Visual Analysis of Spinels with General Line Coordinates

Leandro Luque
Inst. for Computer Science and Eng. (CONICET-UNS)
Department of Computer Science and Engineering (UNS)
Bahia Blanca, Argentina
leandro.luque@cs.uns.edu.ar

María Luján Ganuza
Inst. for Computer Science and Eng. (CONICET-UNS)
Department of Computer Science and Engineering (UNS)
Bahia Blanca, Argentina
mlg@cs.uns.edu.ar

Ernesto Bjerg
INGEOSUR (UNS-CONICET)
Department of Geology (UNS)
Bahia Blanca, Argentina
ebjerg@ingeosur-conicet.gob.ar

Boris Kovalerchuk
Department of Computer Science,
Central Washington University
Ellensburg, United States
Boris.Kovalerchuk@cwu.edu

***Abstract*—In the geological context, the analysis of multidimensional data is a very common task and requires visualization techniques for exploration. General Line Coordinates (*GLC*) is a technique especially intended for lossless representation among the various techniques available for the visualization of multidimensional data. In this study, the application of *GLC* for pattern analysis and identification in mineral datasets is investigated. Furthermore, we develop a web-based tool to explore the possibility of employing *GLC* to leverage these approaches to derive visual discovery rules for pattern recognition.**

***Index Terms*—General Line Coordinates, Multidimensional Data, Geochemical Data**

I. Introduction

In recent years, the use of visualization techniques to facilitate pattern discovery in heterogeneous datasets has emerged as an alternative to traditional methods. The iterative process of visualization-based solutions allows refinement of employed methods and addressing different problem domains from various perspectives. Particularly in the field of geology, the use of visualization has been a significant advancement, condensing and simplifying a large amount of data mostly stored in tabular structures (e.g., spreadsheets) into more straightforward graphical representations.

Regarding data visualization techniques focused on minerals, various alternatives have emerged aiming to exploit these types of data with lossless and lossy approaches. The first approach has been extensively explored in the works of Ganuza et al [1] and Antonini et al. [2], offering visual representations based on geometric constructions such as prisms and their facet views to analyze data according to different combinations of elements. These visualizations not only serve as static representations but also provide suitable interactions to engage with and obtain more precise analyses according to domain experts' requirements.

The lossy approach is oriented towards employing current dimensionality reduction methods to handle datasets where the number of instances and dimensions are considerably larger that can be handled by traditional methods. [3]. While these techniques allow considering all the information, for some scenarios, it becomes confusing for experts because the reference to real dimensions is lost. Therefore, current solutions aim to combine dimensionality reduction representations with auxiliary views using techniques that employ raw dimensions.

Recently, a new family of lossless visualization techniques aiming to represent large datasets has emerged, called Generalized Line Coordinates (*GLC*). These techniques have different derivations listed in Antonini et al. [4] and Kovalerchuck [5], which can be employed in various contexts according to the data characteristics and expert requirements. One of the most important aspects to discern in this technique is whether it utilizes Non-Paired versions (where each dimension has its own geometric representation in space) or Paired versions (where dimensions are taken in consecutive disjoint pairs, each having its own representation in space).

This paper addresses the use of Paired GLCs, specifically the case of Shifted Paired Coordinates (SPC), for three main reasons:

- Its spatial representation is familiar to experts, as it involves scatterplots connected by polylines generated by points from different dimension pairs for each data item.
- They provide an alternative to the issues of occlusion and visual scalability faced by parallel coordinates, which are one of the traditional techniques employed by experts.
- It is a lossless information technique that maintains the context of all dimensions without performing any aggregation or transformation operations on them.

Our work presents a novel tool [1] for exploring mineral data with *GLC* and classifying unknown points with this technique. We also present an approach based on anchor or scope rules

[1] https://icic.uns.edu.ar/sandboxviz/

2375-0138/24/$31.00 ©2024 IEEE
DOI 10.1109/IV64223.2024.00047

to enhance the understanding of the classification proposed by our tool.

The paper is structured as follows: Section II introduces the mathematical formalism of *GLC* and the identified visual mappings. Sections III, IV, and V explain how the tool was designed and the interactions it supports. Sections VI and VII outline the workflow for conducting the analysis and real-world usage scenarios. Finally, we discuss the findings, limitations, and future work.

II. Definitions and Rational Design

In this section, the basic mathematical concepts for the development of *GLC* techniques are introduced, along with how the corresponding visual mapping will be created in the tool. To understand how Shifted Paired Coordinates (SPC) work, it's necessary to introduce the basic algorithm of the techniques from which they derive:

- Normalization of Dimensions: Normalize the dimensions to a range defined by the expert, typically [0, 1], for simplification purposes.
- Pairing Attributes: Group the attributes into consecutive disjoint pairs, such as $(x_1, x_2), (x_3, x_4), ..., (x_{n-1}, x_n)$.
- Plotting: Plot each pair in its corresponding normalized Cartesian system.
- Directed Graph Generation: Generate a directed graph between the generated pairs, for example, $(x_1, x_2) \rightarrow (x_3, x_4) \rightarrow, ..., (x_{n-1}, x_n)$.

For the specific case of *SPC*, each pair of values is plotted in its corresponding orthogonal system with its respective displacement. This shift is achieved by adding a scalar to the value pairs, creating different configurations for constructing directed graphs.

Regarding the development of the tool, the following visual mappings were taken into account:

Axes Layout: In this work, we introduce a layout based on Shifted Paired Coordinates, where dimensions are arranged in consecutive disjoint pairs, each represented by a Cartesian system. Dimensions are previously normalized to enable visual comparisons. Additionally, the positions of the different Cartesian systems can be freely chosen, but for convenience, they are arranged side by side.

Data Items: The items corresponding to each data point are represented by a polyline that connects positions in each Cartesian system.

Slope Change: Given that each segment of a polyline indicates the relationship between pairs of dimensions, its slope serves as a reference and can be employed to enhance visibility in the graph. In this case, we map the slope value to the alpha channel of a visual mark to make the zones where there is no correlation between consecutive pairs more visible.

Dimensional Reordering: A dataset can have multiple dimensions, it is important to identify which ones are most relevant for analysis. Therefore, reordering to optimize the search for visual patterns becomes a necessary task to enhance the quality of representation. For each pair of dimensions, we computed the Spearman rank-order correlation coefficient

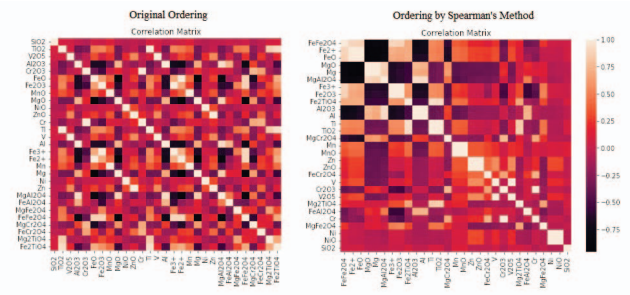

Fig. 1. Correlation Matrix among dimensions.

to evaluate the monotonic relationships, capturing both linear and non-linear associations. These coefficients were organized into a correlation matrix, where each cell (i, j) represents the correlation between the i_th and j_th dimensions(see Figure 1). In this way, the goal is to optimize coordinate systems that provide the most information to the user.

III. Tool

The system was built using D3.js [6], a modified version of npGLC [7], and offline processing methods for computing clusters and rendering tasks.

The tool developed for this work consists of three main views: *Setting View*, *VKD View*, and *GLC View* (see Figure 2).

The *Filter View* allows the user to perform configuration activities before analyzing the data, such as:

- Selecting which validation data will be plotted in the GLC View.
- Choosing which classes of the dataset to visualize.
- Setting the threshold value for HDBSCAN required to eliminate samples not belonging to one of the calculated clusters. These clusters are calculated offline on the server-side using the HDBSCAN [8] library to improve the performance of the tool. By default, the method's base parameters are sufficient for handling the dataset employed.
- Options for downloading and/or saving results obtained after the analysis.

The *VKD View* displays relevant information regarding the VKD rules generated when the user interacts with the GLC View, including the degree of purity of this rule set. The purity degree in this case is calculated as the accuracy value of the selected data by the VKD rules to belong to each class.

The *GLC View* is based on *SPC* representation over a set of predefined features that are of interest to domain experts. In this view, the user can interactively select regions of the data space using the rectangular brushing tool, which are potential candidates to be part of a VKD rule for data classification. The selections made are loaded and displayed in the VKD View, providing the user with continuous feedback on the construction of the rule set used to characterize a class or relevant data subset.

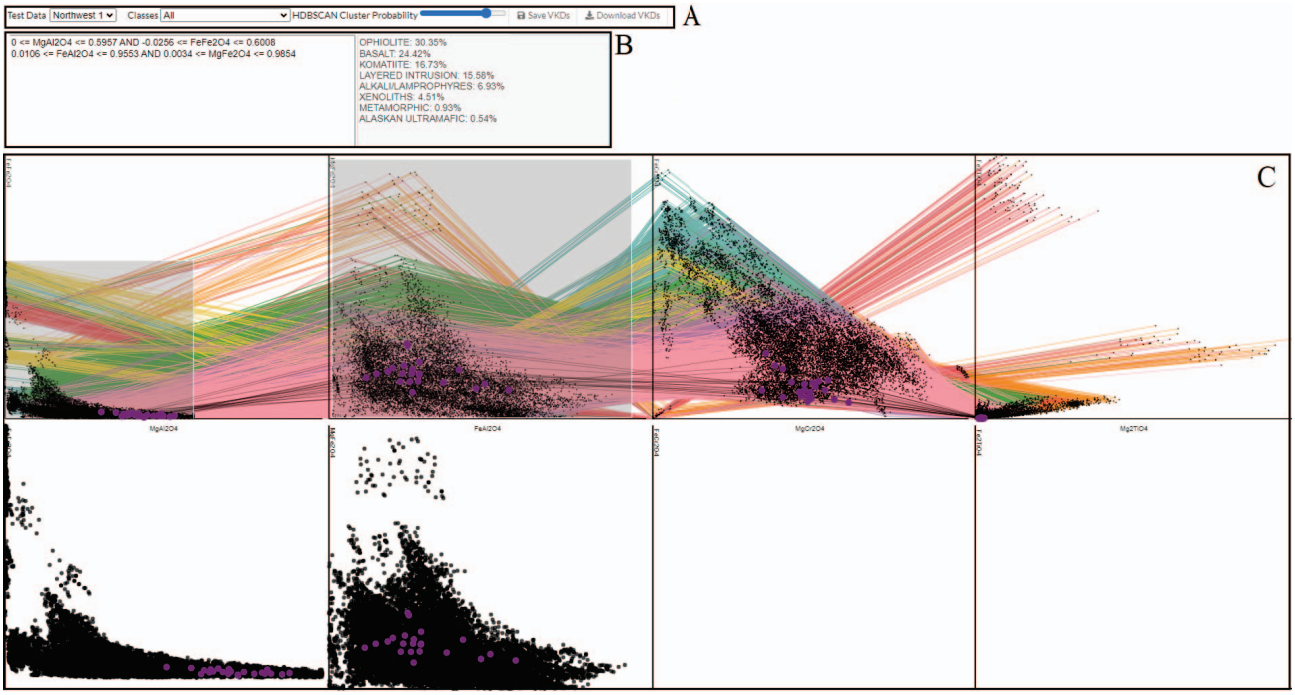

Fig. 2. Overview of our tool. (A) Filter View: Settings and auxiliary options, (B) VKD View: Contains the decision rules generated by the expert and the accuracy value for each class according to the selection, and (C) GLC View: Top row shows the GLC view based on SPC, where in addition to the polylines connecting each Cartesian system, points are represented for each pair of dimensions (black points). Bottom row displays the scatterplot representation of each Cartesian system of the SPC as an auxiliary view to display the data selected in the view above.

IV. Interactions

The developed tool features a set of brushing-based interactions that allow users to manipulate the SPC graph. Users can select regions in each of the defined orthogonal systems to highlight the attribute regions of interest.

When a region is selected in any of the coordinate systems displaying the polylines related to the training dataset and validation data, an auxiliary scatterplot representation is created with the zoomed-in selected points for interaction. In this linked view, users can remove points from the reference dataset to eliminate outliers that do not contribute to the purity of the classes, mostly due to measurement errors or invalid data according to their feature values (see Figure 3).

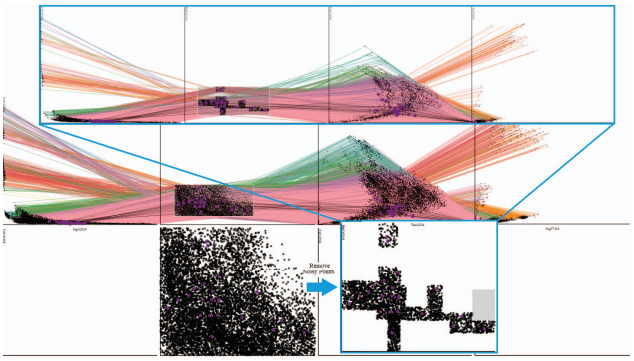

Fig. 3. Through multiple brush interactions, the expert removes instances from the auxiliary view in an attempt to determine which class the purple-colored validation points belong to.

V. Pipeline and Materials

In this section, the working methodology and the datasets used for the usage scenarios are introduced. It details how the developed tool allows the analysis of this data through classification techniques and outlier detection.

We utilized the dataset provided by Barnes and Roeder [9], which defined various tectonic settings based on multiple mineral features such as oxides (SiO2, TiO2, etc), cations (Si, Ti, etc), and end members (MgAl2O4, FeFe2O4, etc). Because an end member is a mineral that represents the extreme end of a mineral series in terms of purity, it is more suitable for analyzing this spinel dataset It is well-known that this dataset serves as a benchmark for the development of new techniques for mineral data analysis, making it particularly useful for the development and validation of new visual tools. This dataset has multiple classes that are relevant for most experts, such as BASALT, OPHIOLITE, XENOLITH, LAYERED INTRUSION, among others.

The validation data used consists of rock mineral datasets from the Northwest and South of Argentina. To maintain consistency with the B&R [9] dataset and to ensure comparability, the validation data went through the same preprocessing steps, which included the removal of samples with measurement

errors or for domain-specific reasons, as well as feature scaling.

The method employed in this work focuses on addressing two significant problems: detecting outliers and noisy instances and generating rules to characterize each class dataset [10]. Regarding the first problem, one of the major challenges in geological data is how to deal with samples that generate noise and do not belong to a defined group, and therefore are outliers. One of the reference techniques for this is HDBSCAN [11], [12], which uses density analysis to identify clusters without the need for user intervention. This method may be used to filter the dataset and provide better visual patterns by calculating the probability value of each sample's membership in a cluster.

To characterize the different classes and determine which ones the validation data belongs to, we employ an approach based on rules such as Scope Rules or Anchors [13]. Using this technique, decision rules based on the values that sample features should have in order to fall into a certain class may be extracted. These rules are extracted using interpretable models such as Decision Rules [14]. Another approach that can be used is the Box Classification Algorithm [15], which seeks to generate these rules by exploring different combinations of variable ranges that maximize sample membership in a class. Decision Trees may be used to guide the search for optimum rules, hence reducing the computation time required to determine optimal boxes.

VI. Usage Scenarios

This section presents two different usage scenarios based on two datasets provided by domain experts. To better understand the nature of the data we perform a preliminary analysis of the correlation, the importance of the features, and which of them are more relevant for the domain experts. After that, the chosen subset of features was only end-members. These features are continuous values that represent a percentage of the total composition of a rock. The pre-processing step involves normalization and removing null values to have a consistent dataset.

Norhtwest Dataset

The first usage scenario focuses on a set of mineral data from Northwest Argentina, divided into two sub-datasets. In this case, the threshold value is defined as 0.7 to remove data that do not contribute to the analysis. After this step, it can be observed that the structural behavior of each class is more compact while still preserving some outliers.

When visualizing the validation data corresponding to each sub-dataset, some particular behaviors can be observed:

Sub-dataset 1: When the user selects a box covering the validation data in each Cartesian system (see Figure 4), it is evident that they may belong to two possible classes: BASALTS (Figure 5-left) and OPHIOLITES (Figure 5-right). If the user relaxes the box boundaries and leaves out some validation points that visually appear to behave differently, the purity degree improves for one class over the other.

Fig. 4. VKD rules on Northwest data (Subdataset 1). At the top, it is possible see the rules generated from the completed sections and the possible classes to which those items correspond.

The VKD rules extracted after the user interacted with GLC View were:

IF MgAl2O4 is in [0.3617, 0.3617] ***AND***
FeFe2O4 is in [0.0023, 0.0328] ***AND***
FeAl2O4 is in [0.1085, 0.4106] ***AND***
MgFe2O4 is in [0.0853, 0.2175] ***AND***
MgCr2O4 is in [0.3468, 0.6213] ***AND***
FeCr2O4 is in [0.0514, 0.1480] ***THEN*: class BASALT**

Sub-dataset 2: In this situation, we have a scenario similar to the previous one, but now the user's established VKD rules help differentiate between two classes: LAYERED INTRUSIONS and BASALT. Due to the nature of the data, these classes have many similar distributions in these attributes leading to confusion in determining the class to which a point belongs. In this case, the user can use the removing filters in the auxiliary scatterplot to remove noisy sub-regions from the main selection over SPC (see Figure 6). When the user finishes removing points, we can see that the percentage purity degree improves in comparison with the original state. For these scenarios where the classes are not well specified due to outliers or points with strange behavior, experts appreciate this interaction for mitigating visual occlusion on *GLC View*.

South Dataset

This second usage scenario is focused on data extracted from South Argentina, where the domain expert needs to identify to which classes the data belongs. Initially, the data are presented as having a high probability of belonging to the Xenolith class. The expert makes selections in the different Cartesian systems to delimit the regions of interest in the multiclass context. In this way, it is found that besides the Xenoliths class (Figure 8), the Basalts (Figure 9) and Ophiolites (Figure 10) classes are also candidates for the validation data.

Although the analysis diverges from the expected behavior, it gives rise to a new type of search for the reason behind this prediction. For this purpose, the VKD View is useful for understanding how the attributes interact with each possible class.

The VKD rules extracted after the user interacted with GLC View were:

IF MgAl2O4 is in [0.4149, 0.9638] ***AND***
FeFe2O4 is in [0.0000, 0.0548] ***AND***

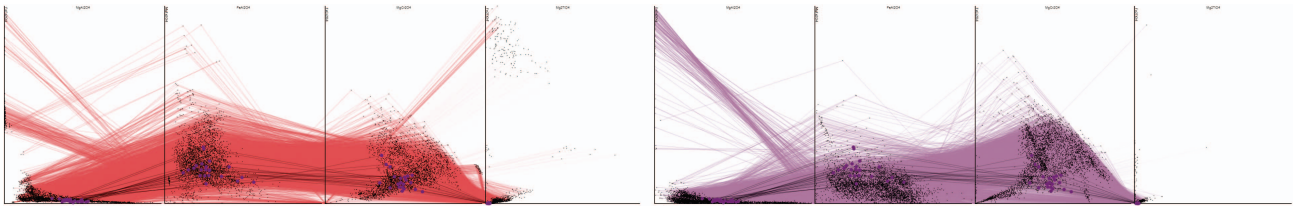

Fig. 5. Class Basalt on Northwest data (subdataset 1) on the left and class Ophiolite on Northwest data (subdataset 1) on the right.

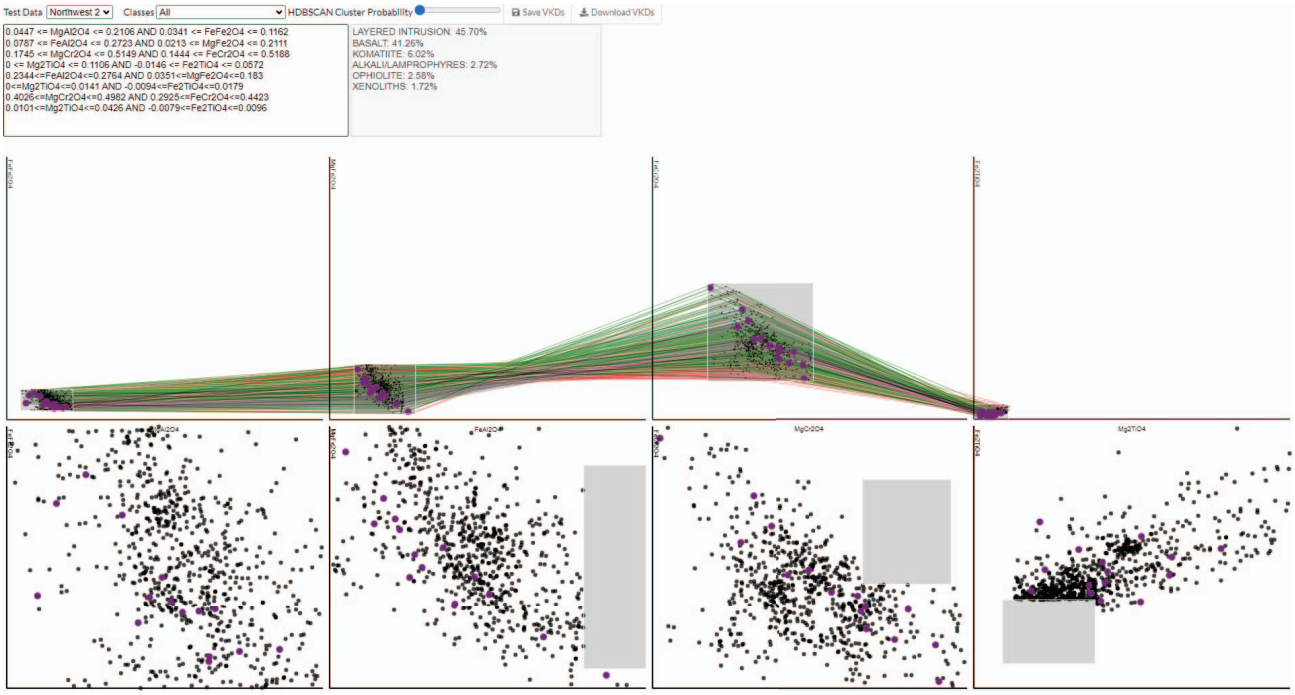

Fig. 6. Selection and filtering on *SPC* to improve the purity of the classes of validation data.

FeAl2O4 is in [0.1957, 0.5277] ***AND***
MgFe2O4 is in [0.0023, 0.1548] ***AND***
MgCr2O4 is in [0.1085, 0.4872] ***AND***
FeCr2O4 is in [0.0068, 0.1881] ***AND***
Mg2TiO4 is in [0.0000, 0.0979] ***AND***
Fe2TiO4 is in [0.0000, 0.0277 ***THEN*: class XENOLITH**

With the rules obtained through interactions on the *GLC View*, the expert can determine the attribute ranges that determine the probability of belonging to a class. Furthermore, since these are interval rules describing how this visual data classification mechanism works, they can be compared with similar analytical models such as Decision Trees, Random Forests, Linear Regression, etc.

VII. Discussion and Conclusions

This paper presents an application of *GLC* techniques in geological sciences to facilitate pattern discovery. While most current geological solutions are based on dimensionality reduction, the use of SPC allows for the preservation of original attributes. This quality is required by domain experts in their analyses, in addition to the ability to interact with the graph to enhance data interpretation.

However, one of the major challenges for domain experts is that utilizing *GLC* in its various forms (*SPC*, etc) requires a deep understanding of them. This is because the best *GLC* representation is obtained by testing different combinations of configurations such as the positioning of Cartesian systems and dimension ordering. Thanks to the use of noisy point removal filters, it is possible to refine the VKD rules to achieve a better interpretation of the predicted classes for the validation data.

As future work, the development of a pipeline for visual analysis assisted by user feedback is proposed, which would allow geologists to find the best GLC configuration to solve a task.

References

[1] M. L. Ganuza, G. Ferracutti, M. F. Gargiulo, S. M. Castro, E. Bjerg, E. Gröller, and K. Matković, "The spinel explorer—interactive visual analysis of spinel group minerals," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1913–1922, 2014.

Fig. 7. Selection of points from the South Dataset to determine possible classes they may belong to according to the Barnes and Roeder Dataset [9]. At the top, the accuracy values for each candidate class can be observed, as well as the associated decision rules. The black lines represent the data from the South Dataset, while the colored lines in the background correspond to the possible classes of spinels.

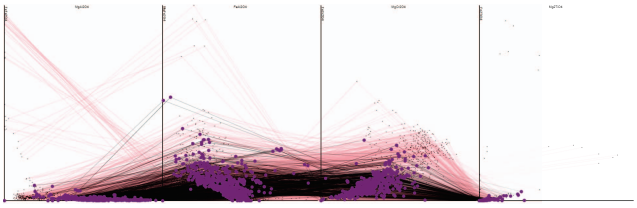

Fig. 8. South data comparing to Xenolith class

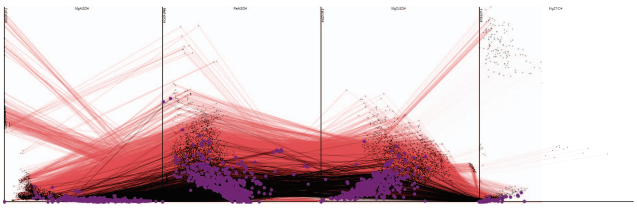

Fig. 9. South data comparing to Basalt class

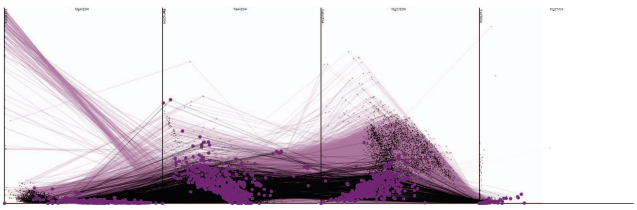

Fig. 10. South data comparing to Ophiolite class

[2] A. S. Antonini, M. L. Ganuza, G. Ferracutti, M. F. Gargiulo, K. Matković, E. Gröller, E. A. Bjerg, and S. M. Castro, "Spinel web: an interactive web application for visualizing the chemical composition of spinel group minerals," *Earth Science Informatics*, vol. 14, no. 1, pp. 521–528, 2021.

[3] T. Horrocks, E.-J. Holden, D. Wedge, C. Wijns, and M. Fiorentini, "Geochemical characterisation of rock hydration processes using t-sne," *Computers & geosciences*, vol. 124, pp. 46–57, 2019.

[4] A. S. Antonini, L. Luque, M. L. Ganuza, and S. M. Castro, "Toward a taxonomy for 2d non-paired general line coordinates: A comprehensive survey," *International Journal of Data Science and Analytics*, vol. 15, no. 2, pp. 133–158, 2023.

[5] B. Kovalerchuk, *Visual knowledge discovery and machine learning*. Springer, 2018, vol. 144.

[6] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[7] L. E. Luque, M. L. Ganuza, A. S. Antonini, and S. M. Castro, "npglc-vis library for multidimensional data visualization," in *Conference on Cloud Computing, Big Data & Emerging Topics*. Springer, 2021, pp. 188–202.

[8] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[9] S. J. Barnes and P. L. Roeder, "The range of spinel compositions in terrestrial mafic and ultramafic rocks," *Journal of petrology*, vol. 42, no. 12, pp. 2279–2302, 2001.

[10] V. Estivill-Castro, E. Gilmore, and R. Hexel, "Constructing interpretable decision trees using parallel coordinates," in *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*. Springer, 2020, pp. 152–164.

[11] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.

[12] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 1, pp. 1–51, 2015.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] J. Fürnkranz, D. Gamberger, and N. Lavrač, *Foundations of rule learning*. Springer Science & Business Media, 2012.

[15] B. Kovalerchuk and H. Phan, "Full interpretable machine learning in 2d with inline coordinates," in *2021 25th International Conference Information Visualisation (IV)*. IEEE, 2021, pp. 189–196.