https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

GeoPDNN: A Semisupervised Deep Learning Neural Network Using Pseudolabels for Three-dimensional Urban Geological Modelling and Uncertainty Analysis from Borehole Data

Jiateng Guo[1]*, Xuechuang Xu[1], Xulei Wang[1], Lixin Wu[2], Mark Jessell[3], Vitaliy Ogarko[3], Zhibin Liu[1] and Yufei Zheng[1]

[1]College of Resources and Civil Engineering, Northeastern University, Shenyang, China; 2000985@stu.neu.edu.cn (X. X.); wxldbdx@163.com (X. W.); 207158587@qq.com (Z. L.); 15333134708@163.com (Y. Z.)

[2]School of Geosciences and Info-Physics, Central South University, Lushan Nanlu 932, Yuelu District, Changsha 410012, China; awulixin@263.net (L. W.)

[3]Centre for Exploration Targeting / Mineral Exploration Cooperative Research Centre / DARE Centre, School of Earth Sciences, The University of Western Australia, Perth, Australia; mark.jessell@uwa.edu.au (M. J.); vitaliy.ogarko@uwa.edu.au (V. O.)

*Correspondence to: Jiateng Guo (guojiateng@mail.neu.edu.cn; Tel.: +86-24-8368-7693)

Abstract. Boreholes are one of the main tools for high-precision urban geology exploration and large-scale geological investigations. At present, machine learning based 3D geological modelling methods for borehole data have difficulty building a finer and more complex model and analysing the modelling results with uncertainty. In this paper, a semisupervised learning algorithm using pseudolabels for 3D geological modelling from borehole data is proposed. We establish a 3D geological model using borehole data from a complex real urban local survey area in Shenyang, and the modelling results are compared with implicit surface modelling and traditional machine learning modelling methods. Finally, an uncertainty analysis of the model is made. The results show that the method effectively expands the sample space, the modelling results perform well in terms of spatial morphology and geological semantics, and the proposed modelling method can achieve good modelling results for more complex geological regions.

1. Introduction

Geological spatial distribution has complexity, fuzziness and uncertainty. To reasonably arrange urban engineering construction, the underground situation of each area of a city needs to be understood and a comprehensive assessment carried out. The establishment of a reasonable 3D geological model, intuitive expression of geological features, display of underground geological structures, and revelation of the spatial distribution law are important foundations to ensure engineering design and implementation. The stratum structure is the result of a long geological process, and its spatial and temporal distribution is uneven and irregular. At present, it is still difficult to summarize a set of reasonable mathematical laws to express the stratum distribution. Deep learning methods can obtain the complex mapping relationship between input and output by relying on the powerful computing power of computers, which has been applied in many complex fields and has increasingly attracted the attention of geological researchers, such as 3D modelling.

At present, underground 3D data acquisition methods include borehole exploration technology and applied geophysical technology. Although the cost of borehole exploration is higher than that of geophysics, its exploration precision is high, making it the main means of high-precision exploration in local areas. Borehole data can intuitively and reliably obtain geological spatial information. 3D geological modelling from borehole data can be divided into explicit modelling and implicit modelling (Wang et al., 2018). Explicit modelling methods more easily add geological semantic constraints during modelling, the boundary control is more accurate, and the modelling results are more in line with the actual geological laws. However, it

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

is difficult to automatically model complex geological structures such as faults, folds and unconformities, and the modelling is not smooth. Examples include automatic modelling methods based on generalized tri-prism volume elements (Wu, L. X., 2004), section connection methods (Yang et al., 2011), etc. Implicit modelling (Caumon et al., 2012; Hillier et al., 2014) solves the implicit equations of the space surface by selecting the appropriate basis functions and using known points in the space to obtain the implicit surface functions and then uses a 3D surface construction algorithm to express them explicitly. Because there is a certain relationship between the implicit surface shape and the selected basis function form, an implicit modelling method has a certain degree of subjectivity for the final model expression. In addition, implicit modelling requires a high data volume, which requires a large amount of borehole data to establish an accurate model, and solving large-scale equations also requires high hardware requirements. Examples include the kriging method (Che et al., 2019), inverse distance weighting method (Liu et al., 2020), Hermite radial basis function method (Guo et al., 2021), etc. Stochastic simulation methods include transition probability-based (Carle and Fogg, 1997), object-based (Lantuejoul, 2002), process-based (Lancaster and Bras, 2002), truncated Gaussian (Matheron et al., 1987), multivariate Gaussian (Armstrong et al., 2011), implicit boundaries (Ferrer et al., 2021), and multipoint statistics (Mariethoz and Caers, 2014) simulations. At present, multipoint geostatistics (MPS) (Guo et al., 2022) has been developed as a method for boreholes. By establishing a grid and defining a random simulation path according to the simulation grid, the stratum attribute values are determined for the grid according to the borehole distribution of a random simulation path.

Machine learning methods have been widely used in 3D geological modelling. Traditional machine learning methods include 3D geological modelling based on support vector machines (SVMs) (Smirnoff et al., 2008; Wang et al., 2014), using the kriging model-based potential field method to implicitly model geological structures (Calcagno et al. ,2008; Goncalves et al., 2017), using Bayesian methods to estimate the uncertainty of geological models, etc. (de la Varga et al., 2016; Wang, H., 2020). These methods are applied to nonimage data.

Compared with traditional machine learning methods, deep learning improves the ability to read mined data and is often combined with complex geophysical and geochemical data for modelling. For example, neural networks are trained to predict geological structures from seismic data (Titos et al., 2018), deep neural networks are used to invert complex binary geological media (Laloy et al., 2017) and generative adversarial networks are used to generate geological models (Zhang et al., 2019). Deep learning is used to comprehensively utilize geological, gravity, and aeromagnetic data to intelligently generate regional 3D geological models, which solves the problem of a long 3D modelling cycle and slow effect (Ran et al., 2020). By designing a targeted U-Net convolutional neural network model, the automatic identification and classification of underground ore minerals based on a deep learning algorithm has been realized (Xu and Zhou., 2018). By designing a geological entity recognition model based on a deep belief network, the problem of structured and standardized processing of geological entity information in text data was solved (Zhang et al., 2018). When performing seismic inversion on different data sets, deep learning methods have the potential to obtain higher resolution results than traditional machine learning in the case of big data (Huang et al., 2020).

Borehole data are common data in geological exploration, and the data are generally sparse. Research has been conducted in the field of geology using machine learning methods from borehole data, which can be divided into those based on spatial data from boreholes and other data from boreholes. There are mainly two kinds of modelling ideas based on borehole spatial data: borehole sequence simulation and borehole spatial point simulation. Borehole sequence simulation is divided into borehole sequence prediction and the prediction of each stratum thickness (Zhou et al., 2019). Borehole spatial point simulation simulates the lithology of the spatial points sampled by the borehole. This method is compared with the borehole sequence simulation method. The borehole sequence simulation method has better continuity in the vertical direction and is not affected by the sampling accuracy, while the lithology simulation of a borehole has higher accuracy in predicting the borehole lithology (Zhang et al., 2021). Models based on borehole spatial point simulation have different advantages and disadvantages due to the different input and sampling methods of the model. The borehole is upsampled according to a certain interval, and each

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

sampling point is used as input (Guo et al., 2019). Strata are generated by randomly selecting B-spline curve functions based on boreholes, and the voxels of each stratum are used as input (Wang et al 2021). The constructed model is more accurate, but the model mainly relies on randomly selected B-spline curves. If the coordinates and starting depth of each stratum drilled are used as input (Kim et al., 2022), although the model accuracy is lower than that of upsampling, it is not easy to overthrow the order of strata. Studies based on other borehole data include lithology classification based on borehole core description data (Bressan et al., 2020) and 3D geological modelling based on described boreholes (Fuentes et al., 2020). In conclusion, among borehole data modelling methods, the lithology prediction method for spatial points is better, but there are still some problems, such as needing 3D geological models established by other methods as references, low modelling accuracy, and difficulty in modelling complex geological phenomena.

In this paper, we propose a semisupervised deep learning algorithm using pseudolabels from borehole data for urban engineering 3D geological modelling. Then, the trained model is used to predict the unlabelled grids, and the pseudolabel data with high confidence are added to the unlabelled grids to expand the sample data space. Finally, a final model is obtained by training the labelled data and the pseudolabel data. This method only uses borehole data and can establish a more accurate and complex 3D geological model. We establish a 3D geological model for a complex real geological project, compare it with the implicit HRBF method and SVM method, and analyse the uncertainty of the model.

2. 3D Modelling Method Based on Deep Learning

2.1. Borehole data preprocessing

In deep learning, the problem of classifying borehole data can be further reduced to a problem of classifying strata. We can take the coordinates of a borehole and the borehole depth as the input vector and the stratum attribute of the borehole as the output vector. For 3D geological modelling, the model at the borehole should be as consistent as possible with the stratum information revealed by the current borehole. The original borehole data include the borehole coordinates X, Y, borehole elevation, stratum thickness, stratum bottom depth, borehole label, borehole stratum label, etc. To increase the amount of data, the borehole data are upsampled. Since the thicknesses of the strata may differ greatly from each other, the data balance will be affected if an equal interval sampling method is used, and the data amount of a thick stratum will be much greater than that of a thin stratum. As a result, the thick strata will dominate the training network, resulting in the classification of unknown regions. The unknown area will be easier to predict as the stratum thickness attribute. The greater the difference in stratum thickness is, the more misclassification will occur.

Based on the above discussion, an unequal interval sampling method is adopted in this paper. Compared with equal interval sampling, unequal interval sampling changes the sampling interval according to the thickness of each stratum to ensure sampling data balance. At the same time, in the interior of each stratum, equal interval sampling is maintained, and the critical point attributes are preserved. Otherwise, the thinner strata may be difficult to predict or be considered as outliers due to too little sampling. The formula for unequal interval sampling can be expressed as follows:

$$Z_{ij} = \frac{(S_{ij} - S_{ij-1})}{n} \quad (1)$$

where Sij is the bottom depth of the jth stratum in the ith borehole, Hij is the thickness of the jth stratum in the ith borehole, n is the number of samples in each stratum, and Zij is the sampling interval of the jth stratum in the ith borehole.

In the borehole stratigraphic data (Fig. 1), different colours indicate different stratigraphic attributes, and the stratigraphic data are displayed in strips, distributed continuously in the vertical direction, with continuous and unique stratigraphic

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

attributes for depth intervals of individual strata and no data gaps between strata. The unequally spaced sampling on the deterministic section is same as the unequally spaced sampling on the borehole, and unequally spaced sampling is also performed in the horizontal direction.

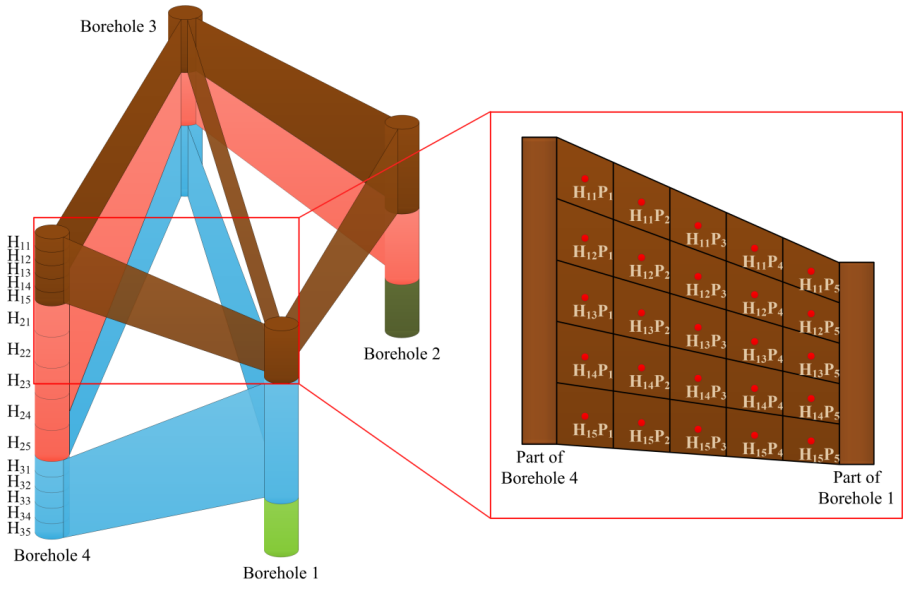


Figure 1. Resampling of borehole data. Upsampling on the boreholes (left); upsampling on the deterministic sections (right).

Borehole data play a direct or indirect role in the generation of the model, and some geological semantic information with high reliability in geology can be obtained through drilling. The borehole data points are inserted point by point according to the Delaunay rule to generate a surface triangular irregular network (TIN), and the basic topological relationship between boreholes is established. Each triangulation consists of three boreholes, and pairs of boreholes with the same attributes are connected to form a deterministic section. At the same time, long and narrow triangles in the TIN are removed to avoid the connection between the long and narrow triangles that are far away and not related to each other. In this way, the distance between boreholes generated by the Delaunay rule is small. This GTP-like section connection method can maintain the internal connection between the three boreholes and can simulate a variety of complex geological phenomena. At the same time, the modelling scope of this study is mainly for a quaternary sedimentary surface, the possibility of a large stratum inversion phenomenon is low, and the strata are deposited in chronological order. After connecting the deterministic sections, they are sampled at unequal intervals in the both horizontal and vertical directions so that the sampling density is consistent with the borehole to avoid oversampling affecting the training of the network.

In the borehole data, the order of magnitude between the coordinates and the depth of each stratum is large. This will affect the results of model training, and to eliminate the dimensional influence between the indicators, data normalization is needed to solve the comparability between the data indicators. After the original data are standardized, each index is on the

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023


Geoscientific Model Development Discussions

same order of magnitude, which is suitable for comprehensive comparative evaluation. To ensure convergence, the data need to be normalized by mapping the resulting values to [0-1]. For any data x, the mapping function is as follows:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

where xmax is the maximum value of the sample data and xmin is the minimum value of the sample data. x' is the normalized result, and x is the input of the model data. Through the normalization method, the convergence speed of the network training model will be improved, the training accuracy will also be improved, and the model will be easier to train.

2.2. Construction of deep neural networks

A single-layer perceptron is one of the simplest feedforward neural networks (Huang et al., 2012), which can be used to simulate partial logic functions and solve linearly separable problems. It cannot classify data sets that are not linearly separable. A multilayer perceptron, by adding N hidden layers between the input layer and the output layer, enhances the model's ability to solve a problem. Multilayer perceptrons have strong robustness, memory ability, and nonlinear fitting ability, can map complex nonlinear relationships, can deal with a large number of data samples, and have simple learning rules that are easy to implement using computers.

A deep neural network uses the input index and output index to form rules and provides the result closest to the expected output value from the input value, which is a multilayer feedforward neural network according to the error backpropagation algorithm. In a deep neural network, the unit output of the first hidden layer is first calculated, and then the output of the unit in the next layer is used to continue to calculate the output of the unit in the next layer until the output layer outputs the result; this process is called forwards propagation. There is a weight between a deep neural network layer and each layer unit, the initial value of the weight is preset, and the weight of the multilayer perceptron can be trained using the back propagation algorithm. The data in the data set are output after the multilayer perceptron, and the output is compared with the expected value to obtain the corresponding error. The error is backpropagated layer by layer, and the weight of each layer is adjusted accordingly. After a number of adjustments, with the result is a weight that fits the model. The relationship between layers can be expressed as follows:

$$Y_j = \sum_{i=1}^{n} W_{ij} X_i + b \quad (3)$$

where Yj is the input of the next layer, Wij is the connection weight from cell X_i of the previous layer to cell Yj of the next layer, and b denotes the offset value.

In the network model (Fig. 2), the coordinate data x, y, z of each upsampled spatial point in the prediction area are taken as the input, and the stratum attribute of the spatial point is the output. Each input represents a dimensional spatial feature, and after four fully connected layers, the result of the dimension expansion is obtained by multiplying the weight matrix. It is considered that the result represents the deep characteristics of the sample, and samples of different categories should have different high-dimensional features. Through a fully connected layer and softmax layer, the output value of the category is

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

normalized to the probability of each class after an exponential function change, and the sum of each class is 1. Finally, the predicted results of each data point are integrated to form the entire 3D geological model.

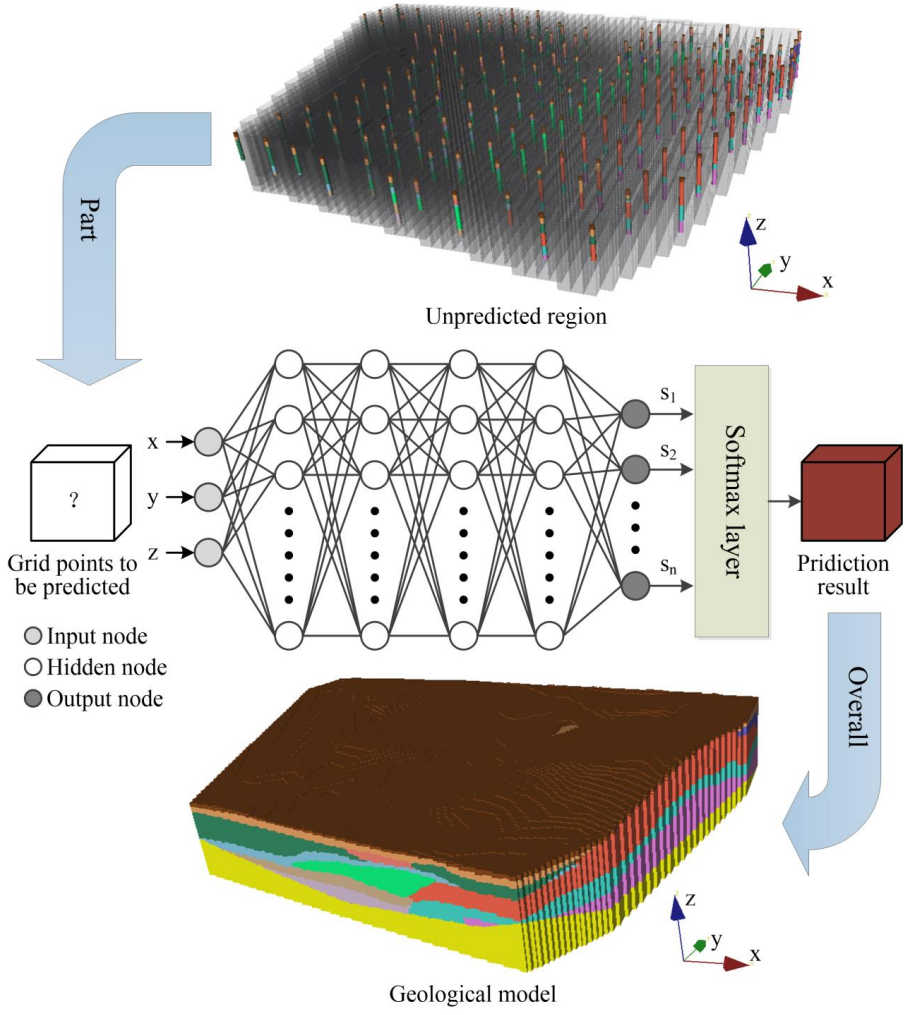


Figure 2. Architecture of a deep neural network. Light grey nodes are input features, dark grey nodes are target outputs, and white nodes are internal network nodes.

2.3 Semisupervised deep learning algorithm using pseudolabels

Compared with images, point cloud data, etc., borehole data tend to be dispersed. Therefore, borehole data can be approximately regarded as a large number of missing point data between points, which makes it difficult to accurately express the variation characteristics, such as the inclination angle of the entire stratum interface. Deep learning requires a large amount of labelled data to improve model performance. Only by upsampling the borehole point data and deterministic borehole section data, for the spatial raster points with high modelling accuracy, the data amount is very small and contains very limited features. To effectively solve the labelling problem, semisupervised learning and deep learning are combined, and a small amount of labelled data and a large amount of unlabelled grids are used to build a model, which is conducive to expanding the sample space and making up for the lack of geological semantic information provided by single borehole data.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

In the geological field, there is no specific mathematical law for the attribute of strata. The borehole vertex data points from the Delaunay rule of the generated surface are irregular triangle net topological relationships of three boreholes, according to the borehole stratum attribute, they are connected into triangular prisms ， then it is considered that the data points in this range have the highest confidence in the stratum attributes determined by the triangular prisms. On this basis, a semisupervised method using pseudolabels is used to enhance learning by generating pseudolabels for unlabelled grids. First, the model is trained using the labelled data, and when the model reaches a high accuracy after a certain number of training rounds, the trained model is used to predict the unlabelled grids, and the prediction result with higher confidence is selected as the pseudolabel. The pseudolabel data and label data are combined for training. After a certain round of training, the above process is repeated until the new pseudolabel data in each round are less than a certain proportion. At this point, it is considered that most of the grids with high confidence have been labelled, and the model has been trained on the data after data augmentation.

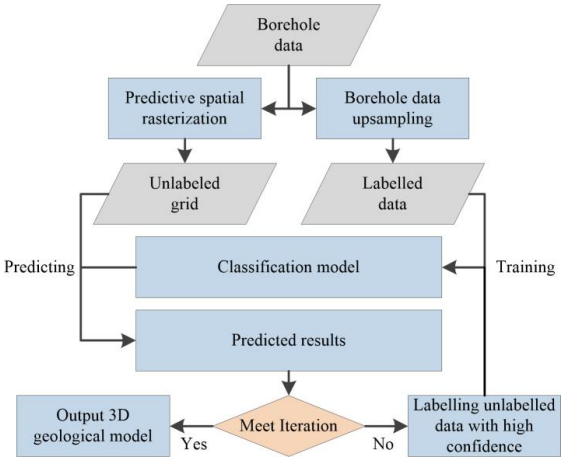


Figure 3. Algorithm flow chart.

2.4. Analysis of model uncertainty

The last layer of the neural network classifier normalizes the probability of the output through the softmax layer, and the softmax normalized result can be approximated as the probability corresponding to each stratum at that data point. Therefore, when analysing the uncertainty of each data point of the raster model, the normalized information entropy can be introduced to quantitatively evaluate the uncertainty of the geological model. The normalized information entropy formula is as follows:

$$H(X) = -\frac{\sum_{x \in S} p(x) \ln(p(x))}{S_{max}} \quad (4)$$

where S is the number of possible geological attributes for each data point, Smax is equal to ln(n), and n is the number of possible geological attributes. The information entropy of each data point is obtained by calculating the probability p(x) of each data point over all geological attributes. The magnitude of information entropy reflects the degree of complexity at a certain location in the geological model. The closer the information entropy is to 0, the higher the certainty of a data point for

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023


Geoscientific Model Development Discussions

a certain stratum attribute, and the closer the information entropy is to 1, the higher the uncertainty of a data point for multiple geological attributes.

In addition, the data can be analysed based on an estimated confusion index (Burrough et al., 1997), and the ambiguity of classification can be evaluated by selecting the results of the two prediction categories with the highest probability for each data point. The confusion index formula is as follows:

$$CI = [1 - (\mu_{max} - \mu_{max-1})] \quad (5)$$

where μmax is the class with the highest predicted probability and μmax-1 is the class with the second highest predicted probability. CI values range from 0-1 to indicate the degree of confusion predicted by the data point, with 0 indicating that the classification result with a low confusion index is not ambiguous and 1 indicating that the classification result with a high confusion index is highly ambiguous.

3. Experimental method and verification

To further illustrate the applicability of the proposed method, this chapter uses a practical geological case to conduct 3D geological modelling and analysis with the proposed method. To verify the rationality of the model, the neural network model is compared with a mature implicit modelling method (HRBF). To illustrate the improvement of the modelling effect of the proposed method compared with the traditional 3D modelling method based on machine learning and the relative reliability of the modelling method in geological semantics, the same section of the 3D geological model established using the proposed method and the SVM method is compared. The proposed algorithm is implemented based on the PyTorch open source machine learning library. The SVM algorithm uses the RBF convolution kernel, the parameters are determined by grid search, and the SVM method in the ThunderSvm library is used for training (Wen et al., 2018). The model established using the algorithm mentioned in the experiment is visualized with the developed visualization platform. All test experiments in this chapter are performed on the same device with the following parameters: Intel(R) Core(TM) i7-10750H CPU @2.60 GHz, NVIDIA GeForce RTX 2060, 16.0 GB RAM, Windows 10 (64-bit).

3.1. Overview of the study area

The study area is located in a region of Shenyang District, Liaoning Province, China, which is located in the middle of the Liaohe Plain, Liaoning Province. The region is mainly plains, mountains, and hills concentrated in the northeast, and the terrain slopes gradually from northeast to southwest. There are four large rivers running through it. This region has a temperate monsoon continental climate and four distinct seasons. The average annual temperature is 8°C, the average precipitation is 628 mm, increasing from north to south, and the precipitation is concentrated in summer.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

3.2. Modelling results and accuracy verification

The area includes data from 167 boreholes distributed over an area of 305 m×264 m, with adjacent boreholes spaced approximately 23 m apart, an average depth of 29.5 m and a minimum thickness of 0.4 m revealed by the boreholes. The ReLU function is used as the activation function in the neural network, the initial learning rate is set to 0.001, and the batch size for training is set to 512. When the model training accuracy reaches 90% and 500 epochs, the unlabelled grids are labelled with a pseudolabel every 100 epochs. When the newly added pseudolabel data are less than 10% of the unlabelled grids evaluated in an epoch, the model continues to train for a total of 2000 epochs before stopping.

The training accuracy and losses in the method process are shown in Fig. 4. In the training process, when the labelled data and pseudolabel data are fused, the boundary demarcation of stratigraphic categories is more finely delineated, the final model training accuracy is above 95%, the loss function is poor, and the precision of the model on the test set is 98.16%. A confusion matrix is obtained from the test set (Fig. 5), which reflects the evaluation result reliability of the model. The classification accuracy of the model is high for all layers. Some strata are more likely to be confused due to thin strata, similar boundaries with other strata, or more geological phenomena, such as depositional termination. The receiver operating characteristic curve (ROC) is another performance indicator that summarizes the performance of the binary classification model in the positive class and thus evaluates the diagnostic ability of the classifier according to the threshold change (Fawcett, 2006). The area under the ROC curve (AUC) (Fig. 6) represents a comprehensive measure of all possible classification thresholds. AUC values greater than 90, 75-90%, 50-75% and less than 50% are considered to represent excellent, good, poor and unacceptable performance, respectively (Ray et al., 2010). The AUC values of the model are all above 90%, indicating that the classification performance of the model is excellent.

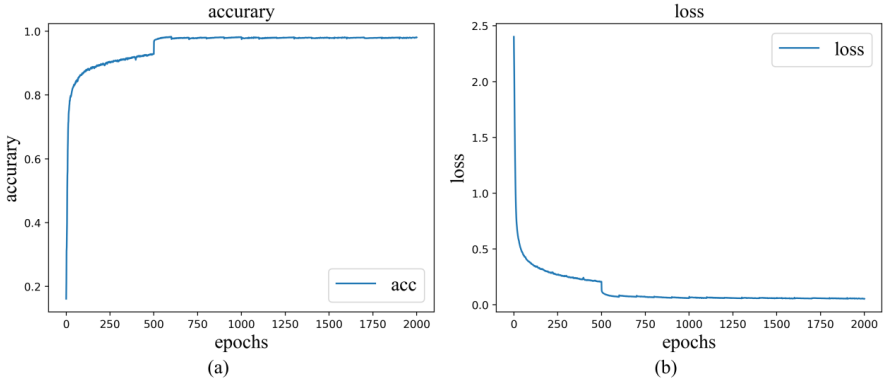
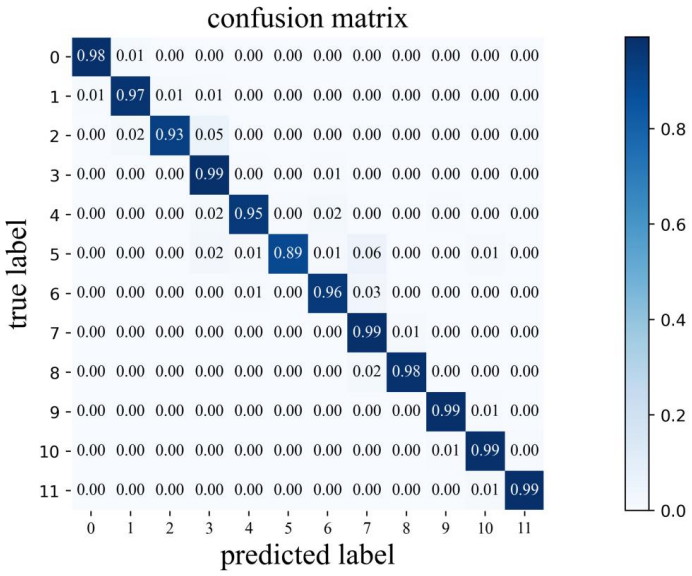


Figure 4. Model training accuracy and loss variation curve.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023



255

Figure 5. Confusion matrix of classification results when the model is applied to the test data set.

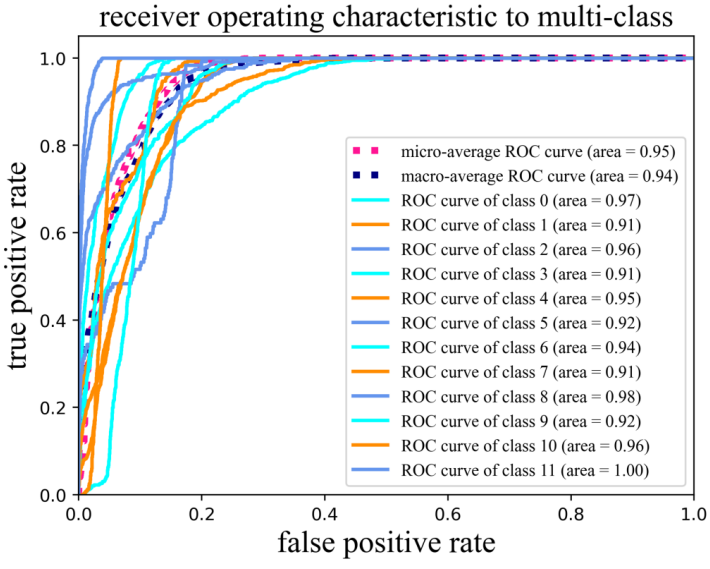


Figure 6. ROC curve for classification.

The grid accuracy used in modelling is 1.5 m×1.3 m×0.3 m. The model uses the Tin mesh constructed from the top of boreholes to restrict the surface. The modelling range is determined according to the convex hull established by the boreholes, and the base of the model is determined according to the convex hull established by the bottom of boreholes. Fig. 7 shows the modelling results of the study area. The model reveals the coverage relationship between the strata and reproduces the contact relationship between the depositional termination and unconformity of the strata.

260

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

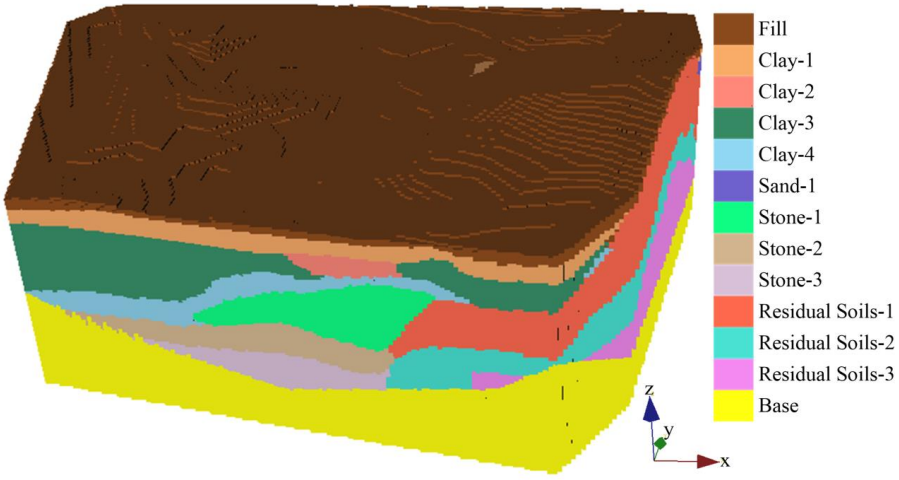


Figure 7. Model built using deep neural networks and model legend.

To test the estimation accuracy of nonborehole positions using the proposed method, the borehole data are divided into a training set and a test set through k-fold cross validation, the training set borehole data is learned, and the test set accuracy is compared and analysed, where K is set to 10.

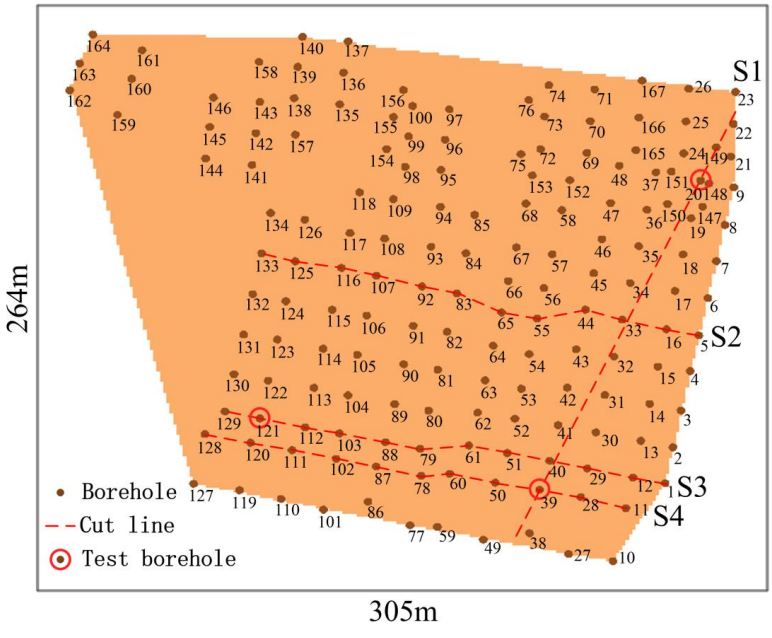


Figure 8. Borehole distribution and experimental analysis of the section line path. The red dotted line is the section route, and the red circled borehole points correspond to the boreholes tested using K1 in the section.

The boreholes in the test set were sampled at equal intervals to determine the data point attributes at the boreholes, and the average accuracy of the k-fold cross validation was calculated to be 71.65%. As there is a certain distance between the

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

boreholes, eliminating an entire borehole will lead to a change in the geological semantic information of the area. When the geological semantic information contained in a borehole is high (Fig. 9), it will be difficult to predict the borehole through the surrounding boreholes, so it is inevitable to obtain poor prediction results when predicting the borehole. Therefore, among the boreholes in the k-fold cross validation, the boreholes that have no more than three depositional terminations between any stratum and the surrounding boreholes and are not at the unconformity boundary are selected for statistical analysis. Among them, the topological relationship of the surrounding boreholes is established using the surface irregular triangulation generated by the Delaunay rule, and the average accuracy is 85.9%.

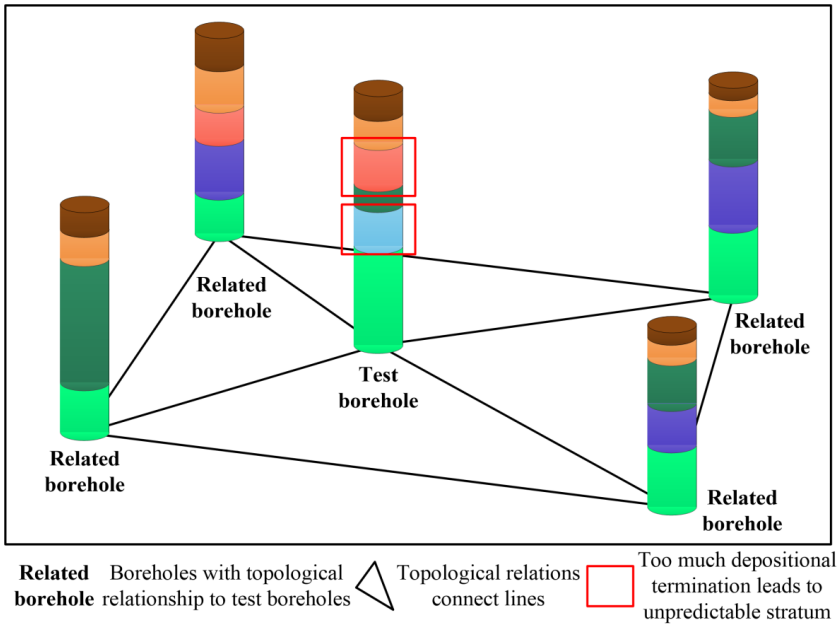


Figure 9. A situation in which too much depositional termination affects the prediction. A related borehole is a borehole that has a topological relationship with the predicted borehole. The red solid line frame is the stratum, which is difficult to predict due to the excessive occurrence of depositional termination.

To further analyse the influence of accuracy on the model, the model with complete borehole data and the model with excluded sample K1 test borehole data were established, and the sections of the model through a test borehole were compared (Fig. 10). The results of straight cutting and cutting along the boreholes are shown. Most areas of the section at the borehole of the test set are consistent with the section established using a complete borehole. Since some test set boreholes are near the depositional termination, there is a certain difference between the model and the test boreholes, but the results are still close to the original model and reasonable. In summary, it can be considered that the modelling method has a strong prediction ability for the neighbouring part of boreholes and can reveal the distribution characteristics of the stratum.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

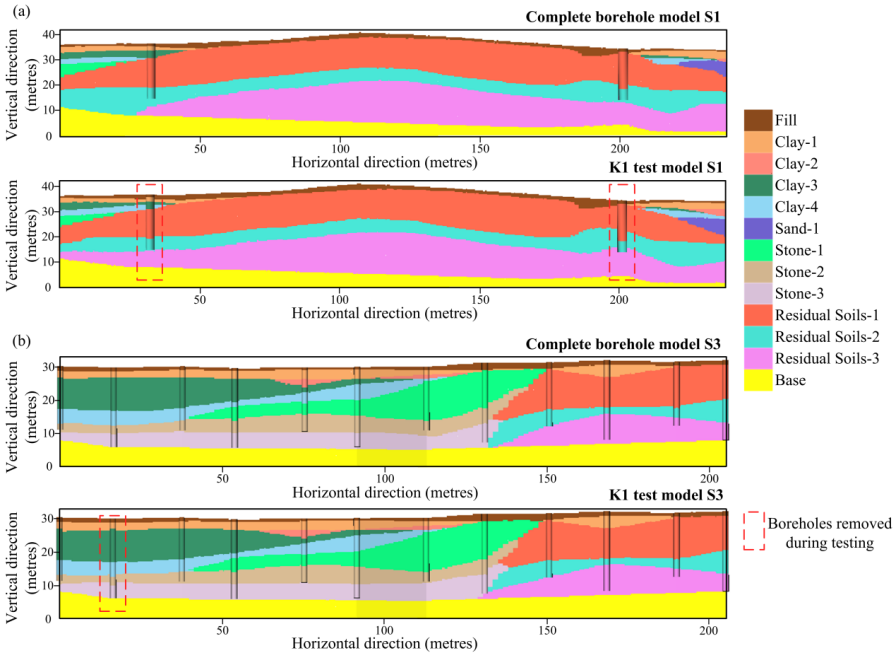


Figure 10. Comparison of the modelling results of sample K1 with the complete drilling results. The dotted box shows the boreholes eliminated during the test.

3.3. Comparative analysis of models

To further verify the rationality of the model, the neural network model is compared with a mature implicit modelling method. The modelling method compared in this study is the implicit Hermite radial basis function (HRBF) 3D geological modelling method. This method uses the implicit Hermite radial basis function to simulate the stratum interface. Since the implicit model is a vector model, the vector model is transformed into a grid model with the same size as the minimum grid cell of the neural network model for comparison. The implicit model has a thicker base than the deep learning model.

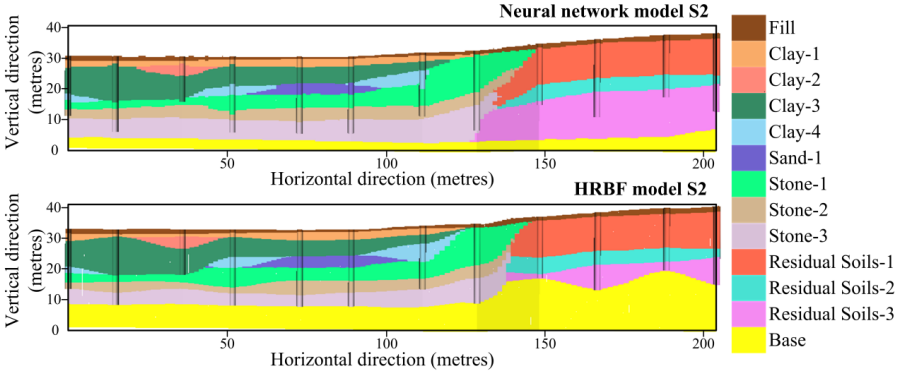


Figure 11. Comparison of the model section between the deep neural network method and the HRBF implicit method.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Geoscientific Model Development Discussions

The deep learning model and the implicit model are visually consistent along the borehole section (Fig. 11) in terms of the thickness and extension angle of the strata. The implicit model constrains the stratum interface through the control points of each borehole and the implicit equation. The deep learning model calculates the labelled and pseudolabelled data loss, trains the neural network through backpropagation to obtain the stratigraphic interface, and predicts the stratum data points in the modelling area. Therefore, when there is a depositional termination or unconformity phenomenon, the deep learning model and the implicit model have certain differences in the depositional termination angle and the thickness change of the stratum. At this time, the shape of the implicit model is mainly determined by the control points determined by the boreholes. However, the machine learning model predicts using upsampled borehole and pseudolabel data with high confidence, which has certain uncertainty.

To illustrate the improvement of the modelling effect of the proposed method compared with the traditional machine learning 3D modelling method and the relative reliability of the modelling method in geological semantics, previous articles have proven (Guo et al., 2019) that the SVM algorithm has the best modelling effect among traditional machine learning 3D geological modelling methods. Therefore, the section of the 3D geological model established using the proposed method and the SVM method is compared along the borehole section. The proposed algorithm is implemented based on the PyTorch open source machine learning library. The SVM algorithm uses the RBF convolution kernel, the parameters are determined by grid search, and the SVM method in the ThunderSvm library is used for training.

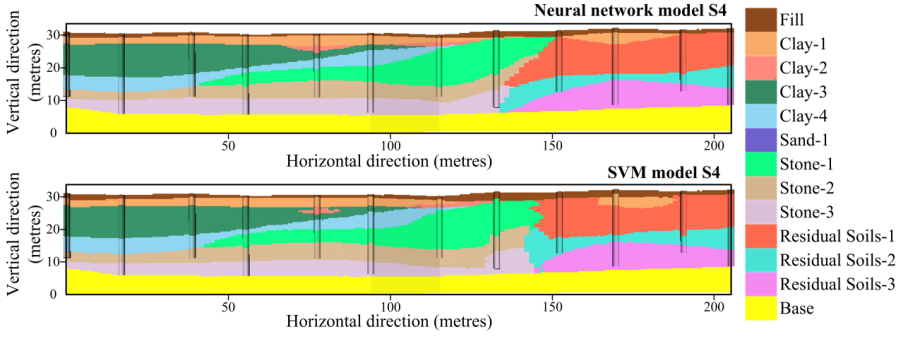


Figure 12. Comparison of the model section between the deep neural network method and the SVM method.

In the study area, the modelling results of the proposed method for complex geological conditions are significantly improved compared with those of the SVM method. By observing the consistency of the attributes of the boreholes and sedimentary strata in Fig. 12, it can be seen that the consistency of the proposed method is higher. In addition, when there is a phenomenon such as depositional termination or unconformity, the variation in thickness and dip angle at the depositional termination or unconformity of the strata modelled using the proposed method is more consistent with the geological semantics. However, the SVM modelling quality decreases significantly when depositional termination or unconformity occurs, and there

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

are many prediction errors or stratum mutation problems. From the model section comparison, it can be concluded that the proposed method has significantly improved model morphology compared with the traditional machine learning method.

3.4. Analysis of model uncertainty

For each data point in the established model, the information entropy is calculated from the normalized probability distribution. Through 3D visualization of the information entropy model with close raster accuracy, the uncertainty of the constructed model can be quantitatively analysed, and the uncertainty of each position in the model can be clearly reflected.

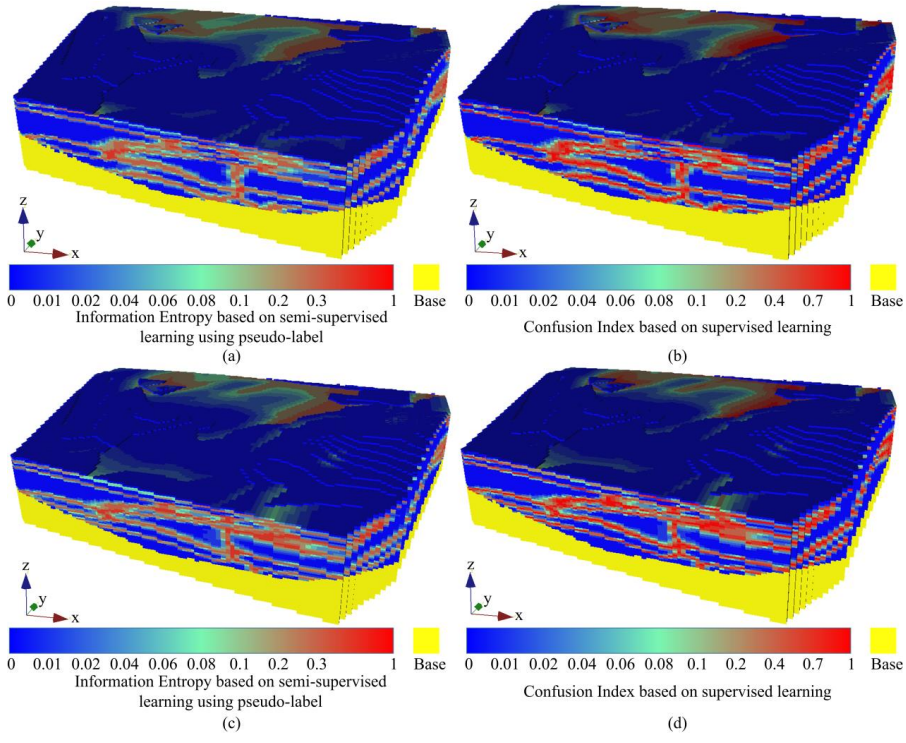


Figure 13. Models of uncertainty: (a) information entropy model based on semisupervised learning using pseudolabels; (b) confusion index model based on semisupervised learning using pseudolabels; (c) information entropy model based on supervised learning; and (d) confusion index model based on supervised learning.

The model (Fig. 13) reflects the uncertainty of the semisupervised learning method using pseudolabels and the supervised learning method to build the model. The blue part of the information entropy model (Fig. 13a, c), where the information entropy is close to 0, means that the uncertainty of the stratum attribute values in the region is low, and the entropy value is small, mainly between the model stratum boundaries. The red part, where the information entropy is close to 1, indicates that the region has a large probability for other stratum attribute values, and the entropy value is large, mainly distributed near the stratum boundary obtained through training. In the confusion index model (Fig. 13b, d), the blue part indicates a low confusion index, and the red part indicates a high confusion index. The overall confusion index of the model is mostly low, and the

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

confusion index increases significantly at the stratum boundary. By comparing the distribution proportions of the two uncertainty models established using the two learning strategies (Fig. 14 and Fig. 15), the model based on the semisupervised learning method using pseudolabels has lower uncertainty than the model based on the supervised learning method, and the semisupervised learning method using pseudolabels can effectively improve the sample space and improve the stability of the model quality.

From the model part, it can be observed that the information entropy and confusion index increase significantly at the boundary of the unconformity or complex depositional termination phenomenon strata, and the stratum boundary has great uncertainty. The uncertainty will increase obviously only when complex geological phenomena such as stratum interface or deposition termination occur, which indicates that the modelling results are stable and the modelling method is reliable.

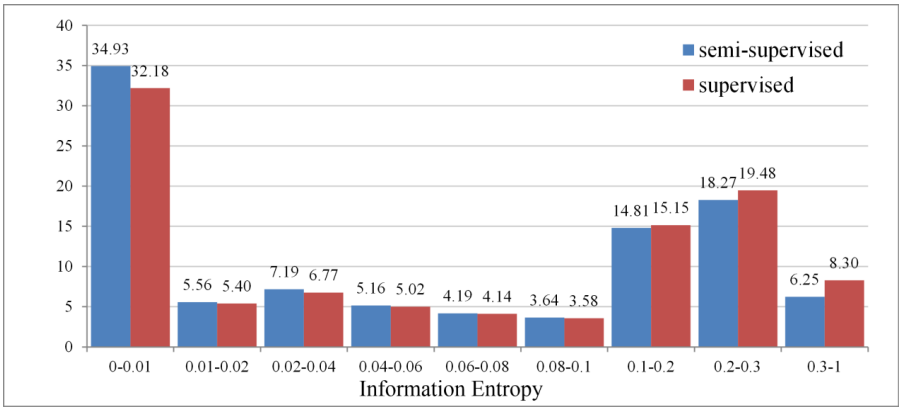


Figure 14. Comparison of the information entropy proportion distribution.

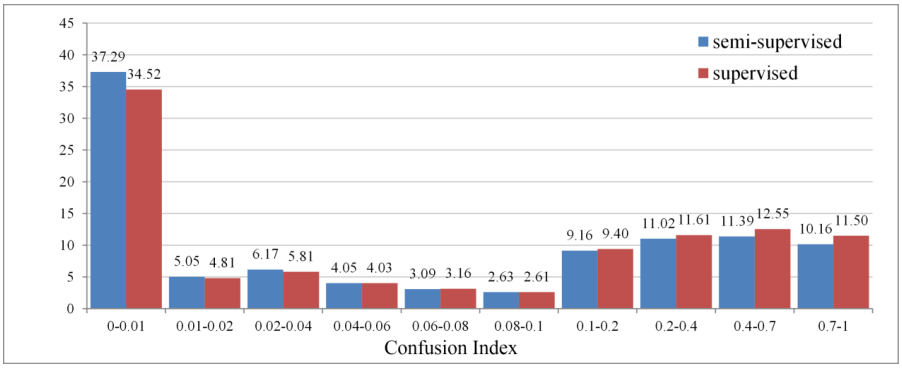


Figure 15. Comparison of the confusion index proportion distribution.

4. Discussion

In this paper, we propose a semisupervised learning algorithm using pseudolabels for 3D geological modelling from borehole data. Because the borehole data sampling density is very sparse relative to the modelling range, it is difficult to obtain a

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

stratigraphic interface with high accuracy through supervised learning. However, the modelling area and modelling accuracy of 3D geological modelling are artificial settings, and the distribution of spatial points that need to be predicted and the distribution of boreholes often lack feature connections, so it is difficult to use unsupervised learning from borehole data. In this paper, the accuracy of the stratigraphic interface obtained through training is improved by adding pseudolabel data with high confidence to the unlabelled grids within the modelling scope. This paper also proves that the modelling method is effective and reliable and can reduce the uncertainty through the 3D geological modelling of the Shenyang complex geological area and the uncertainty analysis, and the modelling results are good and basically in accordance with the geological semantics.

Compared with the MPS method, which builds a grid, defines a random simulation path based on the simulation grid, and determines the stratum attribute values for the grid based on the borehole distribution of a random simulation path, the proposed method trains each stratigraphic interface according to the borehole data and the pseudolabel data predicted between the boreholes and determines the attribute value according to the relationship between the predicted area and the stratigraphic interface. It is not difficult to see from the principle of the method that the MPS method pays more attention to the local borehole distribution, while the machine learning method pays more attention to the macroscopic borehole distribution.

The limitation of the method in this paper is that unequal interval sampling is used, which prevents the problem of severe data imbalance leading to missing stratigraphy, but for thicker boreholes, the interval of borehole sampling increases, which leads to a loss of borehole information to some extent, so how to better reconstruct the borehole data is still a problem worth studying.

Because the surface irregular triangulation network generated using Delaunay's rule is adopted in this paper to establish the topological relationship between three boreholes, the stratigraphic relationship is used to determine the pseudolabel confidence. When the depth of the borehole bottom fluctuates greatly, it is difficult to determine the pseudolabel confidence under a borehole with very shallow fluctuation, which leads to a decrease in local modelling accuracy.

5. Conclusion

In this study, we propose a semisupervised deep learning algorithm using pseudolabels for 3D geological modelling from borehole data. By predicting the pseudolabel for an unlabelled grid within the modelling scope, a 3D geological model is established by expanding the amount of sample data. The proposed method takes the engineering data of Shenyang City as an example to establish a 3D geological model. The accuracy of the deep neural network training model on the test set for sampling data points reaches 98.16%. When the test borehole data without missing geological semantics are predicted through cross-validation, the prediction accuracy of the borehole stratum can reach 85.9%. This shows that the established model conforms to the borehole distribution and has good prediction ability. Compared with the implicit HRBF modelling method and SVM modelling method, the modelling results can express the stratum distribution well, and the modelling results are

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

more accurate than those of the traditional machine learning method. The model uncertainty analysis shows that the pseudolabel method can slightly reduce the uncertainty of the model, which can improve the stability of the 3D geological model and has more advantages in dealing with more complex geological phenomena.

Code availability. The program "GeoPDNN 1.0" was written using Python programming language. The program reads borehole data and preprocesses the borehole data with upsampling and normalization. By using DNN to train and predict the attributes of data points, pseudolabels with high confidence are added to the unlabelled grid points. The code is available for downloading from the following public repository: https://doi.org/10.5281/zenodo.7839508.

Data availability. The model data and terrain data of the case in this paper are available at: https://doi.org/10.5281/zenodo.7535214.

Video supplement. We have provided web links to download the video recordings of our case studies. The real area case verifies the feasibility. The video supplement at: https://drive.google.com/file/d/13VERDXM6YJmP7xMabQy3IjhCExuQSWzk/view?usp=sharing.

Author contributions. Xuechuang Xu and Jiateng Guo conceived the manuscript; Jiateng Guo provided funding support and ideas; Xuechuang Xu was responsible for the research method and program development; Jiateng Guo provided the data used in this research; Xuechuang Xu, Jiateng Guo, Xulei Wang, Lixin Wu, Mark Jessell, Zhibin Liu and Yufei Zheng helped to improve the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could appear to have influenced the work reported in this paper.

Acknowledgements. This work was financially supported by the National Natural Science Foundation of China [grant number: 42172327] and the Fundamental Research Funds for the Central Universities [grant number: N2201022].

References

Armstrong, M., Galli, A., Beucher, H., Loc'h, G., Renard, D., Doligez, B., Eschard, R., and Geffroy, F.: Plurigaussian simulations in geosciences, Springer Science & Business Media, 2011.

Bressan, T.S., de Souza, M.K., Girelli, T.J. and Chemale, F.: Evaluation of machine learning methods for lithology classification using geophysical data, Computers & Geosciences, 139, https://doi.org/10.1016/j.cageo.2020.104475, 2020.

Burrough, P.A., vanGaans, P.F.M. and Hootsmans, R.: Continuous classification in soil survey: Spatial correlation, confusion and boundaries, GEODERMA, 77(2-4): 115-135, https://doi.org/10.1016/S0016-7061(97)00018-9, 1997.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Calcagno, P., Chiles, J.P., Courrioux, G. and Guillen, A.: Geological modelling from field data and geological knowledge Part I. Modelling method coupling 3D potential-field interpolation and geological rules, Physics of the Earth and Planetary Interiors, 171(1-4): 147-157, https://doi.org/10.1016/j.pepi.2008.06.013, 2008.

Carle, S.F. and Fogg, G.E.: Modeling spatial variability with one and multidimensional continuous-lag Markov chains, MATHEMATICAL GEOLOGY, 29(7): 891-918, 1997.

Caumon, G., Gray, G., Antoine, C. and Titeux, M.O.: Three-Dimensional Implicit Stratigraphic Model Building From Remote Sensing Data on Tetrahedral Meshes: Theory and Application to a Regional Model of La Popa Basin, NE Mexico, IEEE Transactions on Geoscience and Remote Sensing, 51(3): 1613-1621, https://doi.org/10.1109/TGRS.2012.2207727, 2012.

Che, D.F. and Jia, Q.R.: Three-Dimensional Geological Modeling of Coal Seams Using Weighted Kriging Method and Multi-Source Data, IEEE Access, 7: 118037-118045, https://doi.org/10.1109/ACCESS.2019.2936811, 2019.

de la Varga, M. and Wellmann, J.F.: Structural geologic modeling as an inference problem: A Bayesian perspective, Interpretation-a Journal of Subsurface Characterization, 4(3): Sm1-Sm16, 2016.

Fawcett, T.: An introduction to ROC analysis, Pattern Recognition Letters, 27(8): 861-874, 2006.

Ferrer, R., Emery, X., maleki tehrani, M. and Navarro, F.: Modeling the Uncertainty in the Layout of Geological Units by Implicit Boundary Simulation Accounting for a Preexisting Interpretive Geological Model, Natural Resources Research, 30: 1-23, https://doi.org/10.1007/s11053-021-09964-9, 2021.

Fuentes, I., Padarian, J., Iwanaga, T. and Vervoort, R.W.: 3D lithological mapping of borehole descriptions using word embeddings, Computers & Geosciences, 141, https://doi.org/10.1016/j.cageo.2020.104516, 2020.

Goncalves, I.G., Kumaira, S. and Guadagnin, F.: A machine learning approach to the potential-field method for implicit modeling of geological structures, Computers & Geosciences, 103: 173-182, https://doi.org/10.1016/j.cageo.2017.03.015, 2017.

Guo, J.T., Liu, Y.H., Han, Y.F. and Wang, X.L.: Implicit 3D Geological Modeling Method for Borehole Data Based on Mechine Learning, Journal of Northeastern University (Natural Science), 40(9): 1337-1342, https://doi.org/10.12068/j.issn.1005-3026.2019.09.021, 2019.

Guo, J.T., Wang, X.L., Wang, J.M., Dai, X.W., Wu, L.X., Li, C.L., Li, F.D., Liu, S.J. and Jessell, M.W.: Three-dimensional geological modeling and spatial analysis from geotechnical borehole data using an implicit surface and marching tetrahedra algorithm, Engineering Geology, 284, https://doi.org/10.1016/j.enggeo.2021.106047, 2021.

Guo, J., Wang, Z., Li, C., Li, F., Jessell, M.W., Wu, L. and Wang, J.: Multiple-Point Geostatistics-Based Three-Dimensional Automatic Geological Modeling and Uncertainty Analysis for Borehole Data, Natural Resources Research, 31(5): 2347-2367, https://doi.org/10.1007/s11053-022-10071-6, 2022.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Hillier, M.J., Schetselaar, E.M., de Kemp, E.A. and Perron, G.: Three-Dimensional Modelling of Geological Surfaces Using Generalized Interpolation with Radial Basis Functions, Mathematical Geosciences, 46(8): 931-953, https://doi.org/10.1007/s11004-014-9540-3, 2014.

Huang, G.B., Zhou, H.M., Ding, X.J. and Zhang, R.: Extreme Learning Machine for Regression and Multiclass Classification, IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 42(2): 513-529, https://doi.org/10.1109/TSMCB.2011.2168604, 2012.

Huang, X.R., Dai, Y., Xu, Y.G. and Tang, J.: Seismic Inversion Experiments Based on Deep Learning Algorithm Using Different Datasets, Journal of Soutwest Petroleum University (Science & Technology Edition), 42(6): 16-25, 2020.

Kim, H.-S. and Ji, Y.: Three-dimensional geotechnical-layer mapping in Seoul using borehole database and deep neural network-based model, Engineering Geology, 297: 106489, 2022.

Laloy, E., Herault, R., Lee, J., Jacques, D. and Linde, N.: Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network, Advances in Water Resources, 110: 387-405, https://doi.org/10.1016/j.advwatres.2017.09.029, 2017.

Lancaster, S.T. and Bras, R.L.: A simple model of river meandering and its comparison to natural channels, HYDROLOGICAL PROCESSES, 16(1): 1-26, https://doi.org/10.1002/hyp.273, 2002.

Lantuéjoul, C.: Geostatistical simulation: models and algorithms (No. 1139), Springer Science & Business Media, https://doi.org/10.1007/978-3-662-04808-5, 2001.

Liu, H., Chen, S.Z., Hou, M.Q. and He, L.: Improved inverse distance weighting method application considering spatial autocorrelation in 3D geological modeling, Earth Science Informatics, 13(3): 619-632, https://doi.org/10.1007/s12145-019-00436-6, 2020.

Mariethoz, G. and Caers, J.: Multiple-point geostatistics: stochastic modeling with training images, John Wiley & Sons, https://doi.org/10.1002/9781118662953, 2014.

Matheron, G., Beucher, H., de Fouquet, C., Galli, A., Guerillot, D. and Ravenne, C.: Conditional Simulation of the Geometry of Fluvio-Deltaic Reservoirs, Soc. Petrol. Eng. (SPE), 16753, https://doi.org/10.2118/16753-MS, 1987.

Ran, X.J. and Xue, L.F.: The research of method and system of regional three-dimensional geological modeling, Doctor Thesis, Jilin University, 2020.

Ray, P., Manach, Y.L., Riou, B. and Houle, T.T.: Statistical evaluation of a biomarker, Anesthesiology, 112(4): 1023-1040, https://doi.org/10.1097/ALN.0b013e3181d47604, 2010.

Smirnoff, A., Bolsvert, E. and Paradis, S.J.: Support vector machine for 3D modelling from sparse geological information of various origins, Computers & Geosciences, 34(2): 127-143, https://doi.org/10.1016/j.cageo.2006.12.008, 2008.

https://doi.org/10.5194/gmd-2023-11
Preprint. Discussion started: 20 April 2023

Titos, M., Bueno, A., Garcia, L. and Benitez, C.: A Deep Neural Networks Approach to Automatic Recognition Systems for Volcano-Seismic Events, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5): 1533-1544, https://doi.org/10.1109/JSTARS.2018.2803198, 2018.

Wang, G.C., Carr, T.R., Ju, Y.W. and Li, C.F.: Identifying organic-rich Marcellus Shale lithofacies by support vector machine classifier in the Appalachian basin, Computers & Geosciences, 64: 52-60, https://doi.org/10.1016/j.cageo.2013.12.002, 2014.

Wang, H.: Finding patterns in subsurface using Bayesian machine learning approach, Underground Space, 5(1): 84-92, https://doi.org/10.2172/1467315, 2020.

Wang, J.H., Li, N., Zhang, X.R. and Su, X.: 3D soil layer reconstruction of deep foundation pit based on machine learning, Journal of Chongqing University, 44(5): 135-145, 2021.

Wang, J.M., Zhao, H., Bi, L. and Wang, L.G.: Implicit 3D Modeling of Ore Body from Geological Boreholes Data Using Hermite Radial Basis Functions, Minerals, 8(10), https://doi.org/10.3390/min8100443, 2018.

Wen, Z.Y., Shi, J.S., Li, Q.B., He, B.S. and Chen, J.: ThunderSVM: A Fast SVM Library on GPUs and CPUs, Journal of Machine Learning Research, 19, 2018.

Wu, L.X.: Topological relations embodied in a generalized tri-prism (GTP) model for a 3D geoscience modeling system, Computers & Geosciences, 30(4): 405-418, https://doi.org/10.1016/j.cageo.2003.06.005, 2004.

Xu, S.T. and Zhou, Y.Z.: Artificial intelligence identification of ore minerals under microscope based on deep learning algorithm, Acta Petrologica Sinica, 34(11): 3244-3252, 2018.

Yang, Y.S., Li, Y.Y., Liu, T.Y., Zhan, Y.L. and Feng, J.: Interactive 3D forward modeling of total field surface and three-component borehole magnetic data for the Daye iron-ore deposit (Central China), Journal of Applied Geophysics, 75(2): 254-263, https://doi.org/10.1016/j.jappgeo.2011.07.010, 2011.

Zhang, T.F., Tilke, P., Dupont, E., Zhu, L.C., Liang, L. and Bailey, W.: Generating geologically realistic 3D reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks, Petroleum Science, 16(3): 541-549, https://doi.org/10.1007/s12182-019-0328-4, 2019.

Zhang, X.Y., Ye, P., Wang, S. and Du, M.: Geological entity recognition method based on Deep Belief Networks, Acta Petrologica Sinica, 34(2): 343-351, 2018.

Zhang, Y. and Chu, B. Z.: Automatic Borehole Comparison Technology Based on Machine Learning, Master Thesis, China University of Geosciences (Beijing), 2021.

Zhou, C.Y., Ouyang, J.W., Ming, W.H., Zhang, G.H., Du, Z.C. and Liu, Z.: A Stratigraphic Prediction Method Based on Machine Learning, Applied Sciences-Basel, 9(17), 2019.