Quality Evaluation Method for Settlement Data Matching Based on Grey Correlation Analysis

Yangyang Jiao¹ ² ³*, Pingzhi Liu² ³ and Peixin Qi² ³

¹ Information Engineering University, Zhengzhou, Henan, 45001, China
² State Key Laboratory of Geo-information Engineering, Xi’an, 710054, China
³ Xi’an Research institute of Surveying and Mapping, Xi’an, Shanxi, 710054, China

*Corresponding author’s e-mail: johnpanther@163.com

Abstract. To solve the problem that the quality of spatial data matching is difficult to evaluate, this paper proposes a method of matching quality evaluation based on grey correlation machine learning analysis, it’s a method that takes settlement data for example and utilizes the weight of matching index. The method can be performed through the following steps: first, measure the location, area, orientation delta of the settlement entity, and analyzes its grey correlation relationship, then the paper gets the weight of each kind of graphic data delta; second, calculate the graphic data delta of identical entity in original matching data to judge whether the matching relationship is correct. Experiments illustrate that the method is more feasible and practical, for it can measure the graphic data delta and weight effectively and practically and detect the mistakes of original matching data.

1. Introduction

After the completed of the fundamental geo-information database, the main work has become the updating and maintenance of fundamental spatial data. Spatial data matching is the key technology of updating and integration [1-2]. Nowadays many scholars have launched in-depth study about spatial data matching [3-10]. Although the spatial data matching is more important in theory, it still need a lot of human-computer interaction and editing work, which cannot satisfy the actual needs. The main reason is that the quality of the matching results cannot be guaranteed and the correctness of the matching results is generally by artificial judgment in most quality evaluation methods. By measuring the graphics data delta of settlement original matching entities, this paper establishes the grey correlation relationship between various kinds of graphic data delta indexes to research the matching quality evaluation methods and processes.

2. Grey correlation analysis in the application of the match

Grey correlation refers to the uncertainty relation between the internal factor and the subject itself. By measuring the geometric proximity between the internal factor sub-vector and main reference vector, the grey correlation method can analyze their correlation degree. As geometric approach degree is higher, the higher the degree of correlation, whereas the smaller.

Spatial data matching quality evaluation need to measure the comprehensive similarity or difference of evaluation indexes between different entities. This paper proposes space entity comprehensive difference measuring method by measuring weighted average of different types of differences, which is expressed as:
\[ d_{gg} = \sum o_i d_i \]  
\( i \)

Normally, the method of obtaining weight of each index only by expert experience is not scientific, because the importance of difference evaluation index is unknown. Therefore, according to grey correlation theory, this paper set various types of evaluation indexes as internal factors, comprehensive evaluation of each evaluation index as the main entity. By measuring the grey correlation, the weight which means the importance of each evaluation index to the final evaluation results can be obtained, and finally the comprehensive evaluation results can be got.

3. The selection of evaluation index and its measurement method

Spatial data matching is mainly based on geometry information, attribute information and spatial relation information. Among them, the difference of attribute and spatial relation information is difficult to measure standardly. Therefore, the paper defines geometric difference between entities as graphic data delta which is the main basis of matching quality evaluation. Generally, graphic data delta indexes are location, area, shape, length, angle, etc, this paper chooses these indexes: location, area and orientation which have mature technology and good accuracy.

The calculation formula of location data delta: 
\[ \Delta_{\text{location}}(O_{1k}, O_{2k}) = \sqrt{(x_{1k} - x_{2k})^2 + (y_{1k} - y_{2k})^2} \],

\( \sqrt{(x_{1k} - x_{2k})^2 + (y_{1k} - y_{2k})^2} \) is the centroid distance between entities \( O_{1k}, O_{2k} \).

The calculation formula of orientation data delta: 
\[ \Delta_{\text{orientation}}(O_{1k}, O_{2k}) = |\alpha - \beta| \], \( \alpha, \beta \) is the orientation angle of minimum circumscribed rectangle spindle of two entities \( O_{1k} \) and \( O_{2k} \).

The calculation formula of area data delta: 
\[ \Delta_{\text{area}}(O_{1k}, O_{2k}) = |\text{Area}(O_{1k}) - \text{Area}(O_{2k})| \], \( \text{Area}(O_{1k}) \) and \( \text{Area}(O_{2k}) \) are the area of two entities \( O_{1k} \) and \( O_{2k} \).

4. The quality evaluation of settlement data matching

4.1. Judging the correctness of the matching relationship between entities by grey correlation analysis

Step 1: Select sample entities.

To obtain weight it needs grey correlation analysis between “matching” and each index. So it is necessary to choose a certain number of sample entities which is matching correctly.

The measurement results of three graphic data delta indexes are divided into three grades:

1. small-difference(\( \Delta^s \)): difference value is small.
2. mid-difference(\( \Delta^m \)): a tolerance grade to make the measurement more comprehensive.
3. huge-difference(\( \Delta^h \)): existing massive difference, no similarity.

Sample selection criteria is as follows: When three indexes measurement results are both small-difference, or two are small-difference and another one is mid-difference, this group of settlement entities was chosen as sample entities. The formal expression: 
\[ (\Delta^s \land \Delta^s \land \Delta^s) \lor (\Delta^s \land \Delta^s \land \Delta^m) \rightarrow \text{select} \]

Step 2: Obtain weights.

Use grey correlation analysis method to determine the weights of three graphic data delta indexes:

1. Normalizing the initial measurements of the location, orientation, area data delta of sample entities. As a result, each kind of graphic data delta is comparable. The normalized formula:
\[ x_{i}(k) = \frac{x_{i}(k) - \min_{i} x_{i}(k)}{\max_{i} x_{i}(k) - \min_{i} x_{i}(k)} \]  
\( i \) is the serial number for each kind of graphic data delta, the value of \( i \) ranges from 1,2,3. \( k \) is the serial number for certain group of sample entities, the value of \( k \) ranges from 1 to \( m \), \( m \) is the number of sample entities. These definitions apply to all the formulas in this paper.

2. Build the sub-vector in three kinds of graphic data delta of sample entities.
\[ x_i = \{x_{i}(1), x_{i}(2), x_{i}(3), \ldots, x_{i}(m)\}, i = 1,2,3 \]
(3) Select the ideal one from each element of different sub-vectors to form the ideal reference vector

$$x_0 = \{x_0(1), x_0(2), \cdots , x_0(k), \cdots x_0(m)\}$$  \hspace{1cm} (4)

In the above formula, the ideal one is the minimum one:  \(x_0(k) = \min_{i \in \mathbb{I}} \{x_i(k)\} \).

When measuring certain group of sample entity \(k\), this paper uses the comprehensive association calculation formula \([11]\), the correlation coefficient \(\xi_{\omega}(k)\) of the ideal reference vector \(x_0(k)\) and the sub-vector \(x(k)\) is

$$\xi_{\omega}(k) = \frac{\min \{\min |x_0(k) - x(k)| + \rho \max \max |x_0(k) - x(k)|\}}{|x_0(k) - x(k)| + \rho \max \max |x_0(k) - x(k)|}$$  \hspace{1cm} (5)

In the above formula, \(\rho\) is the identification coefficient, the value of which is between 0 and 1. It usually sets as 0.5 to reduce the influence caused by too large of \(\max \max |x_0(k) - x(k)|\).

(4) at this moment, the correlation coefficient \(\gamma_\omega\) of the ideal reference vector \(x_0\) and the sub-vector \(x\) is

$$\gamma_\omega = \frac{1}{m} \sum_{i=1}^{m} \xi_{\omega}(k)$$  \hspace{1cm} (6)

Summarize the correlation coefficient of each kind of graphic data delta to reach the respective weight

$$\omega_i = \frac{1}{\gamma_\omega} \sum_{i=1}^{3} \gamma_\omega$$  \hspace{1cm} (7)

**Step 3:** Calculate the comprehensive graphic data delta.

Using the weights obtained in previous step to calculate the weighted average of each kind of graphic data delta to reach the comprehensive graphic data delta.

$$d_{\omega \gamma}(k) = \frac{\sum_{i=1}^{3} \omega_i \gamma_i(k)}{3}$$  \hspace{1cm} (8)

When \(d_{\omega \gamma}(k)\) is less than the threshold, matching relations is correct, otherwise not.

### 4.2. The quality evaluation process of settlement data matching

The quality evaluation of this paper needs three phases, as shown in Figure 1.

**Phase 1:** Measure the location, area and orientation graphic data delta of identical entity in the original matching data. According to the method of 4.1 step 1, select the sample entities.

**Phase 2:** According to the method of 4.1 step 2 and step 3, obtain the weight and calculate the comprehensive graphic data delta of identical entity which has not been selected as sample entity in phase 1, then judge whether the matching relationship is correct. The judging result of identical entity is correct matching when the identical entity is selected as sample in phase 1 or its matching relationship is correct in phase 2; the judging result of identical entity is wrong matching when its matching relationship is wrong in phase 2.

**Phase 3:** Summarize the non-identical entities in the original matching data which need quality evaluation and the identical entities which matching relationship is wrong in phase 2 to judge whether the matching relationship is correct according to the method of 4.1 step 2 and step 3. If the matching relationship is correct, these set of entities is missing matching.
5. **Experiment and Analysis**

To verify the effectiveness of the method, the following experiments has been carried.

![Diagram](image)

**Figure 1. The matching quality evaluation process**

- **Figure 2. Experiment data and matching results**
  - (a) Data A
  - (b) Data B
  - (c) matching data

The experimental data is shown as follows: Figure 2(a), settlement data A in map of a certain domestic urban area, the map scale is 1:25000; Figure 2(b), settlement data B in map of the same scale, same areas and different source. This experiment use the matching method based on degree of unilateral overlap to match data A and B, when the matching threshold \( \delta = 0.16 \), the matching results as shown in Figure 2(c), the identical entities are marked by centroid line. This matching method is widely used, easy to verify the effectiveness of the evaluation method in this paper.

Conduct the procedure of 4.2 phase 1. The sample entities which are selected as shown in fig.4. The identical entities which marked in red in Figure 3(a) and gray in Figure 3(b) are the sample entities of the procedure of 4.2 phase 1. Other identical entities that are need to be evaluated of next phase.

By analyzing sample entities, obtain weights of each kind of graphic data delta, as be shown in Table 1. When the judgment threshold \( \delta = 0.18 \), the judgment results of matching relationship correctness is shown in Figure 4.

| graphic data delta | location | area    | orientation |
|--------------------|----------|---------|-------------|
| weight             | 0.339107 | 0.374439| 0.286454    |

**Table 1. The weights of each kind of graphic data delta**
The identical entities that marked in yellow in the Figure 4(a) and blue in the Figure 4 (b) are the wrong-matching in phase 2. It can be concluded that the majority of wrong matching are the identical entities whose area delta is large in this phase. It turns out that the matching method based on degree of unilateral overlap is less effective and the method proposed in the paper to obtain weights is more scientific.

In phase 3, summarize the wrong matching entities [Figure 5(a)] in phase 2 and non-identical entities [Figure 5(b)] in original matching data to measure the comprehensive graphic data delta. When the data delta threshold $\delta = 0.15$, the evaluation results as shown in Figure 5(d).

As is shown in Figure 5, a pair of non-identical entities is missing matching. To summarize the above experiments, the evaluation results are shown in Figure 6 (a).

In order to verify the effectiveness of the proposed method, the cartographer-set-weight method is used for comparative analysis. According to the experience of 10 professional cartographers, set the weight as $(0.361162, 0.337359, 0.301479)$, when the threshold $\delta = 0.18$, the evaluation results as shown in Figure 6 (b). In Figure 6, the matching entities and missing matching entities are marked in red and grey, the wrong matching entities are marked in yellow and blue, the un-matching entities are marked in no color.
Figure 6. The final evaluation results

The evaluation results of the original matching data in the experiment are shown in Table 2. NC is the number of correct matching entities, NW is the number of wrong matching entities, NM is the number of missing matching entities, NJC is the number of judging correct, NJW is the number of judging wrong.

It can be seen that, the cartographer-set-weight evaluation method is incapable when it applied to judge the wrong matching and missing matching, but the method proposed in this paper works well in this aspect. The matching precision and recall are improved when the original matching results are adjusted by the method proposed in this paper.

| matching direction | evaluation method      | \(N_C\) | \(N_W\) | \(N_M\) | \(N_{JC}\) | \(N_{JW}\) | accuracy        | adjusted matching results | precision | recall |
|--------------------|------------------------|---------|---------|---------|-----------|-----------|----------------|--------------------------|-----------|--------|
| forward            | this paper             | 142     | 7       | 1       | 150       | 7         | 95.54%         | 95.07%                  | 85.98%    |         |
|                    | cartographer-set-weight| 146     | 3       | 0       | 145       | 12        | 92.35%         | 92.41%                  | 85.35%    |         |
| Reverse            | this paper             | 101     | 6       | 1       | 105       | 3         | 97.22%         | 97.03%                  | 90.74%    |         |
|                    | cartographer-set-weight| 104     | 3       | 0       | 101       | 7         | 93.51%         | 94.17%                  | 89.81%    |         |

6. Conclusion

The experimental results show that the features of the matching quality evaluation method based on grey correlation machine learning analysis proposed in the paper are:

1. it can measure the graphic data delta of identical entity in the original matching data quickly and easily;
2. it analyzes the correlation between each type of graphic data delta scientifically and obtains the weight;
3. it is rigorous and comprehensive, and it avoids the missing and missed matching.

The quality evaluation of spatial data matching is a complicated work. There are still many aspects to be improved in this method. In the next research, on the one hand, appropriate evaluation indicators should be added to make the difference measurement more comprehensive and reasonable. On the other hand, grey relational analysis method is used to solve the quality evaluation problem of other types of elements matching.

References

[1] Cooper, A. (2003) The Concepts of Incremental Updating and Versioning. In: Proceedings of the 21st International Cartographic Conference. Durban. pp. 855-857.
[2] Zhang, Q.P., Li D.R., Gong J.Y. (2004) Areal Feature Matching among Urban Geographic Databases. Journal of Remote Sensing, 8: 107-112.
[3] Cobb, M., Chung, M., Foley, H. (1998) A Rule-based Approach for the Conflation of Attributed Vector Data. GeoInformatica, 2: 7-35.

[4] Walter, V., Fritsch, D. (1999) Matching Spatial Data Sets: a Statistical Approach. International Journal of Geographical Information Systems, 13: 445-473.

[5] Beeri, C., Doytsher, Y., Kanza, Y. (2005) Finding Corresponding Objects when Integrating Several Geo-Spatial Datasets. In: Proceedings of the 13th Annual ACM International Workshop on Geographic Information Systems. Bremen. pp. 87-96.

[6] Safra, E., Doytsher, Y. (2006) Using matching algorithms for improving locations in cadastral maps. In: XXIII FIG Congress. Munich.

[7] Albakri, M., Fairbairn, D. (2012) Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. Taylor & Francis, Inc.

[8] Ruiz, J.J., Ariza, F.J., Urea, M.A., et al. (2011) Digital map conflation: a review of the process and a proposal for classification. International Journal of Geographical Information Science, 25: 1439-1466.

[9] Liu, C., Qian, H., He, H., et al. (2017) A Construction Method of Road and Residence Correlation Based on Urban Skeleton Network.

[10] Liu, S.F., Guo, T.B., Dang, Y.G. (1999) Theory and Application of Grey System. Science Press. Beijing.

[11] Luo, D., Liu, S.F. (2005) Study on the Method for Grey Incidence Decision-Making. Chinese Journal of Management Science, 13: 101-106.