Principal component analysis as pre-processing stage for self-organized maps

Olaf J. Cortés Arroyo[1], Annika Steuer[1], and Benedikt Preugschat[1]

[1]*Federal Institute for Geosciences and Natural Resources (BGR), Hannover*

Abstract

We report a new improvement in self-organized maps for geological interpretation of geophysical data. By using a multi-geophysical dataset recorded in the mining area of Thuringia, Germany, we show the results of replacing the typical feature analysis by a principal component analysis. By performing a transformation of the dataset according to a few of the principal components, we obtain a more detailed representation of the local geology than previous works. Results also show a significant improvement in processing time, while also minimizing influence of user´s interpretation.

1 Introduction

New available technologies in geosciences allows us nowadays to perform multi-geophysical measurements in large areas, with levels of resolution not previously possible. This achievement brings as consequence the great challenge of performing a geological interpretation based on very large amount of data, usually with a limited amount of time. Luckily, current technology also brings new tools to help us with the task. In this work we describe recent developments in project DESMEX-II, related to application of the self-organized maps for geological interpretation. These results are then applied in a multi-feature dataset registered in a mining area located in the state of Thuringia, Germany.

2 Theory

2.1 Self-organizing maps

The self-organizing map (SOM) represents a set of high-dimensional data items as a quantized two-dimensional image in an orderly fashion. Every data item is mapped into one point in the map, where spatial distance of items reflect similarities between them. Every single neuron in the map is connected to the data by an individual weight vector, and the selected point (known as Best Matching Unit or BMU) is the neuron with the smallest Euclidean distance between data and each individual cell.

Once a BMU is selected, an optimization process activates a modification of weight values in the neuron. This process also triggers activation of the neighboring neurons, where the amplitude of changes for the neighbors is in function of the distance to the BMU. Therefore, neurons near the BMU will have a more significant weight adaptation, while cells located at a large distance from the BMU will have a very small weight update or none.

In order to evaluate the SOM results we make use of two error values named quantization error (QE) and topographic error (TE), where QE represents an estimation of data dispersion and TE is a measure of topology distortion.

Following Pölzbauer (2004) QE is computed by determining the normalized, average distance of the sample vectors to the cluster centroids that representes them, while TE is calculated as the normalized distance between the best and second-best matching units. Both errors are represented as values between 0 and 1, and is important to keep both values in balance, as also close to zero as possible.

Data samples sharing similar properties will show in the SOM as a cluster of points with very small distance between them. Once we obtain a map with satisfactory QE and TE levels, the next step is to define boundaries between the clusters.

2.2 K-means clustering

K-means (Lloyd, 1957; MacQueen, 1967) is one of the most popular algorithm to define boundaries between the SOM-clusters and is often a standard option in available SOM algorithms. In general terms, first step is to choose the initial centroids, with the most basic method being to choose samples from the dataset. Step two assigns each sample to its nearest centroid. Step three creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. Difference between the old and the new centroids are computed and the algorithm repeats these last two steps until this value is less than a threshold (Scikit-learn developers, 2021).

2.3 Principal Component Analysis

Principal component analysis (PCA) is the new incorporation into our SOM process. PCA finds a new set of orthogonal axes that have their origin at the data mean, rotated so that data variance is maximized. The first principal component (PC) represents the largest percentage of data variance, the second PC band represents the second-largest data variance, and so on (Guo et al., 2009).

3 Schleiz dataset and project DESMEX

3.1 Schleiz dataset

In the framework of project DESMEX (Becken et al., 2020) multiple geophysical surveys were carried out by the DESMEX Working Group in the mining area of Schleiz, located in the state of Thuringia, Germany (Smirnova et al., 2019; Steuer et al., 2020).

The survey area is located in the Berga anticline (Figure 1), which elongates in the SW-NE direction and is bordered by several major faults and synclines. According to the three-dimensional geological model of Müller and Kroner (2019), the sedimentary formations in the anticline are an overthrust unit (Ordovician Weißelster Group, Ordovician Phycodes and Gräfenthal Group) and a Devonian unit (Devonian and Silurian and Lower Carboniferous). The area was selected due to the antimony deposits near the city of Schleiz, extracted among other minerals from 1846 until beginning of the 1950s (Dill, 1985).

A regional-scale reconnaissance survey was performed with BGR's standard helicopter-borne geophysical system. The resulting dataset, complemented with petrophysical and geological information, consists of more than 200,000 samples. Features in the dataset consists of UTM coordinates (X,Y), frequency-domain electromagnetic (1-D, resistivity model, 25 m depth), magnetic (apparent magnetic susceptibility, analytical signal, total magnetic field anomaly) and radiometric data (Uranium, Thorium, Potassium, Total count). Gravimetry (vertical and horizontal gradients) from LIAG (2010) is also included.

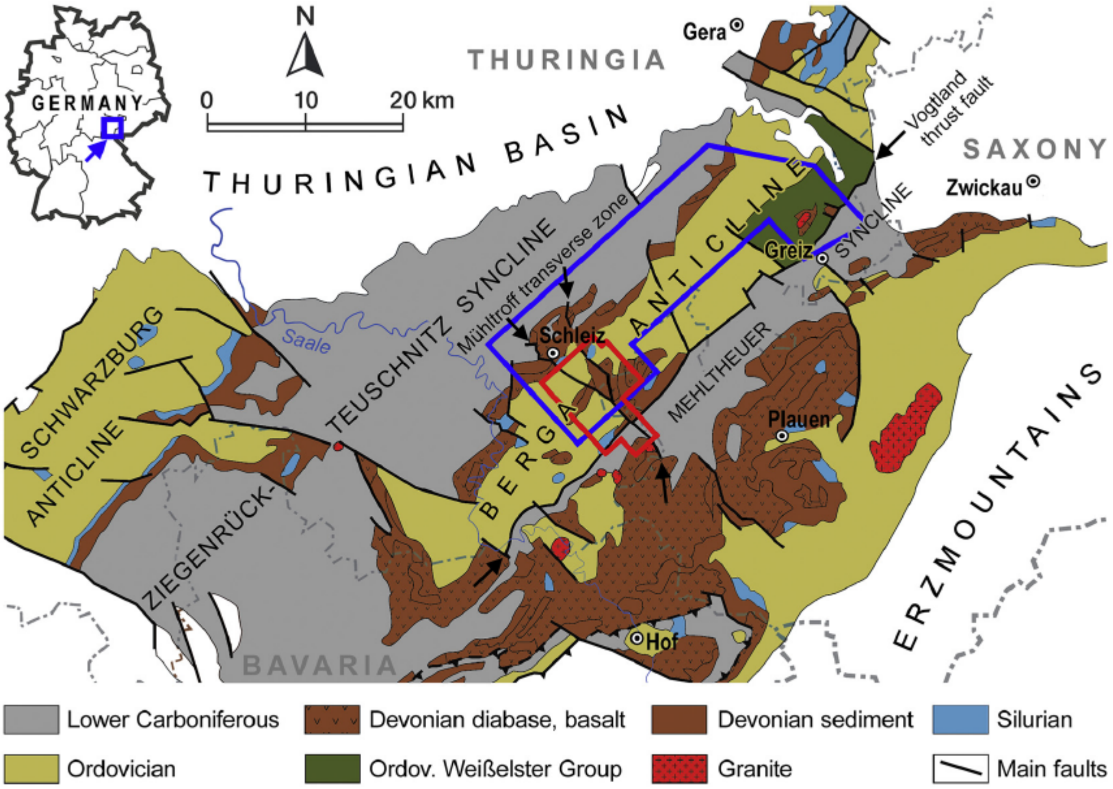

Figure 1: General workflow for the SOM analysis of the Schleiz data: survey dataset is applied as input data for SOM analysis, defining several clusters which, by comparing by geographical position, are expected to represent the different geological structures in the area. Modified from Steuer et al. (2020).

3.2 Workflow of SOM analysis for geological interpretation

We use the SOM analysis with the goal of performing a geological interpretation directly from the dataset. In order to do this, we follow the general workflow shown in Figure 2. The recorded dataset works as input data for the SOM algorithm, which uses the K-means algorithm to define the limits between clusters. User must provide in advance a number of clusters for K-means calculations. Since our goal is geological interpretation, the number of observed rock types in the area should be the preferred quantity of clusters, but usually this value is reduced when the dataset does not possess enough information to discriminate between all of rock types (resolution in data). Finally, by using the geographical position of each sample we obtain a cluster-representation that hopefully resembles the local geology. This result is not unique and is dependent of the personal interpretation of the user. However, in our experience an overall accurate representation of the geology is usually achievable.

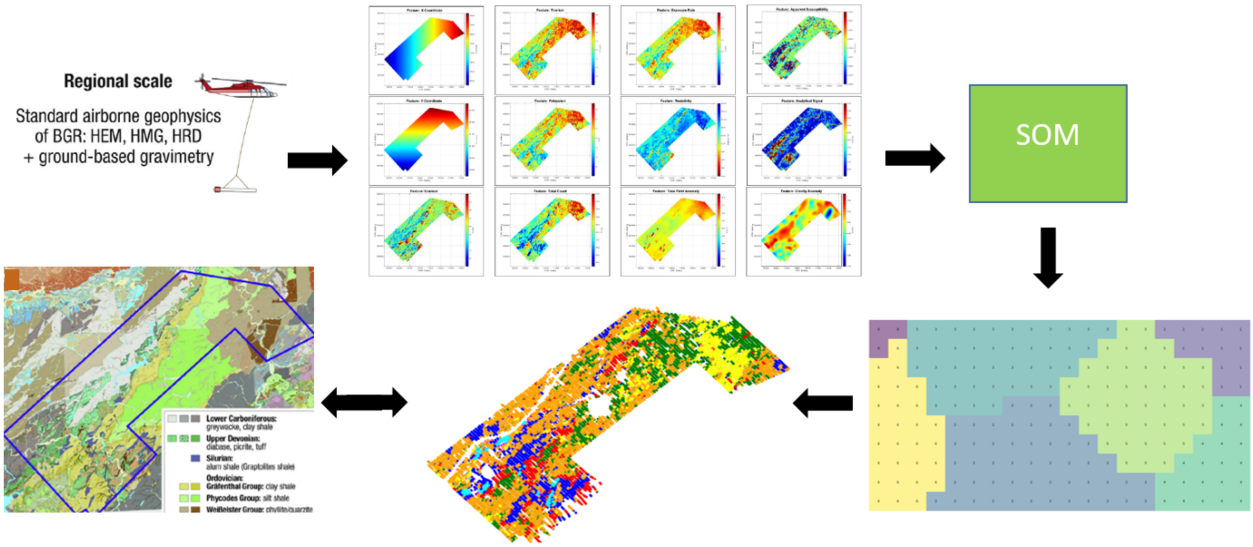
Regional scale
Standard airborne geophysics of BGR: HEM, HMG, HRD + ground-based gravimetry
SOM
Lower Carboniferous: greywacke, clay shale
Upper Devonian: diabase, picrite, tuff
Silurian: alum shale (Graptolites shale)
Ordovician:
Gräfenthal Group: clay shale
Phycodes Group: silt shale
Weißelster Group: phyllite/quarzite

Figure 2: Geology of the Schleiz survey area in Thuringia, Germany. Blue line represents the regional survey area covered in project DESMEX-II from Steuer et al. (2020).

3.3 Feature analysis and SOM model for the Schleiz dataset

Preugschat (personal communication) performed a rigorous statistical analysis of the dataset in order to select the features with valuable information, and discard those who show either a high-correlation level or little information regarding the geology of the area. This defines a reduced dataset of only five features: Uranium, Thorium, Potassium, electrical resistivity and analytical signal. This selected dataset was used first to define the optimal dimensions for the SOM, resulting in a grid of 20 x 20 neurons. This analysis also defines that maximal resolution of the map consists of seven clusters. The SOM analysis uses package SOM Toolbox in Matlab. The analysis of the SOM-clusters shows a good correlation with borehole and petrophysical data, as also with visual comparison of the geological structures in the area. Despite good results, several geological structures are not very well represented, such as the Weißelster group in the northwest-end of the area (Figure 3). A more detailed description of the results is found in Steuer et al. (2020).

4 New developments in DESMEX-II

4.1 SOMPY software and re-evaluation of datasets and models

First, we tested an open source software for SOM written in Python called SOMPY (Moosavi et al., 2014), which is very similar to Matlab´s SOM Toolbox. We use most of the same parameters of Preugschat (personal communication) and Steuer et al. (2020), the only difference is the shape of the neurons grid (rectangular instead of hexagonal) due to better performance in SOMPY. The dataset is the same 5-features dataset previously described. Visual inspection of our results (Figure 5A) shows good agreement with the model of (Steuer et al., 2020). For a second test, we repeated the same process, replacing the 5-feature dataset of Preugschat with the original full dataset. Results (Figure 5B) for most of the area represent poorly the geology of the area. However, a striking feature is the clear representation of the Weißelster group, not visible in the previous model. This

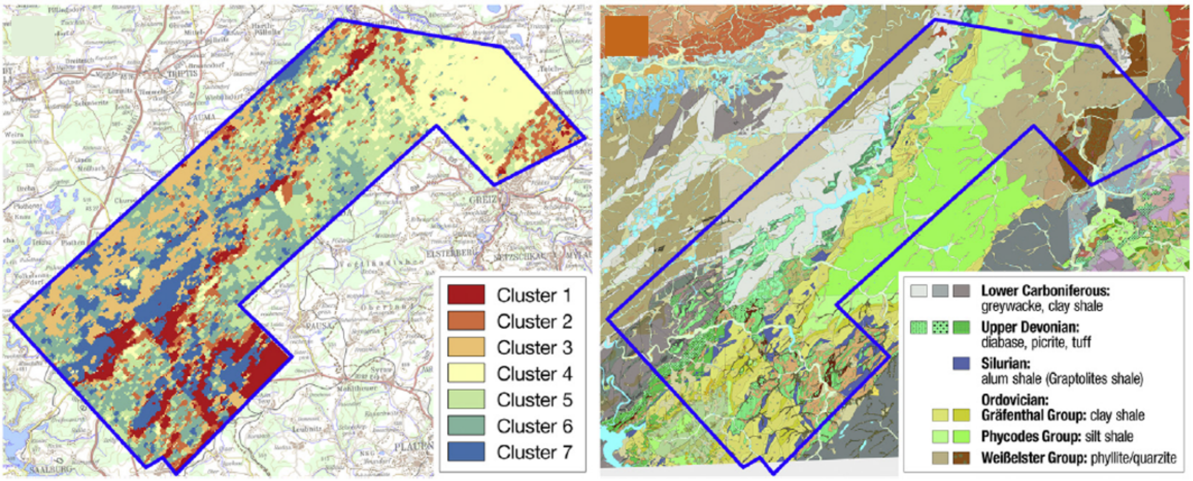
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Lower Carboniferous: greywacke, clay shale
Upper Devonian: diabase, picrite, tuff
Silurian: alum shale (Graptolites shale)
Ordovician:
Gräfenthal Group: clay shale
Phycodes Group: silt shale
Weißelster Group: phyllite/quarzite

Figure 3: Comparison of the SOM-clusters with local geology, resulting from the dataset of five selected features (Steuer et al., 2020).

result shows us that a process of feature selection, where some of the recorded parameters are fully discarded, can provide good overall results for the final model estimation, but it may also cause the loss of some of the information in the process.

4.2 PCA analysis of the Schleiz dataset and new model

In order to improve the model, we replace the feature-selection process by PCA. This process divides in two stages: First, we evaluate the variance percentage for all principal components in the original dataset (Figure 4), using the PCA function of Scikit-Learn´s Python library (Pedregosa et al., 2011). The first three principal components represents more than 50% of the variance, and after several trial-and-error tests we confirm that no meaningful information is recovered by including further components.

The second step consists of performing a transformation of the original dataset. This is possible by obtaining a representation of the dataset, using only the first three principal components. Because of this, we obtain a new dataset with a dimensionality reduction of approximate 75%. Repeating the SOM analysis by using the same parameters of the previous tests and our new dataset, we obtain the model shown in Figure 5C.

4.3 Model-comparison for the Schleiz dataset

Comparison of the three models with the local geology is shown in Figure 5. After a visual inspection, is clear that the “PCA model” provides a more accurate representation for the geology of the whole area, including the Weißelster Group. For the “PCA model”, a quantization error of 0.177 and a topographic error of 0.235 are observed, where the first is the lowest error for the three models, while the second is above the “all features” model and below the “five features” model. These results confirms that the best SOM-based geological interpretation for the Schleiz dataset is obtained from our new PCA-SOM process.

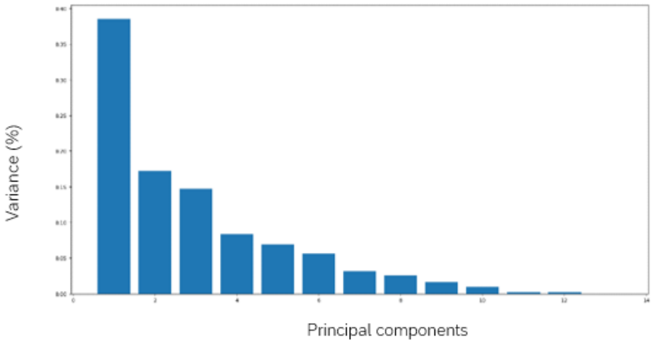

Figure 4: Representation of the percentage of variance of the principal components in the Schleiz dataset. The first three principal components represent more than 50% of the total variance. The rest are discarded (set equal to zero) during transformation of the dataset.

5 Discussion

There is a clear advantage in replacing the feature selection process by a PCA analysis: is faster (no need for time-consuming feature analysis), analysis is very straightforward, influence of user´s interpretation is minimized and a higher dimensional reduction of the dataset (75% instead of 60%) reduces the calculation time. However, despite many tests, including some recommended by Kohonen (2014), we were not as effective as the process developed by Preugschat (personal communication) to define the dimensions of the SOM. Therefore, a merge of both methods is desirable. Application of SOM for geological interpretation is not a novelty (Taner et al., 2001; Bauer et al., 2012; Fraser et al., 2012; Carneiro et al., 2012; Roden et al., 2015). However, to our knowledge this work is the first time that application of a PCA-transformed dataset is reported, at least for this particular goal. Examples of PCA application as a pre-processing filter can be found in the machine learning literature (VanderPlas, 2016).

6 Conclusions

Application of a PCA analysis and transformation of the Schleiz dataset proves to be a significant step forward, compared to previous results. A reduction in time due to removal of the feature-selection process and a larger dimensionality reduction of the dataset translates into a significant reduction of SOM calculations. The resulting model is a better representation of the local geology. To our knowledge, this is the first time that such a PCA approach for SOM analysis is applied for geological interpretation. In a future work, we will explore the implementation of the methodology by Preugschat for map configuration. Finally, this work also shows the power and value of open-source software alternatives.

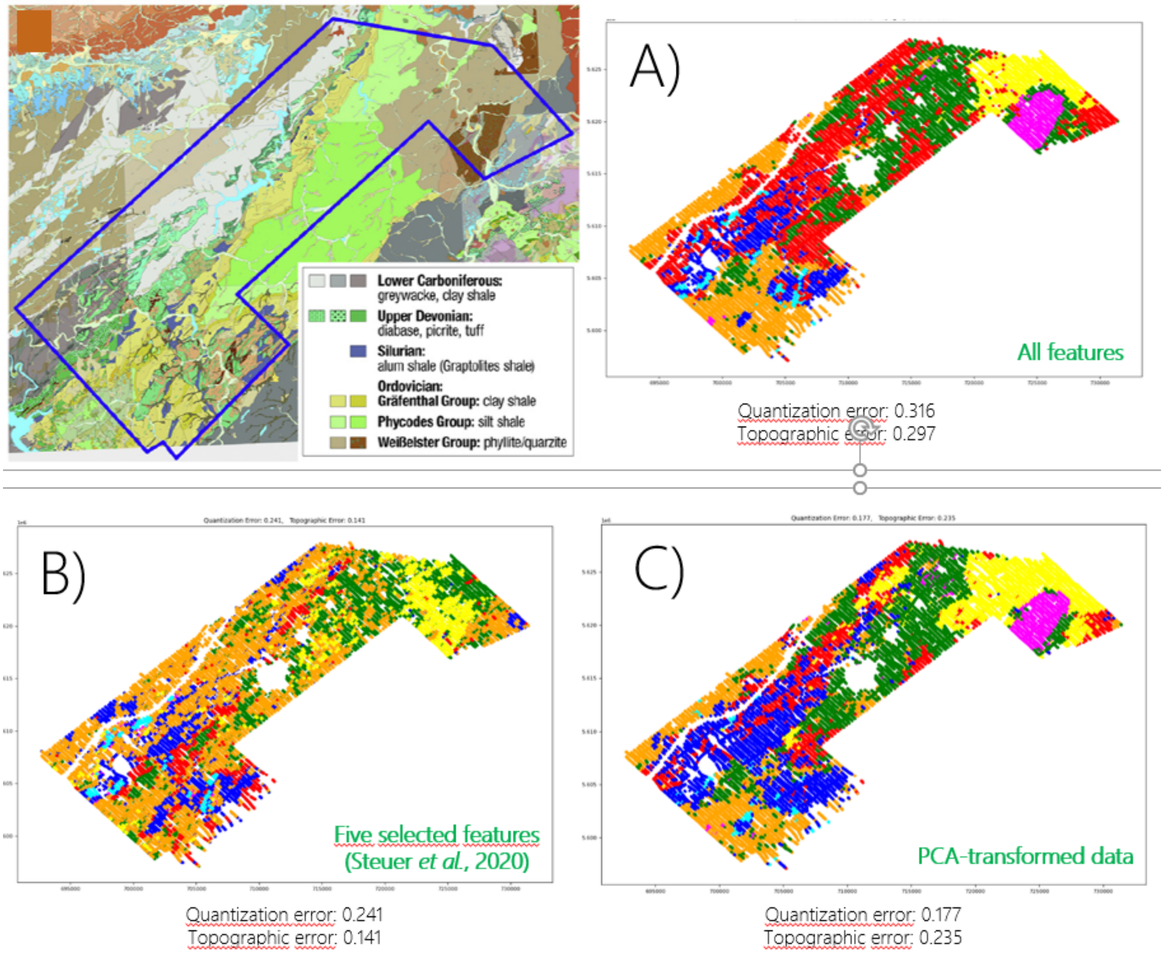
Lower Carboniferous:
greywacke, clay shale
Upper Devonian:
diabase, picrite, tuff
Silurian:
alum shale (Graptolites shale)
Ordovician:
Gräfenthal Group: clay shale
Phycodes Group: silt shale
Weißelster Group: phyllite/quarzite
A)
All features
Quantization error: 0.316
Topographic error: 0.297
Quantization Error: 0.241, Topographic Error: 0.141
B)
Five selected features
(Steuer et al., 2020)
Quantization error: 0.241
Topographic error: 0.141
Quantization Error: 0.177, Topographic Error: 0.235
C)
PCA-transformed data
Quantization error: 0.177
Topographic error: 0.235

Figure 5: Geology of the survey area (up, left). Model A (up, right) model resulting from SOM classification for the original dataset. Model B (down, left) is the five-features, equivalent model reported in Steuer et al. (2020). Model C (down, right) is the SOM model for the PCA-transformed dataset reported in this work. All models were obtained with a map of 20 x 20 neurons, considering 7 clusters during K-means calculations.

References

Bauer, K., Muñoz, G., & Moeck, I. (2012). attern recognition and lithological interpretation of collocated seismic and magnetotelluric models using self-organizing maps. *Geophysical Journal International*, *189*(2).

Becken, M., Nittinger, C. G., Smirnova, M., Steuer, A., Martin, T., Petersen, H., et al. (2020). Desmex: A novel system development for semi-airborne electromagnetic exploration. *Geophysics*, *85*.

Carneiro, C., Fraser, S., Crósta, A., Silva, A., & Barros, C. (2012). Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the brazilian amazon. *Geophysics*, *77*.

Dill, H. (1985). Antimoniferous mineralization from the mid-european saxothuringian zone: mineralogy, geology, geochemistry and ensialic origin. *Geologische Rundschau*, *74*(3).

Fraser, S., Wilson, G., Cox, L., Čuma, M., Zhdanov, M., & Vallée, M. (2012). Self-organizing maps for pseudo-lithological classification of 3d airborne electromagnetic, gravity gradiometry and magnetic inversions. *ASEG Extended Abstracts*, *2012*.

Guo, H., Marfurt, K., & Liu, J. (2009). Principal component spectral analysis. *Geophysics*, *74*.

Kohonen, T. (2014). Matlab implementations and applications of the self organizing map. *Unigrafia Oy, Helsinki, Finland.*

LIAG. (2010). Schwerekarte der bundesrepublik deutschland 1:1 000 000, bouguer – anomalien. *Leibniz-Institute for Applied Geophysics, Hannover.*

Lloyd, S. (1957). Least squares quantization in pcm. *Technical Report RR-5497, Bell Lab.*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. California: University of California Press*, *1*.

Müller, F., & Kroner, U. (2019). Tectonic 3d-model of the berga antiform – saxo-thuringian zone. *GEOMÜNSTER 2019 Conference, Book of Abstracts.*

Moosavi, V., Packmann, S., & Vallés, I. (2014). *Sompy: A python library for self organizing map (som).* (GitHub.[Online]. Available: https://github. com/sevamoo/SOMPY)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*.

Pölzbauer, G. (2004). Survey and comparison of quality measures for self-organizing maps. *Proceedings of the Fifth Workshop on Data Analysis. Sliezsky dom, Vysoké Tatry, Slovakia, Elfa Academic Press.*

Roden, R., Smith, T., & Sacrey, D. (2015). Geologic pattern recognition from seismic attributes: Principal component analysis and self-organizing maps. *Interpretation*, *12*(3).

Scikit-learn developers. (2021). *Scikit-learn user´s guide: K-means.*

Smirnova, M. V., Becken, M., Nittinger, C., Yogeshwar, P., Mörbe, W., Rochlitz, R., et al. (2019). A novel semiairborne frequency-domain controlled-source electromagnetic system: Three-dimensional inversion of semiairborne data from the flight experiment over an ancient mining area near schleiz, germany. *Geophysics*, *84*.

Steuer, A., Smirnova, M., Becken, M., Schiffler, M., Günther, T., Rochlitz, R., et al. (2020). Comparison of novel semi-airborne electromagnetic data with multi-scale geophysical, petrophysical and geological data from schleiz, germany. *Journal of Applied Geophysics*, *182*.

Taner, M. T., Walls, J. D., Smith, M., Taylor, G., Carr, M. B., , et al. (2001). Reservoir characterization by calibration of self-organized map clusters. *SEG Technical Program Expanded Abstracts*.

VanderPlas, J. (2016). Python data science handbook. *O´Reilly Media*.