5TH EURO-MEDITERRANEAN CONFERENCE FOR ENVIRONMENTAL INTEGRATION
2–5 OCTOBER 2023, RENDE (COSENZA), ITALY

Springer | Publishing Partner

The Impact of Compositional Data in Environmental Risk Assessment through Information Theory

María Pazo*[1], Teresa Albuquerque[2,3], Rita Fonseca[3], Joana Araújo[3], Natália Mota[3], Roberto Silva[3], Saki Gerassis[1].

[1] CINTECX, GESSMin Group, Natural Resources and Environmental Engineering, University of Vigo, 36310 Vigo, Spain.
[2] Instituto Politécnico de Castelo Branco, Polytechnic University, CERNAS, Castelo Branco, Portugal
[3] Institute of Earth Sciences, School of Sciences and Technology, University of Évora, Évora Portugal
Corresponding author: maria.pazo@uvigo.gal

Abstract. A water system impacted by mining activities was assessed to determine the extent of contamination, in the Trimpancho River mining system, in Spain. This system is in the Iberian Pyritic Belt, a metallogenic province in the southwest region of the Iberian Peninsula. Related pollution has been studied by multiple authors in recent decades. However, a pollution geochemical signature is not yet defined, even if, a few elements such as Cd, Cr, Cu, Fe, Hg, Mn, Pb, and Zn reach critical values, much above legislation for surface waters. Mercury is responsible for the highest level of hazard and therefore is central to defining water pollution signatures associated with acid drainage. Water samples were collected at the surface level of the streams, acidified with nitric solution, and stored in dark glass (only for Hg) and polyethylene containers at 4°C. Samples were digested with nitric and hydrochloric solutions in a high-pressure microwave unit and analyzed in ICP-OES for the majority of metals. Hg was directly analyzed in a mercury analyzer (NIC MA-3000). Since the chemical element concentration is compositional, an analysis was conducted to quantify how the uncertainty of the states of a to-be-predicted variable (mercury) is influenced by using both raw and centered log-ratio transformation (CLR) data. For that purpose, a methodology based on information theory (IT) and implemented through a Bayesian approach was used to about the obtained results, the normalized entropy decreased from 43% (raw data) to 33% (compositional data), and a Contingency Table Fit of 21% (raw data) was obtained compared to 71% (compositional data).

Keywords: Iberian Pyritic Belt, Pollution geochemical signature, Compositional data, Information Theory.

1 Introduction

The compositional nature of geochemical data and other geo- and environmental sciences has been studied since the 1980s, when J. Aitchison started delving into the development of *compositional data analysis* (CoDa), introducing what is now known as the *log-ratio approach* [1, 2, 3]. Specifically, compositional data is a representation of the parts of a whole, where each of its positive components (D) belonging to a random vector (Z) carries only relative information [4].

In this sense, it is widely recognized that the analysis and interpretation of regionalized compositions treated as raw data, although still commonly practiced, can easily lead to spurious correlations [5, 6]. Hence, the adoption of log-ratio transformation stands endorsed in the realm of environmental sciences, notably within the domains of geology and geochemistry [7, 8, 9]. One particularity of these analysis domains, such as contaminant flow control, is that in the field of uncertainty and probabilistic risk analysis, we can observe that both aleatory or stochastic in-situ uncertainty and epistemic or subjective in-situ uncertainty are entirely dependent on the quantity and quality of the available data [10].

Therefore, the present study introduces a novel methodology based on the Information Theory (IT) introduced by Claude Shannon [11]. The proposed approach recognizes the inherent significance of information theory as a differential tool for uncertainty interpretation, thereby tapping into its potential for informed decision-making a pollution geochemical signature is not yet defined and conducting probabilistic risk analysis. In this context, it also addresses the need to establish a well-informed pollution geochemical signature. For these purposes, a Bayesian machine learning (BayesianML) framework was developed to systematically assess regionalized compositions treated as raw data and establish a comparison with transformed variables using the centered log-ratio transformation (clr).

2 Materials and Methods

2.1 Data collection

The contamination levels of the Trimpancho River mining system in Spain, located in the metal-rich Iberian pyritic belt, were assessed. Water samples were collected from surface streams and tested for various elements including Al, Ca, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Zn, Sulfate, Phosphate, Nitrate, and Hg. For the analysis of the metallic elements, samples were acidified using a nitric solution, stored in polyethylene containers at 4°C, and processed using a high-pressure microwave unit with nitric and hydrochloric solutions. Analysis was conducted using ICP-OES. Mercury was determined in refrigerated samples stored in dark glass containers, using a mercury analyzer (NIC MA-3000) based on thermal decomposition, gold amalgamation, and cold vapor atomic absorption spectroscopy detection. Nitrates, phosphates, and sulfates were analyzed in non-acidified samples, nitrates by a portable photometer, and phosphates and sulfates by UV-Vis spectrophotometry.

2.2 Data transformation

The initial step of the analysis entails the transformation of the raw data to real space (clr-coefficients) based on the centred log-ratio transformation (clr) [3]. To that end, the compositional data transformation was performed using CoDaPack v2 software [12]

$$y = \mathrm{clr}(x) = \left[\ln \frac{x}{g_D(x)}\right] = [\ln \frac{x_1}{g_D(x)}, \ln \frac{x_2}{g_D(x)}, \ldots, \ln \frac{x_D}{g_D(x)}] \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^{D-1}$ and $g_D(x)$ is the geometric mean of the parts involved.

2.3 Entropy and Mutual Information

The objective of knowledge modeling and reasoning with BayesianML is to anticipate and understand the consequences of uncertainty, whether positive or negative. For that, one can express the quantification of normalized uncertainty $H_n(x)$ associated with the probability distribution of a variable X or a set of variables G as [11, 13]:

$$H_n(X) = \frac{H(X)}{\log_2(\phi_x)} = \frac{-\sum_{x \in X} p(x_i)\log_2(p(x_i))}{\log_2(\phi_x)} \tag{2}$$

where $x_i, \ldots, x_n$ represent the potential outcomes of X, each occurring with a corresponding probability of $p(x_i), \ldots, p(x_n)$, while ϕ_x represents the total number of states of a variable X. In addition to Shannon's entropy expression, another essential parameter in information theory is the mutual information (MI). The MI conceptually describes the interdependence between two variables, X and Y, by means of their information content. Mathematically, from an entropy perspective, MI can be expressed as:

$$\mathrm{MI}(X, Y) = H(X) - H(X|Y) \tag{3}$$

where H(X) represents the marginal entropy, and H(X|Y) the conditional entropy.

3 Results

3.1 Exploratory Analysis

In the first stage of analysis, a primary Bayesian model is created to evaluate the association rate of variables in terms of probabilistic relationships between nodes (Fig. 1).

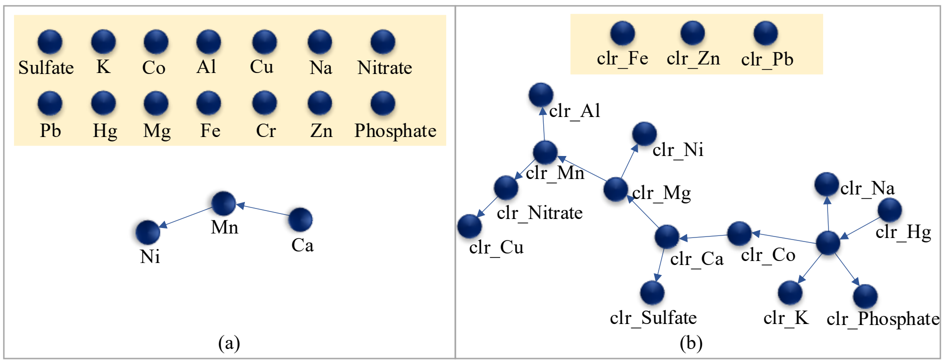

Fig. 1. Association rate analysis between (a) Raw data, and (b) clr data.

In Fig. 1, a significant improvement can be observed in uncovering potential relationships between variables when transformed compositional data is employed. Additionally, the uncertainty of the CoDa model was reduced from 43% (raw data) to 33%. Among these results, the most remarkable findings were seen in the Contingency Table Fit of the compositional data model, with a significant 41% improvement versus the raw data BayesianML network, from 21% to 70%.

3.2 Supervised Analysis on Hg

Fig. 2 shows the analysis of mutual information. The upper number in the box represents the information exchanged relative to the secondary node, while the red number refers to the main node. Otherwise, the symmetric measure of the information exchanged between each node and the target node is graphically represented by the thickness of the arc and its distance from the target node. This symmetric representation means that the amount of information, for example, supplied by node K about Hg is the same as the amount of information Hg supplied about node K. Thanks to this information analysis, it is possible to identify the predictive importance of variables such as Cr, K, Co, Na, and Co for understanding the state of Hg, considering the available data set in February of 2022. A new campaign was conducted in February 2023 and the new dataset will be reflected in the calibration model.

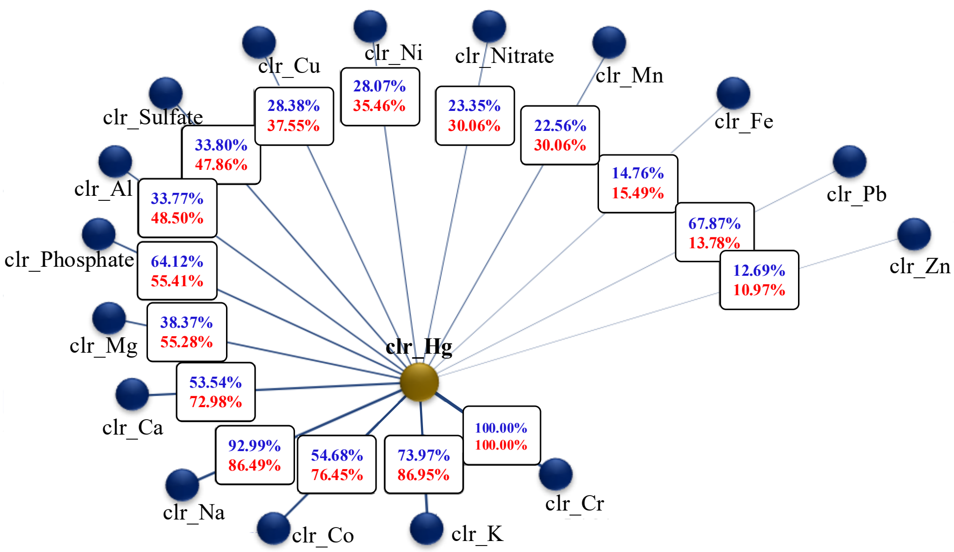

Fig. 2. Radial hierarchy layout from the highest-ranked node to the lowest-ranked information node.

4 Discussion

In mine exploration, environmental impacts, among others, the general focus, concerning geochemical signatures' definition is traditionally based on the uncertainty arising from sparse data and not on uncertainty arising from the model, even though the model is inferred, and its parameters estimated. The total Hg content of the Trimpancho River water in the present survey is the total Hg. The composition of Hg can be divided into different forms or pools, such as "dissolved" Hg, Hg associated with particulate and colloidal matter, volatile elemental Hg0, and labile (or reactive) Hg(II) [14]. The predictive importance of variables such as Cr, K, and Co, for Mercury's fate interpretation, can be explained by the colloid particulate-bound in the Hg forms, which can have identical association with the signalized elements. Divalent mercury, readily soluble as HgCl2, can finally explain the importance of Na, usually associated with Cl-, in the water column of Mediterranean rivers.

5 Conclusions

Considering the definition of future geochemical signatures, for the Trimpancho River's pollution characterization, the relationships between elements must be evaluated considering the Hg contents in its dissolved forms, in the water column, as well as the forms in which this element occurs in the deposited sediments, which represent the largest pool of this element in these polymetallic-sulfide mining areas, which enrich in cinnabar (HgS).

The collected samples were log-centred transformed after which a BayesianML analysis was carried out using the Information Theory fundaments for uncertainty and mutual information quantification. The findings revealed a significant increase in understanding of the study area by exploring transformed analytical data. In addition, CoDa not only facilitated the identification of preferred associations but also provided a comprehensive framework for defining water pollution signatures. The authors believe that this approach will provide valuable insights that will pave the way for more effective management and mitigation strategies in the future.

References

1. Aitchison, J., and Shen, S.M. Logistic-Normal Distributions: Some Properties and Uses. Biometrika 67(2), 261-272 (1980). https://doi.org/10.2307/2335470.
2. Aitchison, J. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society: Series B (Methodological) 44(2), 139-160 (1982). https://doi.org/10.1111/j.2517-6161.1982.tb01195.x
3. Aitchison, J. The Statistical Analysis of Compositional Data. Mono graphs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). pp 416. (1986).
4. Pawlowsky-Glahn, V., and Buccianti, A. Compositional Data Analysis: Theory and Applications (2011). https://doi.org/10.1002/9781119976462.ch17.

5. Pearson, K. Mathematical contributions to the theory of evolution. In the form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society 60(359-367), pp. 489-498 (1897). https://doi.org/10.1098/rspl.1896.0076
6. Pawlowsky-Glahn, V., & Egozcue, J. J. Spatial analysis of compositional data: A historical review. Journal of Geochemical Exploration, *164*, 28–32 (2016). https://doi.org/10.1016/J.GEXPLO.2015.12.010
7. Rollinson, H.R. Using Geochemical Data: Evolution, Presentation, Interpretation. Longman Scientific and Technical Press, 26, pp. 352. London (1993).
8. Schaeben, H. Vera Pawlowski-Glahn and Ricardo A. Olea: Geostatistical analysis of compositional data. Math Geol 39, 435–437 (2007). https://doi.org/10.1007/s11004-007-9105-9
9. Boente, C., Albuquerque, M. T. D., Gallego, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. Compositional baseline assessments to address soil pollution: An application in Langreo, Spain. Science of The Total Environment 812(15), 152383 (2022). https://doi.org/10.1016/J.SCITOTENV.2021.152383
10. Sagar, Daya, Qiuming Cheng, and Frits Agterberg. Handbook of mathematical geosciences: fifty years of IAMG. Springer Nature, (2018). https://doi.org/ 10.1007/978-3-319-78999-6
11. Shannon, C. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3).
12. Egozcue, J.J., Tolosana-Delgado, R., Ortego, M.I., eds. CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant Feliu de Guíxols. (2011)
13. Conrady, S., & Jouffe, L. Bayesian Networks and BayesiaLab – A Practical Introduction for Researches. Bayesia USA. (2015).
14. Bank, M., S. 2012. Mercury in the Environment: Pattern and Process. University of California Press. 1st Edition, 358 pp. ISBN 978-0-520-27163-0