Random Forest Classification and Principal Component Analysis in Geochemical Data of a Sulfide and Silicate Base Metal Mineralization in Vazante District, Brazil

S. Ilkay Cevik[1], Julian M. Ortiz[1], Gema R. Olivo[2]

[1] The Robert M. Buchan Department of Mining, Queen's University, [2] Department of Geological Sciences and Geological Engineering, Queen's University

1. Introduction

- Mineral exploration: deeper targets and higher cost per discovery
- Utilization of all the available data in the decision-making process

Machine learning algorithms and advanced statistical methods are highly effective to identify the **multidimensional relationships** in big databases.

2. Objectives

Use rock geochemistry from Vazante-Paracatu District in Brazil to:

(1) Evaluate the sensitivity of prediction accuracy of **Random Forests** for different types of treatment of the data

(2) Gain insights about the geochemical processes which may have implications in exploration

3. Study Area

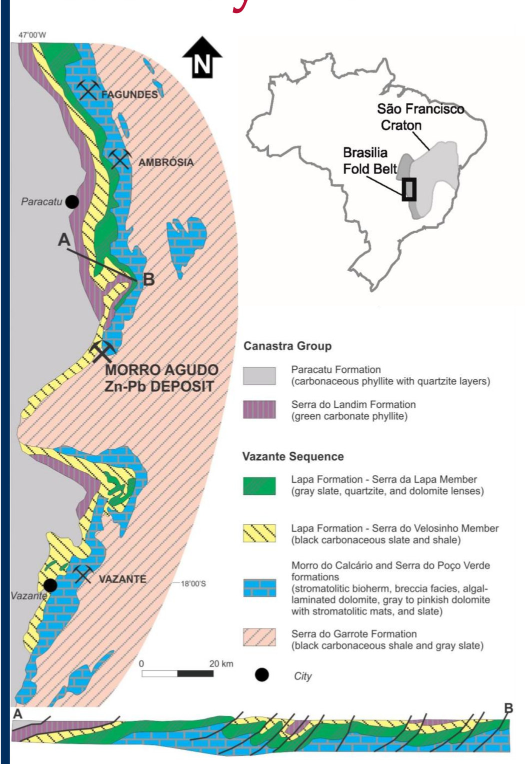


Fig 1. Location map of the study area (Cordeiro [2])

- Southern region of the Brasília Fold Belt, Western part of the São Francisco Craton, Minas Gerais, Brazil

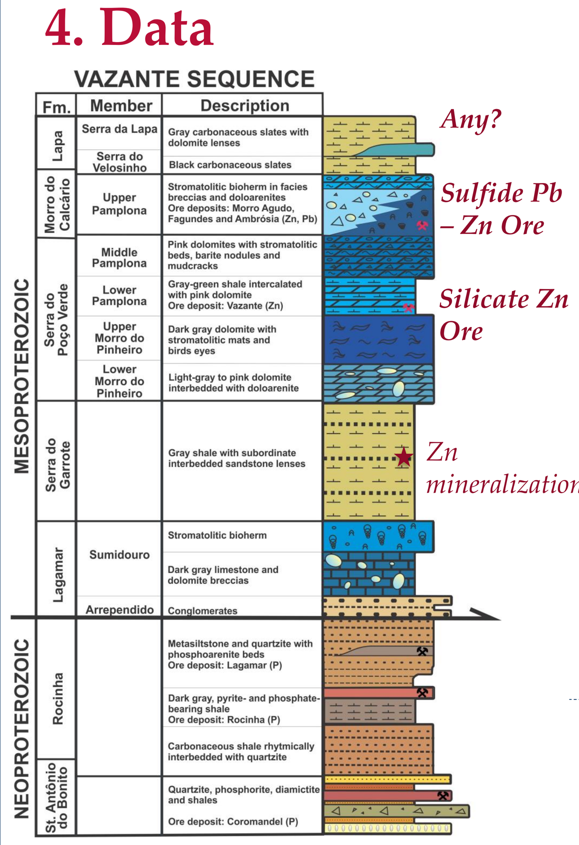
Northern Part – Vazante / Paracatu (VP)

- Sulfide Pb – Zn Deposits: Morro Agudo, pre-production resources, 18.3 Mt @5.08% Zn and 1.75 % Pb [2]

Southern Part – Vazante South (VS)

- Vazante Mine: Hypogene Zinc-silicate resources, 30.59 Mt @ 21 Zn, 0.48 Pb%, 33.57 Ag [3]

4. Data



- *A small size (n = 182) database from* ***Morro Agudo Sector in VP*** *(Sulfide Pb – Zn mineralization)*
- *A larger size (n= 407) database from* ***Vazante Sector in VS*** *(Zinc silicate mineralization)*

Fig 2. Stratigraphic section of the Vazante sequence (Cordeiro [2])

Analytical Methods:
XRF *- X-ray fluorescence analysis*
4ACD8 *- four acid digestion based analysis*
MS *- inductively coupled plasma mass spectrometry*

Table 1. Summary of the available variables in both databases

Data	Details
Analysis Results	Raw and corrected (imputation and log transformation)
Major Oxides (14) XRF	SiO_2, Al_2O_3, Fe_2O_3, CaO, MgO, Na_2O, K_2O, Cr_2O_3, TiO_2, MnO, P_2O_5, BaO, SrO (Only Vazante Sector)
Multi Element - ME-MS81 (31)	Ba, Ce, Cr, Cs, Dy, Er, Eu, Ga, Gd, Ge, Hf, Ho, La, Lu, Nb, Nd, Pr, Rb, Sm, Sn, Sr, Ta, Tb, Th, Tm, U, V, W, Y, Yb, Zr
Multi Element - ME-MS42 (10)	As, Bi, Hg, In, Re, Sb, Sc, Se, Te, Tl
Multi Element - ME-4ACD8 (10)	Ag, Cd, Co, Cu, Li, Mo, Ni, Pb, Sc, Zn

5. Methodology

Data Processing

Geochemical data were treated as described below:

- Censored values were replaced by:

  (1) the half of the lower limit of detection (LLD)
  (2) the values calculated with a k-nearest neighbour approach (***Aitchison distance*** *– KNNA*)

$$d_a = (x, y) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2}$$

- Raw geochemical data transformed by using centered log transformation **(clr)** for closure (all components sum to a constant)

$$clr(x) = \left[\ln\frac{x_1}{g(x)}; \ldots; \ln\frac{x_D}{g(x)}\right]$$

where g(x) is the geometric mean of the composition

Classification

For the **supervised classification** task:

- **Random Forest (RF),** an ensemble tree machine learning algorithm to classify lithology by using geochemical data (Fig 3.)

Table 2. Summary of cases of simulations

Experiment	Data	Imputation	Algorithm
1	Raw	Half of LLD	Decision Tree
1RF	Raw	Half of LLD	Random Forest
2	CLR	Half of LLD	Decision Tree
2RF	CLR	Half of LLD	Random Forest
3	Raw	KNNA	Decision Tree
3RF	Raw	KNNA	Random Forest
4	CLR	KNNA	Decision Tree
4RF	CLR	KNNA	Random Forest

- A decision tree classifier used as a benchmark to test the performance of Random Forest
- Total of 8 different classification experiments were conducted for each database (Table above)
- Each experiment simulated 100 times in order to report more robust accuracy scores
- For each simulation, databases were split into training (2/3) and test groups (1/3) randomly for unbiased estimation of accuracy

For **unsupervised classification**, principal component analysis was conducted on transformed data (clr)

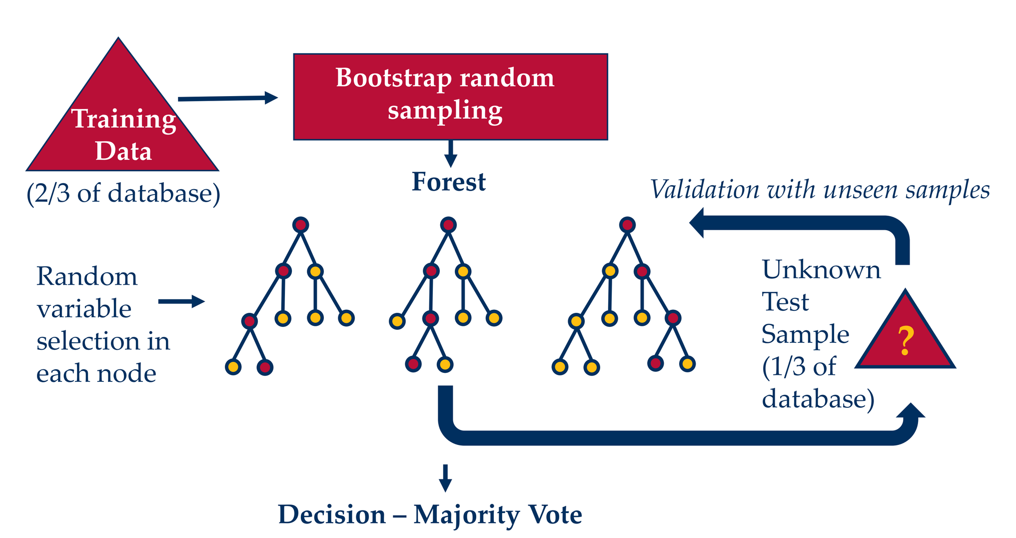


Fig 3. Schematic work flow of the Random Forest classifier. Modified from Harris et al. [4]

6. Results

- Random Forest **improved** the prediction accuracy significantly:

Accuracy

Morro Agudo Sector (VP) (n = 182) & 100 Simulations

	Decision Tree 1	2	3	4	Average Accuracy	Random Forest 1-RF	2-RF	3-RF	4-RF	Average Accuracy
Count	100	100	100	100	40%	100	100	100	100	47%
Mean	42	39	41	37		47	47	49	46	
Std	7	6	6	6		6	6	6	6	
Min	25	23	25	23		33	34	36	30	
Median	42	39	41	38		46	46	48	46	
Max	59	56	52	51		61	59	67	62	

Vazante Sector (VS) (n = 407) & 100 Simulations

	Decision Tree 1	2	3	4	Average Accuracy	Random Forest 1-RF	2-RF	3-RF	4-RF	Average Accuracy
Count	100	100	100	100	71%	100	100	100	100	80%
Mean	73	71	71	69		81	81	79	80	
Std	4	4	4	4		3	3	3	3	
Min	63	59	61	57		73	73	71	73	
Median	73	71	71	69		81	82	80	80	
Max	82	81	80	79		86	87	87	87	

Confusion

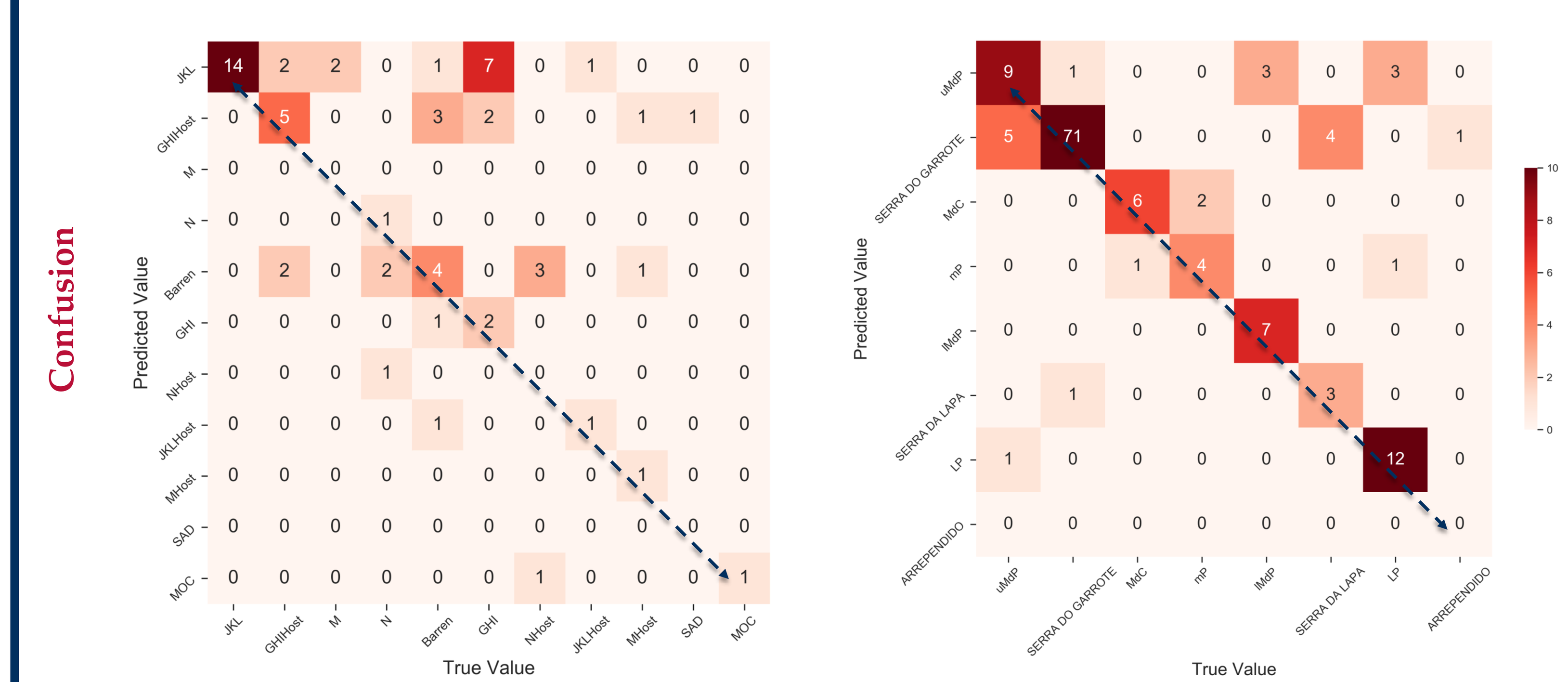
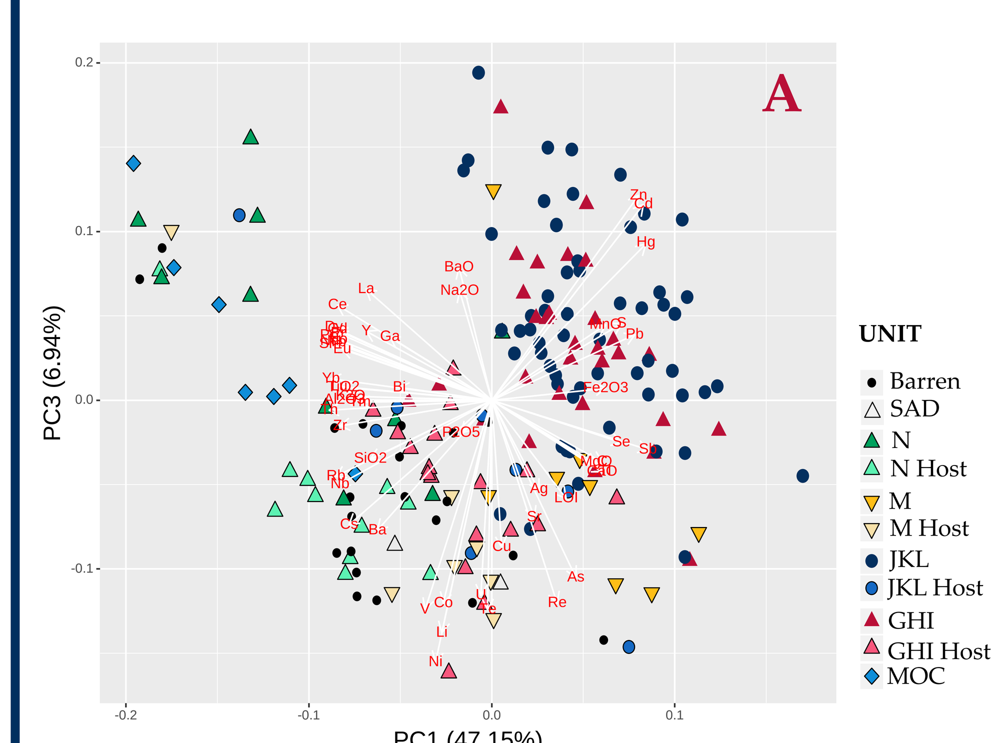


Table 3 and Figure 4. Accuracy table and confusion matrix of the VP database (upper and lower left). ***Most of the confusion:*** *GHI and JKL ore bodies. Accuracy is higher and* ***variance is lower*** *on VS (Right) – due to higher amount of sample*

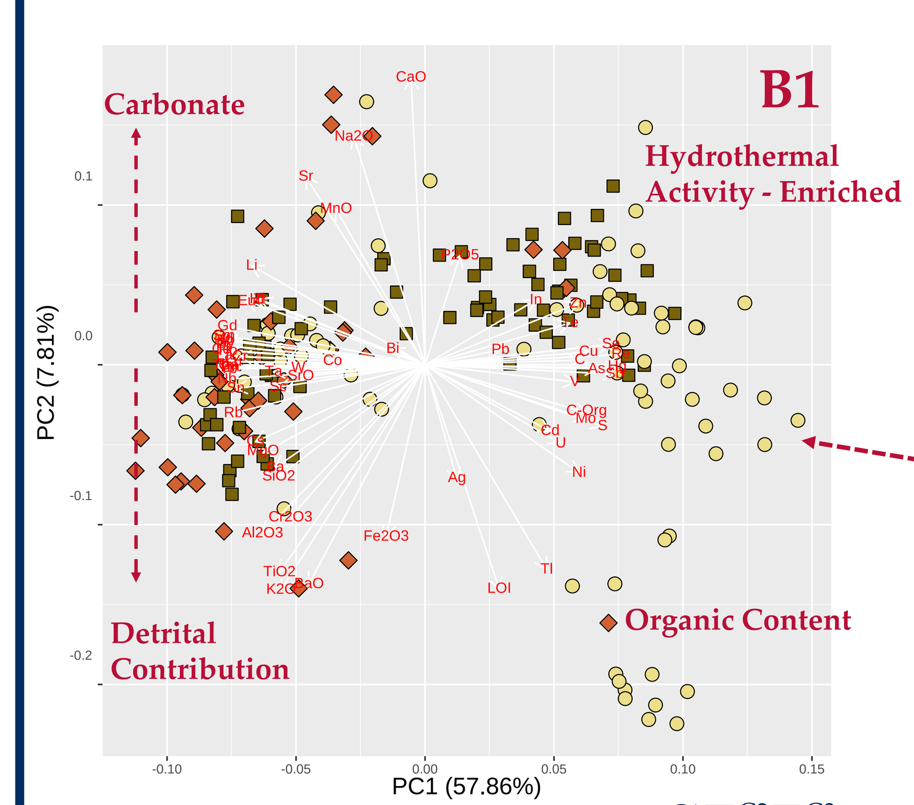
- In **unsupervised classification,** the principal component analysis, following were observed:

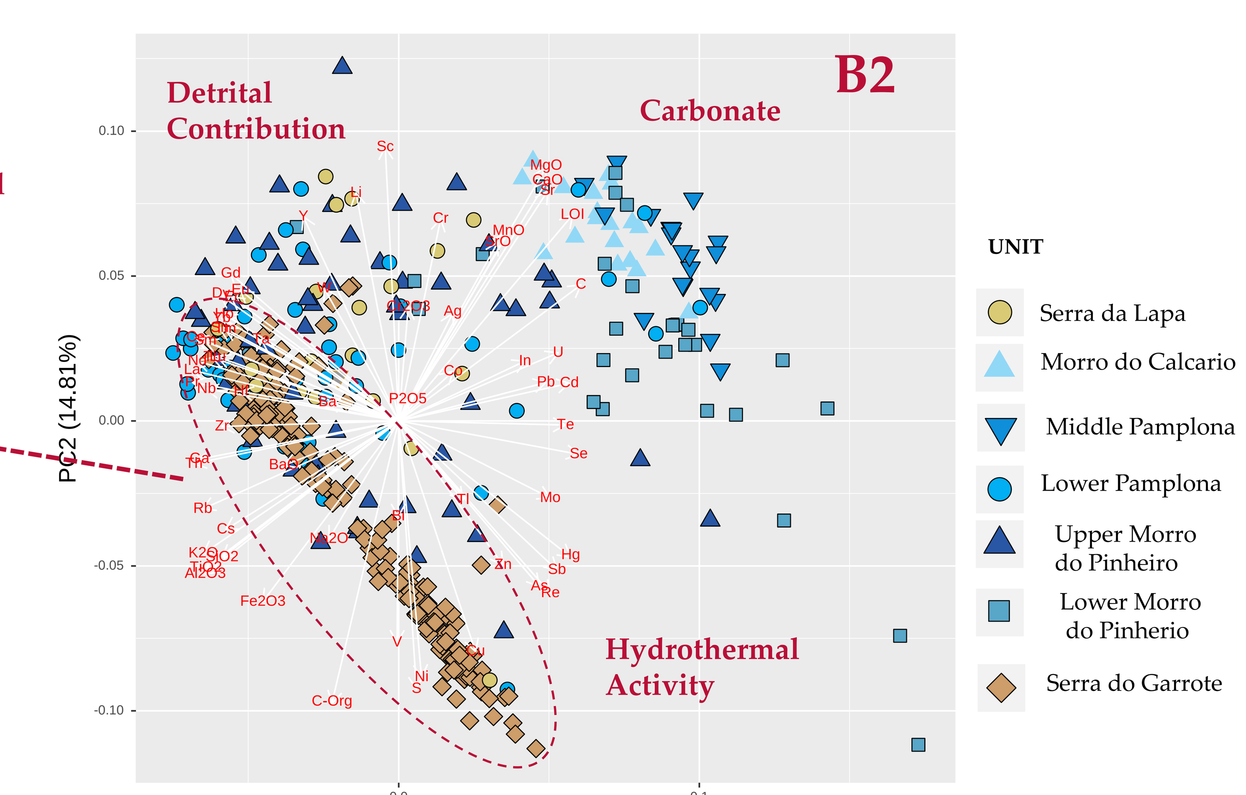


Morro Agudo Sector (VP) (n = 182)

- GHI and JKL ore bodies are hard to distinguish – note that confusion was highest in these two units (Fig 4)
- Other ore bodies are clearly distinct after several PC plots
- Host rocks can be separated

Fig 5. PCA Biplots - biplot of VS database (lower right – ***B2****) with a biplot focused only Serra do Garrote (lower left –* ***B1****) and biplot of VP database (upper left -* ***A****)*





Vazante Sector (VS) (n = 407)

- Lithological differences are clearly identified on PC1 vs PC2 plot
- Different carbonates tend to cluster slightly different
- Mineralized Serra do Garrote – G1 samples have defined "Trend" associated with hydrothermal activity

7. Conclusion

- For both VP and VS database, random forest classifier **outperformed** a basic decision tree classifier regardless of using raw or clr data
- Centered log transformation did not affect the prediction results considerably, however, it did change the cluster shapes in PCA results
- Imputation methods did not affect prediction results significantly
- Random Forest's accuracy and precision increases with increasing sample size
- Confusion patterns might be an indication of the similarities in the genesis of ore bodies or rock types. This can help geologists to understand the underlying process
- In PCA of VS samples, the association of elevated MgO, CaO, Sr, C and loss on ignition characterize Carbonate rocks while elements like Zr, Hf, Ta, Nb and, REE characterize the detrital contribution
- Mineralization within the siliciclastic domain (Serra do Garrote) captured in PCA by the association of As, C, Cu, Hg, Mo, Ni, Pb, Re, Sb, Se, S, Te, U, V, Zn
- In PCA of VP, the association of Zn, Cd, Hg, S, MnO, Pb, Se, Sn, Sb, Fe_2O_3, CaO, MgO, C characterize GHI and JKL ores while Sb, Fe_2O_3, Pb and S differentiates M, and BaO and REE indicates N ore bodies
- In the absence of key horizons in a drill hole, a rock sample can be located in stratigraphy by using its geochemistry
- The latter can be done with surface rock samples and used as proxies to ore bodies

Acknowledgments

The author would like to thank Colin Aldis and Neil Fernandes since they shared the geochemical rock sample databases which belong to their Ph.D. projects and also the Nexa Resources since they allowed to collect and analyse the rock samples and use the results in research purposes. Cevik would like to thank the Ministry of National Education Turkey because of their funding for the master program. Dr. Ortiz acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956. Dr. Olivo also acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Schodde, R. (2017). Challenges of Exploring Under Deep Cover. AMIRA International's 11th Biennial Exploration Managers Conference, Healesville: March 2017, www.minexconsulting.com/publications.html.
[2] Cordeiro, P. F., Oliveira, C. G., Paniago, L. N., Romagna, G., & Santos, R. V. (2018). The carbonate-hosted MVT Morro Agudo Zn-Pb deposit, central Brazil. Ore Geology Reviews, 101, 437-452.
[3] Olivo, G., Monteiro, L., Baia, F., Slezak, P., Carvalho, I., Fernandes, N., ... & Moura, M. (2018). The Proterozoic Vazante Hypogene Zinc Silicate District, Minas Gerais, Brazil: A Review of the Ore System Applied to Mineral Exploration. Minerals, 8(1), 22.
[4] Harris, J. R., Grunsky, E., Behnia, P., & Corrigan, D. (2015). Data-and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geology Reviews, 71, 788-803.

Biography

Ilkay S. Cevik has a B.Sc. in Geological Engineering from Middle East Technical University, Ankara, Turkey. After five years of experience in mineral exploration in West Africa and Europe he became an M.Sc. student at Queen's University in September 2018, under the supervision of Dr. Julián M. Ortiz and Dr. Gema R. Olivo.

E-mail: 17SIC1@QUEENSU.CA