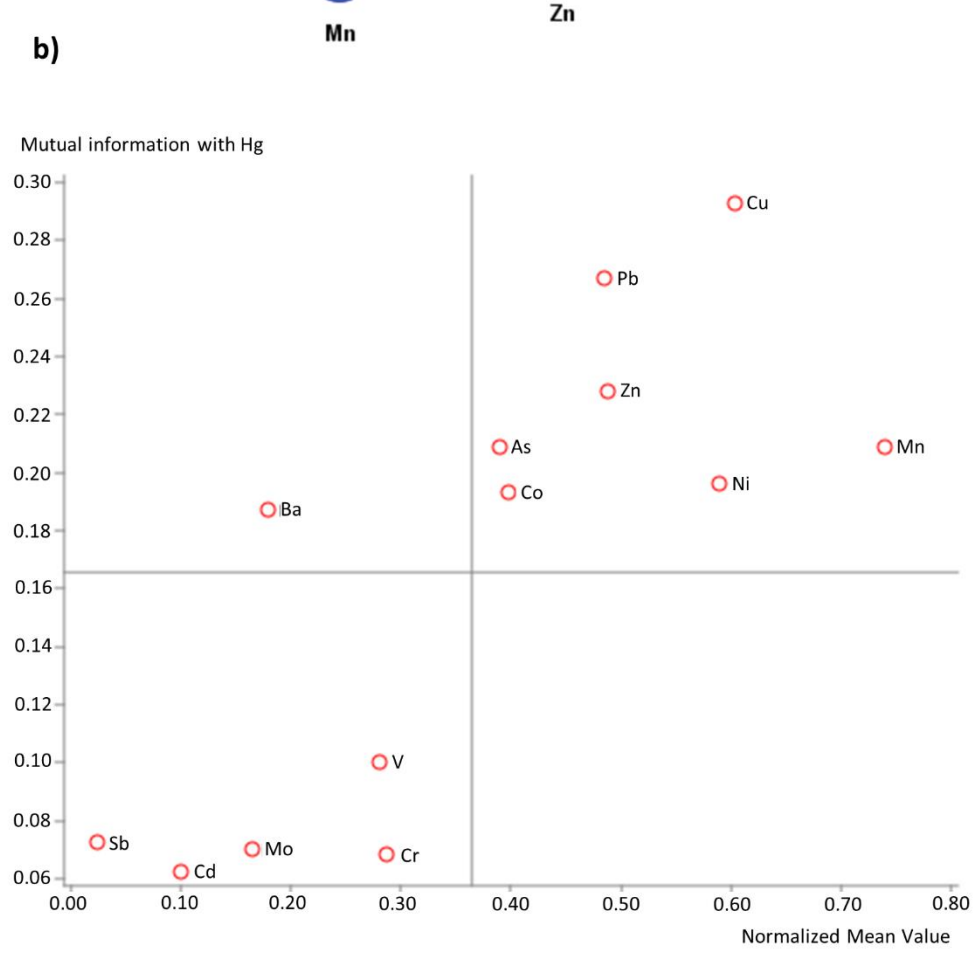
To gain a better understanding of the Hg fate in the soils of Caunedo, a supervised machine-learning procedure was performed. As a result, a supervised Bayesian network was constructed (Fig. 4) to characterize the target node (Hg). To create the model, a package of supervised learning algorithms included in the BayesiaLab software was executed. Upon assessing the network performance for each algorithm using k-fold cross-validation, Augmented Naive Bayes was the algorithm with the highest precision for this problem domain.

The algorithm selected produced a model (Fig. 4a) with a double architecture made up of a naive structure (gray arrows), enriched by an unsupervised search of relations (black arrows) between the remaining nodes in the network that maximize the prediction of the target node (Catal et al., 2011; Webb et al., 2005).

First, the model provides conceptual knowledge about the relationships influencing the presence of high levels of Hg. Second, a mutual information analysis was carried out from the Bayesian network generated, in order to capture the strength of relationship between the Hg and the rest of the PTEs. In Fig. 4b a quadrant map plots the mutual information for each element regarding their normalized mean concentration. As can be appreciated, the elements that present higher mean concentrations also show a higher informative content for Hg. Additionally, if the maximum value that the entropy of Hg can take for this study is considered, normalized mutual information can be obtained, resulting in descending order: Cu (14.64%), Pb (13.37%), Zn (11.41%), Mn (10.46%), As (10.41 %), Ni (9.83%) and Co (9.67%). In other words, knowing the concentration of Cu, the most influential element found, the uncertainty regarding Hg is reduced on average by 0.292 bits, which can be translated to a 14.64% of gain of information. This can be considered a low-to-moderate gain, thus certifying the independent behavior of Hg previously observed. On the other hand, with the exception of As, which is the second contaminant, and Pb, the remaining PTEs belong to the aforementioned factor of natural elements (F1) that appears in low concentrations. The findings of this section demonstrated the

399 independent geochemical behavior of Hg in the study area, and strongly proposed its connection
400 with the Hg-mining works and the spoil heaps.
401


a)
Mo
Cu
Co
Ni
As
V
Hg
Ba
Cr
Cd
Pb
Sb
Zn
Mn


b)
Mutual information with Hg
0.30
0.28
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
Cu
Pb
Zn
As
Mn
Co
Ni
Ba
V
Sb
Mo
Cr
Cd
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
Normalized Mean Value

402

Figure 4. Machine-learning analysis. a) Supervised Bayesian network built with an Augmented Naive Bayes algorithm for Hg characterization. The naive structure is represented by gray arrows, while the secondary (augmented) structure is shown with black arrows. b) Quadrant mapping representing the mutual information supplied by each element related to Hg.

The machine-learning procedure ended with a test of sensitivity (Fig. SM2). Discretization defined three states labeled as the desirable 0 – 1 $mg \cdot kg^{-1}$ (regional RBSSL), 1 –2.5 $mg \cdot kg^{-1}$ and > 2.5 $mg \cdot kg^{-1}$ (local SSL). In brief, the results obtained for a 95% confidence level showed that a random sample collected in Caunedo would carry an average probability of 73.5% of being over the local SSL, whereas the regional RBSSL would be exceeded on average in 88.5% of the cases.

### 3.4. Iso-probability maps

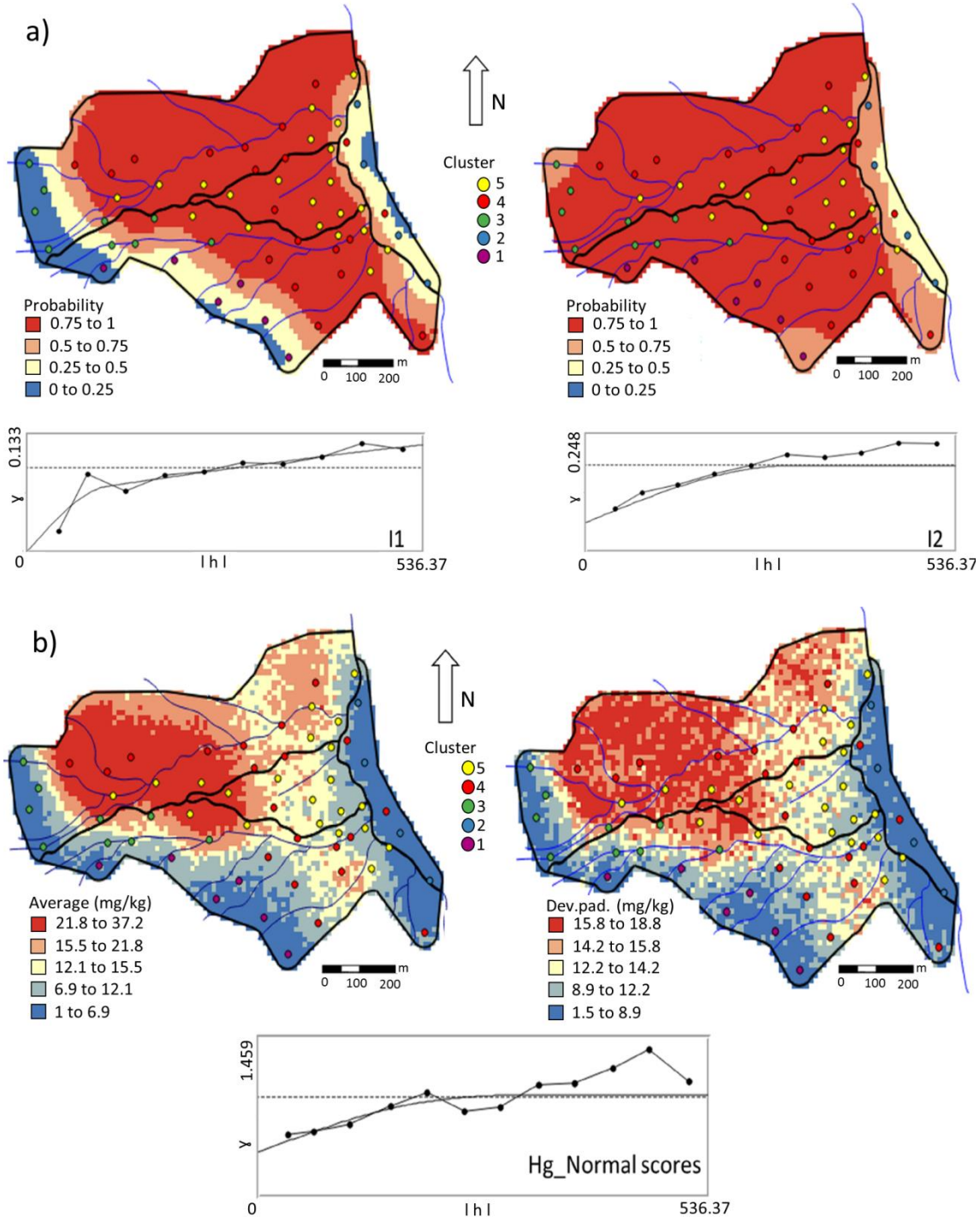
The use of indicator geostatistics allows mapping of the probability of attribute values exceeding a certain cut-off value. These probability maps are useful for decision makers as they are easy to interpret and as many maps as thresholds of interest can be produced (Ribeiro et al., 1997). In the present application, the two adopted cut-off values were 2.5 ppm, corresponding to the local SSL (Indicator I1 – local background), and 1 ppm, corresponding to the regional RBSSL (Indicator I2). Fig. 5 shows the experimental omnidirectional variograms fitted by spherical models for the two indicator variables. For both indicators, the nugget effect was between 0 and 25% of the total variance and the range was between 300 to 350 m, defining the inertia used in the interpolation process (over 75%) and the variable structure influence zone, respectively. The probability maps, generated by indicator kriging, are shown in Fig. 5a. The I1 probability map shows a strong central anomaly corresponding to the main water drainage direction from abandoned mines (Fig. 5a (left)). Concerning I2, the probability of exceeding the regional background (1 ppm) was higher than 50% for almost the entire survey area (Fig. 5a (right)).

These high probabilities for Hg soil contamination indicate that the soil surpasses the RBSSL and, in addition, that it is highly enriched in Hg as a consequence of the human footprint, as mentioned in previous sections.

### 3.5. Hg spatial patterns and associated spatial uncertainty

A hundred simulations were performed using Sequential Gaussian Simulation (SGS) on a 100 x 100 m grid. These simulations were used to generate 100 equiprobable scenarios as a conditional stochastic simulation of the distribution of Hg concentrations (Fig. 5b (left)) and the special uncertainty – Standard Deviation – (Fig. 5b (right)). As no single realization can be taken as a better representation of reality, the average spatial image (AI) is then used to assess the Hg spatial distribution pattern, while the spatial variability image (standard deviation map) allows quantification of spatial uncertainty. This shows the normal score experimental omnidirectional variogram for Hg fitted by two nested spherical models. The nugget effect is about 45% of the total variance and a range of 300 m, defining the inertia used in the interpolation process (65%) and the variable structure influence zone, respectively.

The spatial patterns allowed the classification of two different zones of Hg contamination: The northern area (high associated spatial uncertainty), and the southern (low associated spatial uncertainty). The former is clearly influenced by the western heap, and the degree of dispersion was found to be greater in the northern area. Indeed, this area of Caunedo is where remediation measures should be undertaken, prioritizing actions to address public health concerns associated with the western spoil heap.




450

451 Figure 5. Geostatistical analysis of Hg. a) (Left) Experimental indicator variogram for I1 and
452 the probability map for Hg content exceeding 2.5 ppm (local SSL); (Right) Experimental
453 indicator variogram for I2 and probability map for Hg content exceeding 1 ppm (regional
454 RBSSL). b) (Left) Omnidirectional variogram and fitted spherical models; (Right) Average
455 Image (AI) and Spatial uncertainty (Standard Deviation).

456

457

# 4. Conclusions

The present study sought to provide a group of tools to assess the environmental impact of derelict mines. The methodology used synchronized classical geochemistry, uni and multivariate statistics, machine learning/Bayesian networks, indicator kriging and sequential Gaussian simulations to simplify a complex database of PTEs for the purpose of identifying pollutants and providing details of their patterns, relationships, origins and distribution. Our approach was successfully applied to the abandoned Hg mine of Caunedo, in the Somiedo Nature Reserve.

It was observed that Hg and to a lesser extent As were the principal pollutants. Of the environmental compartments studied, mechanical dispersion via weathering agents (surface run-off, winds, slope) affected only soils, while river sediments showed mild pollution coherent with the values found for soils. Water showed no pollution. The toxicity and bioavailability studies performed on the soils revealed that both Hg and As persisted in the environment mainly in the form of their less toxic chemical species.

After identifying Hg as the main pollutant, we performed a detailed study of this PTE using multivariate statistics and Bayesian networks. The combination of techniques divided the samples into five clusters in which Hg was present in three, while Bayesian network revealed the independent behavior of Hg, for which the elements most contributing (although to a very low degree) to Hg pollution were Cu, Zn, Mn and As.

Finally, the geostatistical study of Hg in soils revealed that almost the entire area surpassed the regional RBSSL (or legal concentration limits – 1 mg/kg for Hg) and even the local SSLs (natural enrichment). The western heap had a greater effect on the environment that the eastern heap, which was better encapsulated, and the geochemical anomalies were delimited for all directions except north. The harsh topography made it impossible to sample in these areas, given that rocky outcrops are prevalent in this direction.

Remediation measures should be applied in the spoil heaps since their influence on the soils of the Nature Reserve has been demonstrated. For the heaps, confinement and/or stabilization techniques should be considered followed by restoration techniques whereas for soils, phytoremediation approaches with autochthonous species could be a sustainable choice.

In conclusion, the advances reported here allowed addressing properly each step of a characterization process: This included the identification of pollution sources (heaps), the delineation of the area impacted (geostatistical study), and the examination of the geochemical behavior of pollutants (by means of both multivariate and Bayesian approaches). This methodology may be extrapolated to other study cases focused on pollution by Hg or other heavy metal(loid)s mining.

**Acknowledgements**

Carlos Boente obtained a grant from the "Formación del Profesorado Universitario" program, financed by the "Ministerio de Educación, Cultura y Deporte de España". The authors thank the Principality of Asturias for co-financing this research, the "Servicio Científico-Técnico de Ensayos Medioambientales" of the University of Oviedo, and also Álvaro Dapía, Nora Matanzas, Diego Baragaño and Nerea García for their support during the sampling works.

# References

Abraham, J., Dowling, K., Florentine, S., 2018. Assessment of potentially toxic metal contamination in the soils of a legacy mine site in Central Victoria, Australia. Chemosphere 192, 122–132. doi:10.1016/j.chemosphere.2017.10.150

Albuquerque, M.T.D., Gerassis, S., Sierra, C., Taboada, J., Martín, J.E., Antunes, I.M.H.R., Gallego, J.R., 2017. Developing a new Bayesian Risk Index for risk evaluation of soil

507 contamination. Sci. Total Environ. 603–604, 167–177.
508 doi:10.1016/j.scitotenv.2017.06.068

509 Bamberger, M., Oswald, R.E., 2015. Long-term impacts of unconventional drilling operations
510 on human and animal health. J. Environ. Sci. Health. A. Tox. Hazard. Subst. Environ. Eng.
511 50, 447–59. doi:10.1080/10934529.2015.992655

512 Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater
513 contamination risk of multiple aquifers using multi-model ensemble of machine learning
514 algorithms. Sci. Total Environ. 621, 697–712. doi:10.1016/j.scitotenv.2017.11.185

515 Benndorf, J., 2013. Application of efficient methods of conditional simulation for optimising
516 coal blending strategies in large continuous open pit mining operations. Int. J. Coal Geol.
517 112, 141–153. doi:10.1016/j.coal.2012.10.008

518 Betrie, G.D., Tesfamariam, S., Morin, K.A., Sadiq, R., 2013. Predicting copper concentrations in
519 acid mine drainage: A comparative analysis of five machine learning techniques. Environ.
520 Monit. Assess. 185, 4171–4182. doi:10.1007/s10661-012-2859-7

521 Biasioli, M., Grcman, H., Kralj, T., Madrid, F., Díaz-Barrientos, E., Ajmone-Marsan, F., 2007.
522 Potentially toxic elements contamination in urban soils: a comparison of three European
523 cities. J. Environ. Qual. 36, 70–9. doi:10.2134/jeq2006.0254

524 Boente, C., Matanzas, N., García-González, N., Rodríguez-Valdés, E., Gallego, J.R., 2017. Trace
525 elements of concern affecting urban agriculture in industrialized areas: A multivariate
526 approach. Chemosphere 183, 546–556. doi:10.1016/j.chemosphere.2017.05.129

527 Boluwade, A., Madramootoo, C.A., 2015. Geostatistical independent simulation of spatially
528 correlated soil variables. Comput. Geosci. 85, 3–15. doi:10.1016/j.cageo.2015.09.002

529 BOPA, Boletín Oficial del Principado de Asturias, 91, April 21, 2014. Generic reference levels for
530 heavy metals in soils from Principality of Asturias, Spain. http://sede.612

asturias.es/bopa/2014/04/21/2014-06617.pdf (accessed November 2018).

Castillo-Eguskitza, N., Rescia, A.J., Onaindia, M., 2017. Urdaibai Biosphere Reserve (Biscay, Spain): Conservation against development? Sci. Total Environ. 592, 124–133. doi:10.1016/j.scitotenv.2017.03.076

Catal, C., Sevim, U., Diri, B., 2011. Practical development of an Eclipse-based software fault prediction tool using Naive Bayes algorithm. Expert Syst. Appl. 38, 2347–2353. doi:10.1016/j.eswa.2010.08.022

Chakraborty, S., Man, T., Paulette, L., Deb, S., Li, B., Weindorf, D.C., Frazier, M., 2017. Rapid assessment of smelter/mining soil contamination via portable X-ray fluorescence spectrometry and indicator kriging. Geoderma 306, 108–119. doi:10.1016/j.geoderma.2017.07.003

de Mahiques, M.M., Figueira, R.C.L., Salaroli, A.B., Alves, D.P.V., Gonçalves, C., 2013. 150 years of anthropogenic metal input in a Biosphere Reserve: The case study of the Cananéia-Iguape coastal system, Southeastern Brazil. Environ. Earth Sci. 68, 1073–1087. doi:10.1007/s12665-012-1809-6

Evers, D.C., Keane, S.E., Basu, N., Buck, D., 2016. Evaluating the effectiveness of the Minamata Convention on Mercury: Principles and recommendations for next steps. Sci. Total Environ. doi:10.1016/j.scitotenv.2016.05.001

Fan, M., Hu, J., Cao, R., Ruan, W., Wei, X., 2018. A review on experimental design for pollutants removal in water treatment with the aid of artificial intelligence. Chemosphere. doi:10.1016/j.chemosphere.2018.02.111

Fernández, S., Cotos-Yáñez, T., Roca-Pardiñas, J., Ordóñez, C., 2018. Geographically Weighted Principal Components Analysis to assess diffuse pollution sources of soil heavy metal: Application to rough mountain areas in Northwest Spain. Geoderma 311, 120–129.

555 doi:10.1016/j.geoderma.2016.10.012

556 Fernández, S., Poschenrieder, C., Marcenò, C., Gallego, J.R., Jiménez-Gámez, D., Bueno, A., Afif,
557 E., 2017. Phytoremediation capability of native plant species living on Pb-Zn and Hg-As
558 mining wastes in the Cantabrian range, north of Spain. J. Geochemical Explor. 174, 10–20.
559 doi:10.1016/j.gexplo.2016.05.015

560 Gallego, J. R., Esquinas, N., Rodríguez-Valdés, E., Menéndez-Aguado, J. M., & Sierra, C., 2015.
561 Comprehensive waste characterization and organic pollution co-occurrence in a hg and as
562 mining and metallurgy brownfield. J. Hazard. Mater. 300, 561-571.
563 doi:10.1016/j.jhazmat.2015.07.029

564 González-Fernández, B., Rodríguez-Valdés, E., Boente, C., Menéndez-Casares, E., Fernández-
565 Braña, A., Gallego, J.R., 2018. Long-term ongoing impact of arsenic contamination on the
566 environmental compartments of a former mining-metallurgy area. Sci. Total Environ.
567 610–611, 820–830. doi:10.1016/j.scitotenv.2017.08.135

568 Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation (Applied Geostatistics).
569 Oxford Univ. Press. New York 496.

570 Hilson, G., Nyame, F., 2006. Gold mining in Ghana's forest reserves: A report on the current
571 debate. Area 38, 175–185. doi:10.1111/j.1475-4762.2006.00670.x

572 Huamain, C., Chunrong, Z., Cong, T., Yongguan, Z., 1999. Heavy Status Metal and in China : in
573 Soils Pollution Countermeasures. Ambio 28, 130–134.

574 Jiménez-Moreno, M., Barre, J.P.G., Perrot, V., Bérail, S., Rodríguez Martín-Doimeadios, R.C.,
575 Amouroux, D., 2016. Sources and fate of mercury pollution in Almadén mining district
576 (Spain): Evidences from mercury isotopic compositions in sediments and lichens.
577 Chemosphere 147, 430–438. doi:10.1016/j.chemosphere.2015.12.094

578 Journel, A., Huijbregts, C., 1978. Mining geostatistics. San Diego: Academic Press.

579 Li, S., Wu, J., Gong, J., Li, S., 2018. Human footprint in Tibet: Assessing the spatial layout and
580 effectiveness of nature reserves. Sci. Total Environ. 621, 18–29.
581 doi:10.1016/j.scitotenv.2017.11.216

582 Lillebø, A.I., Coelho, P.J., Pato, P., Válega, M., Margalho, R., Reis, M., Raposo, J., Pereira, E.,
583 Duarte, A.C., Pardal, M.A., 2011. Assessment of mercury in water, sediments and biota of
584 a southern European estuary (Sado Estuary, Portugal). Water. Air. Soil Pollut. 214, 667–
585 680. doi:10.1007/s11270-010-0457-2

586 Luque, C., Gutierrez-Claverol, M., 2006. La minería del mercurio en Asturias. Rasgos históricos.

587 Matanzas, N., Sierra, M.J., Afif, E., Díaz, T.E., Gallego, J.R., Millán, R., 2017. Geochemical study
588 of a mining-metallurgy site polluted with As and Hg and the transfer of these
589 contaminants to Equisetum sp. J. Geochemical Explor. 182, 1–9.
590 doi:10.1016/j.gexplo.2017.08.008

591 McIlwaine, R., Cox, S.F., Doherty, R., Palmer, S., Ofterdinger, U., McKinley, J.M., 2014.
592 Comparison of methods used to calculate typical threshold values for potentially toxic
593 elements in soil. Environ. Geochem. Health 36, 953–971. doi:10.1007/s10653-014-9611-x

594 Mohmand, J., Eqani, S.A.M.A.S., Fasola, M., Alamdar, A., Mustafa, I., Ali, N., Liu, L., Peng, S.,
595 Shen, H., 2015. Human exposure to toxic metals via contaminated dust: Bio-accumulation
596 trends and their potential risk estimation. Chemosphere 132, 142–151.
597 doi:10.1016/j.chemosphere.2015.03.004

598 Murtagh, F., Legendre, P., 2014. Wards Hierarchical Agglomerative Clustering Method: Which
599 Algorithms Implement Wards Criterion? J. Classif. 31, 274–295. doi:10.1007/s00357-014-
600 9161-z

601 Naves, J., Fernández-Gil, A., Delibes, M., 2001. Effects of Recreation Activities on a Brown Bear
602 Family Group in Spain EFFECTS OF RECREATION ACTIVITIES ON A BROWN BEAR FAMILY

603 GROUP IN. Source: Ursus 12, 135–139.

604 Nores, C., Llaneza, L., Álvarez, Á., 2008. Wild boar Sus scrofa mortality by hunting and wolf
605 Canis lupus predation: an example in northern Spain. Wildlife Biol. 14, 44–51.
606 doi:10.2981/0909-6396(2008)14[44:WBSSMB]2.0.CO;2

607 Nunes, R., Almeida, J.A., 2010. Parallelization of sequential Gaussian, indicator and direct
608 simulation algorithms. Comput. Geosci. 36, 1042–1052. doi:10.1016/j.cageo.2010.03.005

609 Qu, M., Li, W., Zhang, C., 2014. Spatial Distribution and Uncertainty Assessment of Potential
610 Ecological Risks of Heavy Metals in Soil Using Sequential Gaussian Simulation. Hum. Ecol.
611 Risk Assess. An Int. J. 20, 764–778. doi:10.1080/10807039.2013.770352

612 Ransom, K.M., Nolan, B.T., A. Traum, J., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D.C.,
613 Z. Rosecrans, C., Jurgens, B., Schwarz, G.E., Belitz, K., M. Eberts, S., Kourakos, G., Harter,
614 T., 2017. A hybrid machine learning model to predict and visualize nitrate concentration
615 throughout the Central Valley aquifer, California, USA. Sci. Total Environ. 601–602, 1160–
616 1172. doi:10.1016/j.scitotenv.2017.05.192

617 Reimann, C., De Caritat, P., 2005. Distinguishing between natural and anthropogenic sources
618 for elements in the environment: Regional geochemical surveys versus enrichment
619 factors. Sci. Total Environ. doi:10.1016/j.scitotenv.2004.06.011

620 Ribeiro, L., Pina, P., Muge, F., 1997. Contribution of indicator geostatistics and mathematical
621 morphology to the characterisation of aquifer heterogeneities in the vicinities of waste
622 disposal sites, Engineering geology and the environment. Proc. symposium, Athens, 1997.
623 Vol.2.

624 Sánchez-Chardi, A., Ribeiro, C.A.O., Nadal, J., 2009. Metals in liver and kidneys and the effects
625 of chronic exposure to pyrite mine pollution in the shrew Crocidura russula inhabiting the
626 protected wetland of Doñana. Chemosphere 76, 387–394.

627 doi:10.1016/j.chemosphere.2009.03.036

628 Škrbić, B.D., Buljovčić, M., Jovanović, G., Antić, I., 2018. Seasonal, spatial variations and risk
629 assessment of heavy elements in street dust from Novi Sad, Serbia. Chemosphere 205.
630 doi:10.1016/j.chemosphere.2018.04.124

631 Soil Screening Guidance: User's Guide. (2007). US Environmental Protection Agency, 1-39.

632 Sui, H., Li, L., Zhu, X., Chen, D., Wu, G., 2016. Modeling the adsorption of PAH mixture in silica
633 nanopores by molecular dynamic simulation combined with machine learning.
634 Chemosphere 144, 1950–1959. doi:10.1016/j.chemosphere.2015.10.053

635 Syversen, T., Kaur, P., 2012. The toxicology of mercury and its compounds. J. Trace Elem. Med.
636 Biol. doi:10.1016/j.jtemb.2012.02.004

637 Tessier, A., Campbell, P.G.C., Bisson, M., 1979. Sequential extraction procedure for the
638 speciation of particulate trace metals. Anal. Chem. 51, 844-851.
639 doi:10.1021/ac50043a017.

640 Venter, O., Sanderson, E.W., Magrach, A., Allan, J.R., Beher, J., Jones, K.R., Possingham, H.P.,
641 Laurance, W.F., Wood, P., Fekete, B.M., Levy, M.A., Watson, J.E.M., 2016. Sixteen years of
642 change in the global terrestrial human footprint and implications for biodiversity
643 conservation. Nat. Commun. 7. doi:10.1038/ncomms12558

644 Vitousek, P.M., Mooney, H. a, Lubchenco, J., Melillo, J.M., 1997. Human Domination of Earth' s
645 Ecosystems. Science (80-. ). 277, 494–499. doi:10.1126/science.277.5325.494

646 Wang, Q., Yang, Z., 2016. Industrial water pollution, water environment treatment, and health
647 risks in China. Environ. Pollut. 218, 358–365. doi:10.1016/j.envpol.2016.07.011

648 Webb, G.I., Boughton, J.R., Wang, Z., 2005. Not so naive Bayes: Aggregating one-dependence
649 estimators. Mach. Learn. 58, 5–24. doi:10.1007/s10994-005-4258-6

Whyte, A., Ferentinos, K.P., Petropoulos, G.P., 2018. A new synergistic approach for monitoring wetlands using Sentinels -1 and 2 data with object-based machine learning algorithms. Environ. Model. Softw. 104, 40–54. doi:10.1016/j.envsoft.2018.01.023

Zacháry, D., Jordan, G., Völgyesi, P., Bartha, A., Szabó, C., 2015. Urban geochemical mapping for spatial risk assessment of multisource potentially toxic elements - A case study in the city of Ajka, Hungary. J. Geochemical Explor. 158, 186–200. doi:10.1016/j.gexplo.2015.07.015

Zapico, I., Laronne, J.B., Martín-Moreno, C., Martín-Duque, J.F., Ortega, A., Sánchez-Castillo, L., 2017. Baseline to Evaluate Off-Site Suspended Sediment-Related Mining Effects in the Alto Tajo Natural Park, Spain. L. Degrad. Dev. 28, 232–242. doi:10.1002/ldr.2605

672