https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Broken Terrains v. 1.0: A supervised detection of fault-related lineaments on geological terrains

Michał P. Michalak[1], Christian Gerhards[2], Peter Menzel[2]

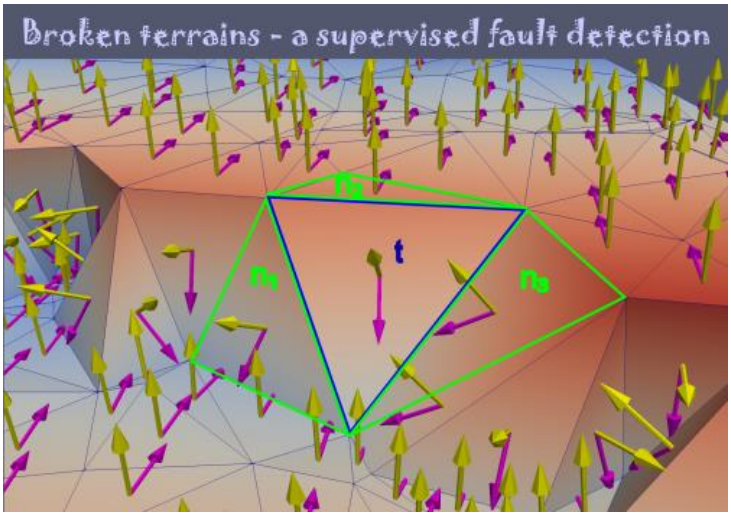
[1]Faculty of Geology, Geophysics and Environmental Protection, AGH University of Science and Technology, Mickiewicza 30, 30-059 Cracow, Poland, ORCID: https://orcid.org/0000-0002-1376-235X

[2]Institute of Geophysics and Geoinformatics, TU Bergakademie Freiberg, Gustav-Zeuner-Straße 12, 09599 Freiberg, Germany

Correspondence to: Michał P. Michalak (michalm@agh.edu.pl)

Abstract.

The study presents a novel approach for fault detection on geological terrains using supervised learning algorithm and careful variable selection. Synthetic faulted terrains are generated using Delaunay triangulation via the Computational Geometry Algorithms Library (CGAL) allowing for adjustments of parameters. We introduce 24 variables, including local geometric features and neighborhood analysis, for classification. Support Vector Machine (SVM) is employed as the classification algorithm, achieving high precision and recall rates for fault-related observations. Application to real borehole data demonstrates the effectiveness of the method in detecting fault orientations, the challenges remain with respect to distinguishing faults with opposite dip directions. The study highlights the need to address 3D fault zone complexities and their identification. Despite limitations, the proposed supervised approach offers significant advancement over clustering-based methods, showing promise in detecting faults of various orientations. Future research directions include exploring more complex geological scenarios and refining fault detection methodologies.



https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Short Summary

This study presents a novel method for fault detection on geological terrains. Using synthetic models, we applied machine learning to classify terrain shape and nearby features. Testing on real borehole data validated its effectiveness across various fault orientations. The supervised approach represents a significant improvement over older methods that relied on simpler clustering techniques which were capable of identifying less orientations of faults.

1 Introduction

Geological engineers and structural geologists aim to identify lineaments or faults on geological terrains. However, current methods are typically tailored for seismic data rather than terrains (An et al., 2021; Kaur et al., 2023). Additionally, supervised methods for fault detection can encounter challenges related to subjectivity, ambiguity, or time-consuming processes such as manual labeling of training data (Mattéo et al., 2021; Vega-Ramirez et al., 2021).

Our goal is to expedite the process of generating ground truth data for faulted triangulated terrains using the Computational Geometry Algorithms Library (CGAL.org, 2023). We will employ supervised machine learning algorithms for binary classification to predict possible lineament/fault presence within faulted terrains (Fig. 1). Our hypothesis posits that while traditional geometric attributes such as normal or dip vectors can still be useful for classification, integrating variables reflecting angular relationships between triangles and their neighbours is crucial for accurate classification, especially for fault detection on homoclines. We assert that analyzing distances for neighbours (Fig. 2) is advantageous due to its insensitivity to terrain rotation, unlike traditional geometric attributes such as dip direction (Hu et al., 2021) or the orientation of normal vectors (Michalak et al., 2022). As such, neighbourhood analysis can be linked, e.g., to curvature in seismic data (de Oliveira Neto et al., 2023) in terms of its insensitivity to terrain rotation.

The main challenges relate to the effectiveness of machine learning algorithms, variable selection, and the applicability of the method to diverse geological structures, potentially impacting classification accuracy and generalizability. To mitigate these challenges, we will conduct optimization of the algorithm's performance and variable selection (see Methodology). Validation across various geological terrains will ensure the method's robustness and applicability for fault detection on homoclines. This

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

structured approach aims to enhance classification accuracy and the method's utility in practical geological applications. The overall workflow of the study is presented in Fig. 3.

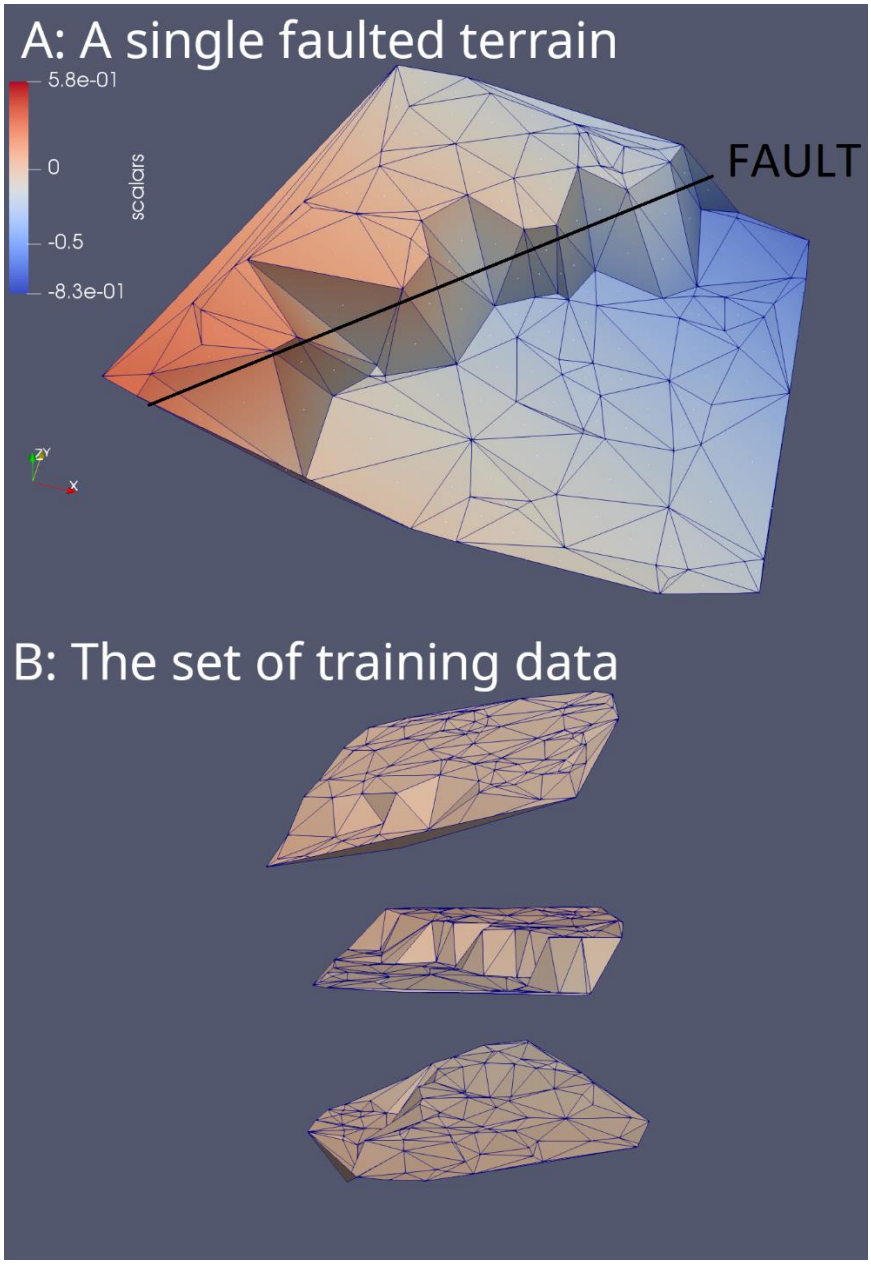


Figure 1. A triangulated model of a faulted geological terrain: (A) we can see an inclined terrain and triangles that intersect a fault line. (B) A set of terrains with different parameters (dip angle and dip direction) can be used as training data in the classification task. In this panel, we showed only three terrains, but in practice an arbitrary number of terrains can be generated.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

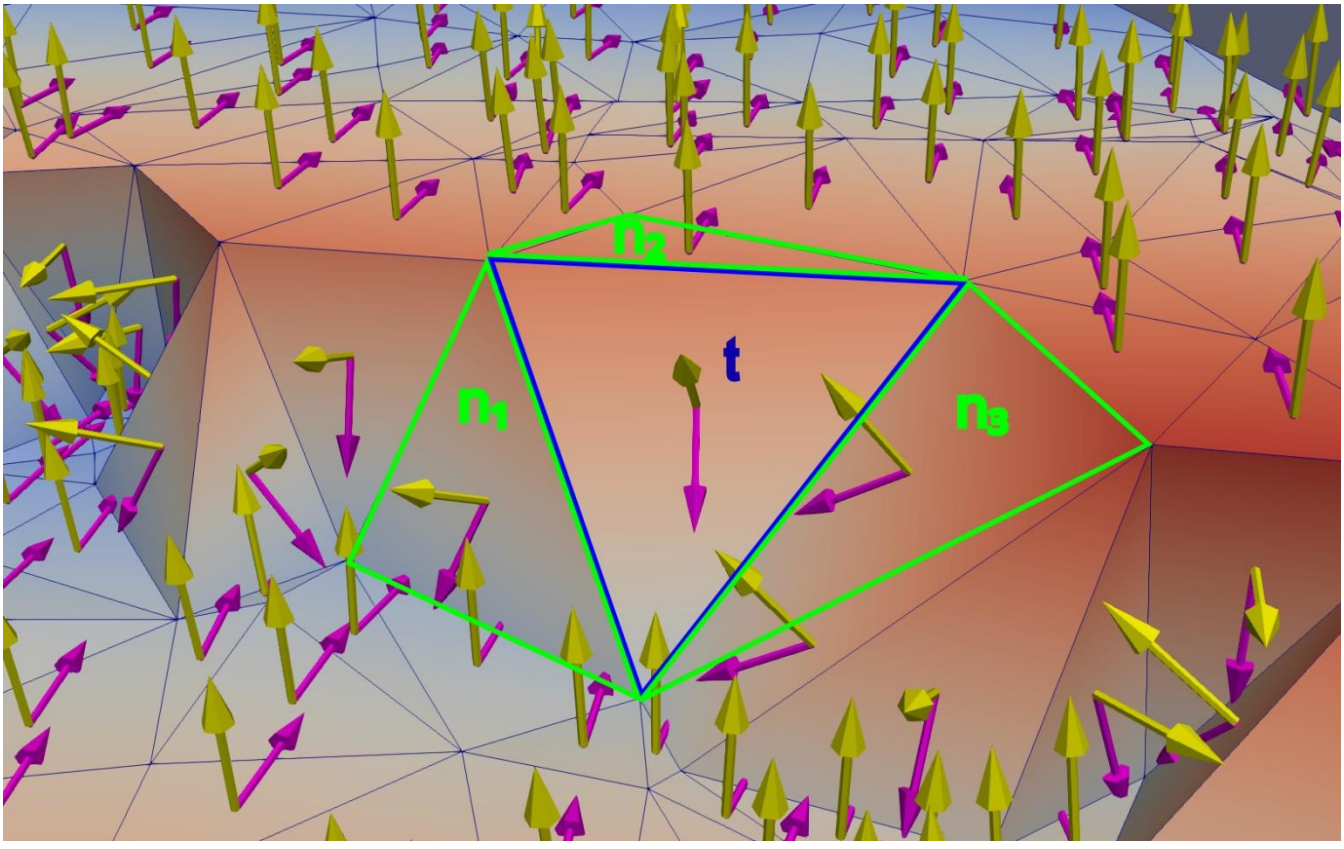


Figure 2. Because the fault introduced changes in the angular relationships between the orientation of fault-related triangles (t) and their neighbours (n1, n2, n3), the analysis of these relationships is essential for a successful classification. For example, we can measure the angular distance between normal vectors for three pairs corresponding to a specific triangle and its neighbours. The resulting value of angular distance can serve as a variable in the classification task.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

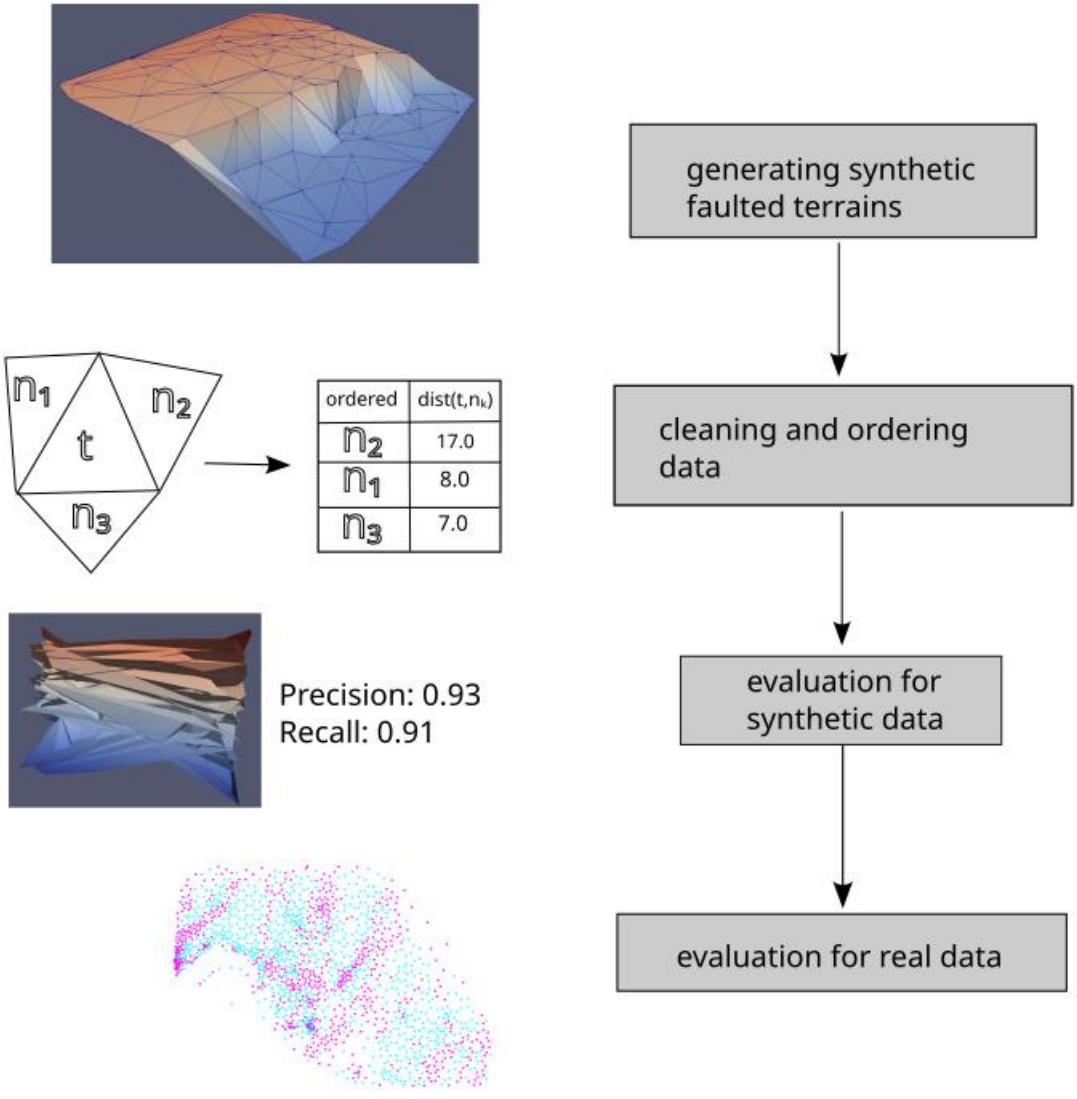


Figure 3 Workflow applied in this study. We create many random faulted geological terrains controlled by random parameters. Then, for each triangle, we sort the distances between neighbours to reduce randomness. In the next step, we test machine learning algorithms for synthetic data. At the end of the procedure, we evaluate the proposed approach for real data to test generalizability.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

2 State of the art

2.1 Related methodological developments

In geological mapping, machine-learning methods have been applied in the supervised lithology classification (Cracknell and Reading, 2014; Kuhn et al., 2018; Xiong and Zuo, 2021; Wang et al., 2020). In geological engineering, unsupervised methods were used to delineate subsets of observations representing discontinuities (Hammah and Curran, 1999; Zhan et al., 2017). In subsurface geological modelling, neural networks were used to delineate paleovalleys using topographic data as input data (Jiang et al., 2021) and convolutional neural network were used to create geological models with structural features controlled by a set of random parameters (Bi et al., 2022). As a specific unsupervised learning method, clustering algorithms find application in triangulated geological terrains (Michalak et al., 2022). These methods generate partitions comprising geometrically similar observations based on cosine similarity (Choi et al., 2014). However, they place the burden on the user to determine whether a specific observation represents a fault. This can pose challenges, as some anomalous orientations may be associated with other structures or measurement errors. Moreover, applying unsupervised learning to 3D orientations reveals sensitivity to the choice of vectorial representation (Michalak et al., 2022), resulting in varying clustering results for dip and normal vectors. From a geological viewpoint, a serious limitation of unsupervised learning is that the orientations of the identified lineaments depend on the partition resulting from the clustering. For example, previous results suggest that for uniformly oriented sub-horizontal terrains (homoclines) the clustering algorithms may find it difficult to distinguish between observations related to the regional trend and observations related to faults striking perpendicular to the regional trend (Fig. 4b, 4c) (Michalak et al., 2022). In the problem of fault/lineament detection, the majority of available supervised methods are primarily tailored for seismic data (An et al., 2021; Kaur et al., 2023). For topographic data, supervised methods were utilized for fault-scarp prediction (Vega-Ramirez et al., 2021) using Fisher Linear Discriminant Analysis. However, this analysis relied on high-resolution bathymetric data based on a small training dataset (163 samples). Another example involves the use of topographic attributes such as DEM, slope, aspect, faults and environmental variables such as vegetation and climate for monitoring of ground deformation (Hu et al., 2021).

2.2 Geological Setting

As a relevant case study, we selected Kraków-Silesian Homocline (KSH) – a geological unit considered to be a slope of the Szczecin-Łódź-Miechów Synclinorium. The formation of KSH is mainly attributed to the inversion of the Permian-Mesosoic Polish Basin (Dadlez et al., 1995; Słonka and Krzywiec, 2019). From a geometric perspective, KSH dips at low angles to NE (Matyszkiewicz et al., 2015; Marynowski et al., 2007; Michalak et al., 2019; Znosko, 1960). It is generally assumed that the faults form a unimodal set of sub-parallel faults trending NE-SW (Fig. 4) (Hermański, 1993; Bardziński et al., 1985). However,

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

later experiments (Michalak et al., 2022) added knowledge about geometric anomalies also aligned with the N-S direction (Fig. 4).

Little is known about faults trending perpendicular to the preferred dip direction. While the results (Fig. 4b, 4c) suggest that they may not exist, we note that this negative effect can be due to limitations of unsupervised learning methods: the spatial distribution of labels depends on the partition induced by clustering algorithms. This dependence may result in visual disintegration of rare structures represented by observations being in different clusters. For example, the boundary between blue and purple labels may be related to faults dipping to SW (Fig. 4). Likewise, it is unlikely that all observations dipping to NE are genetically related to the homocline; instead, observations dipping to NE but having dip angle greater than that of the homocline, may be related to faults dipping to NE.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

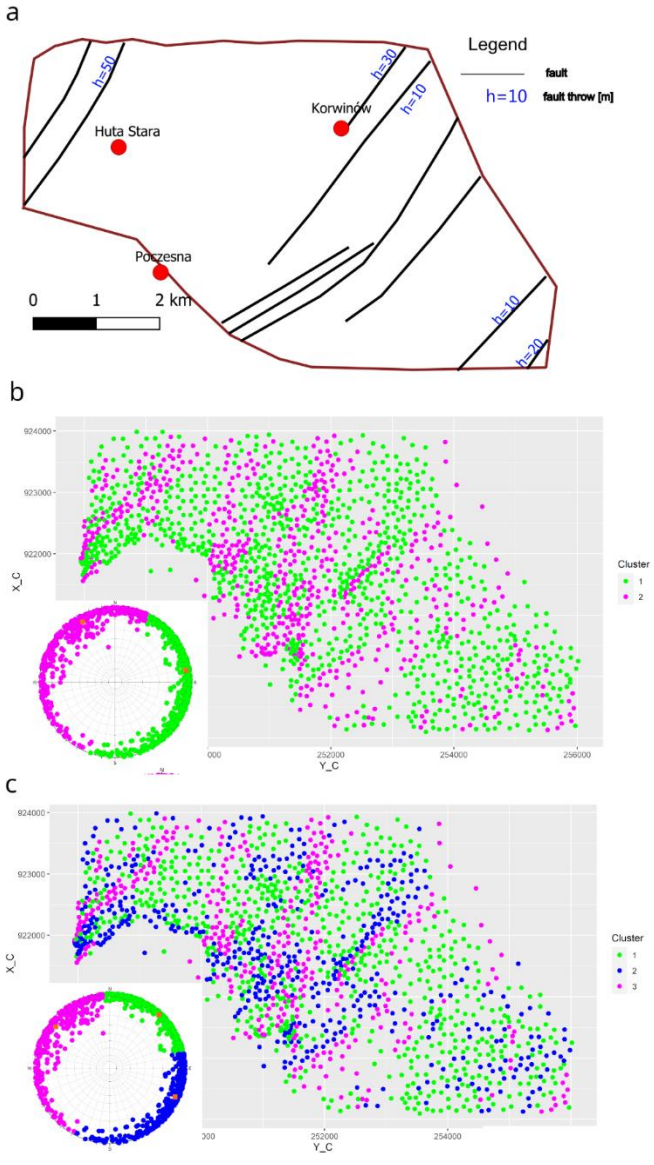


Figure 4 Progressing knowledge about tectonics of the Kraków-Silesian Homocline. (a) due to abandoned mining activity in the area it was possible to confirm some of the faults and their features such as fault throw in underground mines (Hermański, 1993). Later experiments based on cluster analysis (Michalak et al., 2022) of normal and dip vectors provided evidence about the orientation of geometric anomalies. (b) – clustering of dip vectors for two clusters. The spatial distribution of labels suggests presence of geometric anomalies trending from S-N to SW-NE (c) – clustering of dip vectors for three clusters. The spatial distribution of labels in the NW part of the study area suggests presence of more than one fault trending SW-NE with opposite dip direction. However, the partition induced by the clustering makes it impossible to identify faults dipping to NE steeper than the homocline.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

3 Methods

3.1 Generating terrains

In our study, the training data consists of many triangulated terrains using the Delaunay triangulation (De Berg et al., 2008). A user has the flexibility to adjust parameters of the resulting data set in the following fields: the number of files to generate, the lower and upper bound of terrain sizes, the left and right range of the dip direction, the lower and upper bound of the dip angle, the lower and upper bound of the number of points in the triangulation, the lower and upper bound of the surface noise, the lower and upper bound of the fault throw. However, if a constant value of a parameter should be investigated, it is possible to set the same value for the lower and upper bound. In the next step, tools from the C++ random number library generate random numbers from the uniform distribution using the bounds entered by a user. The information about parameters are saved to a text file to allow further inspection.

The faulted triangulated terrains are created in the following sequence which is also summarized in Fig. 5. First, a container with 2D points is generated within a square of a given size**.** Then, a new container of 3D points is created with the Z coordinate corresponding to the random value of dip and dip direction (ranges specified by a user). In the next step, a noise is introduced to the surface defined as a random fraction (ranges specified by a user) of the elevation difference within the terrain. Then, a fault is introduced with the throw defined as a random fraction (ranges specified by a user) of the maximum elevation difference within the generated terrains. The orientation of the fault is determined by two points randomly selected from the boundary of the square. Next, the triangulation of the terrain is performed and the attributes including relationships with neighbours are calculated. Finally, we note that the classification task involves labeling each observation based on whether it is a fault-related observation (label=1) or not (label=-1). Therefore, we use the intersection predicate (CGAL.org, 2023) to test whether a specific triangle intersect the line representing a fault. Following this approach, we are capable of generating great amount of synthetic and labeled ground truth data. To ensure that the training is performed on good quality data, we removed triangles with high degree (0.90 and greater) of collinearity defined as a ratio between the longest triangle's edge and the sum of remaining lengths (Michalak, 2018). This coefficient lies in the interval [0.5, 1] with lower and higher values pointing to equilateral and collinear configurations, respectively (Michalak et al., 2021; Michalak, 2018).

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

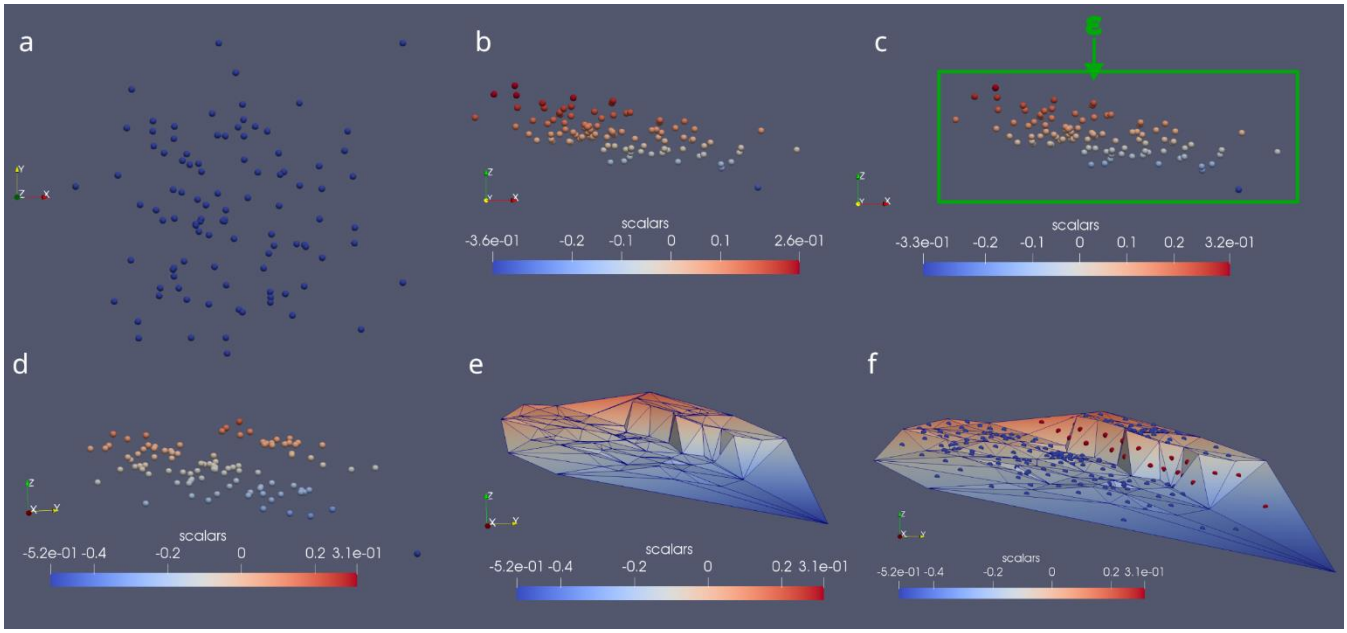


Figure 5. Depiction of sequence of processes applied to generate training data: (a) creating points in 2D space. (b) assigning elevation to the data depending on the randomly generated dip angle and dip direction, (c) adding noise to the data, (d) introducing faults and resulting elevation changes, (e) applying triangulation to the data, (f) labelling the data according to the intersection test with the fault line.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

3.2 Selecting meaningful and consistent variables

There can be numerous geometric variables used for the purpose of classification such as dip angle or dip direction (Hu et al., 2021; Wang et al., 2021). However, including dip direction for classification as a value within the [0, 360] range may not always be successful. This is because northern directions indicate great numerical difference (e.g. 358-2=356) but very small geometric difference (4 degrees). Sometimes the limitations of using dip direction are acknowledged and the variable is removed from the analysis (Yang et al., 2023).

In this study, to predict the correct label (label=1 for fault-related observations and label=-1 for non-fault-related observations), we used 24 variables. The set consists of six local geometric variables and eighteen variables corresponding to the neighbourhood analysis. The first group consists of coordinates of normal and dip vectors. The second group includes variables corresponding to the neighbourhood component of the analysis. The variables are as follows: angular distance, Euclidean distance and cosine distance applied to both normal and dip vector representations. The formulas for angular, Euclidean and cosine distances are given in the below equations, respectively:

$$d_a(x, x') = \text{acos}\left(\frac{|x \cdot x'|}{||x|| * ||x'||}\right) \quad \text{(Eq. 1)}$$

where " $\cdot$ " is the dot product, and $||x||$ is the length of the vector x. In our case, the vectors have unit length. The use of absolute value in the numerator reflects the use of acute angles between vectors.

$$d_e(x, x') = ||x - x'|| \quad \text{(Eq. 2)}$$

$$d_c(x, x') = 1 - x \cdot x' \quad \text{(Eq. 3)}$$

However, in relation to the proposed neighbourhood analysis, an obstacle arises in processing this data due to the lack of a clear distinction between first, second, and third neighbouring triangles (Fig. 2, middle part of Fig. 3). This lack of order introduces randomness/arbitrariness into the analysis and compromises the consistency of data processing, which is crucial for the accuracy and reliability of the results.

To address this, we sort the distances to neighbouring triangles in decreasing order. Sorting these values eliminates randomness from the analysis and ensures consistency in data processing, thereby enhancing the correctness and credibility of the results. Subsequently, several classification algorithms available in the scikit-learn library (Pedregosa et al., 2011) can be tested in terms of precision and recall. However, in this study we work with a single algorithm to keep focus on the new classification method. We selected the Support Vector Machine which is considered a suitable tool for binary classification problems in high-dimensional spaces (Bishop, 2006) and which performed well in terms of precision and recall in our preliminary research.

3.3 Visualization

In our study, we visualize the classification results for real data using spatial clustering (Fisher, 1993; Fisher et al., 1985). The labels of triangles are recorded initially as integers corresponding to fault-related triangles (label=1) or triangles belonging to the homocline (label=-1). Then, the integers are converted to colors and presented on a map.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

3.4 Support Vector Machine

For the purpose of binary classification, we used the support vector machine algorithm, a two class classifier (Bishop, 2006; Vapnik, 2000), available in the scikit-learn library (Pedregosa et al., 2011). The support vector machine algorithm can be considered an optimization algorithm because the decision is based on a hyperplane with the maximum margin. The margin is defined to be the minimal distance between a point in the training set and the hyperplane. The motivation behind the concept of margin is that if a margin is large, then it will be capable of separating the training set even after small perturbation of the instances (Shalev-Shwartz and Ben-David, 2013). Formally, the optimization objective looks as follows (Bishop, 2006):

$$arg \max_{w,b}\{ \min_{n} \left[t_n \left(\frac{w^T f(x_n)+b}{||w||} \right) \right] \}, \qquad \text{(Eq. 4)}$$

where $t_n \in \{-1,1\}$ are target values, $f(x)$ denotes a fixed feature-space transformation. This transformation is expected to facilitate separation of instances which were not linearly separable in the original space. Common choices of transformations (kernel functions) include: linear, polynomial and radial basis functions. Next, w is the vector of weights which determines the orientation of the decision surface, and b is the bias parameter (not to be confused with bias in the statistical sense). The expression $\left(\frac{w^T f(x_n)+b}{||w||} \right)$ denotes the perpendicular distance of a point x_n to the decision surface $y(x) = w^T f(x) + b = 0$. This decision surface separates points with different labels: -1 and 1. The multiplication $t_n y(x_n)$ visible in the optimization task filters solutions for which all data points are correctly classified, i.e. $t_n y(x_n) > 0$. We note that in some formulations of the optimization problem, the vector of weights has unit length (Shalev-Shwartz and Ben-David, 2013). Because all N points lie beyond the margin area, they are at some distance from the hyperplane corresponding to the size of the margin. While the distances of N points relative to the decision boundary can be different, they are all greater than a fixed number corresponding to the size of the margin which can be expressed by a set of N inequalities. Therefore, the optimization objective (Eq. 4) together with the set of N inequalities form a constrained optimization problem which can be solved be using the Lagrange multipliers (Bishop, 2006 - Appendix E).

In the presence of outliers, a soft margin classifier can be applied which allows some samples to be classified incorrectly (Shalev-Shwartz and Ben-David, 2013). For the radial basis function kernel, C and gamma parameters are considered. The parameter C, common to all SVM kernels, is a penalty parameter: a low C tends to make the decision surface simple (thus, avoiding overfitting but possibly affecting the correct classification of the training data), while setting a high C will result in classifying training examples more correctly (possibly leading to poorer generalizability). Gamma defines the radius of the similarity of a single training sample. The lower gamma is, the greater the similarity radius of a sample (Pedregosa et al., 2011).

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

We use the following metrics as evaluation metrics: $precision = \frac{true\ positive}{true\ positive + false\ positive}$ and $recall = \frac{true\ positive}{true\ positive + false\ negatives}$. The definition implies that precision is maximized if there are no false positives and the recall is maximized when there are no false negatives. Based on these definitions the harmonic mean of both can be defined as follows $F_1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} = 2 * \frac{precision * recall}{precision + recall}$.

4 Results

4.1 Synthetic data

For arbitrary selected hyperparameters (C=0.05, gamma=0.042), with radial basis function as kernel function) in the scikit-learn framework we achieved precision and recall for the fault-related observations to be 0.92 and 0.95, respectively (Tab. 1). These results relate to the test data which were not seen by the algorithm during training. To further increase the values of the classification metrics, we tested many combinations of the hyperparameters as a part of the grid search optimization (Pedregosa et al., 2011). The optimal combination of hyperparameters turned out to be as follows: C=100, gamma=0.01 with radial basis function as the kernel function. The classification results change slightly after the grid optimization stage (Tab. 2). We note that after every execution of the code the values of the hyperparameters can change due to random steps of the procedure.

Tab. 1 Results for the classification of test data (unseen terrains) for arbitrarily selected hyperparameters

Class	Precision	Recall	F1-score
Non-fault	0.95	0.92	0.94
Fault	0.92	0.95	0.93

Tab. 2 Results for the classification of test data (unseen terrains) after fine-tuning of the hyperparameters during grid search optimization

Class	Precision	Recall	F1-score
Non-fault	0.96	0.93	0.95
Fault	0.93	0.96	0.94

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

4.2 Real data

To investigate generalizability of the method for real data, we used borehole data (Michalak, 2024a) corresponding to a horizon separating Middle-Jurassic rock units: Kościeliska sandstones from ore-bearing clay deposits (Matyja and Wierzbowski, 2000; Kopik, 1998). The results of supervised classification using SVM are similar to those obtained using unsupervised classification (compare Figs 4 and 6) in that the majority of faults has the SW-NE, SSW-NNE or S-N orientation. However, there are significant differences which relate to visibility of new potential faults trending perpendicular to the preferred dip direction. For example, Fig. 6b in the central part (near coordinates 921500, 251000) shows two potential faults trending NW-SE at the termination of S-N and SSW-NNE trending faults. Another difference is that the unsupervised classification presented the major fault in the NW part of the study area as possibly composed of smaller faults with opposite dip direction (Fig. 4c, near coordinates 922000, 248500). In contrast, the binary classification cannot distinguish between faults with opposite dip directions. Therefore, the zone of fault-related labels near the discussed fault zone appears relatively wide.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

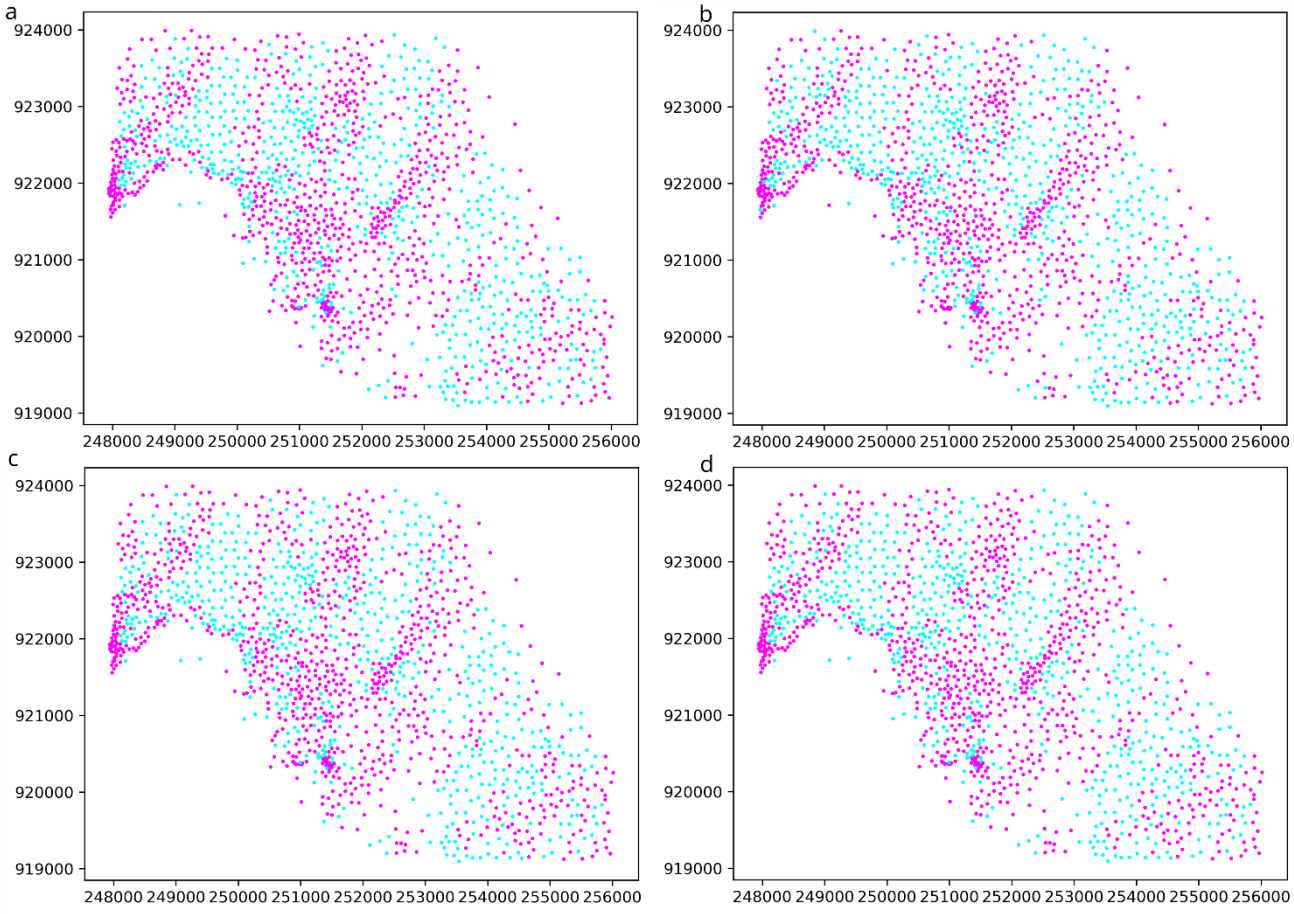


Figure 6 Classification results for the Kraków-Silesian Homocline: (a) the optimal combination of the hyperparameters as confirmed by the grid search optimization (C=100, gamma=0.01, with radial basis function as the kernel function) (b) a custom combination of the hyperparameters (C=1, with linear kernel) (c) a custom combination of the hyperparameters (C=10, with linear kernel) (d) a custom combination of the hyperparameters (C=10, gamma=0.01 with radial basis function as the kernel function)

5 Discussion

In our study, we used local geometric attributes of triangles and neighborhood analysis to predict faults on geological terrains. In subsurface geological modelling, the neighborhood analysis was already applied for individual boreholes of triangulated surfaces to analyze connectivity of strata (Guo et al., 2024). From a viewpoint of graph theory, in our case the neighborhood analysis is performed on finite faces of the triangulation rather than on its finite vertices (boreholes). Because, for every triangulation, with k being the number of points on the edge of the convex hull, the relationship between vertices (n) and triangles (m) is $m = 2n - 2 - k$ (De Berg et al., 2008), our approach will usually (except very small data sets) result in a

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

greater number of observations compared to a potential approach of considering boreholes as observations. Moreover, our approach ensures that every observation has three finite neighbors which testifies that observations are comparable. When neighbors of points are considered, this is not the case because the degree of a vertex usually is not a constant number.

The main assumption for generating the terrain data (Sect. 3.1) is that a fault is always represented by a plane. In our case, we assume that any terrain point lies either on one or on the other side of the fault and it is not possible that any terrain point is located on the fault. When these terrain points are triangulated, each fault-related triangle connects points from both sides of the fault. Every fault-related triangle has at least one neighbour that is not associated with the fault. The fault is represented by a stripe of the fault-related triangles (see triangles with green markers in Fig. 7).

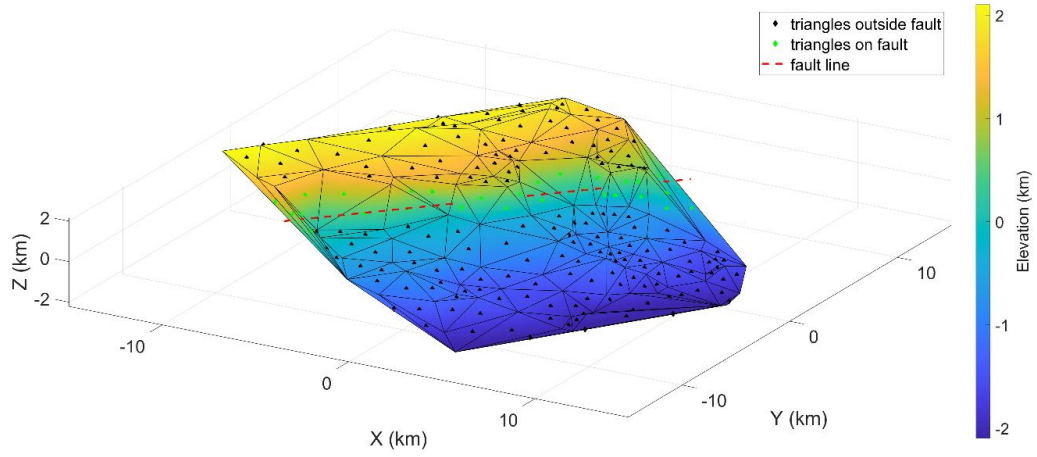


Figure 7: Example for a terrain generated as explained in Sect 3.1 with the fault being represented by a plane.

This main assumption is a simplification. In reality, a fault structure is never a purely planar object. It is mostly a 3D structure that has an extension perpendicular to the fault line direction (Childs et al., 2009). Therefore, it is possible that at least some terrain points may be located inside the fault zone. When triangulating these points, there is the possibility to create fault-related triangles all of whose neighbours are fault-related triangles as well. (see Fig. 8, triangles with red markers).

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

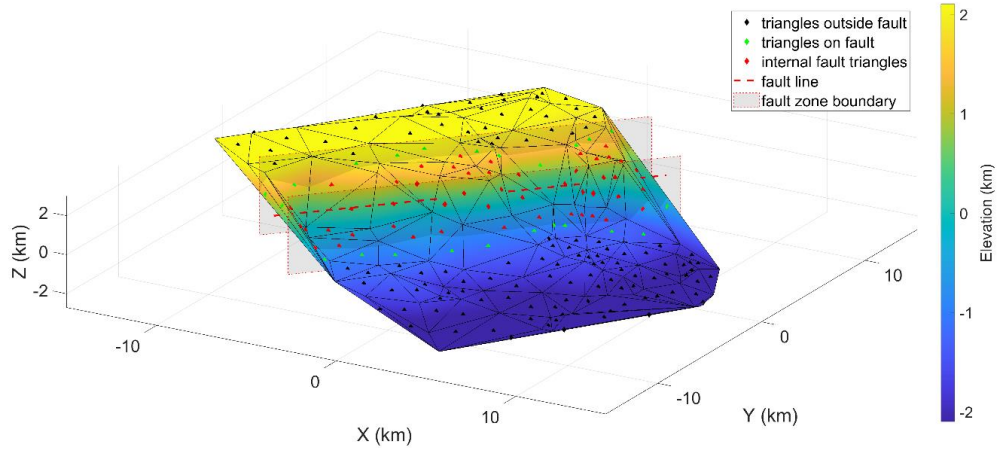


Figure 8: Example for a terrain generated as explained in Sect 3.1 with extended fault zone. Triangles with green markers show at least one non fault-related triangle as a neighbour. Triangles with red markers have only fault-related triangles as neighbours. Grey planes represent the border of the extended fault zone.

These "internal fault-related" triangles show a combination of features that was never trained. Triangles that show features that were actually trained are still present (see Fig. 8, triangles with green markers) but are only located at the borders of the fault zone (grey rectangles in Fig. 8). This leads to the assumption that the presented classification system may classify the extended fault zone not by one sequence of color-coded labels but by two quasi-parallel sequences of this type. Whether the "internal fault-related triangles" can be successfully classified is not clear. In terms of local geometric variables specific to a single triangle such as coordinates of normal vectors, the "internal fault-related" triangles are more similar to the classical fault-related triangles. In contrast, the neighbourhood analysis alone would likely classify these triangles as non fault-related triangles (triangles with black marker in Figs. 7 and 8).

The influence of 3D fault zones for the classification result needs to be further studied. Nevertheless, in the context of this study the data points are assumed to be sparsely scattered. If the fault shows an extension significantly lower than the mean data point spacing, the simplified assumption of purely planar faults is valid. The problem of identifying internal fault triangles could possibly be ameliorated by including not only the direct neighbours of a triangle in the training procedure but also triangles that are 2nd-, 3rd-degree neighbours or even higher degrees. Using this approach, it is more likely that the extended neighbourhood of an "internal fault-related" triangle also contains non-fault related triangles. However, the concept of the suggested approach would stay the same. For illustrational purposes we, therefore, stick to the simplest setup.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

6 Conclusions

The proposed supervised method has the potential to identify fault-related lineaments of any orientation. This is a significant improvement over the clustering-based method whose results depend on the partition generated by clustering algorithms. The latter often struggle to separate regional trend from faults striking perpendicular to the regional trend on homoclines. Compared to the unsupervised method (Fig. 4c, the NW part of the study area), the main drawback of the supervised approach is that the algorithm cannot distinguish between different dip directions of a fault. Therefore, a zone of fault-related labels may consist of many sub-parallel sequences of labels corresponding to more than one sub-parallel faults possibly with opposite dip direction (Fig. 6, the NW part of the study area). The main challenge of the workflow is to eliminate arbitrariness in variable selection in relation to neighbourhood analysis. Sorting distances among neighbours eliminates arbitrariness from the analysis but it is also the most computationally intensive part of the workflow. Further studies can focus on considering more complex geological scenarios including the influence of 3D fault zones and physics-based models (compare with Conclusions in Reichstein et al., 2019).

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Code availability

Name of code: BrokenTerrains. **License**: GNU General Public License v3.0. **Developer**: Michał Michalak. **Contact address**: AGH University of Krakow, Poland. E-mail: michalm@agh.edu.pl . **Year first available**: 2024. **Hardware required**: The computer code was run on a laptop with Intel(R) Core ™ i7-7500U CPU 2.70 GHz, 16 GB RAM. **Software required**: CGAL library (v. 4.8), Microsoft Visual Studio 2022. **Program language**: C++, Python. **Program size**: 738 KB. **How to access the source code**: (Michalak, 2024b) https://doi.org/10.5281/zenodo.12375568 Setup guide: https://github.com/michalmichalak997/BrokenTerrains/blob/main/README.md

Data availability

Datasets for this research (input and processed data) are available in these intext data citation references (Michalak, 2024a)

Author contribution

MM devised the project, wrote the computer code and the manuscript, performed the computations and discussed the results. CG participated in the study conceptualization (sorting distances with neighbours), PM discussed the results.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The study was supported by the AGH University of Science and Technology, grant number 16.16.140.315.

References

An, Y., Guo, J., Ye, Q., Childs, C., Walsh, J., and Dong, R.: Deep convolutional neural network for automatic fault recognition from 3D seismic datasets, Comput. Geosci., 153, 104776, https://doi.org/10.1016/j.cageo.2021.104776, 2021.

Bardziński, W., Lewandowski, J., Więckowski, R., and T, Z.: Szczegółowa Mapa Geologiczna Polski 1: 50 000, ark. Częstochowa (845), 1985.

De Berg, M., Cheong, O., Van Kreveld, M., and Overmars, M.: Computational Geometry: Algorithms and Applications, 3rd

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Ed., Springer, 364 pp., https://doi.org/10.2307/3620533, 2008.

Bi, Z., Wu, X., Li, Z., Chang, D., and Yong, X.: DeepISMNet: Three-dimensional implicit structural modeling with convolutional neural network, Geosci. Model Dev., 15, 6841–6861, https://doi.org/10.5194/gmd-15-6841-2022, 2022.

Bishop, C. M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

CGAL.org: CGAL, Computational Geometry Algorithms Library, https://www.cgal.org, 2023.

Childs, C., Manzocchi, T., Walsh, J. J., Bonson, C. G., Nicol, A., and Schöpfer, M. P. J.: A geometric model of fault zone and fault rock thickness variations, J. Struct. Geol., 31, 117–127, https://doi.org/10.1016/j.jsg.2008.08.009, 2009.

Choi, J., Cho, H., Kwac, J., and Davis, L. S.: Toward sparse coding on cosine distance, in: 2014 22nd International Conference on Pattern Recognition, https://doi.org/10.1109/ICPR.2014.757, 2014.

Cracknell, M. J. and Reading, A. M.: Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, Comput. Geosci., 63, 22–33, https://doi.org/10.1016/j.cageo.2013.10.008, 2014.

Dadlez, R., Narkiewicz, M., Stephenson, R. A., Visser, M. T. M., and van Wees, J. D.: Tectonic evolution of the Mid-Polish Trough: modelling implications and significance for central European geology, Tectonophysics, 252, 179–195, https://doi.org/10.1016/0040-1951(95)00104-2, 1995.

Fisher, N. I.: Statistical analysis of circular data, Cambridge University Press, 277 pp., https://doi.org/10.1017/cbo9780511564345, 1993.

Fisher, N. I., Huntington, J. F., Jacket, D. R., Willcox, M. E., and Creasey, J. W.: Spatial analysis of two-dimensional orientation data., J. Int. Assoc. Math. Geol., 17, 177–194, https://doi.org/https://doi.org/10.1007/BF01033153, 1985.

Guo, J., Xu, X., Wang, L., Wang, X., Wu, L., Jessell, M., Ogarko, V., Liu, Z., and Zheng, Y.: GeoPDNN 1.0: a semi-supervised deep learning neural network using pseudo-labels for three-dimensional shallow strata modelling and uncertainty analysis in urban areas from borehole data, Geosci. Model Dev., 17, 957–973, https://doi.org/10.5194/gmd-17-957-2024, 2024.

Hammah, R. E. and Curran, J. H.: On distance measures for the fuzzy K-means algorithm for joint data, Rock Mech. Rock Eng., 32, 1–27, https://doi.org/10.1007/s006030050041, 1999.

Hermański, S.: Mapa stropu i miąższości warstw kościeliskich. Rejon Żarki-Wieluń. Skala 1:100000, 1993.

Hu, X., Bürgmann, R., Xu, X., Fielding, E., and Liu, Z.: Machine-Learning Characterization of Tectonic, Hydrological and Anthropogenic Sources of Active Ground Deformation in California, J. Geophys. Res. Solid Earth, 126, https://doi.org/10.1029/2021JB022373, 2021.

Jiang, Z., Mallants, D., Gao, L., Munday, T., Mariethoz, G., and Peeters, L.: Sub3DNet1.0: A deep-learning model for regional-scale 3D subsurface structure mapping, Geosci. Model Dev., 14, 3421–3435, https://doi.org/10.5194/gmd-14-3421-2021, 2021.

Kaur, H., Zhang, Q., Witte, P., Liang, L., Wu, L., and Fomel, S.: Deep-learning-based 3D fault detection for carbon capture and storage, Geophysics, 88, IM101--IM112, https://doi.org/10.1190/geo2022-0755.1, 2023.

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Kopik, J.: Lower and Middle Jurassic of the north-eastern margin of the Upper Silesian Coal Basin (in Polish with English summary), Biul. Państwowego Inst. Geol., 378, 67–129, 1998.

Kuhn, S., Cracknell, M. J., and Reading, A. M.: Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia, Geophysics, 83, B183–B193, https://doi.org/10.1190/geo2017-0590.1, 2018.

Marynowski, L., Zatoń, M., Simoneit, B., and Otto, A.: Compositions, sources and depositional environments of organic matter from the Middle Jurassic clays of Poland, Appl. Geochemistry, 22, 2456–2485, https://doi.org/10.1016/j.apgeochem.2007.06.015, 2007.

Mattéo, L., Manighetti, I., Tarabalka, Y., Gaucel, J. M., van den Ende, M., Mercier, A., Tasar, O., Girard, N., Leclerc, F., Giampetro, T., Dominguez, S., and Malavieille, J.: Automatic Fault Mapping in Remote Optical Images and Topographic Data With Deep Learning, J. Geophys. Res. Solid Earth, 126, https://doi.org/10.1029/2020JB021269, 2021.

Matyja, B. A. and Wierzbowski, A.: Ammonites and stratigraphy of the uppermost Bajocian and Lower Bathonian between Częęstochowa and Wieluń Central Poland, Acta Geol. Pol., 50, 191–209, 2000.

Matyszkiewicz, J., Kochman, A., Rzepa, G., Gołębiowska, B., Krajewski, M., Gaidzik, K., and Zaba, J.: Epigenetic silicification of the Upper Oxfordian limestones in the Sokole Hills (Kraków-Częstochowa Upland): Relationship to facies development and tectonics, Acta Geol. Pol., 65, 181–203, https://doi.org/10.1515/agp-2015-0007, 2015.

Michalak, M.: Numerical limitations of the attainment of the orientation of geological planes, Open Geosci., 10, https://doi.org/10.1515/geo-2018-0031, 2018.

Michalak, M.: Broken Terrains v. 1.0: A supervised detection of fault-related lineaments on geological terrains - Input and processed data, https://doi.org/10.5281/zenodo.12209024, 2024a.

Michalak, M., Teper, L., Wellmann, F., Żaba, J., Gaidzik, K., Kostur, M., Maystrenko, Y. P., and Leonowicz, P.: Clustering has a meaning: optimization of angular similarity to detect 3D geometric anomalies in geological terrains, Solid Earth, 13, 1697–1720, https://doi.org/10.5194/se-13-1697-2022, 2022.

Michalak, M. P.: michalmichalak997/BrokenTerrains: v. 1.0 - Initial release, https://doi.org/10.5281/zenodo.12375568, 2024b.

Michalak, M. P., Bardziński, W., Teper, L., and Małolepszy, Z.: Using Delaunay triangulation and cluster analysis to determine the orientation of a sub-horizontal and noise including contact in Kraków-Silesian Homocline, Poland, Comput. Geosci., 133, 104322, https://doi.org/10.1016/j.cageo.2019.104322, 2019.

Michalak, M. P., Kuzak, R., Gładki, P., Kulawik, A., and Ge, Y.: Constraining uncertainty of fault orientation using a combinatorial algorithm, Comput. Geosci., 154, 104777, https://doi.org/10.1016/j.cageo.2021.104777, 2021.

de Oliveira Neto, E. R., Fatah, T. Y. A., Dias, R. M., Freire, A. F. M., and Lupinacci, W. M.: Curvature analysis and its correlation with faults and fractures in presalt carbonates, Santos Basin, Brazil, Mar. Pet. Geol., 158, 106572, https://doi.org/10.1016/j.marpetgeo.2023.106572, 2023.

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine

https://doi.org/10.5194/egusphere-2024-2004
Preprint. Discussion started: 8 July 2024

Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, F.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Shalev-Shwartz, S. and Ben-David, S.: Understanding machine learning: From theory to algorithms, Cambridge university press, 397 pp., https://doi.org/10.1017/CBO9781107298019, 2013.

Słonka, Ł. and Krzywiec, P.: Upper Jurassic carbonate buildups in the Miechów Trough, Southern Poland – insights from seismic data interpretation, Solid Earth Discuss., 11, 1097–1119, https://doi.org/10.5194/se-2019-178, 2019.

Vapnik, V. N.: The nature of statistical learning theory. Statistics for Engineering and Information Science, Springer-Verlag, New York, 2000.

Vega-Ramirez, L. A., Spelz, R. M., Negrete-Aranda, R., Neumann, F., Caress, D. W., Clague, D. A., Paduan, J. B., Contreras, J., and Peña-Dominguez, J. G.: A new method for fault-scarp detection using linear discriminant analysis in high-resolution bathymetry data from the alarcón rise and pescadero basin, Tectonics, 40, e2021TC006925, https://doi.org/10.1029/2021TC006925, 2021.

Wang, H., Zhang, L., Yin, K., Luo, H., and Li, J.: Landslide identification using machine learning, Geosci. Front., 12, 351–364, https://doi.org/10.1016/j.gsf.2020.02.012, 2021.

Wang, Y., Ksienzyk, A. K., Liu, M., and Brönner, M.: Multi-geophysical data integration using cluster analysis: Assisting geological mapping in Trøndelag, Mid-Norway, Geophys. J. Int., 225, 1142–1157, https://doi.org/10.1093/gji/ggaa571, 2020.

Xiong, Y. and Zuo, R.: A positive and unlabeled learning algorithm for mineral prospectivity mapping, Comput. Geosci., 147, 104667, https://doi.org/10.1016/j.cageo.2020.104667, 2021.

Yang, J., Xu, J., Lv, Y., Zhou, C., Zhu, Y., and Cheng, W.: Deep learning-based automated terrain classification using high-resolution DEM data, Int. J. Appl. Earth Obs. Geoinf., 118, 103249, https://doi.org/10.1016/j.jag.2023.103249, 2023.

Zhan, J., Xu, P., Chen, J., Wang, Q., Zhang, W., and Han, X.: Comprehensive characterization and clustering of orientation data: A case study from the Songta dam site, China, Eng. Geol., 225, 3–18, https://doi.org/10.1016/j.enggeo.2017.01.010, 2017.

Znosko, J.: Tektonika obszaru częstochowskiego, Przegląd Geol., 8, 418–424, 1960.