International Journal of Applied Science and Technology
ISSN: 2221-0997 (Print), 2221-1004 (Online)

BRPI

Original Article | Open Access

Using Multivariate Statistical Analysis to Validate a Geochemical Typology Model: A Case Study of the New Model Proposed for Paraná Igneous Province in the State of Paraná, Brazil

Alexandre Cancian Bajotto[1], Anselmo Chaves Neto[2], and Otavio Augusto Boni Licht[3]

[1]Graduate Program in Numerical Methods in Engineering – PPGMNE, Federal University of Paraná – UFPR, Rua Francisco Heráclito dos Santos, 100, Centro Politécnico, Jardim das Américas, CEP 81531-980, Curitiba, Paraná, Brazil; alexandre.bajotto@ufpr.br; 55 41 98889 9223.

[2]Graduate Program in Numerical Methods in Engineering – PPGMNE, Federal University of Paraná – UFPR, Rua Francisco Heráclito dos Santos, 100, Centro Politécnico, Jardim das Américas, CEP 81531-980, Curitiba, Paraná, Brazil; anselmo@ufpr.br.

[3]Graduate Program in Geology, Federal University of Paraná – UFPR, Rua Francisco Heráclito dos Santos, 100, Centro Politécnico, Jardim das Américas, CEP 81531-980, Curitiba, Paraná, Brazil; otavio.licht@gmail.com.

Address for Correspondence:

Alexandre Cancian Bajotto, Graduate Program in Numerical Methods in Engineering – PPGMNE, Universidade Federal do Paraná – UFPR, Rua Francisco Heráclito dos Santos, 100, Centro Politécnico, Jardim das Américas, CEP 81531-980, Curitiba, Paraná, Brazil. (alexandre.bajotto@ufpr.br; 55 41 98889 9223)

Article History:

Received: 1 April 2025; Accepted: 26 April 2025; Published: 29 April 2025

Abstract In 2018 a new rock classification model for Paraná Igneous Province (PIP) was proposed using four discriminant variables: SiO_2, Zr, TiO_2, and P_2O_5. This research introduces the Multivariate Analysis as a technique to verify and check this new model. The study used a subset with samples from boreholes in the state of Paraná, Brazil. The Subset is a matrix of 1,030 observations x 73 geochemical variables and Factor Analysis reduced the subset to a Factor Scores matrix of 1,030x24. Then, Cluster Analysis grouped the Factor Scores in 5 clusters using K-Media Method, followed by a Pattern Recognition and Classification Analysis using Neural Networks to corroborate the clustering. The Principal Component Analysis extracted Principal Component Values and Weights from the geochemical types proposed by the 2018 model. Finally, the Canonical Correlation Analysis correlated vector U1 of Factor Scores with V1 of Principal Component Values and Weights and the result was the validation of the new classification model.

Keywords Lithogeochemistry; Multivariate Analysis; Factor Analysis; Cluster Analysis, Pattern Recognition and Classification Analysis, Principal Component Analysis; Canonical Correlation Analysis.

Volume 14, 2025

Publisher: The Brooklyn Research and Publishing Institute, 442 Lorimer St, Brooklyn, NY 11206, United States.

DOI: 10.30845/ijast.vol14p2

Reviewers

Dr. Tapas Pal, Assistant Professor of Geography, Raiganj University, India; International Collaborator, GEITEC, Federal University of Rondônia, Brazil; Email: Geo.drtapaspal.in@gmail.com.

Caroleen Auma Barasa, Kenyatta University, Kenya; Email: caroleenbarasa@gmail.com.

Citation: Bajotto, A. C., Chaves Neto, A., & Licht, O. A. B. (2025). Using Multivariate Statistical Analysis to Validate a Geochemical Typology Model: A Case Study of the New Model Proposed for Paraná Igneous Province in the State of Paraná, Brazil. *International Journal of Applied Science and Technology*, *14*, 11-24. https://doi.org/10.30845/ijast.vol14p2

1. Introduction

Over the last decades many authors have produced geochemical datasets and have conducted research regarding Paraná Igneous Province (PIP) characterization. Peate et al. (1992) introduced the currently most accepted grouping method based on the concentration of the chemical elements and ratios of oxides and trace elements proposing a model with six clusters of basaltic rocks: Esmeralda, Gramado, Ribeira, Paranapanema, Pitanga and Urubici, and two groups of acidic rocks: Palmas and Chapecó. Figure 1 presents the organization chart of Peate et al. (1992) proposed model.

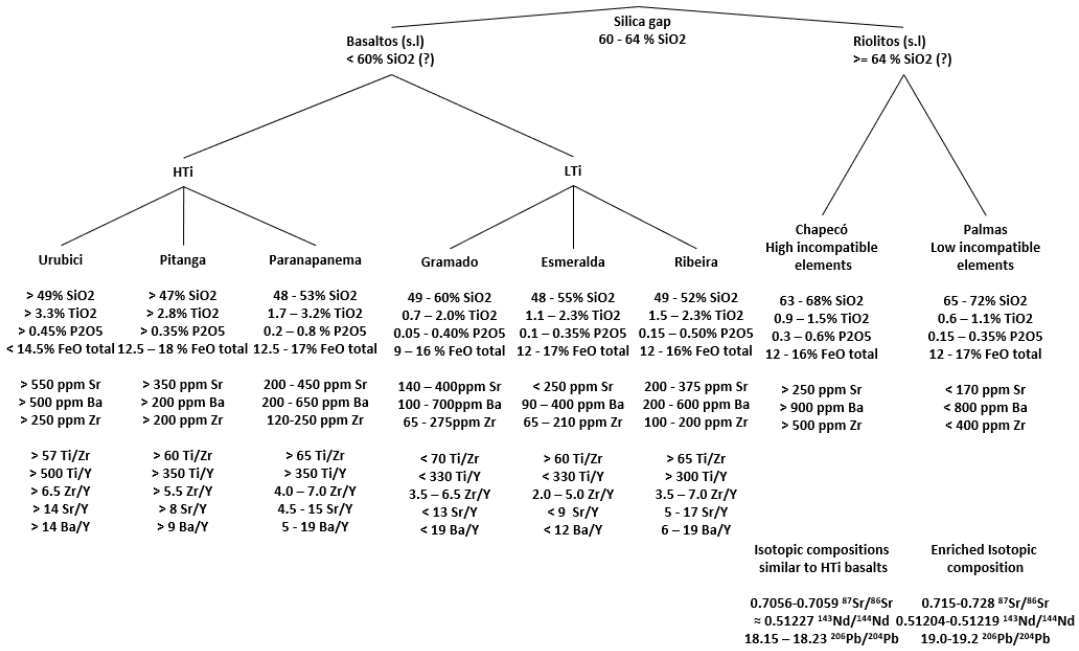

Figure 1 – Organization chart for PIP rock classification in magma types. Source: Peate et al. (1992).

Figure 1 shows the criteria to define into which category a sample must be allocated. However, Licht (2018) pointed out that the model contained criteria overlaps among the categories, making the classification process very difficult to replicate. For example: since rhyolites are the rocks with SiO_2>=64% and 'basalts' (latu sense, i.e., basalts and *andesibasalts*) are the rocks with SiO_2<60%, how would a rock be classified if its SiO_2 concentration is between 60% and 64%?

Figure 2 shows the natural gaps of four variables that were later used by Licht (2018) to build his classification model. The gaps allowed the researcher to split the elements into high and low concentrations.

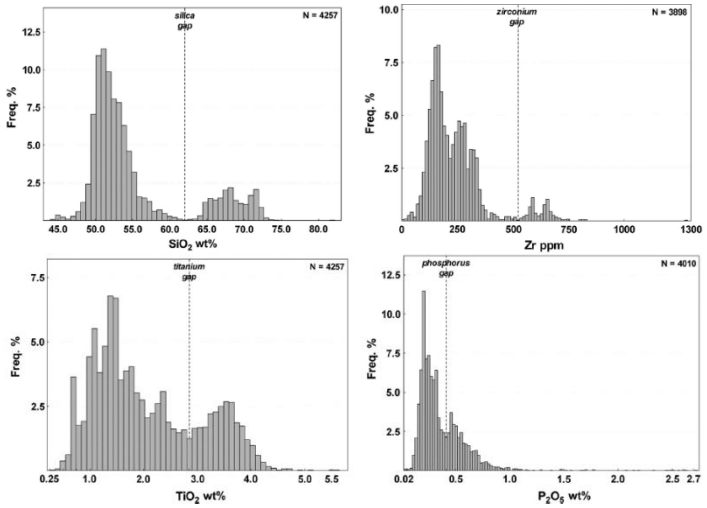

Figure 2 – Natural gaps of the four discriminant variables. Source: Licht (2018).

The selection of these variables was not by chance, as previous authors had proposed using them as criteria for rock classification. Licht (2018) cited that Rüegg (1969, 1970, 1975, 1976), Rüegg and Amaral (1976), Bellieni et al. (1983, 1984 a, b), Atalla et. a. (1983, 1984a, b), Sousa (1983), Marques (1983), Mantovani et al. (1985, 1988), Bellieni et al. (1986), Fodor (1987), Petrini et al. (1987), also adopted SiO_2, Zr, TiO_2 and P_2O_5 as a geochemical indicator to discriminate PIP rocks groups.

This research innovates in the application of multiple Multivariate Statistical Analysis techniques to check and verify Licht (2018) model, which is a geochemical rock-type classification model originally built using descriptive statistics and probability plots of chemical element's concentration.

The goal of applying multiple Multivariate Analysis techniques was to demonstrate that these techniques are robust statistical analysis to check and verify research hypotheses and conclusions, to present a more comprehensive statistical study of the dataset and analyze if Licht (2018) model is a valid and consistent model to be applied for PIP rock classification.

Then, the objectives of this article are: (1) to present the exploratory study performed to build the subset; (2) to show the application of Factor Analysis and Principal Component Analysis in geochemical datasets to reduce matrix dimensions and extract data in a different form without losing important data information; (3) to demonstrate that the Ward and K-Means Cluster Analysis Methods are effective for grouping Factor Scores datasets; (4) to use the Pattern Recognition and Classification Analysis of Neural Networks to validate the Cluster Analysis; and (5) to verify if Licht (2018) proposed model is a valid model for PIP rock classification by applying the Canonical Correlation Analysis between (a) the Factor Scores, grouped using the Cluster Analysis techniques, and (b) the Principal Component Values or Weights, extracted from the types proposed by Licht (2018).

Subsection 2 of this paper introduces Multivariate Analysis techniques and previous research related to their use in geochemistry or geology. Section 3 shows the materials and methods used in this research, followed by section 4 that presents the results and the discussion, and section 5 with the conclusion.

It can be considered that the results presented in this article may have petrological implications for PIP petrology. In spite of that, this paper is not a petrology research and, therefore, it does not examine important issues such as the origin of the basaltic magma or the processes of crustal contamination or the evolution of magma from its source to surface. Some petrological aspects of this innovative approach and evolutionary model are addressed in Gomes et al. (2018).

2. Multivariate analysis techniques and their use in geology and geochemistry

2.1 Multivariate analysis

2.1.1 Factor analysis

During the analysis of a dataset in which the observations are classified according to different variables, there may be a relationship between these variables. The Factor Analysis describes the covariance structure of the relationships among many variables in terms of a few fundamental random but unobservable (latent) variables, called Factors. All variables within a particular group are highly correlated with each other but have relatively low correlations with variables from a different group. Each group of variables represents a Factor, which is responsible for the observed correlations (Johnson and Wichern, 1998). Then, Factor Analysis reduces the dimensions of the data and, by doing this, statistically creates new data – the Factor Scores – that contain the same information as the original data. This research used the Factor Scores as the first vector, or first Canonical Variable (U1), in the Canonical Correlation Analysis.

2.1.2 Principal component analysis

Similarly to Factor Analysis, Principal Component Analysis seeks to explain the relationship between variables of a dataset using the variance-covariance structure of the data matrix through uncorrelated linear combinations of the p original variables. The maximum number of new variables (Principal Components) that explain the data variability is equal to the number of original variables. However, the variability can often be explained by a smaller number k of Principal Components (Sharma, 1996).

The k Principal Components can replace the p initial variables with minimal loss of information, and the original dataset consisting of n measurements of the p variables is reduced to one formed by n measurements of the k Principal Components (Chaves Neto, 2023).

The dataset received of 1,374 samples was previously classified according to the 16 types proposed by Licht (2018). After selecting only 1,030 samples of boreholes, the exploratory analysis showed that this subset was formed by 5 geochemical types. Then, the samples were grouped by each of the 5 types, and for each one of these types, both the Principal Component Values and Weights were extracted.

The objective of the Principal Component Analysis was to reduce the dimensions of the matrix of each subset of the 5 types and statistically generate a second vector, the second Canonical Variable (V1), to be used in the Canonical Correlation Analysis and this time, the vector derived from Licht (2018) classification model and contained the same information of the original data.

2.1.3 Cluster analysis

For Johnson and Wichern (1998), Cluster Analysis provides information about the dimension, identifies outliers, and suggests hypotheses about the relationship between the samples in each group, since it is the data itself, or the similarities between the data, which form the groups.

Therefore, Cluster Analysis is a technique that produces groups of objects or individuals, represented by vectors, which are approximately homogeneous within each group, but heterogeneous from group to group. Thus, the formation of groups is based on similarities or distances (Johnson and Wichern, 1998). Then, Cluster Analysis divides the data into groups so that the set of observations that form a group are more similar to each other than the observations from other groups (Jolliffe, 2002).

When items (units or cases) are grouped, proximity is usually indicated by some sort of distance. The Euclidean Distance or the Squared Euclidian Distance is preferred in Cluster Analysis.

This research used the Ward Hierarchical Method and Squared Euclidian Distance for the initial grouping and K-Means Non-Hierarchical Method and the Squared Euclidian Distance to optimize the grouping results to form the groups of Factor Scores.

2.1.4 Patter recognition and classification

Discrimination and Classification Statistical techniques are incorporated into a broader context, which is Pattern Recognition. Together, Mathematical Programming and Neural Networks are part of a set of procedures used in the recognition and classification of objects and individuals. Pattern Recognition and Classification techniques aim to recognize items or individuals based on their characteristics and classify them into one of the previously defined categories following rules constructed using the features of the categories. By treating each observation as a pattern, it was possible to create rules for recognizing and allocating the 1,030 patterns into groups.

Then, considering the 24 Factor Scores of the first subset as observable variables of the 1,030 samples, a Neural Network Rule was applied to verify whether the classification into 5 groups through the K-Means Method was compatible.

2.1.5 Canonical correlation analysis

Canonical Correlation Analysis evaluates the correlation between the linear combination of a group with p variables X and the linear combination of another group Y with q variables and seeks to identify how strong the correlation between these groups of variables is (Johnson and Wichern, 1998).

The objective of the Canonical Correlation is to determine the linear combinations $\underline{c}_1' x$ and $y_2' x$ that have the highest possible correlation. Such correlations can provide valuable information about the relationship between the two sets of variables (Chaves Neto, 2023).

In the context of this research, the Canonical Correlation related the Factor Scores, extracted from Factor Analysis and grouped by Cluster Analysis, as the Canonical Variable U1 and the Principal Components Values or Weights, extracted from Litch (2018) classification model, as the Canonical Variable V1. If the results present a strong correlation between Canonical Variables, the Licht (2018) classification model would be verified by Multivariate Analysis techniques, and it would be considered a valid model for PIP rock classification.

2.2 The use of multivariate analysis in geochemical research

It is relevant for this paper not only to verify whether the classification model proposed by Licht (2018) is a valid model or not, but also to demonstrate that (a) the Multivariate Analysis techniques applied to geochemistry and geology can also serve to validate both hypothesis and tests carried out in the laboratory, and (b) that other studies have used one or two Multivariate techniques for analysis, different from this research that used multiple techniques to analyze its dataset and check Licht (2018) model.

For example, Giuseppe et al. (2014) used Factor Analysis to analyze the density of sediment particles in 8 samples from the Reno River and 8 samples from the Po River in the province of Ferrara, Italy: in both cases, 4 high-density samples and 4 low-density samples. The authors applied the Varimax rotation and Kayser-Meyer-Olin (KMO) test to verify if the Factor Analysis was applicable to the case study. The original data contained 19 geochemical variables that were reduced to 4 Factors. The authors considered that Factor Analysis was a good discriminator to identify density differences between sediment origins.

Zhang et al. (2022), for example, used Multivariate Analysis techniques to analyze samples from 52 sphalerite (PbS) deposits in southern China. The authors concluded that (1) overlapping geochemical effects imposed severe restrictions on determining the influence of a single factor on the distribution of elements in rock formations, and (2) the combination of Factor Analysis and Principal Component Analysis provided an efficient approach to highlight the relationships between elements and to reveal the potential causes that contribute to the significant differences in trace elements that compose the southern China's block minerals.

Additionally, Ghorbani et al. (2022) used Principal Component Analysis to study the concentration of chemical components in samples of bark, branches, and leaf tips of black spruce (*piceamariana*). The analysis resulted in 2 Principal Components that explained 79% of the sample variability and the Regression using Principal Components as variables showed that the association of gold (Au) and arsenic (As) represented an indicator of the presence of gold in the Twin Lakes deposits in Manitoba, Canada.

Furthermore, Ragland et al. (1997) used Principal Component Analysis and Discriminant Analysis to analyze the Martinsville igneous complex in Virginia, United States of America. In the study, the authors compared the results of multivariate analyses with more traditional geological approaches and found that multivariate approaches not only grouped and classified the data in the same way as more traditional approaches but also captured and portrayed the petrogenic significance of several mineralogical groupings.

Another example of Multivariate Analysis in geology was the usage of Cluster Analysis for the characterization of rock formations, or their boundaries. The K-Means Method was used by Fonteles et al. (2020) to group 1,384 samples taken from an iron mine in Bonito, in Rio Grande do Norte, Brazil. The researcher considered concentrations of Fe_2O_3, SiO_2, Al_2O_3, P, and Mn as variables. The researchers showed that 8 clusters provided the most appropriate typological arrangement in terms of balance between the geochemical profiles of the samples and statistical restrictions.

In another clustering study, Ghezelbash et al. (2020) used the traditional K-Means Method and the Hybrid Genetic K-Means Clustering (GKMC) in Varzaghan District, northwestern Iran, to delineate target areas for copper exploration in porphyry/skarn deposits.

These studies have shown that Multivariate Analysis can be of great use in geology and geochemistry by bringing statistical techniques to validate data collected in the field and analyzed in the laboratory. However, none of the mentioned studies applied more than two Multivariate techniques and none was developed in the Paraná Igneous Province, Brazil. Therefore, this research innovates in applying various Multivariate techniques to a large sample of rocks from PIP.

2.3 Licht (2018) classification method

Using Peate et al. (1992) approach, Licht (2018) tried to classify a dataset composed of 5,974 samples, collected from all over PIP through the years and reached a success rate of 53%. The low success rate can be explained by the following reasons: (1) the classification model proposed by Peate et al. (1992) was built using a restricted database; (2) the criteria the researchers used were too complex to be applied to large databases; (3) the "step-by-step" application contains ambiguous and overlapping limits; and (4) the criteria consider all samples contained in the database, not discarding those with LOI>2% (Licht, 2018).

Therefore, these caveats were considered sufficient justifications for deepening the research to establish a classification system based on well-defined and clear numerical criteria with no overlaps between categories and that could be applied to large and small datasets (Licht, 2018).

Then, Licht (2018) proposed a classification method based on the natural gaps identified in the frequency of the concentration distribution curves of four discriminant variables: SiO_2 (62.02%), Zr (522.15 μg/g), TiO_2 (2.85%) and P_2O_5 (0.413%). The gaps allowed the researcher to combine these variables and, for each one, propose a Low and High band. Then, the combination of the four variables and two concentration bands resulted in the creation of 16 geochemical possible types of rocks for PIP.

The first criterion applied to the dataset with 5,974 samples was to drop the samples with LOI>2%. The remaining 3,296 samples were classified into 16 types. Figure 3 shows the classification tree proposed and the number of samples allocated to each type of rock. Of the 16 statistically possible types, only 8 types presented samples, meaning that only 8 types are geochemically valid: types 1, 2, 3, 4, 9, 10, 13, and 14. Type 10, despite being a valid geochemical combination, had only so few samples (10) which raised questions about the real incidence of this type. Then, for the purpose of this research, only 7 types were considered valid: 1, 2, 3, 4, 9, 13, and 14.

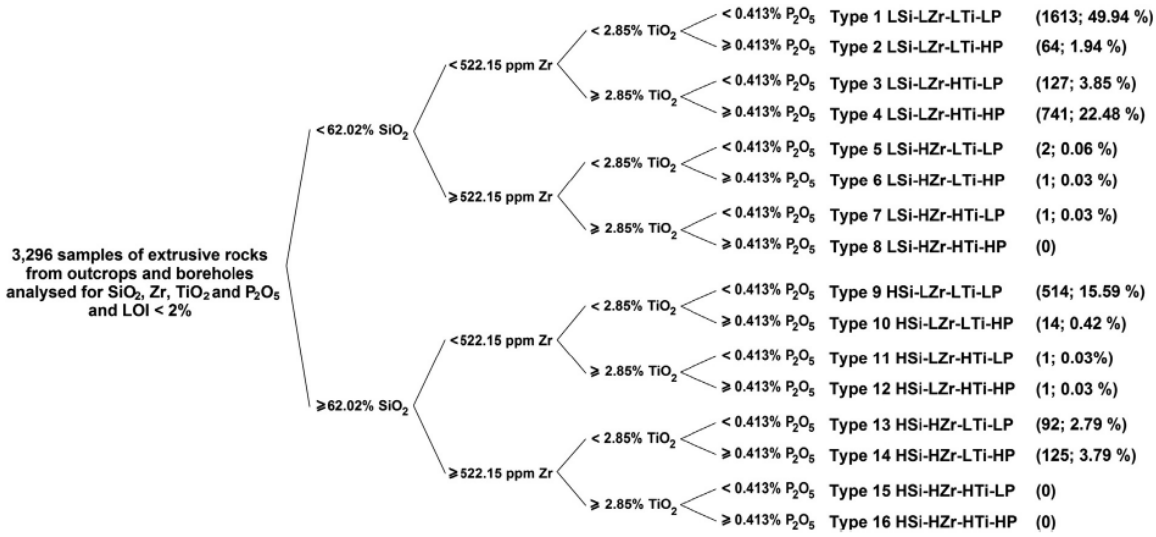

Figure 3 – Organization chart of the geochemical types based on the Si-Zr-Ti-P gaps and results of its application to the geochemical database; inside parentheses are the number of samples and their relative frequency. Source: Licht (2018).

3. Materials and methods

3.1 The database and the exploratory analysis

The research analyzed a database consisted of 1,374 samples and 73 variables – a 1,374x73 matrix – extracted from the dataset used by Licht (2018). The original set was built using the information collected from the analysis of samples from all over PIP using a combination of techniques, such as: FRX, ICP-MS, cold vapor generation and Fire-Assay. The subset with 1,374 samples studied in this research contain samples collected from outcrops and drillings of vertical water production boreholes in the state of Paraná, Brazil.

The concentration of the geochemical variables, elements and oxides, was expressed in parts per billion (ppb), parts per million (ppm) or percentage (%). The exploratory analysis of the 1,374x73 identified that there were 1,030 samples from drilling boreholes and this subset contains these samples classified in 5 geochemical types.

Figures 4 and 5 present a SW-NE drilling line that started in Foz do Iguaçu and ended in Bandeirantes and the cross-section of the drillings that generated some samples.

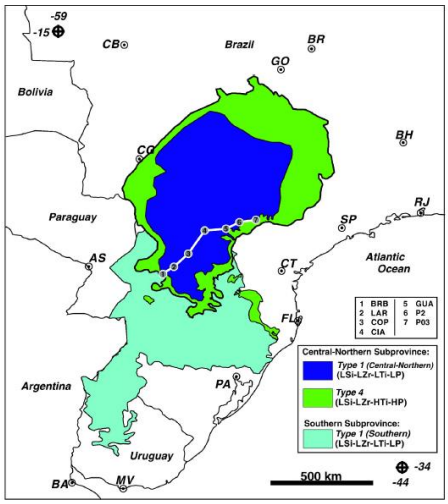

Figure 4 – Map of the distribution of the main low silica geochemical types in the PIP with the location of the boreholes plotted. Source: Gomes et al. (2018).

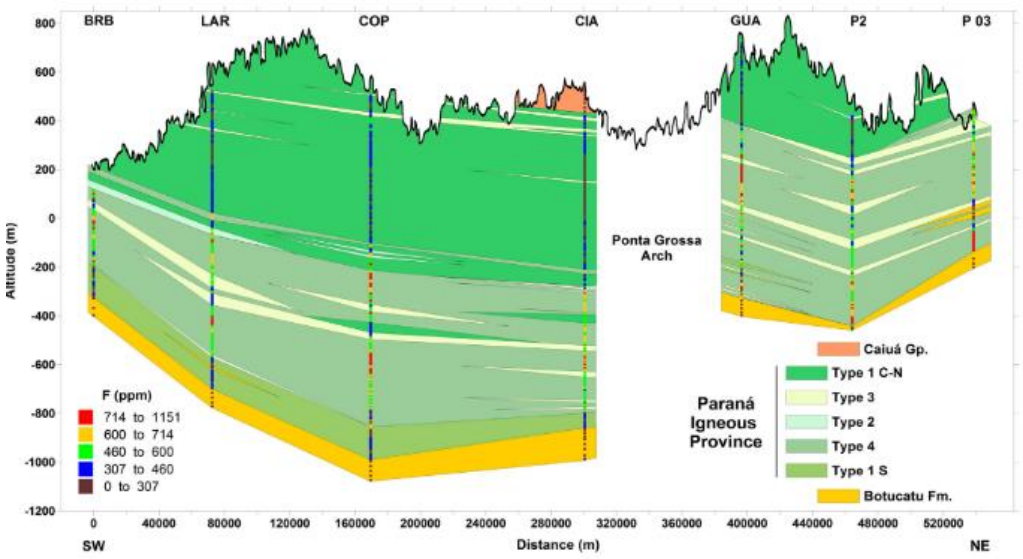

Figure 5 – SW-NE cross section shows the detailed chemo-stratigraphy of the PIP. Despite types 1 C-N and Type 1 S represent different stratigraphical situations, they are geochemically similar, what reduces the geochemical types in the section above to four. Source: Licht (2018).

3.2 Multivariate analysis techniques

The research focused on the statistical analysis of the subset using multiple Multivariate techniques. The Factor Analysis extracted the Factor Scores to be used as statistical variables throughout the analysis. Then these Factors were grouped by the Ward Hierarchical Method and K-Means Non-Hierarchical Method, both with Squared Euclidean Distance for clustering the Factor Scores into 5 clusters. Additionally, a Pattern Recognition and Classification Analysis using Neural Networks was performed to corroborate the Cluster Analysis. Finally, the Canonical Correlation Analysis was used to test and verify Licht (2018) classification model. The Canonical Variables were Factor Scores grouped by the K-Means Method (U1) and Principal Components Values and Principal Component Weights (V1), extracted from Principal Component Analysis grouped according to Licht (2018) proposed geochemical types. Therefore, if the Canonical Correlation Analysis shows a strong correlation between U1 (generated by Multivariate Analysis) and V1 (generated by Licht (2018) grouping), the new model proposed by Licht (2018) can be considered a valid model to be used for PIP rock classification.

3.3 Software used in the research

The authors received the database with 1,374 observations in Microsoft Excel® and used Excel native functions and formulae for the exploratory analysis described above. The Multivariate Analysis was conducted using Statgraphics 19® Centurion.

4. Results and discussion

4.1 Factor analysis

Factor Analysis was applied to the Correlation Matrix of the 73-dimensional vector observed 1,030 times. It was extracted m=24 Factors, with a degree of explanation of 9019%, which means that by working with a vector of dimension 24 instead of 73, the loss of explanation of the data variability is 9.81%. Thus, the other 49 Factors corresponded to a dispersion of variability, a "dust", not important. An important highlight is that none of the 73 geochemical variables was discarded since each Factor is a linear combination of all of them.

The following table presents the eigenvalue, the percentage of variance explained, and the cumulative percentage of variance explained by each Factor.

Table 1 – The Factors extracted from the Factor Analysis.

Factor Number	Eigenvalue	Percentage of Variance	Cumulative Percentage
1	26.37	36.12	36.12
2	11.80	16.17	52.29

3	4.03	5.53	57.82
4	2.53	3.47	61.28
5	2.06	2.82	64.10
6	1.78	2.44	66.55
7	1.55	2.12	68.67
8	1.40	1.92	70.59
9	1.28	1.75	72.34
10	1.20	1.65	73.99
11	1.14	1.57	75.56
12	1.10	1.50	77.06
13	1.07	1.47	78.52
14	0.97	1.32	79.85
15	0.92	1.27	81.11
16	0.89	1.22	82.33
17	0.85	1.17	83.50
18	0.83	1.14	84.64
19	0.77	1.06	85.70
20	0.69	0.95	86.65
21	0.68	0.93	87.58
22	0.67	0.92	88.50
23	0.64	0.88	89.37
24	0.60	0.82	90.19
25	0.57	0.78	90.97
...	...	...	...
73	0.00	0.00	100.00

Table 1 shows that the first Factor explained 36.12% of the data variability and that the twenty-fourth first Factors explained 90.19% of the data variability. It also shows that the eigenvalue of the twenty-fourth Factor was 0.6, which is considered a good measure of importance.

The Kaiser-Meyer-Olkin (KMO) test measures the adequacy of data for Factor Analysis, and in this analysis, it resulted in a KMO=0.88>0.6. This value indicates how much of the common variation is extracted by Factors. The minimum acceptable is 0.6. Therefore, Factor Analysis was adequate for the subset.

The Bartlett Sphericity Test tests the hypothesis that the correlation matrix is an identity matrix. The result was a Chi-Square of 133185 with 2628 degrees of freedom (D.F.) and a p-value=0.00. Thus, the null hypothesis of an identity matrix for the correlation matrix was rejected, and Factor Analysis was adequate for the subset.

4.2 Principal component analysis

The Principal Component Analysis was performed for each of the rock types proposed by Licht (2018).

The Principal Component Analysis of the rocks type 1 resulted in table 2. The eigenvalue of the twenty-third Component was 0.6, which is considered a good measure of importance, and table 2 also shows that the first Component explained 28.41% of the data variability and that the twenty-third first Components explained 89.36% of the data variability.

Table 2 – The Principal Components extracted from the Principal Component Analysis of the rocks type 1.

Component Number	Eigenvalue	Percentage of Variance	Cumulative Percentage
1	20.74	28.41	28.41
2	10.02	13.73	42.14
3	5.17	7.09	49.23

4	3.84	5.26	54.49
5	3.09	4.23	58.72
6	2.54	3.49	62.21
7	2.40	3.29	65.50
8	2.15	2.94	68.44
9	1.70	2.32	70.76
10	1.64	2.25	73.01
11	1.34	1.84	74.85
12	1.26	1.72	76.57
13	1.12	1.53	78.10
14	1.03	1.42	79.52
15	0.95	1.31	80.83
16	0.92	1.26	82.08
17	0.91	1.24	83.32
18	0.89	1.22	84.53
19	0.80	1.09	85.63
20	0.76	1.05	86.67
21	0.70	0.96	87.63
22	0.66	0.91	88.54
23	0.60	0.82	89.36
...	...	...	...
73	0.00	0.00	100.00

The KMO=0.81>0.6, Chi-Square of 51647,60 with 2628 degrees of freedom (D.F.) and a p-value=0.00. Thus, the Principal Component Analysis was adequate for the first subset. In order to obtain the vectors for each type proposed by Licht (2018), this same analysis was repeated for the rock types 2, 3, 4, 9.

4.3 Cluster Analysis

The Cluster Analysis using Ward Method was applied to the Factor Scores and produced the following Dendogram. Then, the K-Means Method was performed to optimize the grouping, and the results are in table 3.

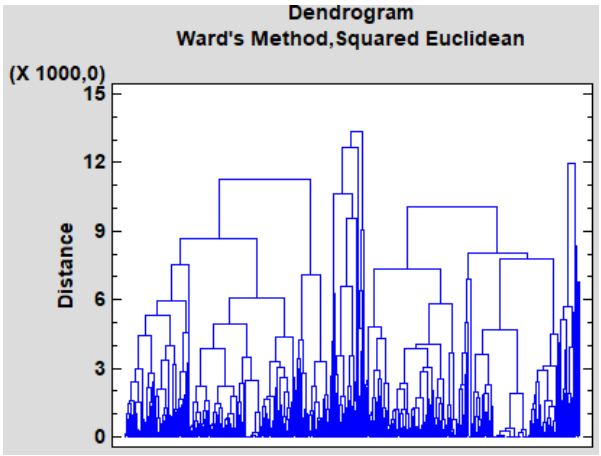

Figure 6 – Dendogram of 24 Factor Scores in 5 groups– Ward's Method and Squared Euclidian Distance.

Table 3 – Clusters created by the K-Means Method and Squared Euclidian Distance for the Factor Scores.

Cluster	Members	Percent
1	44	4.27
2	57	5.53
3	14	1.36
4	477	46.31
5	438	42.52

4.4 Patter recognition and classification

Table 4 presents the results of the Neural Network Recognition and Classification technique. The analysis shows that the allocation of Factor Scores into 5 groups by the K-Means Method was 97.67% correct. This can be considered a very good result for the Cluster Analysis allocation process and provides evidence of the robustness of the Cluster Analysis grouping technique.

Table 4 – Recognition and classification of Factor Scores into 5 groups using Neural Network technique.

Actual Cluster	Group Size	Group Predicted by Neural Network				
		1	2	3	4	5
1	44	44	0	0	0	0
		(100.00%)	(0.00%)	(0.00%)	(0.00%)	(0.00%)
2	57	0	45	1	5	6
		(0.00%)	(78.95%)	(1.75%)	(8.77%)	(10.53%)
3	14	0	1	10	1	2
		(0.00%)	(7.14%)	(71.43%)	(7.14%)	(14.29%)
4	477	0	0	0	470	7
		(0.00%)	(0.00%)	(0.00%)	(98.53%)	(1.47%)
5	438	0	0	0	1	437
		(0.00%)	(0.00%)	(0.00%)	(0.23%)	(99.77%)
Total	1,030	44	46	11	477	452

4.5 Canonical correlation

The check and validation of Licht (2018) classification model was obtained by the application of the Canonical Correlation Analysis between the first pair of Canonical Variables.

To perform the Canonical Correlation, the pairs of Canonical Variables were defined as follows:

- Cluster 1 (U1) and Type 9 (V1).
- Cluster 2 (U1) and Type 3 (V1).
- Cluster 3 (U1) and Type 2 (V1).
- Cluster 4 (U1) and Type 1 (V1).
- Cluster 5 (U1) and Type 4 (V1).

To demonstrate this process, table 5 presents the results of the Canonical Correlation for the first pair of Canonical Variables (U1, V1), U1 being the Factor Scores of cluster 1 and V1 being the Principal Component Weights of samples classified as type 9 and table 6 presents U1 the Factor Scores of cluster 4 and V1 the Principal Component Weights of samples classified as type 1.

Table 5 – Canonical Correlation of the first pair of Canonical Variables (U1, V1) - the Factor Scores of cluster 1 (U1) and the Principal Components Weights of type 9 (V1).

Number	Eigenvalue	Canonical Correlation
1	0.98	0.99
2	0.96	0.98

3	0.92	0.96
4	0.90	0.95
5	0.83	0.91
6	0.67	0.82
7	0.62	0.78
8	0.54	0.74
9	0.43	0.66
10	0.38	0.62
11	0.36	0.60
12	0.24	0.49
13	0.11	0.33

Table 5 shows a strong Canonical Correlation of ρ_1^* = corr(U1, V1) = 0.99. Figure 7 shows this relationship between these two vectors (U1,V1).

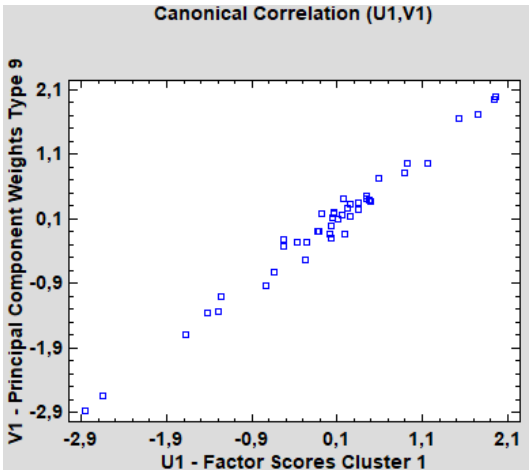

Figure 7 – Canonical Correlation of the first pair of Canonical Variables (U1, V1) - the Factor Scores of cluster 1 (U1) and the Principal Components Weights of type 9 (V1).

Table 6 – Canonical Correlation of the first pair of Canonical Variables (U1, V1) - the Factor Scores of cluster 4 (U1) and the Principal Components Weights of type 1 (V1).

Number	Eigenvalue	Canonical Correlation
1	0.89	0.95
2	0.85	0.92
3	0.82	0.90
4	0.73	0.85
5	0.65	0.81
6	0.60	0.77
7	0.56	0.75
8	0.54	0.74
9	0.49	0.70
10	0.47	0.68
11	0.34	0.59
12	0.31	0.55
13	0.22	0.47

14	0.18	0.43
15	0.13	0.35
16	0.10	0.32
17	0.09	0.31
18	0.06	0.25
19	0.06	0.24
20	0.03	0.16
21	0.02	0.15
22	0.01	0.08
23	0.00	0.06

Table 6 shows a strong Canonical Correlation of ρ_1^* = corr(U1, V1) = 0.95. Figure 8 shows this relationship between these two vectors (U1,V1).

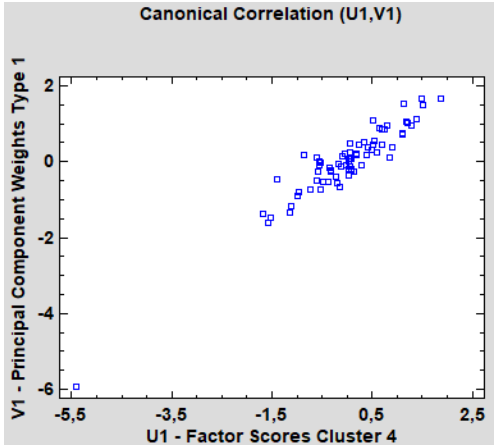

Figure 8 – Canonical Correlation of the first pair of Canonical Variables (U1, V1) - the Factor Scores of cluster 4 (U1) and the Principal Components Weights of type 1 (V1).

4.6 Results and discussion

The findings have shown that Multivariate Analysis techniques can be applied to studies from different areas of knowledge and can be a sound statistical method for validation of hypotheses and conclusions.

The Factor Analysis and the Component Principal Analysis proved to be efficient in reducing the dimensionality of matrices with low loss of information, and these techniques created combinations between the variables that can be useful for evaluating the relationship between these variables.

Another important technique, the Cluster Analysis, was especially useful for analyzing the combinations that can be extracted from the data and identifying the clusters that best represent the data. Additionally, the Pattern Recognition and Classification can not only corroborate the data arrangement generated from the clustering but also identify different patterns existents in the data and correct the clustering if needed.

Finally, the Canonical Correlation results presented valid evidence that the Licht (2018) classification model is (a) valid for PIP rock classification, (b) effective in large sets of data, (c) easy to understand and to apply in different datasets, and (d) statistically sound.

Therefore, the goals established for this research were achieved, and the hypotheses were validated.

5 Conclusion

The research had five specific objectives to achieve. All objectives were achieved and will be addressed specifically according to the results obtained, and the work performed during the research.

The first objective was to present the exploratory analysis performed to treat the data. The paper showed that the exploratory analysis was effective in identifying the characteristics of the original dataset and creating a subset of

1,030 observations to run the necessary analyses to test whether Licht (2018) classification model was a valid model PIP rock classification.

Following the subset creation, the second objective regarded Factor Analysis and Principal Component Analysis. Regarding the Factor Analysis, it extracted the Factor Scores from the 1,030x73 matrix and these Factor Scores kept the same information as the original data. The analysis showed that Factor Analysis was a Multivariate Analysis technique that can be an effective technique to analyze geochemical datasets not only to reduce dataset information, the 1,030x73 matrix was reduced to a 1,030x24, but also to extract valuable information from the dataset with low loss of information.

Additionally, the Principal Component Analysis, executed in the subsets formed by the samples classified according to Licht (2018) proposed model, reduced the size of the subsets' individual matrices, For example, rock type 1 formed a 456x61 matrix and was reduced to a 456x23 matrix, with low loss of information and the technique was also effective in analyzing geochemical datasets and extracting valuable information from datasets of distinct characteristics.

The third research objective was to perform the Cluster Analysis and show it was an effective technique for grouping the Factor Scores. The analysis showed that both the Ward Hierarchical Method and the K-Means Non-Hierarchical Method extracted 5 clusters of the Factor Scores, similarly to the 5 groups geochemically identified from the sample analysis. This evidence proved that Cluster Analysis is an efficient technique in grouping the Factor Scores that replicate the same information presented in the geochemical types proposed by Licht (2018).

Related to the previous goal, the research also aimed to show that Pattern Recognition and Classification using Neural Networks Method could be applied to assure the effectiveness of the Cluster Analysis in grouping the data. The results showed that the Cluster Analysis allocation was assertive, with a precision of 97.67%.

Finally, the Canonical Correlation results showed that this Multivariate Analysis technique can be applied in the geochemistry and geology fields, and it is a sound statistical test to check the validity and assertiveness of research hypotheses and conclusions. Moreover, the Canonical Correlation Analysis results also presented evidence that Licht (2018) proposed classification method for PIP rocks classification can be effective and can be used by scholars and professionals of geochemistry, geology, and related areas in their studies.

Conflict of Interest: None declared.

Ethical Approval: Not applicable.

Funding: None.

References

Chaves Neto A (2023) MNUM7006 - Análise Multivariada Aplicada à Pesquisa. Notas de Aula.

Fonteles H R da N et al. (2020) Hybrid multivariate typological model for the banded iron formations from the Bonito mine, Northeastern Brazil. Applied Geochemistry, v. 123.

Ghezelbash, R. et al. (2020) Optimization of geochemical anomaly detection using a novel genetic K-means clustering (GKMC) algorithm. Computers and Geosciences, v. 134, 1 Jan.

Ghorbani Z et al. (2022) Application of multivariate data analysis to biogeochemical exploration at the Twin Lakes Deposit, Monument Bay Gold Project, Manitoba, Canada. Chemical Geology, v. 593.

Giuseppe, D. DI et al. (2014) The use of particle density in sedimentary provenance studies: the superficial sediment of Po Plain (Italy) case study. Geosciences Journal, v. 18, n. 4, p. 449–458, 1 Dez.

Gomes A S et al. (2018) Chemostratigraphy and evolution of the Paraná Igneous Province volcanism in the central portion of the state of Paraná Southern Brazil. Journal of Volcanology and Geothermal Research, v. 355, p. 253–269.

Johnson R A, Wichern D W (1998) Applied Multivariate Statistical Analysis. Prentice Hall Inc., Second Edition, Englewood NJ.

Jolliffe I T (2002) Principal Component Analysis. Springer, Second Edition.

Licht O A B (2018) A revised chemo-chrono-stratigraphic 4-D model for the extrusive rocks of the Paraná Igneous Province. Journal of Volcanology and Geothermal Research, v. 355, p. 32–54.

Peate D W et al. (1992) Chemical Stratigraphy of the Paraná Lavas (South America): classification of magma types and their spatial distribution. Bulletin of Volcanology, p. 119–139.

Ragland, P. C. et al. (1997) Use of principal components analysis in petrology: an example from the Martinsville igneous complex, Virginia, U.S.A. Mineralogy and Petrology, v. 60, p. 165–184, 11 Apr.

Sharma S (1996) Applied Multivariate Techniques. John Wiley & Sons, Inc.

Zhang J et al. (2022) Sphalerite as a record of metallogenic information using multivariate statistical analysis: Constraints from trace element geochemistry. Journal of Geochemical Exploration, v. 232.

Author Biography

Bajotto, A. C. Doctorate candidate, Master of Business Administration, Master in Civil Engineering, Civil Engineer. Areas of interest: business, economy, finance, statistics, multivariate statistical methods, time series forecasting, and numerical methods. Researcher at the Graduate Program in Numerical Methods in Engineering at the Federal University of Paraná https://orcid.org/0009-0006-9069-8028

Chaves Neto, A. Doctor in Electrical Engineering, Master in Statistics, Civil Engineer and Mathematician. Areas of Interest: multivariate statistical methods, time series forecasting, quality engineering, computational intensive methods (Bootstrap and Jacknife), pattern recognition, product reliability, structural reliability, and educational statistics. Professor at the Graduate Program in Numerical Methods in Engineering at the Federal University of Paraná. https://orcid.org/0000-0003-1071-9601

Licht, O. A. B. Geologist, Post Doctor in Geology, Doctor and Master in Geology, and Geologist. Areas of interest: geochemistry in mineral exploration, environmental geochemistry, medical geology, and volcanology. Associate Professor at the Graduate Program in Geology at the Federal University of Paraná. https://orcid.org/0000-0001-6550-9121