Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070404
ISSN 2371-8838 Vol. 7 Num. 4

Automatic Labeling Method of Geological Codes Based on Multi-factor Optimization

Hao Wang[1,a,*], Jian Lin[1,b], Bin Gao[2,c]

[1]School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan, China
[2]Hunan Hangrui Digital Technology Co. Ltd, Changsha, Hunan, China
[a]13973847910@163.com, [b]694586970@qq.com, [c]gisboss@126.com
**Corresponding author*

Keywords: Automatic labeling, multi-factor optimization, multi-position labeling, evaluation method, particle swarm optimization

Abstract: Automatic labeling of geological codes is an important part of automatic mapping of geological maps. A multi-factor optimization based automatic labeling method is proposed to address the issue of existing polygon feature label placement methods being unable to achieve multi-position placement in geological codes. Firstly, classify geological bodies based on whether they can accommodate geological codes within the area; Subsequently, sufficient candidate positions are obtained for the geological body and a candidate position evaluation method that integrates multiple factors is proposed; Finally, based on the candidate position evaluation method, sorting and particle swarm optimization algorithms are used to achieve single position labeling and multi-position labeling. The simulation experiment compares it with existing polygon feature placement methods from the perspectives of coverage, shape, and method. The method proposed in this paper can automatically placement non-conflicting geological codes for geological bodies of different shapes and areas, and has better performance than existing methods in complex geological bodies.

1. Introduction

The automatic labeling of geological codes is an important part of the automatic drawing of geological plan maps, which can accurately visualize map information and has important research significance.

Geological code labeling belongs to the polygon feature label placement in map feature label placement. The method for polygon feature label placement includes the center point method, which is efficient and suitable for circular regions. In complex regions, there is an error in placing labels outside the area using the center point method. In order to avoid such errors, various methods for extracting the skeleton lines of polygon features have emerged[1]. Choosing a suitable position on the skeleton line to placement label effectively solves the problem that labeling in the outside region, and the skeleton line method is suitable for various types of regions. In addition, some scholars believe that the label position should be located at the visual center of the feature[2], and geometric methods

are used to determine the label position of the polygon feature, such as the center of the maximum inscribed rectangle within the feature and the center of the maximum inscribed quadrilateral. Li believes that the label position should maximize the internal area of the Voronoi shape formed by its competition with the boundary of the polygon feature[3].

The shape and area of geological bodies in the geological plan are inconsistent, and there are two situations: labeled outside and labeled inside, and there are label conflicts and multi -position labeling problems. Label conflict refers to the intersection or coverage of label with feature or other label, which affects the reading and use of maps. Multiple positional labeling refer to the simultaneous placement of multiple codes in geological bodies to jointly express geological information. In response to the problem that existing automatic labeling methods for polygon features cannot achieve multi-position labeling of geological codes, this paper proposes a geological code automatic labeling method based on multi-factor optimization, guided by the geological map clearing regulations.

2. Principles for geological code labeling

In order to ensure that geological code labeling conforms to standard specifications, combined with the geological map mapping regulations and the label placement principles proposed by Yoeli[4] and Imhof[5], the principles applicable to geological code labeling are summarized as follows:

Principle 1: Geological codes should be listed in horizontal characters, and the code of structural names should be determined based on the direction of the structure.

Principle 2: Geological codes should not conflict with other feature or feature's label.

Principle 3: The geological code should be clearly expressed and aesthetically pleasing.

Principle 4: If the geological body has a large area or a complex shape, multiple geological codes should be marked simultaneously within the same geological body.

3. Candidate locations and evaluation methods

3.1. Candidate position generation

The geological body is a three-dimensional structure in space, but it appears as a vertical projection of a horizontal plane on the geological plane, which can be abstracted as a polygon Q in the Euclidean plane, as shown in Figure 1.

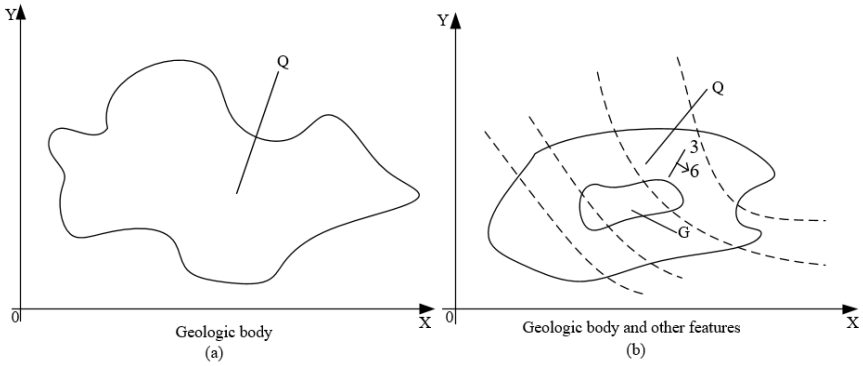


Figure 1: Examples of strata.

The geological code is represented by a rectangle Z with a length of l and a width of w. The coordinates of the midpoint of the rectangle are the coordinates of the geological code on the geological plan, and the coordinates are represented as: $Z_i = (x_i, y_i)$. The relationship between the geological code and the geological body is:$Z \subset Q$.

There are many elements in the geological plan, and there are more occurrence symbols, contour

lines, and other geological bodies in the geological body area in Figure 1 (b) than in Figure 1 (a). According to principle 2, geological codes cannot conflict or overlap with other feature or feature's label, and for geological codes, these elements will affect their location selection. Therefore, these elements that affect geological code labeling inside and outside the geological body area are collectively referred to as obstacle elements, the obstacle elements are represented as H.

According to principle 4, labeling multiple geological body codes can help readers quickly understand geological information. Large-scale geological bodies may include smaller geological bodies with different lithology, as shown in Figure 1 (b) where geological body Q includes geological body G. Therefore, the candidate area for geological code in large-scale geological bodies is the geological body itself, represented as $\mathrm{A} = \mathrm{Q} - (\mathrm{Q} \cap \mathrm{G})$.

3.2. A Method for Evaluating Candidate Positions Based on Multiple Factors

The quality of code is determined by the code position, and the visual effect of code varies at different positions. It is necessary to quantitatively or qualitatively evaluate the quality of candidate positions in order to select suitable positions from these candidate positions. Therefore, the rationality of the candidate position evaluation function determines the final quality of the code. Based on the principle of geological code labeling, the following three factors that affect the quality of code were considered:

Label conflict: Label conflict refers to the conflict between geological codes and geological boundaries or obstacle elements, which can affect the map reading experience and violate principle 2. Set the boundary of the candidate region as B, and the label conflict as shown in formula (1).

$$S_1 = \begin{cases} 1, Z \cap B \neq \emptyset \\ 2, Z \cap H \neq \emptyset \wedge Z \cap B = \emptyset \\ 3, Z \cap H \neq \emptyset \wedge Z \cap B = \emptyset \end{cases} \tag{1}$$

$\mathrm{S}_1 = 1$ indicates a conflict between the geological code and the geological boundary at this location, $\mathrm{S}_1 = 2$ indicates that the geological code at this location only conflicts with obstacle elements, $\mathrm{S}_1 = 3$ indicates that the geological code at this location does not conflict with all elements. If the geological code is located at the candidate location $\mathrm{S}_1 = 1$, then this position does not participate in subsequent steps.

Regional coverage: regional coverage refers to the ratio of the maximum area that can be radiated by the location of the geological code in the geological body to the area of the geological body, which can reflect the visual influence range of the geological code in the geological body. The higher the coverage, the closer the location of the geological code is to the visual center. The initial circle is the largest inscribed circle of the minimum bounding rectangle of the geological code, and the center of the circle is the center of the geological code. The radiation area is expanded by morphological expansion until it collides with the geological boundary, as shown in Figure 2. Initial radius of circle r, maximum expansion factor c_a. Maximum expansion radius$\mathrm{r_a} = (c_a * \Delta r) + r$, Δr is the radius increment, and the candidate area is S_A.The regional coverage rate is shown in formula (2).

$$S_2 = \frac{\pi * {r_a}^2}{S_A} \tag{2}$$

If there is a conflict between the geological code and the obstacle element in the candidate location, a penalty mechanism will be added to halve its score.

Blank area: There are numerous obstacles in the geological plan, which affect the visual perception of geological codes. The blank area refers to the unobstructed element area radiating from the location of the geological code. The larger the area, the clearer the expression of the geological code. Similarly,

circular expansion is used to calculate the area of the blank area, as shown in Figure 2. The area of the blank area is shown in formula (3).

$$S_3 = \pi * r_b^2 \tag{3}$$

r_b is the expansion radius of the blank area, $r_b = (c_b * \Delta r) + r, c_b$ is the expansion factor. When the geological code on the candidate location conflicts with the obstacle element, i.e. when $S_1 = 2$, $c_b = 0, S_3 = 0$.

The evaluation method for candidate positions in large-scale geological bodies is obtained by weighting the three factors of conflict, regional coverage, and blank area, as shown in formula (4).

$$S_4 = w_1 * c_a + w_2 * c_b \tag{4}$$

S_4 is an evaluation method for candidate locations of large-scale geological bodies .The weight w_1、w_2 of the area coverage and blank area corresponding to the geological code is determined by the coverage threshold to determine the number of geological codes labeled on a large area of geological body. The coverage is the decisive factor, the easier it is to attract attention. The blank area is an aesthetic factor, so $w_1 > w_2$.

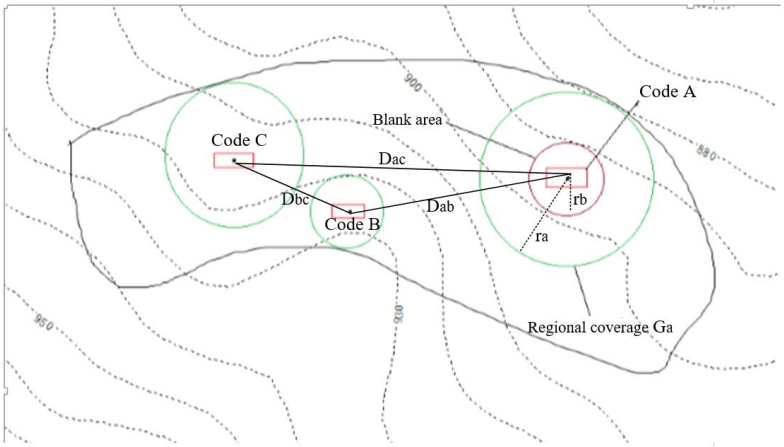


Figure 2: Code coverage area and blank area.

4. Automatic labeling of geological codes

In this section, the goal is to find one or more suitable location to label within the area of large-scale geological bodies, with a focus on selecting a single location or multiple locations simultaneously, and how to determine the location. Given the coverage threshold F, traverse candidate positions in a large geological body that do not conflict with the boundary. If $\max(S_2) \geq F$, there are candidate positions that meet the coverage constraint. For candidate positions that meet the constraint, follow S_4 sort from top to bottom and select the candidate position where $\max(S_4)$ is located for labeling.

4.1. Multi-position automatic labeling method based on particle swarm optimization

If $\max(S_2) < F$, the candidate positions in the geological body do not meet the coverage constraint, it is considered that the geological body needs to be labeled with multiple geological codes to express its meaning. The multi-location label problem is described as finding a set of candidate locations that meet coverage constraints, are evenly distributed, and have the least number. If the number of geological codes is too large, it can actually cause reading difficulties for readers.

Labeling multiple geological codes simultaneously in a large-scale geological body not only needs

to consider the quality of the annotation position, but also the positional relationship between multiple codes. Therefore, in response to the issue of uneven distribution of codes, increasing codes dispersion is used to evaluate the distribution of geological codes.

Code discreteness: Code discreteness represents the distribution of multiple geological codes in a large geological body, as shown in Figure 2. Code B and C are dense than Code A and C. By reflecting the density of the two codes through their distance, and combining with the coverage area radius of the candidate location where the annotation is located, the density between codes is divided into five levels. Label the candidate positions a and b with code Z, where the initial radius of code Z is r and the coverage area radius corresponding to positions a and b is r_a, r_b. The positional relationship between any two points is shown in formula (5).

$$S_7 = \begin{cases} 0{,}0 < D_{ab} \leq 2r \\ 0.25{,}2r \leq D_{ab} < \dfrac{r_a + r_b}{2} \\ 0.5{,}\left(\dfrac{r_a + r_b}{2}\right) \leq D_{ab} < r_a + r_b \\ 0.75{,}r_a + r_b \leq D_{ab} < 3 * \dfrac{r_a + r_b}{2} \\ 1{,}3 * \dfrac{r_a + r_b}{2} \leq D_{ab} \end{cases} \tag{5}$$

D_{ab} is the distance between positions a and b.

S_7 represents the density of two geological codes marked in a large geological body. The lower the score, the denser the two geological codes are. Therefore, the distribution of multiple geological codes is represented by the average density between any two codes, as shown in formula (6).

$$S_8 = \frac{\sum_1^n S_{7i}}{\frac{n(n-1)}{2}} \tag{6}$$

S_8 represents the distribution of multiple geological codes simultaneously labeled, with smaller scores and denser distribution, n represents the number of geological codes, and $n \geq 2$.

A comprehensive evaluation method for multiple geological codes is obtained by combining the distribution of geological codes with the quality of candidate locations where geological codes are located, as shown in formula (7).

$$S_9 = w_3 * \left(\frac{\sum_{i=1}^n f(S_{4i})}{n}\right) + w_4 * s_8 \tag{7}$$

S_9 is a comprehensive evaluation method for multiple geological codes, where $f(S_{4i})$ is the normalized candidate location evaluation score, w_3, w_4 are the weight of correspond to the average quality of the candidate positions where multiple geological codes are located, and the degree of dispersion corresponding to multiple geological codes. If multiple geological codes are gathered together, the significance of labeling multiple geological codes is lost. Add coverage constraints to establish a mathematical model for multi-position labeling based on formula (9), as shown in formula (8).

$$\begin{cases} max: S_9 \\ s.t.\, sum(S_{2i}) \geq F \end{cases} \tag{8}$$

Solve this problem using Particle Swarm Optimization (PSO)[6], an intelligent biomimetic algorithm that mimics birds searching for food. The algorithm runs iteratively, updating the position

of the population by updating the speed and position of the individual each iteration. Individual and population updates are influenced by individual optimal position (pbest) and global optimal position (gbest). In the algorithm, the update of particle position is determined by the current position and velocity of the particle. The velocity of the particle consists of the momentum part, its own cognitive part, and social cognitive part. The velocity change and particle position update are shown in formulas (9) and (10).

$$V_k(t+1) = w_k(t)V_k(t) + c_1 r_1\big(pbest_k(t) - U_k(t)\big) + c_2 r_2\big(gbest_k(t) - U_k(t)\big) \quad (9)$$

$$U_k(t+1) = U_k(t) + \mathrm{V_k}(t+1) \quad (10)$$

Where $\mathrm{V_k}(t+1)$ is the velocity of particle k in the t+1st iteration; $U_k(t)$ is the position of particle k in the second iteration; $\mathrm{c_1}$and $\mathrm{c_2}$ is the actual acceleration coefficient that controls the impact of global and individual optimal positions on particle velocity; $w_k(t)$ is the inertia weight of particle k in the t-th generation; r_1 and r_2 is a random number evenly distributed within the range of 0 and 1, used to maintain sufficient diversity in the population.

The detailed steps for multi-location labeling of geological codes based on PSO are as follows:

(1) Initialize particle swarm parameters: initialize population size N, algorithm iteration times T, initialize individual learning factor $\mathrm{c_1}$ maximum value $\mathrm{c_{1max}}$ and minimum value $\mathrm{c_{1min}}$, social learning factor $\mathrm{c_2}$ minimum value $\mathrm{c_{2min}}$ and maximum value $\mathrm{c_{2max}}$, maximum value w of inertia weight $\mathrm{w_{max}}$, minimum value $\mathrm{w_{min}}$, $\mathrm{c_1} = \mathrm{c_{1max}}, c_2 = \mathrm{c_{2min}}, w = \mathrm{w_{max}}$.

(2) Constructing the geological code group: The theoretical optimal value for the number of geological codes in this problem is 2, with the minimum number as the greedy strategy, and sort candidate positions based on the size of $\mathrm{c_a}$, select n candidate position that satisfy $sum(S_{2i}) \geq F$ and $\mathrm{G_i} \cap G_j = \emptyset$, $\mathrm{G_i}$、G_j refers to the coverage area of the i-th and jth geological codes. The locations of these n geological codes are added to the population as individuals, represented by $\mathrm{Z} = [\mathrm{z_1(x_1, y_1)}, \mathrm{z_2(x_2, y_2)}, \dots, \mathrm{z_n(x_n, y_n)}]$. Randomly select $\mathrm{n} \in [\mathrm{n}, \mathrm{n}+1]$codes that meet the constraint conditions from the candidate positions as new individuals to join the population until the upper limit of population size is reached. Initialize the individual's movement speed, expressed as: $\mathrm{V_{lk}} = \big[v_{l1}\big(v_{l1x}, v_{l1y}\big), v_{l2}\big(v_{l2x}, v_{l2y}\big)\big], \dots, v_{ln}\big(v_{lnx}, v_{lny}\big)$, $\mathrm{V_{lk}}$ represents the movement speed of the kth annotation in the jth individual, including the abscissa velocity and the ordinate velocity.

(3) Calculate individual fitness: Calculate the fitness and coverage of an individual according to formula (9). If the individual coverage does not meet the constraint, reduce the individual's fitness as a penalty, and update the population's gbest and pbest.

(4) Update geological code location: The individual in the population is a set of codes with coordinates, and the geological code position in the individual is updated according to formulas (9) and (10). After updating the location, the geological code may need to be adjusted again to become a new individual. The reasons and practices are shown in Table 1.

Table 1: Reason and practice of label adjustment after updating.

Order	question	Solution
1	The new location is no longer within the candidate area	Randomly select a new location
2	$\mathrm{sum(S_{2i})} - \mathrm{min(S_{2i})} \geq \mathrm{F}$	Remove the geological code with the lowest coverage rate

Adjust the position of geological codes in the order shown in the table, and if there are no occurrences in the table, keep them as new individuals. At this step, update the geological code location and reduce the quantity.

(5) Update particle swarm algorithm parameters and code movement speed: The updated parameters include inertia weight and learning factor, which reflect the individual's ability to inherit

the previous speed and directly affect the algorithm's search ability. This article uses a nonlinear decreasing function to control the change of inertia weight, as shown in formula (11).

$$w(t) = w_{max} - (w_{max} - w_{min}) * \left(\frac{t}{T}\right)^2 \tag{11}$$

Individual learning factor c_1 and Social Learning Factor c_2. Both jointly determine the direction and speed of individual movement, reflecting the information exchange between individuals. If $c_1 > c_2$, the particles move in the direction of the individual's optimal position, and vice versa, they move in the direction of the global optimal position. In order to comply with the idea of algorithm improvement, an asymmetric linear change learning factor is used, as shown in formula (12).

$$c_1(t) = c_{1max} - t * \frac{c_{1max} - c_{1min}}{T}, c_2(t) = c_{2min} + t * \frac{c_{2max} - c_{2min}}{T} \tag{12}$$

According to formulas (11) and (12), update the inertia weight and learning factor, and combine pbest and gbest to update the individual's speed. To prevent falling into local optima, when the speed drops to 0, a small speed is randomly assigned to maintain position updates.

(6) Automatic labeling: If the algorithm reaches the maximum number of iterations, the algorithm stops, and the individual with no conflict and the best fitness is selected as the result in the population. The corresponding group of positions are the number of geological codes to be marked and the location to be marked. If the algorithm termination condition is not met, return to step 3 to continue execution.

5. Experiments

We will compare the labeling results of our method under different conditions and the labeling results of different methods under the same conditions from three perspectives: method and shape.

1) Different methods

Select external surface elements with an area of 1501.42 m², the area of internal surface elements is 446.73 m². The geological body is labeled using the center point method, maximum quadrilateral method, and the method described in this article, as shown in Figure 3.

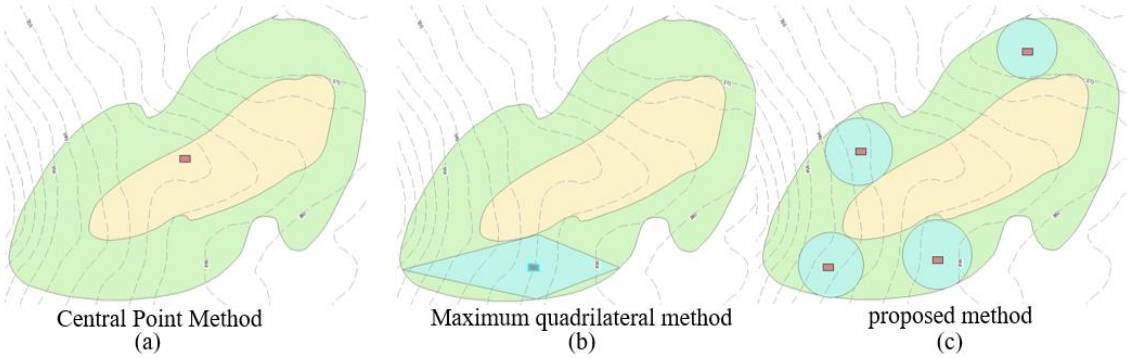
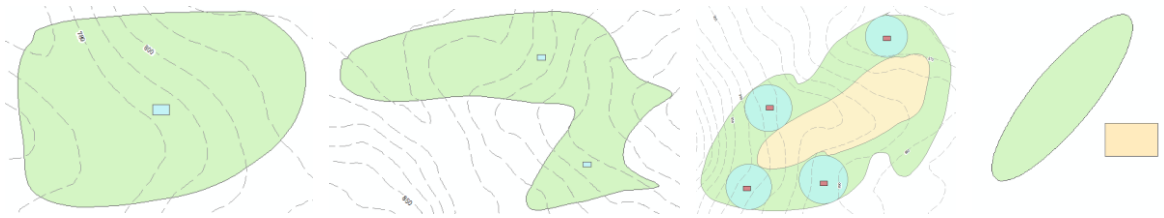


Figure 3: Labeling results by different methods.

Observing Figure 3, it can be seen that the center point method is labeled incorrectly in other geological bodies; the maximum inscribed quadrilateral area calculated by the maximum quadrilateral method is 166.16 m². The coverage rate is 15.7%. Due to not considering obstacle factors, in this case, the code intersects with contour lines, resulting in annotation conflicts, and this method does not consider labeling multiple codes; the method in this article sets the coverage rate to 30%, and labeling four codes can meet the coverage constraint, with a total area of 327.31 m². The coverage rate is 31.03%.

2) Different shapes

Using the method described in this article, four geological bodies with different shapes were labeled with a coverage threshold of 30%, and the results are shown in Figure 4.



(a) (b) (c) (d)

Figure 4: Labeling results of different shapes.

The coverage rates of geological bodies in the three shapes of (a), (b) and (c) are 51.8%, 32.6%, and 31.03%, respectively. (d) is a small area geological body label.

6. Conclusion

In order to achieve automatic labeling of geological codes in geological plans, this article proposes an automatic labeling method based on multi-factor optimization. This method classifies geological bodies and automatically labels code according to the process of map feature label placement. Simulation experiments have shown that this method can label geological bodies of different shapes and areas without conflicting geological codes, and its labeling effect is better than other methods in complex geological bodies.

References

[1] Abe N, Kuroda K, Kamata Y, et al. Implementation and evaluation of a fast area feature labeling method using auxiliary lines [J]. ISPRS International Journal of Geo-Information, 2020, 9(9): 529.

[2] Wu C, Ding Y, Zhou X, et al. A grid algorithm suitable for line and area feature label placement [J]. Environmental Earth Sciences, 2016, 75(20): 1368.

[3] Li L, Li B, Wu Z, et al. Automatic placement of annotation in area feature by map spatial geometry information measurement [C]//Geoinformatics 2007: Cartographic Theory and Models. SPIE, 2007, 6751: 512-519.

[4] Yoeli P. The logic of automated map lettering [J]. The Cartographic Journal, 1972, 9(2): 99-108.

[5] Imhof E. Positioning names on maps [J]. The American Cartographer, 1975, 2(2): 128-144.

[6] Marini F, Walczak B. Particle swarm optimization (PSO). A tutorial [J]. Chemometrics and Intelligent Laboratory Systems, 2015, 149: 153-165.