# Methodology for calculating the geological structure complexity index using remote sensing data to improve the efficiency of machine learning

Volodymyr Hnatushenko<sup>1</sup>, Serhii Nikulin<sup>1</sup>, Vita Kashtan<sup>1</sup> and Olga Korobko<sup>1</sup>

<sup>1</sup> Dnipro University of Technology, 19 av. Dmytra Yavornytskoho, Dnipro, 49005, Ukraine

#### Abstract

The article proposes a methodology for calculating a new Geological Structure Complexity Index (GSCI) based on a joint analysis of contract brightness, tone or color boundaries of images represented by raster maps of geophysical fields, digital elevation models, and space images. The calculation of the GSCI maps includes two stages - detecting contrast boundaries in the image using the Canny method and calculating the total length of the boundaries in a sliding window of a certain size. Thus, the index has a simple meaning and is easy to calculate. The information content of the obtained index maps was tested on three real sites when solving the problem of forecasting new ore and oil and gas deposits. As shown, maps of this index are more informative compared to the initial remote sensing data and can be effectively used as additional data set when forming a feature subset for classification with supervised machine learning.

#### Keywords

Remote sensing, feature subsets, geological structure complexity, machine learning, brightness boundaries, histogram

## 1. Introduction

Remote sensing data has significant potential for geosciences, serving as a source of spatial information, including for machine learning operations. Remote sensing data can be used in various fields, including geological research [1], land cover mapping [2], climate change detection, and environmental monitoring.

In machine learning, the effective solution to practical problems depends on the ability to generate such a set of remote sensing data that would be sufficiently complete and informative from the point of view of a particular task. The usually available set of satellite images and other remote sensing data most often does not meet these conditions. In this connection, various methods of processing initial images are applied.

The basic idea of image processing is to obtain additional geospatial information from interpreted remote sensing data, which depends on the texture, composition, and structure of the objects that form the Earth's surface [3].

Developing new approaches and algorithms for extracting information from remote sensing data is a popular trend in Earth sciences [4-6]. One such approach is the study and assessment of the complexity of the landscape structure, which directly depends on the complexity of the geological structure of the territory.

#### 2. Statement of the problem

To effectively solve applied problems, the Earth's surface can be accurately represented by a grid of *NxM* size, comprised of square cells that align with the pixels of the base satellite image. The solution to most practical tasks using machine learning methods requires the whole volume of available data, including multiscale satellite images, digital elevation models, physical field maps,

© 024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>&</sup>lt;sup>1</sup>ICST-2024: Information Control Systems & Technologies, September, 23 – 25, 2024, Odesa, Ukraine

hnatushenko.V.V@nmu.one (V. Hnatushenko); nikulin.s.l@nmu.one (S. Nikulin); kashtan.v.yu@nmu.one
 (V. Kashtan); korobko.o.v@nmu.one (O. Korobko)

<sup>©0000-0003-3140-3788 (</sup>V. Hnatushenko); 0000-0003-1795-3599 (S. Nikulin); 0000-0002-0395-5895 (V. Kashtan); 0000-0002-7491-9162 (O. Korobko)

etc. As a consequence, each grid cell with coordinates (m, n) can be matched with a set of measurements forming a vector of features:

$$(m,n) \leftrightarrow X^{(p)}_{(m,n)} \tag{1}$$

where p is the dimensionality of the vector  $X^{(p)} = (X_1, X_2, ..., X_p)$ , depending on the number of available features.

The ability of a set of attributes to adequately describe the phenomena, objects, or processes under study determines the quality and meaningfulness of the results obtained. Despite the growth of diversity and volumes of remote sensing, measured attributes, and even their combinations in lots of cases do not have sufficient informativeness for obtaining qualitative solutions to complex problems. This problem can be solved by calculating transformants of the initial data, which would reflect to a greater extent those or other essential aspects of the studied phenomenon or process [7]. A feature set (the initial data and their transformants) can be selected a set of attributes, a subset that is potentially "best" in terms of the chosen method for solving a particular problem. [8].

This paper considers the computer technology of construction and estimation of informativity of one such transformant, reflecting the measure of complexity of geological structure of the territory. Its calculation is based on the detection and processing of contrasting boundaries identified on the source data - satellite images, DEM and geophysical fields.

The transformant should be able to improve the accuracy of the results of procedures applied to data sets when performing supervised classification [9, 10] using some known machine learning method - using neural networks, support vector machines, linear discriminant analysis, decision trees or others.

#### 3. Related works

In recent decades, the complexity of the geological structure has also been recognized as a positive indicator for localizing mineral deposits, including both ore and oil/gas. Thus, [11] considered the effect of the scale of geological maps on the strength of the relationship between geological complexity and gold mineralization. It is shown that geologic complexity proves valuable as an initial predictor map for analyzing prospects and identifying gold exploration targets. Article [12] points to the confinement of gold deposits to fault zones created by groups of subsidiary faults. The articles [13-15] used geologic complexity as a positive indicator of the presence of ore deposits. The authors of this paper have shown that the complexity of the geologic structure is in direct correlation with the probability of discovering oil and gas and ore deposits [16-18]. A common problem of most of the above works is the widespread use of geological maps to assess the complexity of the geological structure. Such maps are constructed by experts and depend heavily on their subjective assessments and preferences. As a result, complexity assessments are also influenced by the subjective factor. Next, geological maps are simplified models of the surface and lack many details that are present in reality. In addition, the complexity indexes used often have a controversial and cumbersome method of calculation and depend on parameters and coefficients that are also assigned subjectively by experts.

Therefore, it seems important to find such a geological structure complexity index that 1) could use any objective images of the earth's surface, 2) would have a simple and understandable meaning, and 3) would be easy to calculate.

This paper considers the calculation of a Geologic Structure Complexity Index (GSCI) that satisfies the above requirements and presents a methodology for the preparation of GSCI maps based on digital remote sensing data.

#### 4. Experimental data

The computer technology described was tested experimentally by us in three known mineral deposit areas.

Area 1. The area is about 800 km<sup>2</sup> and is located within the Turan plate (Uzbekistan). Several ore occurrences and individual points with elevated gold content have been discovered within the area, which were used as reference objects (points of interest). Initial data are represented by the results of observations of 6 geophysical fields at a scale of 1:50000 (vertical derivative of the gravitational field, two derivatives of magnetic fields -  $\Delta Ta$  and  $\Delta Za$ , *y*-rays field, two natural electric fields), as

well as synthesized Landsat satellite image (channels 2,3,4) with 30 m resolution. The training set of data included the listed initial data, some of their traditional transformants (e.g., contrasting or smoothing in a sliding window), and constructed maps of this index. The centers of known ore occurrences were used as reference points.

Area 2. The area is about 17,000 km<sup>2</sup> and is located in the central part of the oil and gas-bearing Dnipro-Donetsk Depression (Ukraine). Scores of deposits are known in the territory of the area, mainly gas condensate fields. The input data are represented by magnetic and gravimetric surveys on a 500x500m grid, and a radar satellite image obtained as a result of the SRTM mission [19].

Area 3. The area is located within the Azov block of the Ukrainian crystalline shield. The goldbearing Sorokinskaya granite-greenstone structure and the promising Berestovetskaya structure are located on the territory of the area. Within the Sorokinskaya structure, several gold ore bodies have been identified. They served as reference objects in the experiments. The input data are represented by magnetic and gravitational field surveys at the scale of 1:50000 (partially 1:10000 and 1:25000) and SRTM radar images. Examples of baseline maps are shown in Fig. 1.



Figure 1: a) Landsat-8 satellite image (area 1); b) SRTM data (area 2); c) Vertical derivative of gravitational field (area 3)

#### 5. Methodology

The proposed methodology of the GSCI calculation is presented in Fig. 2. The construction of the GSCI maps is based on the use of regular data – two-dimensional potential fields, digital elevation models, and satellite images, both panchromatic and multichannel [20].

Regular data, due to its pixel structure, allows the use of all standard methods for processing halftone raster images.

In this case, we encounter the challenge of choosing some index of geological structure complexity that can be easily computed from the raster image and accurately depicts the density of boundaries within geological features and linear disjunctive structures. Such an indicator may be the length of image tone (color or brightness) boundaries (or, in the terminology of satellite image processing, contrast brightness boundaries).

In this case, we can simplify the task of assessing geological structure complexity to 1) delineation of boundaries on the image and 2) calculation of the total length of boundaries within a sliding window of size  $p \times p$ , where p=3, 5, 7, ... – some odd number of image pixels or physical field grid cells.

At present, the most popular extraction of contrast boundary methods are various mask algorithms and, first and foremost, the so-called optimal Canny detector [24]. The result of the Canny detector is a binary map, where ones correspond to the presence of boundaries and zeros - to their absence. Counting the total length of boundaries within the window is reduced to simply finding the number of pixels with unit values.



Figure 2: Flowchart of proposed methodology

The index calculated in this way reflects not only the total number of boundaries within the window, but also the complexity of their shape (Fig. 3).



Figure 3: Boundaries of simple (a) and complex (b) shapes in a binary image (length - 9 and 16 pixels, respectively)

The problem is that geologic boundaries may be reflected differently in certain physical fields and landscapes, including not being reflected at all. Therefore, to confidently identify boundaries, it is necessary to use the widest possible range of source data, selecting boundaries on separate images and combining the resulting binary maps using the binary pixel disjunction operation: C = A OR B,(2)

where A and B are two binary maps; C is the final map obtained by combining A and B.

This allows minimizing the error and increasing the reliability of the GSCI determination.

Below are the results of computational experiments on a calculation of the GSCI and evaluation of their predictive capabilities on several real gold and oil and gas deposits.

For each experimental area, binary maps reflecting the tone (brightness) boundaries of the available raster maps of potential fields and satellite images were constructed using the Canny detector. Further, the binary maps for a certain area were combined by pixel disjunction, and a set of raster maps of GSCI, representing the total length of boundaries inside sliding window 19x19 grid cells, was constructed using the obtained data (Fig. 4).



Figure 4: The obtained GSCI maps for a) area 1; b) area 2; c) area 3

To assess the informativeness of the GSCI as a feature that may be useful for new deposit prospecting, we analyzed the degree of coincidence of the histograms of the distribution of GSCI map values, plotted a) for the entire area and b) for points of interest (known deposits) – pixels located above known deposits. The lower the degree of histogram overlap, the easier it is to separate promising points, similar to reference points, from the rest of the area.

Root Mean Square Error (RMSE) was used to assess the degree of histogram overlap, calculated as [21]:

$$RSME = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \hat{x}_i)^2}{N}},$$
(3)

where *N* is the number of histogram intervals; *i* is the variable, i=1...N;  $x_i$  is the value of the *i*-th column of the first histogram;  $\hat{x}$  is the value of the *i*-th column of the second histogram.

The *N* value for the sample was calculated using the Sturgess formula [22]:

$$N = 1 + 3.322 \log(n), \tag{4}$$

where *n* is the number of values in the sample.

The calculated value was also used for the sample made up of reference points to make both samples comparable. In addition, to increase reliability, another indicator of histogram similarity was calculated – the shift of their values, measured in the number of intervals (histogram bars). Fig. 5 displays the histograms of the GSCI maps for areas 1–3. The corresponding GSCI map is shown in Fig. 4.



Figure 5: Histograms plotted by the values of the GSCI in the whole territory (orange bars) and in the reference points (blue bars) of areas 1 (a), 2 (b), 3 (c)

The reference points have higher GSCI values in comparison to the whole area, enabling the utilization of the obtained map as an additional search feature for forecasting new mineral deposits. The RMSE values of GSCI histograms plotted for the entire map and for points of interest are given in Table 1.

Histogram overlap indices of the histograms of the GSCI maps for areas 1-3			
Slide window size, cells	RMSE	Shift in mod, intervals	
Area 1	0.0442	2	
Area 2	0.0258	5	
Area 3	0.0534	4	

Table 1

diago of the histographic of the CCCL many for anose 1 Lista

## 6. Discussion

The obvious question is: are GSCI maps more informative than the initial data from which they were calculated?

To answer this question, histograms of the values of some initial data sets were additionally calculated.

Fig. 6 shows the histograms for the electric field, magnetic field, and LandSat-8 image for area 1.

It is easy to notice that the histograms constructed for the whole area and separately for the reference objects generally repeat each other, which indicates their low individual information content for predicting new deposits (Table 2).

For areas 2 and 3, histograms were also calculated for the values of individual source datasets – physical fields, satellite images, and digital elevation models. As in the case of Area 1, the histograms plotted for values across the entire area and in areas above the reference areas have greater overlap than the GSCI and, therefore, have low predictive power.

Table 2	
---------	--

Histogram coincidence indices of the values of the original data sets for area 1

	ÿ	
Data set	RMSE	Shift in mod, intervals
Field of electric resistance	0.0212	0
Magnetic field ∆Ta	0.0226	0
Landsat image	0.0357	0

The information content (potential usefulness) of the obtained maps of the geological structure complexity indicator (GSCI) was analyzed in terms of their effectiveness in separating reference points from others in the multidimensional dataset space, including initial physical fields, satellite images, DEMs, and the results of their various transformations.

To assess the information content of individual datasets, we used criteria based on Kendall's [23] and Bhattacharya's [24] distances.

To expand the feature subsets, we calculated maps of the GSCI using different sizes of the sliding window. As shown in Fig. 7, the obtained maps of the GSCI are generally surpass in information content to the original data sets on the basis of which they were calculated.



Figure 6: Histograms plotted against the values of the original data sets on the whole territory and at the reference points of area 1



Figure 7: a) the information content (according to Bhattacharya) of the datasets for area 1; b) the information content (according to Bhattacharya) of the datasets for area 2; c) the information content (according to Kendall) of the datasets for area 3

### 7. Conclusions

As the conducted studies have shown, the new Geological Structure Complexity Index described in the paper has demonstrated its practical usefulness. The obtained GSCI maps shown in Fig. 4 show the confinement of known ore and, to a lesser extent, oil and gas objects to zones of higher GSCI values, which allows us to consider the index as an additional dataset for machine learning procedures that is more informative than the initial data - physical fields, DEMs and satellite images.

The GSCI maps themselves should not be considered as definitive, predictive maps, since the overlap area of the histograms presented in Fig. 7 in some cases reaches 40–50% (although these values are smaller than those of the initial datasets). It is necessary to use all available data and machine learning tools. The GSCI maps are much more effective when applied in combination with other remote sensed data in supervised classification procedures.

Overall, the calculations performed indicate the prospects of the approach to calculating GSCI maps based on the identification and analysis of contrasting boundaries on raster maps of physical fields, digital elevation models and satellite images. Additional advantages of the presented index are the simplicity of its calculation and clear physical meaning.

GSCI maps can be used independently to highlight promising areas, but they are much more effective when used for supervised classification in a multidimensional space formed from the original datasets and their transformants.

The results obtained, despite their usefulness, can probably be improved by further research. First, it is necessary to study in more depth the extent and manner of influence of the sliding window size on the information content of the map of the GSCI. It is also necessary to find spectral ranges and channels of satellite images that provide the least influence of seasonal factors on the results accuracy. Finally, it is necessary to improve the methods of highlighting contrast boundaries so that they take into account the specificity of the image used, differently processing images and geophysical fields of different accuracy and scale.

## Acknowledgements

The work is supported by the state budget scientific research project of Dnipro University of Technology "Models and information technologies of data processing and analysis in complex computer systems and networks" (state registration number 0121U114523).

#### References

- [1] N.M. Ngom, M. Mbaye, D. Baratoux, L. Baratoux, K.E. Ahoussi, J.K. Kouame, G. Faye, E.H. Sow. Recent expansion of artisanal gold mining along the Bandama River (Côte d'Ivoire). Int. J. Appl. Earth Obs. Geoinf. 2022, 112, 102873.
- [2] V.Yu. Kashtan, V. V. Hnatushenko, S. Zhir. Information Technology Analysis of Satellite Data for Land Irrigation Monitoring, in: 2021 IEEE International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), Kyiv, Ukraine, November 29 – December 3, 2021, pp. 12-15. doi:10.1109/ UkrMiCo52950.2021.9716592.
- [3] V.J. Kashtan, V.V. Hnatushenko, Y.I. Shedlovska, Processing technology of multispectral remote sensing images, in: 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF), 2017. doi:10.1109/ysf.2017.8126647.
- [4] D. Uhryn, Yu. Ushenko, V. Lytvyn, Z. Hu, O. Lozynska, V. Ilin, A. Hostiuk, Modelling of an Intelligent Geographic Information System for Population Migration Forecasting, International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.4, pp. 69-79, 2023. doi:10.5815/ijmecs.2023.04.06.
- [5] V. Vysotska, K. Smelyakov, N. Sharonova, E. Vakulik, O. Filipov, R. Kotelnykov, Fast Color Images Clustering for Real-Time Computer Vision and AI System, (2024) CEUR Workshop Proceedings, 3664, pp. 161 – 177.
- [6] V. Gnatushenko, The use of geometrical methods in multispectral image processing. Journal of Automation and Information Sciences 35 12 (2003) 1-8. doi: 10.1615/JAutomatInfScien.v35.i12.10.
- [7] J.A. Richards, Supervised Classification Techniques, in: Remote Sensing Digital Image Analysis. Springer, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-30062-2\_8.
- [8] M. Mehmood, F. Shahzad, B. Zafar, A. Shabbir, N. Ali, Remote Sensing Image Classification: A Comprehensive Review and Applications. Mathematical Problems in Engineering, Hindawi (2022) 1-24. doi:10.1155/2022/5880959.
- [9] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebar., J. Saeed, A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends 1 (2020) 56-70. doi:10.38094/jastt1224.
- [10] S.H. Huang, Supervised feature selection: A tutorial. Artificial Intelligence Research 4 22 (2015). doi: 10.5430/air.v4n2p22.
- [11] W. Mingyang, E. Wang, X. Liu, C. Wang, Scale-space effect and scale hybridization in image intelligent recognition of geological discontinuities on rock slopes. Journal of Rock Mechanics and Geotechnical Engineering (2023). doi: 10.1016/j.jrmge.2023.08.015.
- [12] D.I. Groves, M. Santosh, R.J. Goldfarb, L. Zhang, Structural geometry of orogenic gold deposits: Implications for exploration of world-class and giant deposits. Geoscience Frontiers 9 4 (2018) 1163-1177. doi:10.1016/j.gsf.2018.01.006.

- [13] Z. Liu, L. Han, C. Du, H. Cao, J. Guo, H. Wang. Fractal and Multifractal Characteristics of Lineaments in the Qianhe Graben and Its Tectonic Significance Using Remote Sensing Images. Remote Sensing 13 4 (2021) 587. doi:10.3390/rs13040587.
- [14] M. Marghany, Structural geology of mineral, oil and gas explorations. Advanced Algorithms for Mineral and Hydrocarbon Exploration Using Synthetic Aperture Radar (2022) 31-79. doi:10.1016/B978-0-12-821796-2.00003-3.
- [15] Augustin, D. Gaboury, Multi-stage and multi-sourced fluid and gold in the formation of orogenic gold deposits in the world-class Mana district of Burkina Faso – Revealed by LA-ICP-MS analysis of pyrites and arsenopyrites. Ore Geology Reviews 104 (2019) 495-521. doi:10.1016/j.oregeorev.2018.11.011.
- [16] K. Anıl Tözün, A. Özyavaş, Automatic detection of geological lineaments in central Turkey based on test image analysis using satellite data. Advances in Space Research 69 9 (2022) 3283-3300. doi:10.1016/j.asr.2022.02.026.
- [17] W. Viveen, P. Baby, C. Hurtado-Enríquez, Assessing the accuracy of combined DEM-based lineament mapping and the normalized SL-index as a tool for active fault mapping. Tectonophysics 813 (2021). doi:10.1016/j.tecto.2021.228942.
- [18] B.S. Busygin, S.L. Nikulin, O.V. Korobko, Concentration of contrast borders of different-scale satellite images and their interconnection with geological objects. in: 16<sup>th</sup> International Conference Geoinformatics – Theoretical and Applied Aspects, 2017. Doi:10.3997/2214-4609.201701871.
- [19] S. M. Mudd, Chapter 4 Topographic data from satellites, Editor(s): P. Tarolli, S. M. Mudd, in: Developments in Earth Surface Processes, Elsevier, Volume 23, 2020, pp. 91-128. Doi:10.1016/B978-0-444-64177-9.00004-7.
- [20] V. Hnatushenko, V. Kashtan, Automated pansharpening information technology of satellite images. Radio Electronics, Computer Science, Control 2 (2021) 123-132. doi:10.15588/1607-3274-2021-2-13.
- [21] Q. Zheng, L. Zeng, G. Karniadakis, Physics-informed semantic inpainting: Application to geostatistical modeling, George, 2019. doi:10.1016/j.jcp.2020.109676.
- [22] I. Zlateva, N. Nikolov, M. Alexandrova, V. Raykov, Constructing an Algorithm for Selecting the Number of Histogram Bins in Statistical Hypothesis Testing for Normal Distribution of Sample Data. International Journal of Engineering Research & Science (IOER) 4 11 (2018). doi:10.5281/zenodo.1745060fatcat:u2nlpalccnadvcqc3 m3ffdjgl4.
- [23] R. Kumar, S. Vassilvitskii, Generalized distances between rankings, in: WWW '10: Proceedings of the 19th international conference on World wide web, 2010, 571–580 pp. doi: 10.1145/1772690.1772749.
- [24] E. Choi, C. Lee, Feature extraction based on the Bhattacharyya distance. Pattern Recognition 36 8 (2003) 1703-1709. doi:10.1016/S0031-3203(03)00035-9.