Geology differentiation by applying unsupervised machine learning to multiple independent geophysical inversions

Aline Melo and Yaoguo Li

Center for Gravity, Electrical & Magnetic Studies, Department of Geophysics, Colorado School of Mines, Golden, CO 80401, USA. E-mail: aline.melo@ucd.ie

Accepted 2021 August 4. Received 2021 July 29; in original form 2020 December 18

SUMMARY

Effective quantitative methods for integrating multiple inverted physical property models are necessary to increase the value of information and advance interpretation further to produce interpretable geology models through geology differentiation. Geology differentiation is challenging in greenfield exploration areas where specific a priori geological information is scarce. The main problem is to identify geological units quantitatively with appropriate 3-D integration of these models. The integration of multiple sources of information has been conducted with different unsupervised machine learning methods (e.g. clustering), which can identify relationships in the data in the absence of training information. For this reason, we investigate the performance of five different clustering methods on the identification of the geological units using inverted susceptibility, density, and conductivity models that image a synthetic geological model. We show that the correlation-based clustering yields the best results for the geology differentiation among those investigated by identifying the correlation between physical properties diagnostic of each unit. The result of the differentiation is a quasi-geology model, which is a model that represents the geology with inferred geological units and their spatial distribution. The resulting integrated quasi-geology model demonstrates that individually inverted models with minimal constraints have sufficient information to jointly identify different geological units.

Key words: Electrical properties; Magnetic properties; Magnetic anomalies: modelling and interpretation; Persistence, memory, correlations, clustering.

1 INTRODUCTION

The increasing use of geophysical inversions to exploration problems has resulted in great success in imaging and understanding geology at depth. The prospects generated through inversion in exploration often become new deposit discoveries, and the chance of success increases with the amount of geological knowledge available to incorporate in the inversion and interpretation. In brownfield exploration, the a priori geological information from producing mines and known prospects can be used to improve geophysical models of surrounding targets. However, such a priori information is rarely available for greenfield exploration prospects. In such cases, one may have only regional geological maps that do not contain much information, especially if the area is covered by overburden. Extracting geological meaning from geophysical inversions is challenging and the interpretation becomes more difficult in areas with little auxiliary geological knowledge. Consequently, the geophysical models may not be well constrained, which increases the risk when used directly for subsequent decision making such as selecting drilling targets. To deal with this challenge, multiple geophysical methods and associated models are needed for interpretation. The simultaneous interpretation of multiple physical property models also increases the complexity of the process and, therefore, automated and objective means are needed. For this reason, we apply unsupervised machine learning to identify meaningful relations among physical properties and to map different geological units within the model domain, that is to carry out geology differentiation.

Geology differentiation and characterization is the interpretation of geophysical data and models to identify associations between inferred physical properties and different geological units. Li et al. (2019) demonstrate that the integration of multiple physical property models enables us to differentiate geological units and derive quasi-geology models. A quasi-geology model is the representation of the geology or an approximation of the geological units. It is subject to the resolution of the geophysical data and inverted models but has the advantage of presenting the geophysically derived information in a directly interpretable geological format.
In the general definition, geology differentiation can be performed in the data or model domain. To our knowledge, Garland (1951) was the first to formally characterize different geological units in the data domain by calculating the ratio of magnetization magnitude to density from magnetic and gravity data through Poisson’s relation. Kanasewich & Agarwal (1970) explore this method using wavenumber-domain properties of gravity and magnetic data. Using a similar approach, Dransfield et al. (1994) calculate the ratio of apparent susceptibility to density and build a pseudo-lithology map for interpretation. Those approaches are limited to results in 2-D maps, and the depth dimension is only explored through the analysis of different frequency content in the data.

With the routine use of geophysical inversions, interpretations have transitioned from the data domain to the model domain. Correspondingly, geology characterization is better achieved in the model domain. When using multiple geophysical property models, the differentiation process consists of two steps. The first step divides or segments the crossplot of the inverted physical properties to identify different combinations of physical property-value ranges that could correspond to different geological units. The second step maps each spatial area of the model domain into different geological units based on the inverted physical property values and the segmentation identified in the first step.

There have been many published works in this area. These works in the model domain can be grouped into three categories based on the amount of a priori geological information about the physical properties used in the first step: (i) site-specific information, (ii) general geological information or (iii) no a priori information. Once the differentiation is done, one may further characterize the identified geological units based on additional information available (e.g. lithology or alteration types).

In the category of methods that requires site-specific a priori geological information, Bosch (1999) presents a formulation to invert for lithology types using density and susceptibility models jointly. The method is termed lithologic tomography and requires petrophysical, geostatistical and structure information to constrain the lithology models, for these reasons it is only applied to areas with high level geological knowledge. Bauer et al. (2003) define physical property classes based on natural clusters formed by the inverted P velocity and the Poisson’s ratio, and associate these classes with petrophysical data of the study area to differentiate lithologies. Guillen et al. (2004) and Lane & Guillen (2005) extend the scheme developed by Bosch (1999) from 2-D to 3-D, refer the method as litho-inversion, and apply it to field data. Fullagar et al. (2004) invert for density and susceptibility variations inside pre-defined lithological units, and use this information to refine the lithological contacts. In one of their two approaches, Martinez & Li (2015) perform an end-member analysis based on a known geological cross-section to establish a mapping of inverted density and susceptibility to different types of iron formation and then apply the mapping to the two 3-D property models to achieve a 3-D lithology characterization. Melo et al. (2015, 2017) identify copper ore through the association between geological units that were identified in a few drill holes and patterns in the crossplot of inverted susceptibility versus conductivity values, guided by the theoretical relation between physical properties. Sun et al. (2020) perform geology differentiation using jointly inverted susceptibility and density models, where the differentiation is guided by the basement geology map of the study area underground.

In the category in which only general a priori physical property information is available, Hanneson’s (2003) uses empirical relations between physical properties and the percentage of magnetite and haematite minerals to classify the mineralogical equivalence in susceptibility and density models. Williams et al. (2004) use Hanneson’s (2003) empirical relations and known geology information to identify alteration zones from inverted density and susceptibility models that are obtained from separate constrained inversions on a regional scale. Further exploring this method, Williams & Dipple (2007) apply a mineralogy unmixing technique through linear programming to estimate mineral abundances and define alteration zones. Martinez et al. (2011) and Martinez & Li (2015) apply specific ranges of density and magnetic susceptibility obtained from literature to the crossplot of physical properties to define classes that correspond to lithology types. These methods are highly relevant for geology differentiation in greenfield exploration.

When there is a lack of directly usable prior information linking physical properties to different geological units, a variety of approaches have been used. Bedrosian et al. (2007) apply a non-linear least-squares fitting to the probability density function of the crossplot of resistivity and velocity model values from magnetotelluric and seismic data so that different classes and, consequently, lithology types, can be identified. Kowalczyk et al. (2010) directly partition the crossplot of density versus susceptibility, derived from inversions, to define classes of different lithologies and apply this classification to produce a 3-D regional geological model of inferred lithological categories without specifying the lithology types. Fraser et al. (2012) apply self-organizing maps (SOM) to physical property models from multiple geophysical inversions to produce a pseudo-geological model. SOM is applied by Giraud et al. (2020) as a post-inversion classification technique by enforcing geological principles to the recovered lithological model. Devriese et al. (2017), Fournier et al. (2017) and Kang et al. (2017) use density, susceptibility, conductivity and chargeability models derived only from airborne geophysical data to build a model of two kimmerlite pipes, which are referred to as the petrophysical model.

Regardless the amount of a priori information available, crossplots of physical properties are a common starting point used in the interpretation of geophysical models to characterize or differentiate between geological units. Manually segmenting the crossplots directly is feasible when two physical properties are being used. The task becomes considerably more challenging and prone to subjectivity when three or more physical properties are involved, which is common in modern exploration for targets at depth and under cover. For these reasons, machine learning methods have the potential to improve the integration and interpretation of varied data sets since it can rapidly evaluate large amounts of data quantitatively to transform data into information, which could collaborate to improve the rate of new discoveries of mineral deposits.

Machine learning (ML) is a subset field within artificial intelligence, which is responsible for developing algorithms capable of learning with experience to improve decisions (Samuel 1959). It involves the development and application of algorithms that can extract information from data without being explicitly programmed or, equivalently, automatically recognize patterns. Learning from experience can be: (i) learning from established knowledge, as in supervised machine learning or (ii) learning from the data structure, as in unsupervised machine learning. In supervised learning, a known data set with labels is used to train the algorithm and build the ability to classify new data. In unsupervised learning, there is no labelled data set for training. Therefore, the algorithm explores the structure of the data to discover meaningful information.
In mineral exploration, data from existing deposits can be used for training supervised machine learning algorithms and potentially generating new targets for brownfield and greenfield exploration. Conversely, greenfield exploration of areas distant from known mines or under cover, will not have enough data for training algorithms. In such cases, unsupervised machine learning can be a powerful tool for integrating multiple geophysical data for geology differentiation and identifying potential new deposits. To meet this challenge, we seek to investigate one aspect of geophysical interpretation in mineral exploration by machine learning: How can we extract directly interpretable geological information from multiple physical property models obtained through independent geophysical inversions?

We address this question by investigating the applicability and effectiveness of a set of unsupervised data classification methods (specifically clustering) because of the importance of a method that is capable of finding patterns associated with geological units in areas where outcrops and drill holes are sparse or unavailable. We emphasize that machine learning tools do not replace the human decisions, but would provide the necessary tools to assist human interpreters and may reduce human biases in the analyses of multiple types of data and increase speed and accuracy.

ML has been applied to the interpretation of geophysical data in petroleum exploration for several decades. Different ML methods have been applied for facies classification using well log data and seismic attributes. For example, principal component analysis and hierarchical clustering (e.g. Serra & Abbott 1982), modal distribution analysis (e.g. Wolf & Pelissier-Combes 1982), k-means clustering and discriminant analysis (e.g. Delfiner et al. 1987), neural networks (e.g. Baldwin et al. 1990; Rogers et al. 1992), graph-based clustering (e.g. Ye & Rabiller 2000) and many others. These works have resulted in a wide range of applications (e.g. Abreu et al. 2016; Schlanser et al. 2016).

In geological mapping, ML tools have been applied to airborne geophysical data, such as radiometric, magnetic, gravity and electromagnetic, focusing on the construction of 2-D pseudo-lithology maps for interpretation. For example, fuzzy c-means clustering (Paasche & Eberle 2009a, b), fuzzy Gustafson-Kessel clustering (Eberle & Paasche 2012) and self-organizing maps (Carneiro et al. 2012). Other works also focus on 2-D geological mapping applying ML to multiple geophysical data (e.g. Ranjarbar et al. 2001; Martelet et al. 2006). Geological mapping can also be achieved using 3-D physical property models (e.g. Paasche et al. 2006, 2010; Di Giuseppe et al. 2014, 2018).

In general, applications of unsupervised ML are increasing in many different subjects in geosciences, such as geothermal exploration (e.g. Garcia-Yeguas et al. 2017), nuclear magnetic resonance (e.g. Kitzig et al. 2017) and geophysical inversions (e.g. Sun & Li 2011, 2015; Li & Sun 2016; Giraud et al. 2017; Singh & Sharma 2018; Astic & Oldenburg 2019; Astic et al. 2020).

In mineral exploration, known gold deposits have been used to train a neural network and construct a favourable map on the regional scale using multiple data sets (Barnett & Williams 2006). Fuzzy c-means is applied to characterize rock types and mineralization in down hole data of sulphide deposits (Mahmoodi et al. 2014; Kitzig et al. 2017). The presence of gold is predicted by applying supervised machine learning algorithms to geophysical logs acquired in drill cores of a volcanogenic massive sulphide deposit in Canada (Cate et al. 2017). Although ML algorithms are being extensively applied for reservoir characterization and facies recognition from seismic data (e.g. Meldahl et al. 1999; Barnes & Laughlin 2005; Zhao et al. 2015; Qi et al. 2016), little research has been focusing on applying ML algorithms for characterization of mineral deposits using physical properties recovered from geophysical inversions.

Overall, much of the existing works using ML either focus on large-scale applications such as those on regional geology and prospectivity mapping, or on formation scale such as lithofacies classification in drill holes. Only limited work is available on deposit scales, where ultimately geology differentiation is necessary to support the planning of drilling targets. Additionally, as in any form of geology mapping, the result of geology differentiation carries with it inherent uncertainties and it must be assessed or quantified. Aiming to make advances in this field and being faced with the above three challenges, namely, ML-based geology differentiation at deposit scale, prediction of potential drilling targets, and assessment of uncertainty, we examine different unsupervised ML algorithms. Therefore, we propose to integrate the inverted models to produce a quasi-geology model by applying unsupervised ML as outlined in Fig. 1.

We introduce a geology differentiation algorithm that relies on the correlation of multiple physical property models for clustering analyses, which has the advantages of requiring the least amount of implicit assumptions about the nature of the clusters. In this paper, we first evaluate different clustering algorithms on a synthetic model and develop the methodology of geology differentiation from the correlation between multiple minimally constrained inversions, where the smoothness of the models poses a challenge for identifying geological units. Then we apply the geology differentiation method to field data and to understand the practical challenges, such as finding the number of identifiable geological units (clusters) in the data.

In the following, we first introduce the study area, Cristalino iron-oxide copper gold (IOCG) deposit in northern Brazil and the 3-D synthetic geological model constructed based on the simplified version of the geology of the deposit. We next simulate and invert the synthetic geophysical data sets individually to obtain inverted models of susceptibility, density and conductivity. Then, we evaluate the performance of different clustering algorithms used for geology differentiation. We then apply the method to the field data of Cristalino deposit after determining the optimal number of clusters.

2 STUDY AREA

Our study focuses on the Cristalino copper deposit located in the Carajás Mineral Province, which is a highly mineralized metallogenec region in northern Brazil. This class of iron oxide copper gold (IOCG) deposits contains economic grades of copper and gold, and are associated with iron oxides such as magnetite or haematite, or both. They are formed by hydrothermal fluids that rise through deep crustal faults as conduits (Hitzman et al. 1992). Exploring for new IOCGs is challenging because fixed exploration models are not always appropriate or applicable. For this reason, Cristalino deposit serves as a highly relevant example for our study.

2.1 Geological setting

Cristalino deposit contains 482 Mt @ 0.65 per cent Cu and 0.06 g/t Au (NCL Brasil 2005). It is hosted by a splay of the Carajás Fault, which is a major crustal fault. This splay fault cuts through a
volcano-sedimentary sequence composed of iron formation interlayered with mafic and felsic volcanic rocks (Fig. 2). This sequence is dipping approximately 50° to southwest, parallel to the fault plane that acted as the conduit for the hydrothermal fluids, and the whole sequence is intruded by a younger gabbro dyke. The main ore minerals are chalcopyrite and gold. The chalcopyrite occurs in the form of stockwork, stringers, breccias and dissemination in the host rock (Huhn et al. 1999). The iron formation unit is not continuous; instead, it pinches out where the copper ore is thicker (Fig. 2) because the hydrothermal alteration process replaced magnetite by chalcopyrite. The ore zone is subdivided into high and low-grade zones.

2.2 Synthetic geology model

Based on the known geology of Cristalino deposit, we constructed a simplified synthetic geology model to evaluate the performance of the geology differentiation methods. The features in the synthetic model simulate the characteristics of the main geological units present in Cristalino: (i) copper ore, (ii) iron formation and (iii) mafic volcanic host rock (Fig. 3). The physical property values were defined based on the values for similar rocks in the literature (Table 1). The conductivity and density values were based on Telford et al. (1990) (page 16 for density and page 289 for the conductivity of the copper ore). The susceptibility values were based on Clark & Emerson (1991) because they have measurements specifically for the iron formation. Although the rock layers in Cristalino dip 50° to west, our synthetic models have vertical bodies for simplicity (Fig. 3c). For the conductivity model, a few shallow conductive cells were added to represent near-surface heterogeneities.

3 GEOPHYSICAL DATA AND INVERSIONS

IOCG deposits are likely to occur in areas of magnetic anomalies because of the association of the copper ore with magnetite; although not necessarily coincident with specific magnetic anomalies, depending on the degree of the replacement of magnetite by chalcopyrite. This association makes the magnetic method an important geophysical tool for IOCG exploration. Associated with the magnetic method, gravity methods are important for selecting specific targets because of the high density of the association of chalcopyrite and magnetite. Once more specific targets have been selected, they are often evaluated for drilling by applying methods for direct detection of chalcopyrite, such as the resistivity method. For this reason, we focused on these three geophysical methods to study geology differentiation at Cristalino deposit.

3.1 Synthetic case

The synthetic susceptibility, density, and conductivity models were used to forward model magnetic, gravity gradient and DC resistivity data (Fig. 4), respectively. Without loss of generality, we simulate ground surveys. The magnetic and gravity gradient data are co-located. The data separation is 50 m in the east direction, 75 m in the north direction and 1.5 m above ground. The Earth’s magnetic field was assumed to be the same as that in the low-latitude region with field strength of 25 000 nT and zero inclination and declination. The DC resistivity data were simulated on flat topography and have line spacing of 100 m in the north direction and station spacing of 25 m in the east direction. The survey uses a dipole–dipole array.
with a 50-m electrode separation and 8 n-spacings. Uncorrelated Gaussian noise was added to the magnetic, gravity gradient and resistivity data (see Appendix A for further details).

The data corresponding to each geophysical method were independently inverted to recover susceptibility, density and conductivity models. No specific prior geological information was used to constrain the inversions because our goal is to simulate exploration in greenfield areas. We used the 3-D potential field inversion algorithm developed by Li & Oldenburg (1996, 2003) to invert the magnetic data, Li (2001) to invert the gravity gradient data, and Li & Oldenburg (2000) to invert the DC resistivity data (more information about the inversion methods is in Appendix B).

---

**Figure 2.** 3-D geological model of Cristalino copper deposit constructed from 303 drillholes (adapted from Vale 2004).

**Figure 3.** Synthetic geological model based on Cristalino deposit showing the geological units in the (a) susceptibility and density models, (b) conductivity model and (c) a sectional view at north = 700 m in the conductivity model.

**Table 1.** Physical property values for the synthetic geological model. The conductivity and density values were based on Telford et al. (1990) (pages 16 and 289), and the susceptibility values on Clark and Emerson (1991).

| Geologic Unit            | Susceptibility (SI) | Density (g cc⁻¹) | Conductivity (x 10⁻⁴ S m⁻¹) |
|--------------------------|---------------------|------------------|-----------------------------|
| Copper ore               | 0.08                | 3.2              | 20000.0                     |
| Iron formation           | 0.2                 | 3.6              | 2.0                         |
| Mafic volcanic           | 0.06                | 2.5              | 0.02                        |
| Near-surface heterogeneity | 0.06              | 2.5              | 100                         |
The data of each geophysical method were independently inverted using the same mesh to ensure the spatial compatibility among the models for subsequent analyses. The mesh is composed of cubic cells of $25 \times 25 \times 25$ m, and padding cells were used in the north, south, west, and east directions as well as at depth. The padding cells with sizes gradually increasing from 50 to 500 m extend the mesh by 2 km beyond the 1.0 km by 1.5 km of the core area in all directions.

The recovered susceptibility model (Fig. 5a) shows two magnetic bodies that are associated with the two segments of the iron formation unit. The recovered density model (Fig. 5b) also shows two main anomalies that are coincident with the two segments of the iron formation. In addition, there is one anomaly of moderate density values associated with the copper ore. The recovered conductivity model (Fig. 5c) has the main anomaly located in the central part of the model, which is spatially coincidental with the copper ore. The

---

**Figure 4.** The simulated data from the physical properties of the synthetic model (Fig. 3), (a) total field anomaly, (b) $T_{zz}$ component of the gravity gradient data and (c) apparent resistivity pseudo-section for north equals to 825 m.
other anomalies of high conductivity over the area are related to the conductive overburden and are limited to the shallow layer only.

3.2 Cristalino deposit

The magnetic and DC resistivity data, and corresponding susceptibility and conductivity models, were first presented by Melo et al. (2015, 2017) in a limited study. Subsequently, a set of full-tensor gravity gradient data was inverted to obtain a 3-D model of density contrast. The feasibility of integrating the density model with susceptibility and conductivity models was first presented by Melo & Li (2016) using k-means clustering analysis. We present the essentials of the three data sets and corresponding inversions in this section. The acquisition parameters of all three data sets used in this work are specified in Table 2. For the airborne magnetic data we removed the International Geomagnetic Reference Field (IGRF) and performed a regional-residual separation in the data using an inversion-based method (Li & Oldenburg 1998). This step was applied due to the strong background magnetic trend in the region. The study area is located in a low latitude near the Equator, with an inducing field inclination of $-3.5^\circ$, declination of $-19^\circ$ and strength of 25 500 nT. The magnetic data have two main anomalies (Fig. 6a) with overlapping patterns. Similarly to the magnetic data, the gravity gradient data show two main anomalies (Fig. 6b), which are more intuitively shown by the $T_{zz}$ component of the data. Another interesting feature is the anomaly with intermediate amplitude values located between the two main anomalies. The DC resistivity lines are approximately perpendicular to the topographic ridge in the area, which is coincidental with the structure that hosts the mineralization. There are two types of high conductivity anomalies (Fig. 6c). One has a large volume, starts near the surface, and extends to deep levels; and the other is of small volume near the surface.

Melo et al. (2017) report the presence of magnetic remanence in the data. However, the anomaly pattern and the estimation of the direction of magnetization (inclination of $0^\circ$ and declination of $-18^\circ$) demonstrate that the direction of total magnetization is sufficiently close to the direction of the inducing field. Therefore, they perform the magnetic inversion by using the inducing field direction as the magnetization direction. Consequently, the recovered model represents an effective susceptibility that includes the remanent effect and will have higher than expected susceptibility values, but the spatial distribution of the effective susceptibility is valid. For the geology differentiation presented here the effective susceptibility is sufficient.

All three data sets were inverted using the same mesh of model discretization to ensure the spatial compatibility between the models. The mesh is composed of cubic cells of 50 m × 50 m × 50 m in the centre, and padding cells were used in the north, south, west, and east directions as well as at depth. The padding cells of the mesh extended 3 km beyond the 1.1 km × 1.5 km of the study area in all directions, increasing the cell sizes gradually from 50 to 800 m. We used the same inversion methods that were applied in the synthetic study.

For the magnetic inversion, we used a lower bound of $b_l = 0$ and a large upper bound $b_u$ well above the expected susceptibility values. For the gravity gradient inversion, we used $b_l = -0.5$ g cm$^{-3}$ and $b_u = 2.0$ g cm$^{-3}$ to allow a wide range of physically possible density contrast variations. The DC inversion was done using the...
Table 2. Acquisition parameters of the magnetic, gravity gradient and DC resistivity surveys.

| Method       | Acquisition year | Line spacing | Line direction | Observations                     |
|--------------|------------------|--------------|----------------|----------------------------------|
| Magnetic     | 2000             | 250 m        | east–west      | draped, average terrain clearance 200 m |
| Gravity      | 2007             | 200 m        | east–west      | draped, average terrain clearance 200 m |
| Gradient     |                  |              |                |                                  |
| DC           | 1999             | 200 and 400 m| east–west      | 60 m dipole-dipole and 5 n spacings |

Figure 6. Geophysical data over Cristalino deposit, (a) total field anomaly, (b) Tzz component of the gravity gradient data and (c) apparent resistivity pseudo-section for north equals to −1800.

The inversion process requires the standard deviations of the noise in the data (see Appendix A for further details of the estimation method used in this study). The estimated standard deviation of the magnetic data corresponding to the final susceptibility model is 20 nT, which is equal to 0.9 per cent of the magnetic-data range.

Logarithm of the conductivity values, which naturally takes care of the lower bound and we did not impose an upper bound. A zero-reference model was used for the magnetic and gravity gradient inversions and a best-fitting half-space as the reference model for the DC resistivity inversion (Li & Oldenburg 2000).
The recovered susceptibility model (Fig. 7a) shows two magnetic bodies dipping approximately 50° to the southwest and they are associated with the iron formation. The northern body has a larger volume than the one to the south, and this difference is possibly associated with differences in the magnetite and haematite contents in the iron formation, or the presence of hydrothermal magnetite. Additionally, the eastern area of the northern anomaly is partially overlapping with the copper ore. The large recovered susceptibility values are judged to be due to the presence of the remanent magnetization in the magnetic rocks in the same direction of the inducing field.

For the density inversion, the estimation of the standard deviation of the noise was done for each component and the final values for the components are: \( \sigma_{xx} = 1.79 \ \text{Eo} \), \( \sigma_{yy} = 1.44 \ \text{Eo} \), \( \sigma_{zz} = 2.42 \ \text{Eo} \), \( \sigma_{xy} = 2.67 \ \text{Eo} \), \( \sigma_{xz} = 3.00 \ \text{Eo} \) and \( \sigma_{yz} = 2.96 \ \text{Eo} \). The standard deviation estimations are between 2.11 and 2.9 per cent of the data range of the components. These estimated standard deviations are reasonable because the error has several different sources, such as aircraft movement, acquisition system and pre-processing steps (Martínez et al. 2012). The recovered density contrast model (Fig. 7b) shows two main anomalies. The high density contrast anomaly in the north is spatially coincident with the inverted susceptibility anomaly, and the one in the south is only partially coincident with the other susceptibility anomaly. This difference is likely related to a larger concentration of haematite in the southern anomaly. In the area between the two high density anomalies is an intermediate density body that is spatially associated with the copper ore.

For the gravity gradient and magnetic data inversions, an estimated constant value of standard deviation was used for each data component as is the common practice. For DC resistivity data, however, the noise characteristics are different. Because the signal level depends on the separation between current and potential electrodes, one constant standard deviation would not be appropriate for data from different electrode separations. The noise level is estimated to be 0.406 mV A\(^{-1}\) plus 2 per cent of the absolute value of the normalized potential differences. The recovered conductivity model (Fig. 7c) has the main anomaly located in the central part of the model, which is spatially coincident with the known copper ore. The other anomalies of high conductivity over the region are related to the conductive overburden and are limited to the shallow layer only. The depth of investigation (DOI) of the DC resistivity data was estimated by altering the reference models in the inversions and identifying the region of similarity between features in the models (Oldenburg & Li 1999).

At Cristalino, the nearby iron formation unit has the strongest density and susceptibility anomalies whereas the copper ore has moderate values of both physical properties. These moderate values are sufficient to distinguish the ore unit from the host rock, but its high conductivity is important to differentiate from the iron formation. It is also noteworthy that these three physical properties appear to have coarse-scale similarity in their structure and boundaries, but the difference in details are precisely what carries the information for geology differentiation.

For an interpreter who is not familiar with the specific geology of the deposit (which is usually the case in greenfield exploration), the interpretation of the susceptibility and density models would be mainly focused on the two anomalous bodies that are partially coincident. However, the conductivity model shows a different structure. Then, the final interpretation would depend on the individual geoscientist’s personal experiences and inclinations. The interpretation of inverted physical properties would commonly stop here at this qualitative analysis, which does not fully support decision making for drilling. This is the point where geology differentiation takes the interpretation one step further by transforming qualitative interpretations of geophysical models into a model that is a direct representation of geology.

### 4 Geology Differentiation Through Clustering Analyses

The basis for geology differentiation is the assumption that each geological unit, may it be a lithology or a zone resulted from a geological process such as the hydrothermal alteration, has well defined ranges and combinations of physical properties. Thus, a key step in the process is the identification of clusters in the crossplots of inverted physical properties in the parameter domain. We examine the use of unsupervised machine learning for this task and develop an objective integration approach that is effective in higher dimensions such as when three or more physical properties must be integrated.

In unsupervised ML, clustering is the process of identifying patterns by grouping similar objects according to their attributes based on a generalized distance. In our study, the objects are the cells of the 3-D models and the attributes are the physical property values of each cell (i.e., susceptibility, density and conductivity). Ultimately, we seek the segmentation of the recovered physical properties in the parameter domain to identify correspondent geological units in the space domain. To use different classification algorithms, we applied feature rescaling with linear transformation, and the physical properties were scaled to vary in a range from 0 to 1 to avoid scaling-related bias.

Among the plethora of published clustering methods, we first evaluated which clustering algorithms could potentially produce the desired classification. Since the physical properties in the objects are characterized by a distance variation, where they are closer to each other near the centre of each cluster associated with a geological unit, we selected the connectivity-based clustering. Additionally, the distances gradually increase and the shape of the clusters are slightly curved. Given these characteristics, density-based clustering is a strong option for the classification, as well as distribution-based clustering. We also evaluated a centroid-based algorithm, k-means clustering, a widely used algorithm for geology differentiation to evaluate its performance considering that the clusters in this study do not have convex geometry. These distances among objects in the parameter domain translate well to geological units in the space domain as discussed extensively in the Introduction section. However, our objective is to apply an algorithm that can find the combination of physical properties that characterize each geological unit, even though some units are not distinguishable in some physical properties. For this reason, we evaluated the performance of correlation-based clustering. Its measure of similarity is more suitable for geological scenarios with units having variable cluster shapes. This method has not been applied to geophysical data to our knowledge.

Our main objective in this section is to evaluate the performance of the algorithms on our synthetic model, which has three geological units (Fig. 8). Therefore, for all algorithms we set the number of clusters as three. For field applications, however, the number of clusters must be estimated and we will discuss that in the next section. For brevity, the main characteristics of the algorithms applied in this study are summarized in Table 3 and more information about the method and results on the synthetic example can be found in the supplementary material.

The resultant quasi-geology model from correlation-based clustering for the synthetic example exhibits the most coherent
structure among those from the methods which we have examined. The method successfully finds the subspaces of dominant correlation between inverted physical properties for each geological unit. It is clear that this result shows the least influence by inversion smoothness. More information about the algorithm is given in Appendix C. Furthermore, the main features in this quasi-geology model are also the most consistent with the true model. Thus, the correlation-based clustering is a superior choice for this application.

4.1 Evaluation of the quasi-geology model

The correlation-based clustering method is able to identify the correlation among susceptibility, density, and conductivity model values and has produced a quasi-geology model consisting of the copper ore, iron formation and volcanic host rock. Visual inspection indicates that each identified geological units are spatially coherent and have good correspondence with the position and extent of the true units in the synthetic model. Here we evaluate the performance of the algorithm and the fidelity of the derived quasi-geology model using a standard criterion used in the data classification field.

The confusion matrix (Fig. 9) compares each predicted cell of the 3-D model with the known unit which they belong to in the synthetic model. In a perfect classification, the predicted classes for all cells have 100 per cent of correspondence with the true classes. In practical applications this is not possible because of limitations in data such as limited depth resolution and spatial coverage, assumptions and approximations of modelling techniques, and the presence of noise. The degree of matching between classified and true units would be higher if prior geological information is used in the classification process. However, our objective here is to evaluate the performance of our classification in an highly challenging situation, that is data with lower spatial resolution than the features in the prospects and lack of prior site-specific geological information. Therefore, a classification method that is able of identifying the ore body is considered successful. Our study shows that 60 per cent of the copper ore cells were classified as ore, while 40 per cent as the adjacent iron formation. Part of the cells of copper ore were classified as iron formation because the smooth susceptibility model overlaps this unit in the interface with iron formation in the northern and southern parts, where the ore is thin. A total of 93 per cent of the iron formation and 92 per cent of the volcanic host rock cells were correctly predicted by the quasi-geology model.

The histogram plots of each physical property for all geology units (Fig. 10a) show that the distributions of properties are similar to the quasi-geology model distributions (Fig. 10b). The susceptibility distribution of the copper ore (Fig. 10a) is bimodal. Cluster 1 of correlation-based clustering reproduces the same two susceptibility groups (Fig. 10b). Additionally, the other clusters have the same physical property patterns as the corresponding geology units. Therefore, the histograms of the classified geological units are consistent respectively with those of the true units.

5 DIFFERENTIATION RESULTS AT CRISTALINO DEPOSIT

We now proceed to the geology differentiation at Cristalino deposit. This deposit has been drilled extensively, and the large amount of
geological information available enables us to compare and validate our geology differentiation results. For the purpose of this study, however, we do not use the geological information in the inversion and differentiation steps. Thus, we simulate a greenfield scenario with the geology differentiation, and use the available information to evaluate the performance of the method.

We apply a combination of k-means and correlation based clustering for the geology differentiation process. The former enables the estimation of the optimal number of clusters and the latter completes the construction of the quasi-geology model as described in the preceding methodology section. We use the three physical property models obtained in Section 3.

5.1 Optimal number of clusters

Clustering requires the number of clusters as an input. In our geology differentiation, specifying the number of clusters is equivalent to specifying the number of distinct geology units that can be identified from the inverted physical property models. When we do not assume prior geological knowledge, this number must be determined from the inverted physical property models. We use a common method for estimating the number of clusters in computer science, known as the elbow method (Raschka 2016), to achieve this goal. This method is similar to the L-curve criterion used in the linear inversion theory (Hansen 1992). For simplicity, we use the k-means clustering because it is the simplest conceptually and fastest to evaluate.

The method is based on evaluating the variability of the clustering performance as a function of the number of clusters. This variability is associated with the measure of similarity of the clustering method and is commonly represented by some measure of distance between data objects and cluster centres. We use the k-means objective function \( E \) (MacQueen 1967) as the measure of variability:

\[
E = \sum_{j=1}^{k} \sum_{i=1}^{n} \| p_i^{(j)} - \mu_j \|^2
\]

where \( k \) is the number of clusters, \( n \) is the number of objects, \( p \) are the objects and \( \mu_j \) is the centroid for cluster \( j \). In our application \( p = (m^{(1)}, m^{(2)}, m^{(3)}) \), where the elements are the scaled physical properties.

With too small a number of clusters, the objective function will be large. On the other hand, the objective function is zero when the number of clusters is equal to the number of data. The optimal number is chosen to be the value beyond which adding more clusters does not improve modelling of the data nor, equivalently, reduce the variability significantly.

Since we seek the number of clusters, \( k \), that minimizes the objective function while keeping a number of clusters consistent with
Table 3. Summary of the clustering algorithms applied in the geology differentiation of the synthetic model.

| Method            | Main idea                                                                 | Measure of similarity                                      | In our study                                                                 | Algorithms used                              | References                      | Observation                                      |
|-------------------|---------------------------------------------------------------------------|-------------------------------------------------------------|-----------------------------------------------------------------------------|---------------------------------------------|----------------------------------|--------------------------------------------------|
| Connectivity-based| Objects are more related to nearby objects than to objects farther away  | Distance between the attributes of the objects              | Objects are the cells of the inverted models and attributes are the trio of physical property values | CLINK (complete-linkage clustering)         | Defays (1977)                   | Takes into account the linkage among attributes for each object |
| Density-based     | Defines clusters as connected regions of high density of objects          | Measures the ‘reachability’ between objects based on a radius distance | The density in our study would be the number of points per unit volume in the parameter domain | DBSCAN (Density-based spatial clustering of applications with noise) | Ester et al. (1996) | Clusters can have arbitrary shapes |
| Distribution-based| Clusters are defined as objects belonging to the same statistical distribution | Measures the maximum likelihood of the statistical distribution of the objects | The number of Gaussian distributions corresponds to the number of clusters | Expectation-maximiation using Gaussian mixture models | Dempster et al. (1977) | Physical properties commonly fit a Gaussian distribution |
| Centroid-based    | Clusters are defined as spherical distribution of objects                 | Measures the distance between the objects and the cluster centre | K-means                                                                     | MacQueen (1967)                             |                                  | Simple initial approach to explore most obvious structures in the data |
| Correlation-based | Identify clusters based mainly on the correlation of the objects          | Uses the correlation among attributes of objects to identify clusters in arbitrarily oriented subspaces |                                                                            | ORCLUS (arbitarily ORiented projected CLUSter generation) | Aggarwal & Yu (2000) | Correlation is a strong geological criteria because some physical properties of units are usually correlated |
the physical property models, we choose as the optimal number of clusters the point of maximum curvature on the curve of $E$ as a function of $k$. For the physical property models at Cristalino, this optimal value is $k = 4$ (Fig. 11). The curve shows that including more than 4 clusters has a small impact in reducing the $k$-means objective function (around 4 per cent). Once the optimal $k$ is defined, we test the consistency of the clustering results with different random initialization of cluster centres that resulted in clusters with 98 per cent of similarity.

5.2 Quasi-geology model

Upon defining the optimum number of clusters, we applied the correlation-based clustering, which is less susceptible to irrelevant attributes because of its measure of similarity based on correlation. The spatial distribution of the correlation-based clustering with $k = 4$ (Fig. 12) shows the same general cluster distribution as $k$-means clustering, but with more compact and coherent clusters. For example, in the southwestern area of the model, $k$-means clustering with $k = 4$ classifies some cells as belonging to cluster 1 (potentially corresponding to ore) while correlation-based clustering does not.

The histogram plots of all clusters for each physical property (Fig. 13) summarizes the ranges that show correlation for each cluster. There is overlapping of susceptibility ranges between clusters 1, 2 and 3, while cluster 4 is clearly associated with the lowest values. The overlapping is a consequence of the smooth transition between anomalies in the inverted models, which is a common characteristic of such physical property models. Although there is a superposition of values, the histogram plot shows that, for susceptibility, cluster 1 is mostly associated with moderate-high values, cluster 2 with highest values, and cluster 4 with moderate-low susceptibilities. The density model has sharper contacts; as a consequence, the clustering shows more defined ranges for clusters 1, 2 and 3, while cluster 4 gets a small subset in the same range as cluster 3. Similar to the susceptibility model, the conductivity model is also smooth. In addition, the conductivity model clearly has two groups of anomalies: shallow ones spread over the mesh and a continuous anomaly that reach greater depths and has a defined geological strike. The shallow anomalies are potentially associated with conductive overburden, while the other anomaly with sulphides. The histogram plot for conductivity shows that cluster 1 has the highest conductivity values, cluster 2 mostly the moderate values, cluster 3 low to moderate values, and cluster 4 gets a small subset in the same range as cluster 3.

The proposed method for selecting the optimum number of clusters provides an approximation for $k$ and it is important to examine the results for different numbers of clusters. In the case of Cristalino Deposit, $k = 4$ is chosen as it objectively provides the best trade-off between physically plausible model and data misfit. This also agrees with the expected geological structure for the region. Using $k = 5$ also provides physically plausible model, with strong consistency between results (Appendix D).

5.3 Evaluation of the quasi-geology model

We now proceed to assessing the reliability and, thereby, the value of information in the quasi-geology model through geology differentiation by comparison with the 3-D geological model of Cristalino built from 303 drill holes (Vale 2004). This detailed geological model captures variations in the geology within meters. However, the geophysical data does not have the resolution to image such small-scale variations from the detailed drilling program and subsequent geology modelling. We downsampled the 3-D geological model to the same mesh used in the geophysical inversions. In the geological model, each unit was represented by a closed volume. These volumes were discretized to the 50-m cubic cells used in the inversion mesh. The result is a simplified geology that represents the geometry and volume of the main units (Fig. 2). In order to simplify the comparison between drilling derived geological model and quasi-geology model through clustering, the known geological units were grouped into three main categories: (i) ore unit, which comprises the high grade ore, (ii) the iron formation unit, which comprises iron formation and the gabbro dyke and (iii) the host rock unit, composed of the mafic volcanic, felsic volcanic, low grade ore and other non-mineralized rocks. The low-grade ores would appear as the host rock, because their geophysical signature of in the inverted models is similar to that of the host rock.

The histogram plots of all geology units for each physical property (Fig. 14) shows that the overall distributions are similar to the clustering histograms. However, the main difference is that the quasi-geology model has fewer cells as host rock than there are in the geological model. This difference happens because the hydrothermal copper deposit, which occur mainly as stockwork veins (Melo et al. 2017), is hosted by a volcano-sedimentary sequence composed of iron formation interlayered with mafic and felsic volcanic rocks. The geophysical data and, consequently, the inverted models, do not have the resolution to differentiate the interlayering of iron formation with the volcanic rocks and mineralized veins. Therefore, the region composed mostly by iron formation, but with intercalations of volcanic rocks, will be highly magnetic in the susceptibility model and classified as one cluster. Thus, the differences between the geological model and the quasi-geology model are due to limitations in the geophysical data resolution and inversions without site-specific constraints.

A visual comparison of spatial patterns in the geological model and the quasi-geology model with $k = 4$ shows the following correspondence: (i) high grade ore and cluster 1, (ii) iron formation and cluster 2 and (iii) host rocks and cluster 3. The irrelevant attributes associated with cluster 4 has no correspondence in the geological model and appear scattered and incoherent.
Figure 10. Histogram plots for each physical property showing the range of values (a) for the geological units of the synthetic model and (b) for the clusters of the quasi-geology model obtained through correlation-based clustering.

Figure 11. Objective function versus number of clusters, the red circle shows the point of maximum curvature $k = 4$, which is the optimum number of clusters.

To quantify the comparison, we compute the confusion matrix shown in Fig. 15. There is a 62–64 per cent match between the three known geology units from drilling and those in the quasi-geology model. 62 per cent of the known ore was classified as such, also 62 per cent of the host rock was classified correctly (cluster 3) and 64 per cent of the iron formation was also correctly predicted (cluster 2). The main spatial differences are that 26 per cent of the known host rock was predicted as ore because of the interlayering between these units, the host rock was misclassified. 28 per cent of the known ore region was predicted as iron formation because parts of the ore rich in magnetite are highly magnetic, therefore, were classified as iron formation. 18 per cent of the known iron formation was predicted as host rock probably because of the small iron formation bodies that occur in the west of the geological model but does not have significant expression in the magnetic and gravity gradiometry data. In addition, 14 per cent of the known iron formation was classified as ore, probably because of the presence of chalcopyrite veins that increase the conductivity.

6 DISCUSSION

In our investigation, to develop an objective and automated geology differentiation approach, the first step was to examine clustering techniques that can explore the structure of, and extract information from, multiple geophysical inversion models in a quantitative and integrated manner. We have shown that clustering methods with different metrics are influenced to different degrees by smoothness of the inverted models when segmenting the crossplot of recovered physical property values.

Each clustering algorithm has different parameters that can be altered in order to improve classification such as using other types of distance metrics instead of Euclidean distance among many others. Here we have carried out an initial study with the objective of comparing the performance of different clustering algorithms on differentiating the units of the synthetic model using multiple physical property models. To present the primary idea and examine the first-order feasibility, we choose not to evaluate the impact of the variation of all parameters on clustering. We remark that additional studies are necessary for a comprehensive evaluation. For instance, connectivity- and density-based clustering could not identify clusters that correspond to the geological units (Fig. 8), and perform geology differentiation, because their classification is based on the distance between groups. Therefore, these methods require that different groups should be separated by a gap, and this is not a characteristic of the physical property models from $L_2$.

Another option of measuring similarity is the statistical distribution of the clusters, which is the basis for distribution-based clustering algorithms. The statistical distribution needs to be known a priori, otherwise it becomes too strong an assumption for the data. The result of the classification shows that assuming a Gaussian distribution worked well to identify the volcanic host rock and iron formation of the synthetic model, because their recovered physical property distributions are close to Gaussian. On the other hand, the assumption of a Gaussian distribution did not work for identifying
Figure 12. 3-D spatial distribution of correlation-based clustering with $k = 4$ and the scatterplots of physical properties, showing less influence of irrelevant attributes and an improved spatial association with the geological units.

Figure 13. Histogram plots for each physical property showing the range of values for all classes of correlation-based clustering with $k = 4$.

Figure 14. Histogram plots for each physical property showing the range of values for the simplified geological units in the 3-D geological model.

the copper ore unit that has a bimodal distribution and, as a result, the differentiated unit became noisy. The main insight is that the statistics of the physical properties of geological units in the field is not known unless petrophysical studies have been conducted, which is commonly not true in greenfield exploration. The application of centroid-based clustering also requires another strong assumption, because its good performance depends on the sphericity of clusters in the data. Therefore, the result will not be accurate if the clusters...
confirm the capability of producing quasi-geology models that are informative approximations of the true geology valuable for exploration purpose.

7 CONCLUSION
We have developed an objective geology differentiation method that enables the integrated interpretation of multiple geophysical inversions in greenfield exploration. Correlation-based clustering successfully identified geological units using the complementary information content in the multiple independent physical property inversion models and showed success in finding patterns in a complex geological setting with minimum influence of near-surface heterogeneities. For the field data application at Cristalino copper deposit in Brazil, we show that k-means clustering is a good initial approach to explore the most obvious structures in the data. We demonstrate how this simple clustering method supports the identification of the optimal number of clusters through the application of the L-curve criterion to its objective function for different number of clusters. However, the results are heavily influenced by irrelevant attributes, and we show that correlation-based clustering identified the copper ore, iron formation, and host rock with a high degree of similarity with the known geological model.

This work shows that the application of unsupervised machine learning is feasible on deposit scales for the identification of potential drilling targets. In addition, we move one step further our understanding in how different clustering algorithms explore the structure of the data for the type of models we use in greenfield exploration. The work flow presented here can be applied to any type of target in greenfield and brownfield exploration. The method presented has the potential to increase the understanding and confidence in the drilling planning stages using a quasi-geology model. The process of extracting information from multiple sources of data with an unbiased, quantitative, and integrated method empowers explorationists and geoscientists in the decision making.

ACKNOWLEDGEMENTS
The first author acknowledges the CNPq (Brazilian National Council for Scientific and Technological Development) for the scholarship. We thank Vale S.A. for providing the data used in this study and for permission to publish results obtained from them. This work is supported in part by the Gravity and Magnetics Research Consortium (GMRC).

DATA AVAILABILITY
The synthetic data underlying this paper will be shared on reasonable request to the corresponding author. The field data underlying this article were provided by Vale S.A. by permission and cannot be shared publicly for reasons of confidentiality.

REFERENCES
Abreu, E., Castagna, J. & Praxedes, E.D.S., 2016. Well-log lithology discrimination using elastic attributes only, in SEG Technical Program Expanded Abstracts 2016, pp. 622–626.
Aggarwal, C.C. & Yu, P.S., 2000. Finding generalized projected clusters in high dimensional spaces, SIGMOD Rec., 29(2), 70–81.
Astic, T. & Oldenburg, D.W., 2019. A framework for petrophysically and geologically guided geophysical inversion using a dynamic Gaussian mixture model prior, Geophys. J. Int., 219(3), 1989–2012.
Figure S4. Expectation-maximization clustering model using Gaussian mixture models (dotted grey lines), and crossplots of the normalized values of density, log susceptibility and log conductivity of the inverted models classified by the corresponding cluster.

Figure S5. K-means clustering model showing the segmentation Voronoi-cells (dotted grey lines) and crossplots of the normalized values of density, log susceptibility and log conductivity of the inverted models classified by the corresponding cluster.

Figure S6. Flowchart of the correlation-based clustering with ORCLUS (arbitrarily ORiented projected CLUSter generation) algorithm showing the three main parts of the clustering process (based on Aggarwal & Yu 2000).

Figure S7. Correlation-based clustering model and crossplots of the normalized values of density, log susceptibility and log conductivity of the inverted models classified by the corresponding cluster. This result shows the good correspondence between this model and the geological model (Fig. 2).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

APPENDIX A: NOISE ESTIMATION

Uncorrelated Gaussian noise was added to the simulated magnetic, gravity gradient, and resistivity data. For the magnetic data, noise with standard deviation of 1 per cent of the datum magnitude plus 1 nT was added to the anomalous magnetic field data. For the gravity gradient data, noise with a standard deviation of 1 per cent of the datum magnitude plus a fixed noise level was added. The fixed noise level was computed for each component separately and it is equal to 0.5 per cent of the minimum value for the component. For the DC resistivity data, noise has a standard deviation of 1 per cent of the datum magnitude plus a fixed noise level. The forward model was used for the magnetic and gravity gradient inversions and a best-fitting half-space as the reference model for the DC resistivity inversion (Li & Oldenburg 2000). Although the synthetic model has sharp contacts between units, we could not, and should not, make this assumption when simulating interpretation in an unexplored area. For this reason, we opted for using L2 inversion to make the method general and evaluate any limitations of this assumption on the classification.

The standard deviations of the Gaussian noise added to the data were used in the inversion process. Therefore, the models were obtained by using the discrepancy principle Parker (1994), where the target misfit equals the number of data when assuming Gaussian noise with zero mean,

\[ \phi_d = \sum_{i=1}^{N} \frac{(d_{obs}^i - d_{pre}^i)^2}{\epsilon_i} = N, \]  

where \( d_{obs}^i \) is the observed data, \( d_{pre}^i \) is the predicted data, and \( \epsilon_i \) is the standard deviation. This is a valid approach as we focus on the performance of ML based geology differentiation using simulated noisy geophysical observations. In field application, an addition step is required that either estimates the standard deviations of data errors or, equivalently, chooses the optimal regularization parameter through criteria such as L-curve (Hansen 1992) and generalized cross-validation (Golub et al. 1979).

APPENDIX B: INVERSION METHOD

Each of the three inversions relies on a respective relationship between the data and model:

\[ \mathbf{d} = F \mathbf{m}, \]  

where \( \mathbf{d} = [d_1, d_2, \ldots, d_N]^T \) is the data vector (magnetic, gravity gradient, or DC), \( F \) is the forward mapping operator and \( \mathbf{m} = [m_1, m_2, \ldots, m_M]^T \) is the model vector containing the physical property (susceptibility, density, or logarithmic conductivity) of each cell in the model. \( N \) is the number of data and \( M \) the number of model cells.

For gravity gradient and magnetic data sets, the forward operator \( F \) becomes a linear matrix system:

\[ \mathbf{d} = G \mathbf{m}. \]  

where \( G \) is the \( N \times M \) sensitivity matrix, whose elements quantifies the gravity gradient or magnetic produced at a data point by a model cell, respectively, with a unit density contrast or susceptibility value. For the DC resistivity problem, the forward mapping is nonlinear and given by a numerical solution of the partial differential equation that governs the electric potential in a conductive medium (Dey & Morrison 1979).

The inverse solution is obtained by solving the following constrained minimization problem using Tikhonov regularization:

\[ \min \phi = \phi_d + \beta \phi_m, \quad \text{subject to} \; b_l \leq \mathbf{m} \leq b_u, \]  

where \( \phi \) is the objective function, \( \phi_d \) is the data misfit function, \( \phi_m \) is the model objective function, \( \beta \) is the regularization parameter, and \( b_l \) and \( b_u \) are the lower and upper bounds, respectively, of the model values. For the magnetic inversion, we used \( b_l = 0 \) and a large \( b_u \) to simulate a situation with no upper bound. For the gravity gradient inversion, we used \( b_l = -1.0 \) and \( b_u = 4.0 \) to allow a wide range of possible density contrast variations. The DC inversion was done using the logarithm of the conductivity values, which naturally takes care of the positivity constraint. A zero-reference model was used for the magnetic and gravity gradient inversions and a best-fitting half-space as the reference model for the DC resistivity inversion (Li & Oldenburg 2000).

The noise values are usually unknown for field data. To ensure an adequate estimation of the standard deviations, we used the method described by Melo et al. (2017). The method starts first by assuming a standard deviation of 1 and running a set of inversions using regularization parameters (\( \beta \)) over a large range of values. Then, we apply the L-curve criterion (Hansen 1992) to the Tikhonov curve of misfit versus model norm values and select the point of maximum curvature of the curve as the optimum regularization parameter. The misfit corresponding to this optimum regularization parameter is then used to estimate the adjusted standard deviation. In the following step, a new Tikhonov curve is built using the parameter. The misfit corresponding to this optimum regularization parameter is then used to estimate the adjusted standard deviation. The procedure of adjusting the standard deviation value is repeated a second time to fine tune the estimation, and the L-curve of the third set of inversions was used to select the final model.

APPENDIX C: CORRELATION-BASED CLUSTERING ALGORITHM

Correlation-based clustering was developed for clustering high-dimensional data. When multiple dimensions are involved the clusters are difficult to visualize, the concept of distance becomes less precise. Furthermore, certain attributes are more relevant to some
clusters than others, and these attributes are also likely to be correlated in arbitrarily oriented subspaces (Kriegel et al. 2009). In our study we used the algorithm ORCLUS (arbitrarily ORiented projected CLUSter generation, Aggarwal & Yu 2000; Schubert et al. 2015) because it is a hybrid approach that identify arbitrary subspaces based mainly on the correlation of the objects. Although ORCLUS is a hybrid method, the main criteria for forming the clusters is the correlation (direction of minimum variance) among objects of a cluster. Therefore, here we refer to it a correlation-based clustering for simplicity. The concept of correlation among objects is a strong geological criteria for similarity considering that some combinations of physical properties can be a strong criteria for identifying a geological unit while other combination can not. Therefore, identifying the physical properties that have a high correlation for each unit is the mathematical criteria that mostly approximates to the geological criteria. This method has not been applied to geology differentiation of geophysical data, and its potential for identifying clusters with unknown and varying shapes is worthy of investigation.

The algorithm initializes with a large number of initial clusters and uses the smallest eigenvectors (smallest variances) of the covariance matrix of the objects within each cluster to find a set of vectors that define the subspace of each cluster. Then, the pairs of clusters are evaluated to decide if two clusters fit into the same pattern of behaviour, if so they are merged into a single cluster. The merging decision is a two step process. First it finds the subspace that defines the pair of clusters. Then, it projects the objects into this subspace and compute the distances of these objects to the centroid of the cluster, if the distance is the smallest compared to the other clusters, the two clusters are merged into one. The algorithm iteratively merges clusters based on their projected distances until the user input number of clusters is reached (Fig. C1). The main idea is to transform each group of the data into a new coordinate system in which the physical properties with high geophysical contrast for each group are the axes of the subspaces. Then, the second order correlations are minimized, meaning that the physical properties that do not show contrast for the specific geological units are not used in their differentiation.

**APPENDIX D: CORRELATION-BASED CLUSTERING OF CRISTALINO WITH 5 CLUSTERS**

This appendix presents the geology differentiation result from applying correlation-based clustering with \( k = 5 \) (Fig. D1), the corresponding histogram plots (Fig. D2), and the confusion matrix (Fig. D3).

With \( k = 5 \) in correlation-based clustering, cluster 1 from \( k = 4 \) becomes smaller in spatial extent and has a higher spatial correspondence with the high-grade ore. This cluster is mainly associated with large conductivity values, showing that the method was able to identify a cutoff value that has higher visual association with the copper ore conductivity. The confusion matrix that compares the known geology with the quasi-geology model with \( k = 5 \) shows a lower match than the results with \( k = 4.48 \) per cent of the ore is predicted as ore, and 57 per cent of both, host rock and iron formation, are correctly predicted. The main difference is that 14 per cent of the ore is classified as this new cluster (cluster 5 or host rock 2). This percentage difference is the primary change between the two quasi-geology models. Five clusters captures transitional physical property values in the inversions, and one third of it corresponds to known host rock, one third to ore, and one third to iron formation. Although the five-cluster model has a lower performance compared with the four-cluster model, the results for the three units of interest are not drastically different.

The histogram plots for five clusters is similar to that for four clusters. The histograms of susceptibility and density show that the ranges stay the same, while for conductivity two different groups are created. Clustering with \( k = 5 \) found a new correlation where the governing attribute is conductivity and the new cluster 1 is associated with the most anomalous conductivity values. Thus, only one cluster from \( k = 4 \) is altered, while the other clusters remain the same.

The consistency in the identified ore and iron formation in the two quasi-geology models gives us confidence that these regions identified for these two units are credible and does not critically depend on the choice of the cluster number. Based on Occam’s razor, the quasi-geology model using cluster number \( k = 4 \) can be considered the simplest interpretation in this case. If additional information is available to support the result from \( k = 5 \) in practice, the corresponding result would also be valid and useful if it yields more interpretable information about the underlying geology.

**Figure C1.** Flowchart of the correlation-based clustering with ORCLUS (arbitrarily ORiented projected CLUSter generation) algorithm showing the three main parts of the clustering process (based on Aggarwal & Yu 2000).
Figure D1. 3-D spatial distribution of correlation-based clustering with $k = 5$ and the scatterplots of physical properties, showing the subdivision of cluster 1.

Figure D2. Histogram plots for each physical property showing the range of values for all classes of correlation-based clustering with $k=5$.

Figure D3. Confusion matrix showing the comparison between the known geology from the 3-D geological model and the predicted units from correlation-based clustering with $k = 5$. 