Spatial multivariate morphing transformation on geochemical data augmentation[1]

Tong Li (tong.li@queensu.ca)
Julian M Ortiz (julian.ortiz@queensu.ca)

Abstract

Geochemical data plays an important role in supporting mineral deposits explorations in various ways. With the growth of deep learning methods utilized in mineral explorations, the demand for more geochemical data with more granularity keeps increasing. In contrast, the acquisition of geochemical data is usually expensive and time-consuming. In this work, we attempt to apply the spatial multivariate morphing transformation on geochemical data for data augmentation. This method decorrelates geochemical data both spatially and statistically by mapping the data into a multi-Gaussian space. Results show that this method is effective on geochemical data, but there are still problems that might be caused by the low stationarity and high dimensionality of geochemical data.

1. Introduction

Geochemical data is one of the most critical data that are used in mineral deposit exploration, especially in regional prospectivity mappings. Using geochemical data can provide vital information for discovering unknown mineral deposits like the spatial association of geochemical patterns, inter-elemental relationships, and geochemical anomalies that are caused by mineralization (Zuo and Xiong, 2020). Various methods had been adopted to process geochemical data to extract mineralization information. Among them, deep learning methods show a strong ability to intelligently find hidden patterns and features in geochemical data. The promise of such ability comes from large amounts of training data, and existing geochemical data usually cannot meet the demands and need augmentation. Traditional data augmentation methods designed on image data could result in severe problems when applied to geochemical data, like creating patterns in the wrong direction or adjusting all elements in the same scale.

To augment geochemical data, spatial multivariate morphing transformation (SMMT) (Avalos and Ortiz, 2022; Avalos et al., 2022) is utilized in this study. SMMT takes randomly sampled data from the initial dataset and statistically decorrelates them by mapping them from the initial multivariate space into a multi-Gaussian space. The decorrelated variables then are simulated with the spatial structure of random Gaussian values by geostatistical simulation algorithm. The simulated data are back-transformed by interpolating from the multi-Gaussian space back into the initial space. The interpolated data have a great reproduction of the multivariate features and relationships of the initial dataset and hence can be used as augmented data. SMMT provides a method to augment geochemical data that not only enlarges the variability of the initial data, but also honors the histogram and spatial variabilities of the initial data.

[1] Cite as: Li T, Ortiz JM (2022) Spatial multivariate morphing transformation on geochemical data augmentation, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2022, paper 2022-08, 119-131.

In this article, we summarize the application of SMMT in geochemical data augmentation, providing the workflow of SMMT in augmentation, we show results under different conditions, and discuss about problems we are facing and some implementation details.

2. A brief introduction to the geochemical data

The geochemical data used in this study is collected from the National Geochemical Surveying and Mapping Project of China (Wang et al., 2007; Xie et al., 1997). Standardized stream sediment samples from southern Jiangxi in China with an average sampling density of 1 sample/km^2 were taken and considered as the average geochemical concentration within this sample area (Xie, 1978). A total of 25 elements can be used in public studies, containing 7 major elements and 18 trace elements. All these elements are preprocessed by removing void values and negative values. All the data is located in a 335 × 335 grid, each grid cell represents 1km × 1km. Fig.1 shows the concentration of iron in the study area.

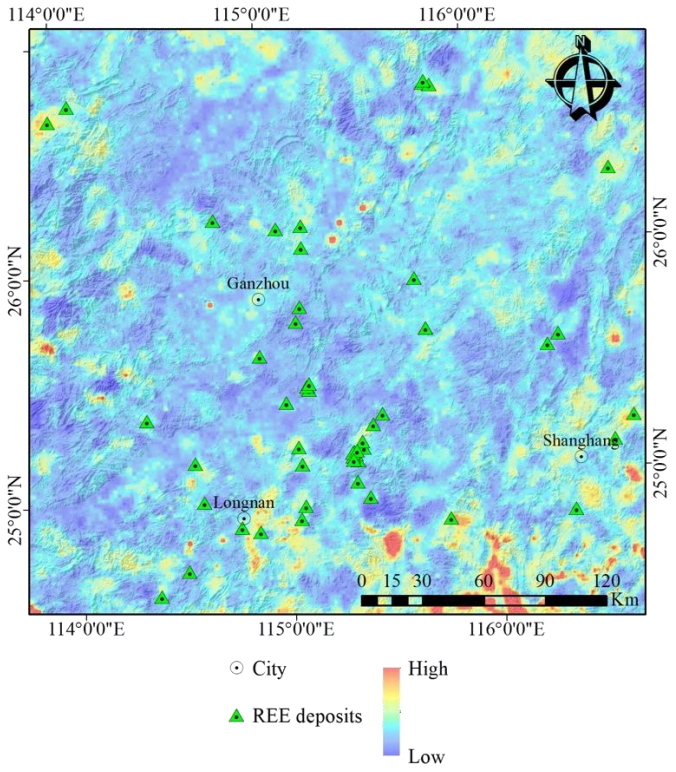

Figure 1: Geochemical distribution of iron (Li et al., 2022).

3. Spatial Multivariate Morphing Transform

In this section, we mainly focus on the workflow of applying SMMT to geochemical data, the theoretical background of SMMT can be reached in Avalos and Ortiz (2022).

1. **Choosing landmark points randomly from the original data.** Generate unduplicated locations from the grid of the study area as the landmark points for one iteration. Geochemical data on landmark points are normalized with an average of 0 and with a standard deviation of 1. Calculate the omnidirectional direct-variogram of each element and the cross-variogram between elements.

2. **Generating Morphing factors and pairing with landmark points.** Draw morphing factors for each element independently from a standard multi-Gaussian distribution with the same amount of landmark points. Compute the empirical cumulative distribution functions (CDF) of both landmark points and morphing factors. Pair the cumulative probability values of landmarks and the morphing factors with optimal transport. Optimal transport computes the Euclidean distances of every pair of samples on each dimension and attempts to find the optimal pair with the shortest distance over all dimensions. Calculate the omnidirectional variogram of paired morphing factors.

3. **Calculate the average of variograms of generated morphing factors.** Repeat Step 2 *n* times (100 times in this study). Calculate the average of all the omnidirectional variograms of morphing factors. The average variogram will be taken as the variogram model for sequential Gaussian simulation in Step 4.

4. **Sequential Gaussian simulation.** Simulate each set of generated morphing factors independently using the sequential Gaussian simulation (Goovaerts, 1997) for *m* times (100 times in this study). Compute the direct- and cross-variograms of the simulated data and make sure such variograms characterize the spatial structure of the morphing factors.

5. ***K*-nearest thin plate spline interpolation.** Map the simulated data from Gaussian space of value range (-∞, +∞) into logit space of value range (0, 1), to be conditioned to the cumulative probability values of landmark points of the value range of (0,1). For each simulated point, *k*-nearest landmark points are utilized as the control points in the thin plate spline interpolation (Bookstein, 1989). After interpolation of every point except for the landmark points, the result is considered as one augmentation of the original geochemical data. Repeat from step 1 for further augmentation until all the original points are used as landmark points.

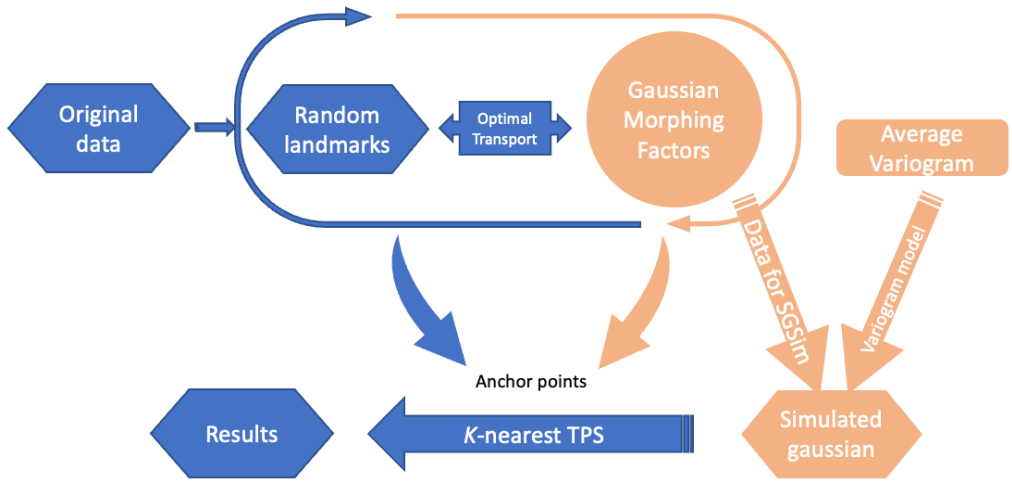

Figure 2: Schematic diagram of the workflow of the SMMT.

4. Results

The workflow in section 3 has been applied to three different sets of geochemical data for specific reasons. The subset with two dimensions contains two elements, iron and manganese, to intuitively illustrate the effectiveness of the SMMT; the subset of 25 dimensions contains the whole geochemical data, to test the SMMT on high dimensional data without extra preprocessing; and the subset with 17 dimensions is the production of preprocessing that removes the elements that are not correlated with other elements. Next, we show the results of the two-dimensional and 17-dimensional cases. Results from the 25-dimensional case are similar to those of the 17-dimensions.

4.1. Subset with 2 dimensions

Only 2 elements of the geochemical data are selected for illustration of the application of the SMMT on geochemical data. The omnidirectional direct-variograms and cross-variogram of two elements are shown in Fig.3. Spatial structures on both elements and between elements are observed.

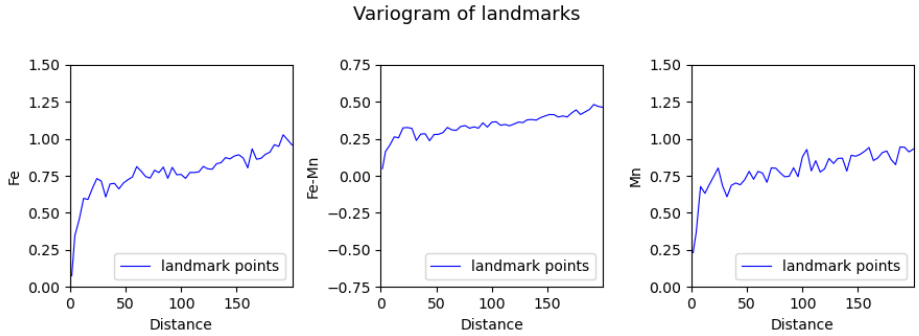

Figure 3: The omnidirectional direct-variograms and cross-variogram of two elements.

Fig.4 shows the landmark points (blue points) in the original space and the empirical cumulative distribution space and one set (out of 100 sets) of morphing factors (red points) in the Gaussian space and the empirical cumulative distribution space on the left side. The right side of Fig.4 shows part of the pairing of landmark points and morphing factors.

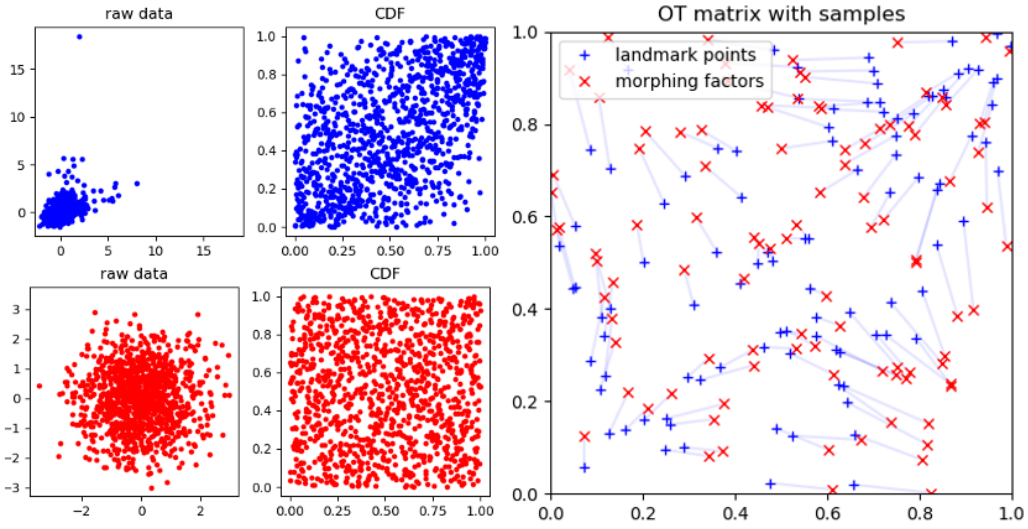

Figure 4: Pairing of the landmark points and morphing factors with 2-dimensional data.

After pairing with landmark points, the sequential Gaussian simulation is applied to every set of morphing factors with the average variogram model of the morphing factors. We use a maximum of 60 conditioning points with data assigned on nodes, and a search radius of 50 km. One realization of the simulation results is displayed in Fig. 5 (a). The spatial structure of the simulated data can be observed. The reproduction of variograms is shown in upper Fig.6. The direct variogram of iron and the cross variogram between iron and manganese are well-reproduced, but the direct variogram of manganese gets higher variance than the average variogram model.

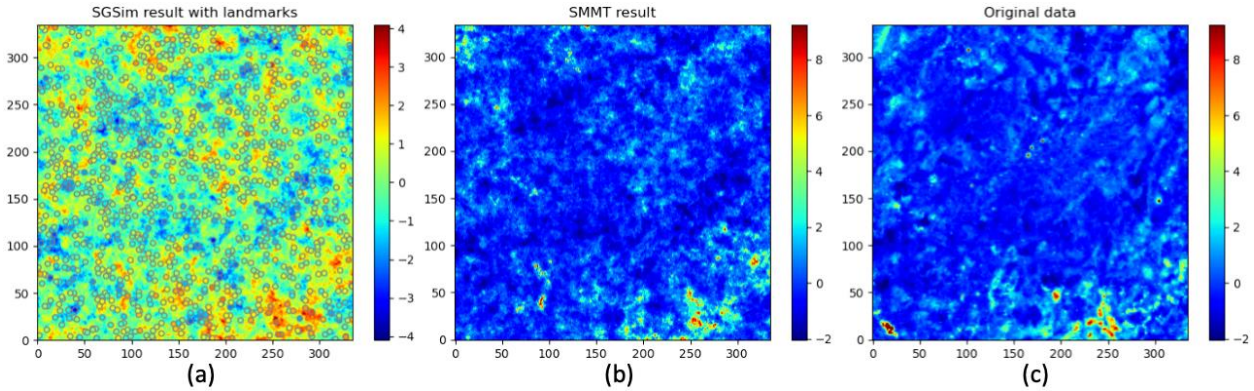

Figure 5: One realization of sequential Gaussian simulation, landmark points (dots in (a) with values), the respective result of SMMT and original data.

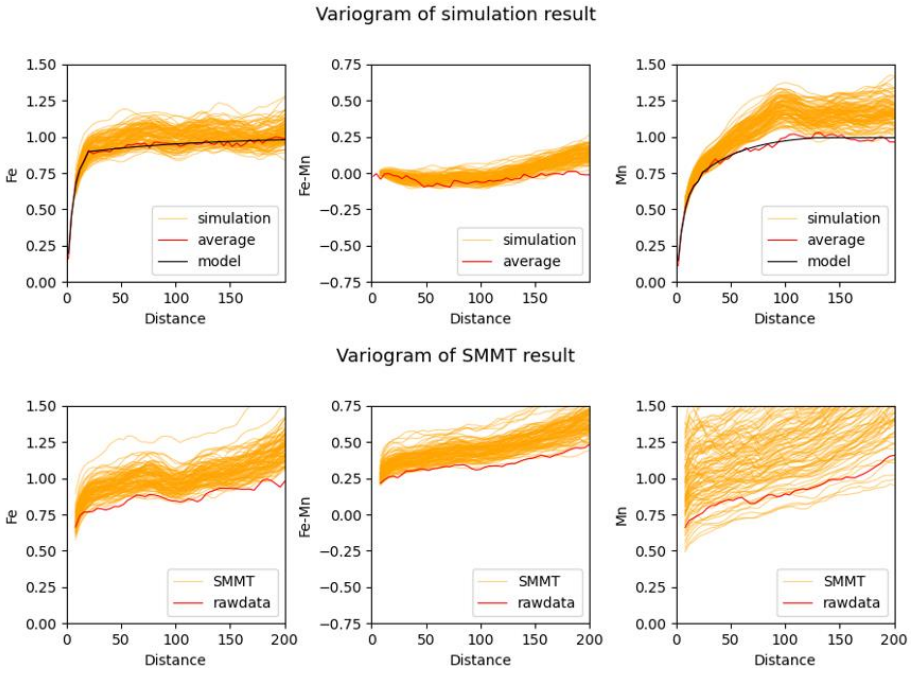

Figure 6: Variograms of the TPS results of 2-dimensional data.

The morphing factors are mapped into the original space via thin-plate spline interpolation (TPS). All points in the grid (except landmark points) are interpolated via TPS based on the 30 nearest landmark points. One of the results of the TPS is displayed in Fig.5 (b), which shows a similar spatial structure to the original data. The lower part of Fig. 6 shows the variogram of original geochemical data and the variogram

of the results of TPS. The variograms of TPS results show higher variances at most lag distances and the variograms of manganese are showing a similar structure but with unstable variances.

4.2. Preprocessed data with 17 dimensions

One of the purposes of utilizing the SMMT is to decorrelate the original geochemical data before simulation. However, high-dimensional data may disturb the optimal transport and affect the pairing. We reduce the dimension of geochemical data by removing some elements that are considered not correlated to other elements. 17 elements are left and used for augmentation with the SMMT. The omnidirectional direct-variograms and some of the cross-variograms are displayed in Fig.7 and Fig.8. All the elements are showing spatial structures with different variances.

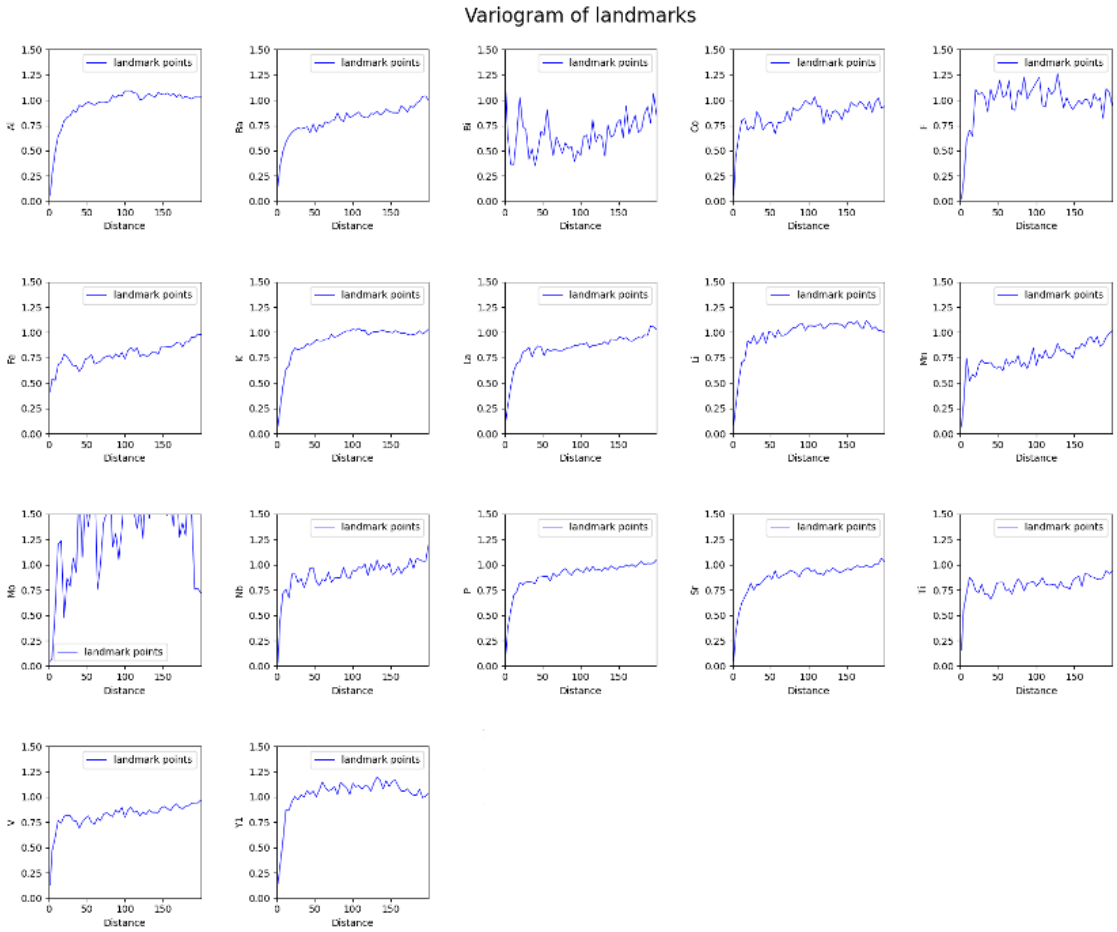

Figure 7: Omnidirectional direct variogram of landmark points.

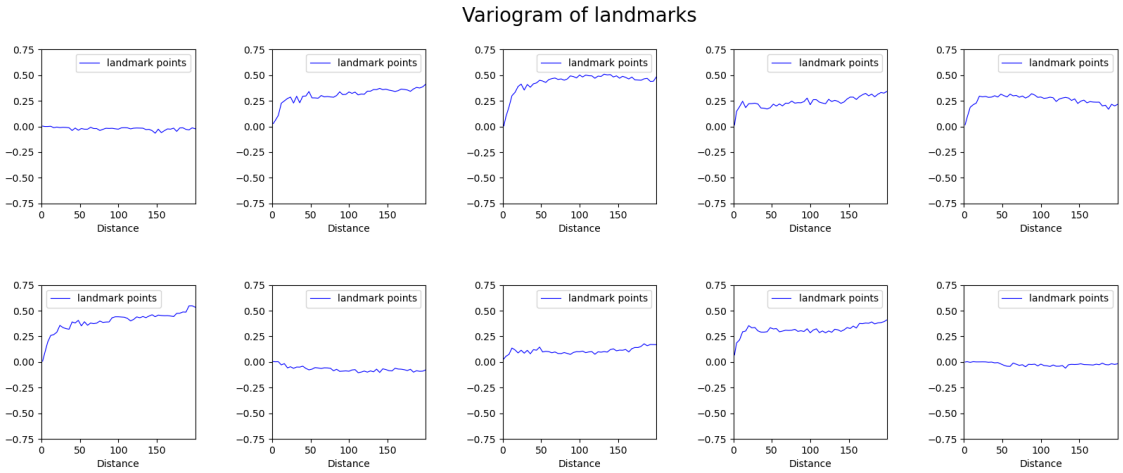

Figure 8: Omnidirectional cross variograms of landmark points (showing 10 out of a total of 136).

The pairing result of the iron and manganese of 17-dimension data is displayed in Fig. 9. With the higher dimension, the optimal transport of middle rankings is more cluttered than the top/last ranking in the cumulative probability values.

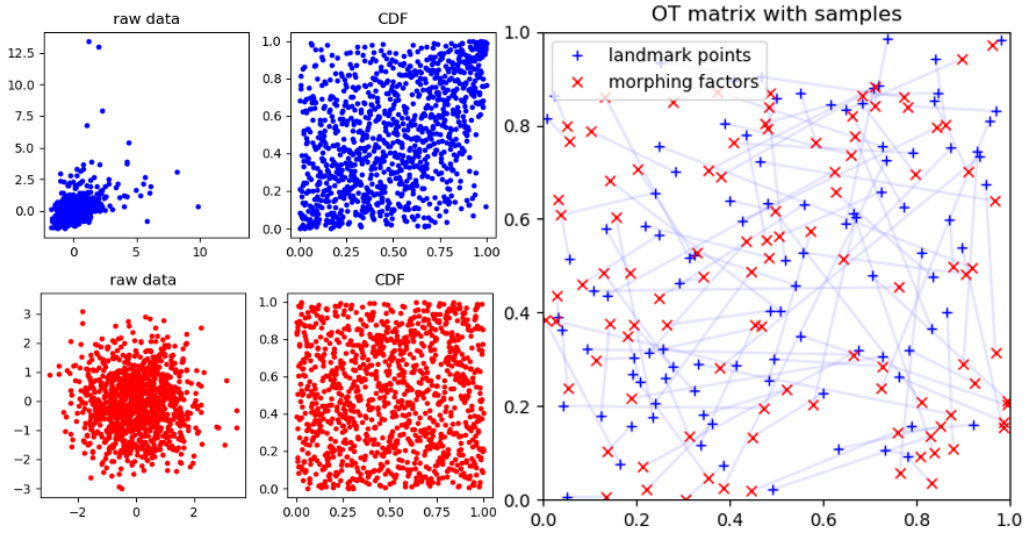

Figure 9: Pairing of the landmark points and morphing factors with 17-dimension data.

Fig. 10 (a) shows one realization of the sequential Gaussian simulation of the 17-dimension data. The simulation result exhibits poor spatial continuity with a fragmented spatial structure, which also resulted in high nugget effects in the direct-variograms shown in Fig. 11. The cross-variograms which are shown in Fig.12 indicate that all the elements are decorrelated.

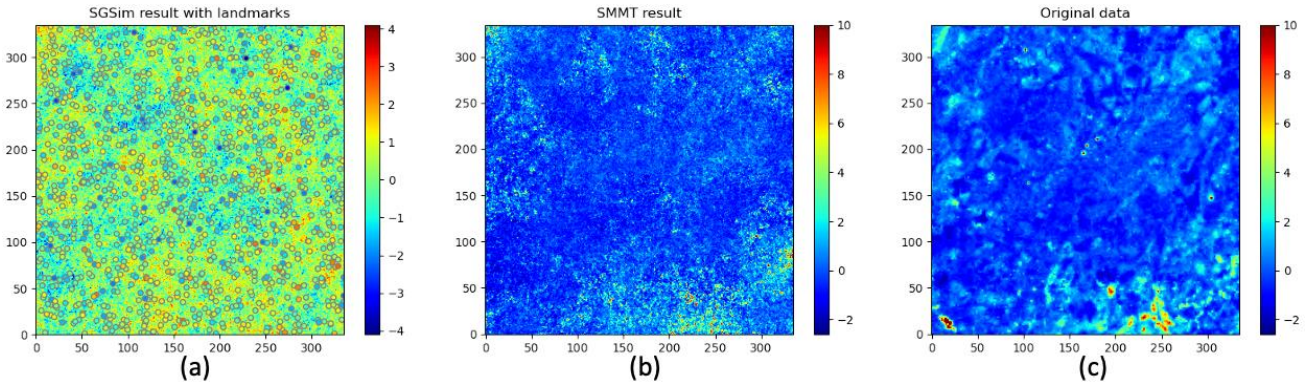

Figure 10: One realization of sequential Gaussian simulation, landmark points (dots in (a) with values), the respective result of SMMT, and original data.

Consequently, the TPS interpolations on such simulation results shown in Fig.10 (b), reproduce the trending of geochemical distribution on a large scale, but the spatial structure is not well-reproduced. Fig. 13 and Fig. 14 show the direct-variograms and cross-variograms of the results of the TPS interpolation. The spatial structure of the original data is reproduced but the average variances of results are mostly higher than those of the original data, which is similar to the 2-dimensional data.

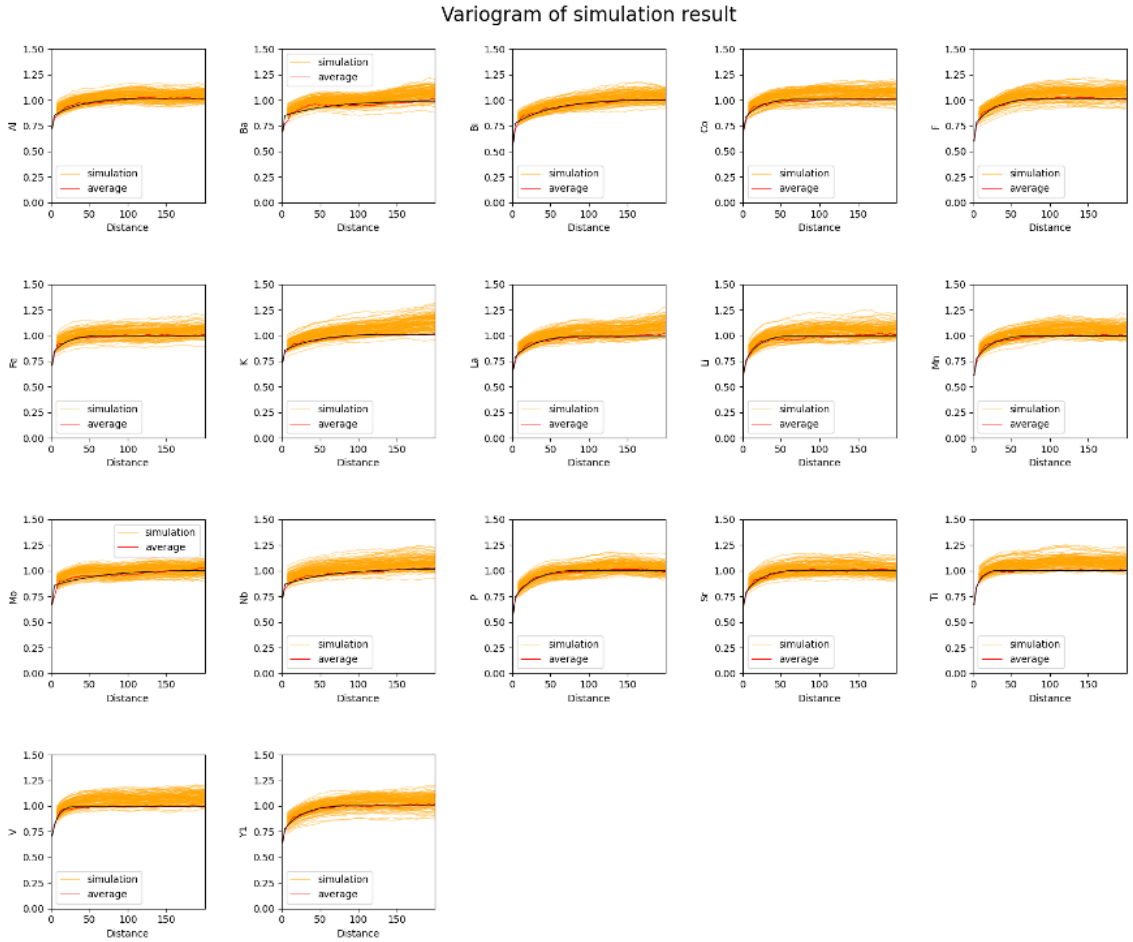

Figure 11: Direct-variograms of sequential Gaussian simulation results of 17-dimension data.

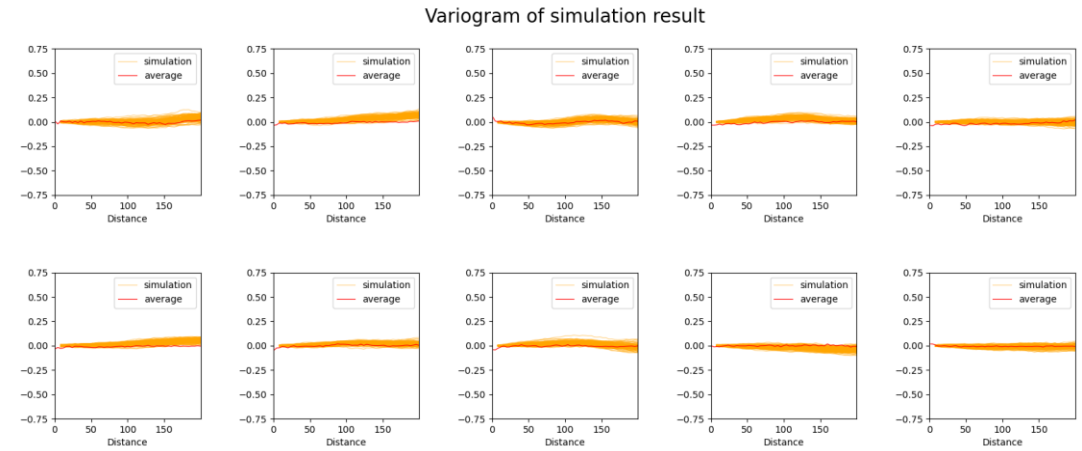

Figure 12: Cross-variograms of sequential Gaussian simulation results of 17-dimension data (showing 10 out of a total of 136).

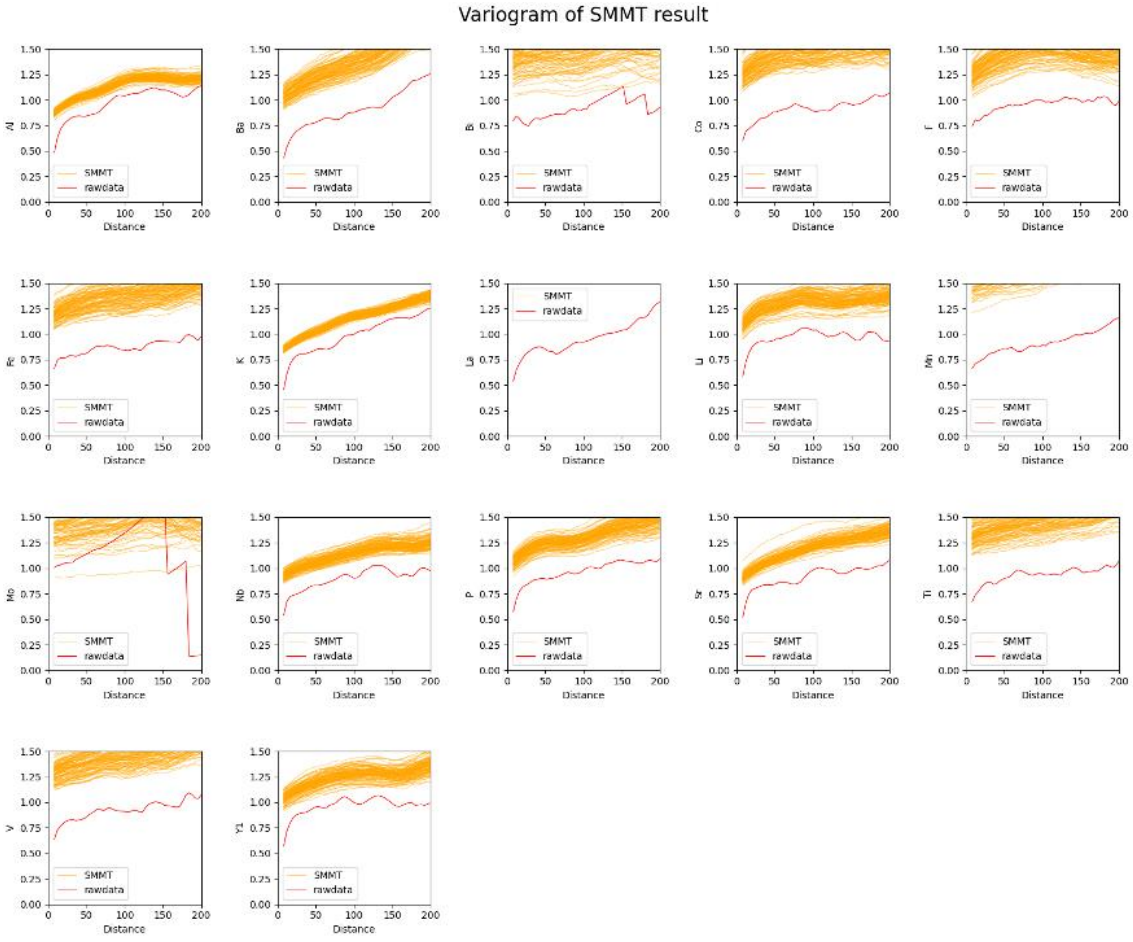

Figure 13: Direct-variograms of the TPS results of 17-dimension data.

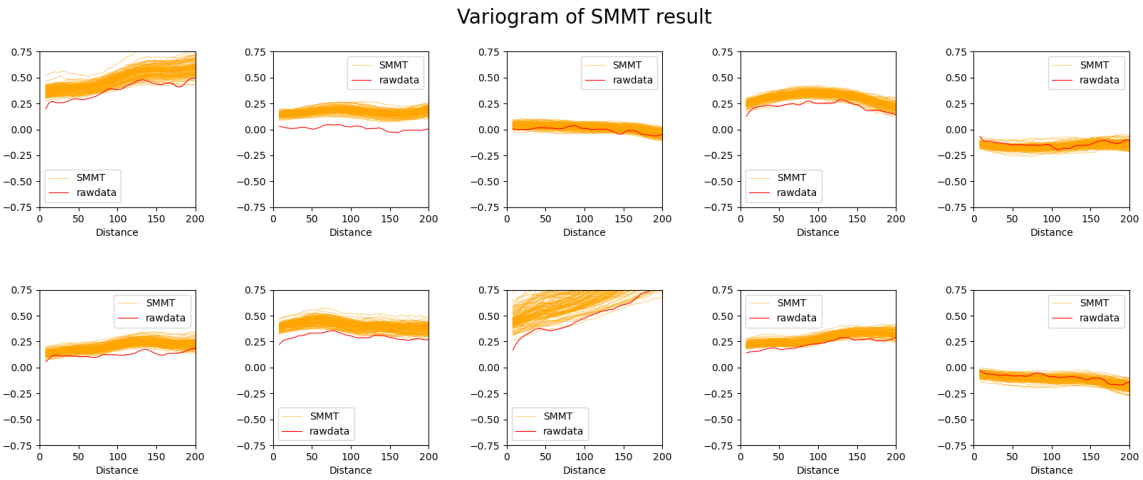

Figure 14: Cross-variograms of the TPS results of 17-dimension data (showing 10 out of a total of 136).

5. Discussion

Application of the SMMT on the different settings of geochemical data demonstrates that such a method can decorrelate the geochemical data and reproduce the spatial structures separately with a univariate geostatistical simulation algorithm. This section provides a discussion on several issues in applications on high-dimension data and some implementation details.

5.1. The curse of dimensionality

There are usually more than ten dimensions in geochemical datasets. Such high dimensionalities lead to a huge volume of data space and the available dataset is often inefficient to promise the effectiveness of algorithms.

In this study, the pairing in the optimal transport is sensitive to dimensionality. The optimal transport is trying to find the shortest distance to map data from the original space to a decorrelated Gaussian space while preserving their univariate relationships, which are their cumulative probabilities in this study. The shortest distance is the sum of distance on every dimension. When dimensions increase, the global optimal transportation on the whole dataset cannot preserve the univariate relationships on each dimension. Fig.15 shows the pairing results of a landmark point and its corresponding morphing factors with different dimensions. The ranking of the cumulative probability of landmark points becomes more unstable when the dimension increases.

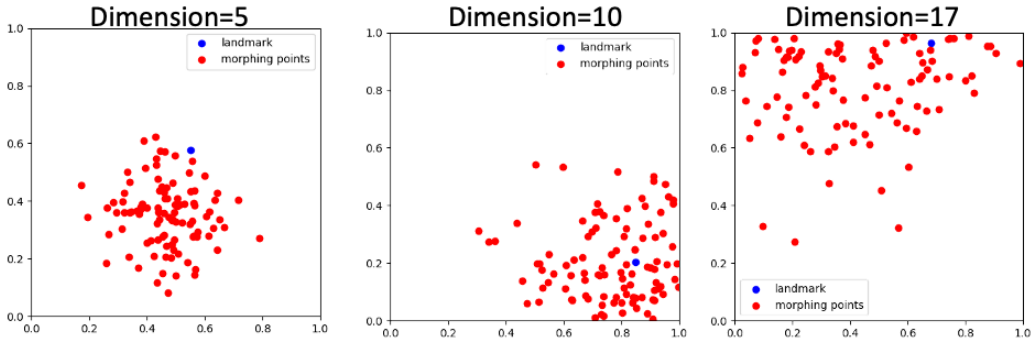

Figure 15: Pairing results of different dimensions.

The curse of dimensionality can be handled in two directions: The first is to increase the amount of available data, which is infeasible in this study; The other one is to restrict the data space by preprocessing the data to remove uncorrelated dimensions and various dimension reduction methods.

5.2. Landmark points selection

In geochemical data augmentation, the algorithm is expected to reproduce all the values and spatial structures of the original data. Therefore, all the points in the original data should be selected at least once as landmark points. The number of landmark points in the augmentation of geochemical data by the SMMT is a paradoxical parameter to consider. The more landmark points selected, the more values and spatial structures will be captured and reproduced, but the variability of augmented data becomes less.

5.3. Mapping simulated values

To project simulated Gaussian values back into the original data space, we use spatially *k*-nearest neighbors of the interpolated point as the anchor of projection in the TPS. This spatial interpolation of the simulated Gaussian values follows the 'First Law of Geography', which infers that near things are more related than distant things (Tobler, 1970). However, spatial interpolation may cause the high variances observed in Fig.6 and Fig.13 for not considering the statistical pattern of the landmark points. There is an alternative way as the statistical interpolation that projects the Gaussian values by their statistically *k*-nearest landmark points. The influences of statistical interpolation on geochemical data are worth exploring in future studies.

In the spatially *k*-nearest TPS interpolation, a different number of anchor points, *k*, produces a different result of interpolations (Fig.16). Theoretically, the number of anchor points must be larger than dimensions+1 while dealing with 2-D data (Rohr et al., 2001). With further increase of the anchor points, high values of landmark points show a stronger influence on neighborhoods. Therefore, the number of anchor points in TPS can be decided either based on the variogram of landmark points or domain knowledge of the geochemical data.

5.4. Implement details

In implementing the workflow of the SMMT, there are serval tricks to accelerate the procedures.

1. **Multi-processing of the sequential Gaussian simulation.** The sequential Gaussian simulation in GSLib can only simulate one dimension at one time. When simulating high-dimensional geochemical data, using the same executable program of the sequential Gaussian simulation with different parameter files will reduce the time on simulation.

2. **Finding *k*-nearest points in an adjustable window.** When projecting simulated Gaussian values back into the original data space, new interpolated points will be added into landmark points for better reproduction of the local structure. However, finding the k-nearest points of the simulated data requires the distances with all the other points which is a heavy calculation for the algorithm with a huge number of landmark points. A window of adjustable size is used to deplete such calculation. The size of the window is decided by the on-time density of the landmark points and to make sure there are more than *k* points inside.

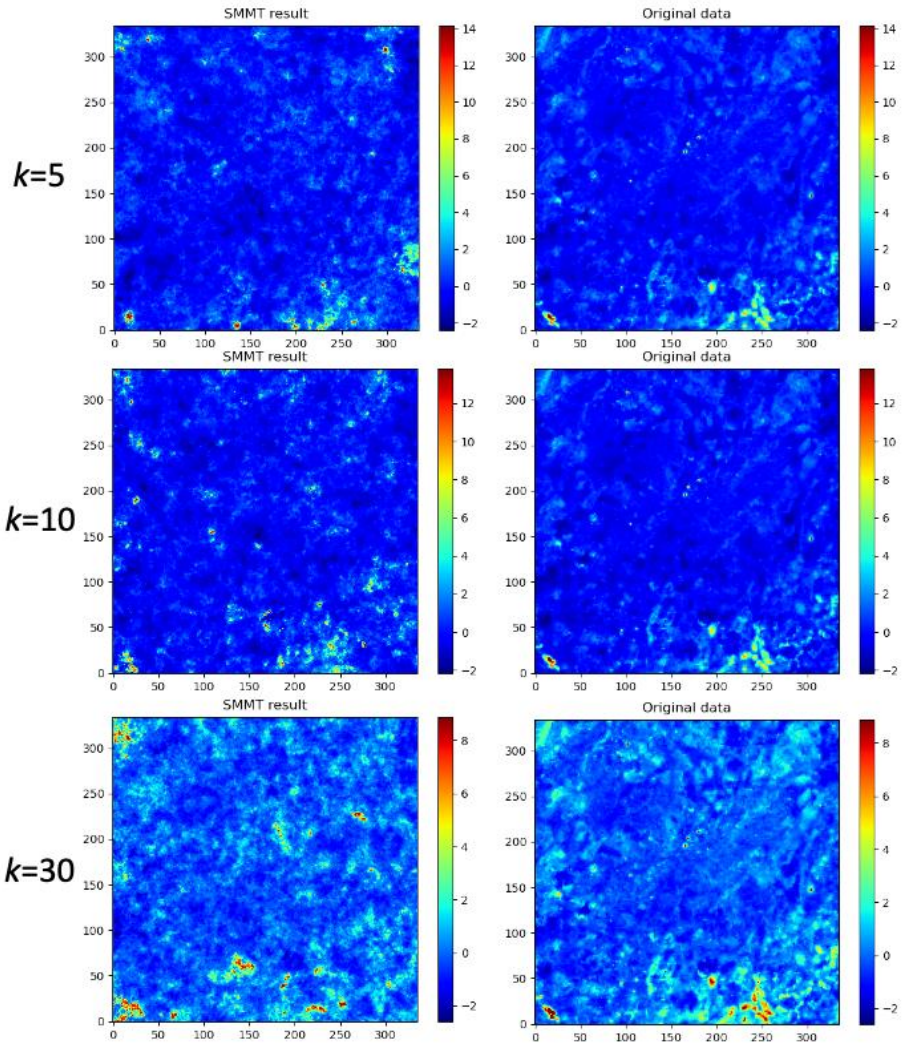

Figure 16: TPS results of different numbers of anchor points (k).

6. Conclusion

In this paper, we presented the workflow of the SMMT on geochemical data for data augmentation. Two sets of geochemical data with a different number of elements are utilized in the SMMT. Results of both sets of geochemical data have shown that the SMMT could decorrelate the data and reproduce the spatial structure of the geochemical data, but the variances of the augmented data were higher than the original data. Several problems that might cause this problem were discussed and future studies on improving the pairing in the optimal transport by reduction of data space and different ways of interpolation were issued to be done.

7. Acknowledgments

We acknowledge the support of the National Natural Science Foundation of China (No. 42172326).

8. References

Avalos S, Ortiz JM (2022) Spatial multivariate morphing transformation applied to geometallurgical attributes, in Geomet-Procemin 2022, Santiago, October 5-7 2022.

Avalos S, Ortiz JM, Leuangthong O (2022) Multivariate morphing transformation: Fundamentals and challenges, in 21st Annual Conference of the International Association for Mathematical Geosciences – IAMG 2022, Nancy, France, Aug 29-Sep 3, 2022.

Bookstein FL (1989) Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence 11(6):567–585.

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press on Demand.

Li T, Zuo R, Zhao X, Zhao K (2022) Mapping prospectivity for regolith-hosted REE deposits via convolutional neural network with generative adversarial network augmented data. Ore Geology Reviews, 142, 104693.

Rohr K, Stiehl HS, Sprengel R, Buzug TM, Weese J, Kuhn M (2001) Landmark-based elastic registration using approximating thin-plate splines. IEEE Transactions on medical imaging 20(6):526–534.

Wang X, Zhang Q, Zhou G (2007) National‐scale geochemical mapping projects in China. Geostandards and Geoanalytical research, 31(4), 311-320.

Xie X (1978) Regional Geochemistry—National Reconnaissance Project. Bulletin of Geophysical and Geochemical Exploration, 3, 28.

Xie X, Mu X, Ren T (1997) Geochemical mapping in China. Journal of Geochemical Exploration, 60(1), 99-113.

Zuo R, Xiong Y (2020) Geodata science and geochemical mapping. Journal of Geochemical Exploration, 209, 106431.