RESEARCH ON THE METHOD OF USING MULTI-SOURCE DATA TO EVALUATE THE QUALITY OF GEOGRAPHICAL NAMES

Xu Yongmin[1,*], Zhang Jixian[1], Wang Xiaodi[1], Zhao Haitao[1], Chen chunxi[1], Mao Wenjuan[1]

1 National Quality Inspection and Testing Center for Surveying and Mapping Products , Number 28 Lianhuachixi Road, 100830 Beijing,China-121176049@qq.com

Commission III, ICWG III/IVb

KEY WORDS: Geographical Names, Multi-Source Data, Confusion Matrix, Weight, Consistence, Quality Evaluation.

ABSTRACT:

Based on the epidemic situation in recent years, it has become an urgent need to study how to use many open-source or existed geographical names to evaluate the quality of newly produced geographical names. By analysing the quality characteristics of geographical names data, we can find the key quality elements of geographical names data. We obtained a method of using multi-source data to evaluate newly produced geographical names. This method based on adaptive weight distribution of multi-source data. This method can not only evaluate the whole quality of geographical names, but also grade the specific quality of geographical names. The experiment shows that this method is effective and has strong reference for users and producers.

1. INTRODUCTION

With the development of Internet technology，positioning and navigation technology, people have more and more needs for positioning and navigation. As the primary information of this demand, the quality of geographical names directly affects the quality of navigation services. At present, there are many technologies to study how to obtain and produce geographical names, and there are few studies on how to use multi-source data to evaluate the quality of geographical names. Due to the impact of the Coronavirus pandemic and the reduction of production costs, it is becoming more and more urgent to use open-source and Internet geographical names to carry out quality evaluation on the produced geographical names. In fact, the quality of open-source data and Internet geographical names is uncertain. We have to use these data to evaluate the newly produced geographical names. How to use these existing data to evaluate the newly produced geographical names? It is a worthy topic to study. Therefore, we studied these data from the perspective of data quality evaluation benchmark. We obtained a method of using these data to evaluate newly produced geographical names. This method uses confusion matrix and weight. The weight is adjusted at any time according to the data. The conditions of weight adjustment are set in this paper.

2. QUALITY CHARACTERISTICS OF GEOGRAPHICAL NAMES DATA

Geographical names are the proper names of natural or human geographical entities in a specific spatial location. Geographical names are given by people[1]. The geographical entities it represents have a certain spatial location and scope on the earth's surface. Therefore, geographical names have the following characteristics:

Geographical name is a kind of name. Therefore, from the perspective of geographical information data, the attribute information of geographical name is the essential feature of geographical name.

Geographical names are the proper names of geographical entities in a specific spatial location. Therefore, from the perspective of geographical information data, geographical names have location and positioning.

The location of geographical names is integrated with the names of geographical names. When the location of geographical names is wrong, even if the names of geographical names are correct, they are still wrong geographical names; When the location of the geographical names is correct, but the name of the geographical names is wrong, it is also the wrong geographical names.

The location accuracy requirements of geographical names are different from other geographic information. Geographical names represent the names of geographical entities within a certain range and have a certain spatial location. Their location accuracy does not distinguish between terrain and landform, but is related to the economic development. The more economy developed, the smaller the scope represented by geographical names. Therefore, the higher the requirements for the location accuracy of geographical names.

Based on the above characteristics of geographical names, as a kind of data of geographic information data, geographical names data not only has the basic quality characteristics of geographic information data, but also has special quality characteristics. The basic quality characteristics of geographical names data include: location accuracy, attribute accuracy, time accuracy, logical consistency, data integrity, representation quality and attachment quality. Geographical entity name is a kind of geographic information data, so it is an important feature of geographical names. Therefore, the name accuracy is

* Corresponding author: Xu Yongmin, E-mail: 121176049@qq.com.

a key quality characteristics of geographical names. At the same time, the consistency between geographical names and location is the key quality.

3. QUALITY EVALUATION METHOD

As a kind of geographic information data, some quality elements evaluation of geographical names can refer to the evaluation methods of other geographic information data, for example: time accuracy, logical consistency, data integrity, representation quality and attachment quality. However, the quality evaluation of the name accuracy and the consistency between its name and geographical location has its own characteristics. This paper mainly studies the quality evaluation method of the name accuracy and the consistency between names and geographical locations. This method is based on multi-source data.

3.1 Evaluation process

Using multi-source data to evaluate the quality of geographical names includes multi-source data collection, multi-source data pre-processing, geographical names classification, geographical names matching, geographical names comparison, confusion matrix establishment, consistency calculation, weight adjustment and quality evaluation of geographical names. The specific evaluation process is shown in Figure 1.

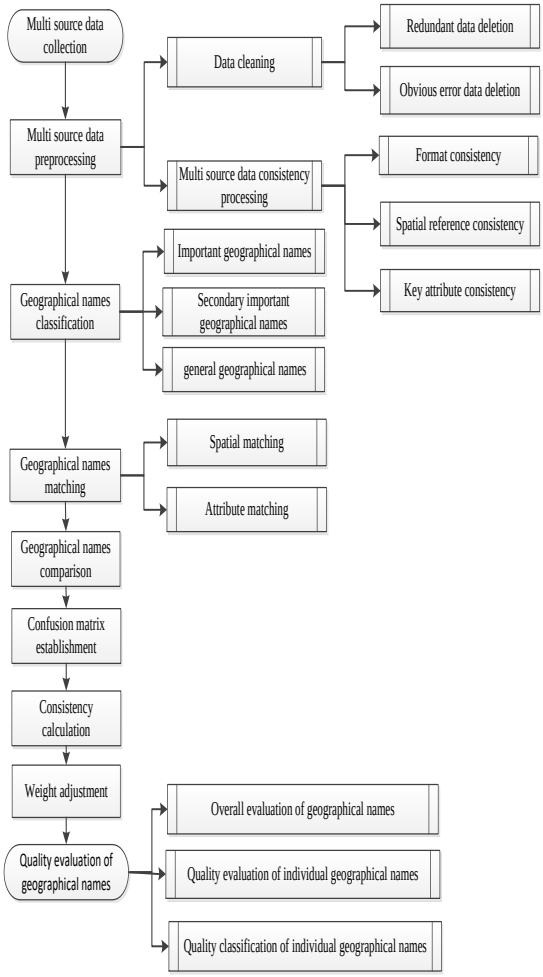

Figure 1. evaluation process of geographical names attributes

3.1.1 multi-source data Collection: We collected five types of geographical names data: GD, SW, SG, BD and CD. These data are reference data, they were used to evaluate newly produced data.

3.1.2 Multi source data pre-processing: including data cleaning and multi-source data consistency processing. We deal with the repetition and obvious error items of each type of data collected. We unified format conversion, coordinate conversion, and coordinate correction. We handled key attribute item names consistently.

3.1.3 Geographical names Classification: According to the importance of Geographical names in the application of navigation services, Geographical names is classified into important Geographical names, less important geographical names and general geographical names. Important geographical names refer to those that constitute the centre of gravity or support point of a city or region, such as administrative division geographical names; less important geographical names refer to those with geographic reference point significance, such as those that are often chosen as destinations by car travellers, and those that provide car travel services. General geographical names are small in size, unstable, highly parasitic, and weak in visibility, and they are not important geographical names and less important geographical names.

3.1.4 Geographical names matching: Geographical names matching includes spatial matching and attribute matching. Spatial matching uses the Euclidean distance of geographical names. Euclidean distance of geographical names is a matching factor to match geographical names. Attribute matching is based on spatial matching. Name, classification codes and other attribute information are the matching factor. In the process of geographical names data matching, the search radius and the matching attribute items are more important. The size of the search radius is related to the density of the distribution of geographical names and the types of geographical names. For administrative division geographical names, the smaller the administrative division level, the smaller the search radius; for other geographical names, the greater the place name density, the smaller the search radius. In this paper, the search radius range of 1 meter, 5 meters and 10 meters are used in the experiment, and 10 meters is finally search radius. In attribute matching, the name of geographical names and geographical names classification codes are selected as matching factors.

3.1.5 Geographical names comparison: There are two types of geographical names comparison: complete consistency and fuzzy consistency. Complete consistency means that the number of words and names of geographical names are completely consistent with the reference data, which can be achieved by machine matching; fuzzy consistency means that the number of words or spelling of the geographical names are not completely consistent with the reference data, but there is a certain proportion of consistency, that is, there is similarity. It can be determined as the same geographical names through similarity screening, manual investigation, etc.

3.1.6 Establish confusion matrix: respectively establish a confusion matrix for the comparison of geographical names and each type of reference data. Consistency or fuzzy agreement is represented by 1, and inconsistency or fuzzy inconsistency is represented by 0.

3.1.7 Consistency calculation: Calculate the complete consistency and fuzzy consistency between the geographical names to be evaluated and each type of reference data.

3.1.8 Weight adjustment: Determine the weight of each type of reference data source in the evaluation based on complete consistency, and determine the evaluation weight of different types of geographical names to be evaluated based on project experience and expert experience.

3.1.9 Quality evaluation of geographical names data: According to the consistency and the evaluation weight of each type of data, we can evaluate the comprehensive quality of the geographical names. According to the consistency between a single piece of geographical names and the reference data source, and the weight of reference data source, we can determine the quality of a single piece of geographical names data. We also can grade the quality of a single piece of geographical names in different quality levels. We can also know which geographical names are contained within each quality level.

3.2 Evaluation method

3.2.1 Statistics and calculate: According to the confusion matrix of the geographical names to be evaluated and each type of reference data, the complete consistency between a single piece of geographical names and each type of reference data can be obtained. The fuzzy consistency of a single geographical names and each type of reference data can be obtained. The statistical method of complete consistency is the same as the statistical method of fuzzy consistency. Next, we will illustrate it through the statistics of fuzzy consistency. For example: a single geographical names A is Fuzzy consistency with 4 types of reference data, then fuzzy consistency is 4 (all fuzzy consistency is 5). As shown in Table 1.

	A	B	C	D	E	fuzzy consistency
GD	1	1	1	0	1	4
SW	0	1	1	0	1	3
SG	1	0	1	1	0	3
BD	1	1	1	1	1	5
CD	1	1	0	1	1	4
fuzzy consistency	4	4	4	3	4	

Table 1. Example of fuzzy consistency for single geographical names.

Through the content of Table 1, we can get the number of geographical names that are fuzzy consistent with a certain type of reference data among the geographical names to be evaluated, so that we can calculate the fuzzy consistency ratio between the geographical names to be evaluated and a certain type of reference data, which can be called As the fuzzy consistency rate. The fuzzy consistency rate is obtained by dividing the total number of geographical names with fuzzy consistency by the total number of geographical names to be evaluated, and then multiplying by 100. The calculation formula is shown in Equation 1.

$$R_{fc} = \frac{N_{fc}}{N_{tl}} \quad (1)$$

where R_{fc} = fuzzy consistency rate
N_{fc} = number of geographical names that are fuzzy consistent with a certain type of reference data
N_{tl} = f number of total geographical names

Taking the content of Table 1 as an example, through the calculation of formula (1), we can obtain that the fuzzy coincidence rates of the reference data GD, SW, SG, BD, and CD are 80, 60, 60, 100, 80, and 76. As shown in Table 2.

type of reference data	fuzzy consistency
GD	80
SW	60
SG	60
BD	100
CD	80

Table 2. Example of fuzzy consistence rates.

In the evaluation process in Section 3.1, we can see that we have a link for classifying geographical names. We divide the geographical names to be evaluated into important geographical names, less important geographical names and general geographical names. Through the contents of Table 1 and Table 2, we can calculate again the fuzzy coincidence rate of important geographical names, less important geographical names and general geographical names with respect to each type of reference data. To better illustrate this issue, let us assume that A is an important geographical name, B and C are less important geographical names, and D and E are general geographical names. At this time, the total number of geographical names to be evaluated is 5, the number of important geographical names is 1, the number of less important geographical names is 2, and the number of general geographical names is 2. Through Table 1 and Formula 1, we can calculate the important geographical names, the less important geographical names and the general geographical names. The fuzzy consistence rates relative to each type of reference data, respectively, are shown in Table 3.

	important	less important	general
GD	100	100	50
SW	0	100	50
SG	100	50	50
BD	100	100	100
CD	100	50	100

Table 3. Fuzzy consistence rates for classified data.

3.2.2 Weight adjustment: Determine the weight of each type of reference data based on the complete consistency between the data to be evaluated and each type of reference data. Since open source geographical names data is also an important reference for producers, the weight of each type of reference data in this paper is inversely proportional to the complete consistency, that is, the higher the complete consistency, the smaller the weight of the reference data in the evaluation process.

3.2.3 Classification evaluation: According to the fuzzy consistence rate between important geographical names and each type of reference data, and the weight of reference data, we can evaluate the important geographical names quality. According to the fuzzy consistence rate between less important geographical names and each type of reference data, and the weight of reference data, we can evaluate the less important geographical names quality. According to the fuzzy consistence rate between general geographical names and each type of reference data, and the weight of reference data, we can evaluate the general geographical names quality. See Equation 1 for details. At the same time, the quality of a single geographical names can also be obtained by using Equation 2.

$$S_i = \sum_{j=1}^{n} R_{fcj} P_j, \quad (2)$$

where S_i =the score for each type of geographical names; or the score for a single geographical name
R_{fcj} = the fuzzy consistence rate of category i names with category j reference data; or fuzzy consistence rate of category i names with category j reference data
P_j = the weight of the j-th reference data
n=the number of the reference data

3.2.4 Overall evaluation: according to the scores of important geographical names, less important geographical names and general geographical names, and the weight of each type of geographical names , We can evaluate the overall quality of the geographical names data, see Equation 3.

$$S = S_{im} \times P_{im} + S_{le} \times P_{le} + S_{ge} \times P_{ge}, \quad (3)$$

where S = the comprehensive score of whole geographical names
S_{im} = the score of important geographical names
S_{le} = the score of less important geographical names
S_{ge} = the score of general geographical names
P_{im} = the weight of important geographical names
P_{le} = the weight of less important geographical names
P_{ge}= the weight of general geographical names

The weight of each type of geographical names adopts the empirical value. According to the project and experts based on experience, in this paper, the weight of important geographical names is 0.7, the weight of less important geographical names is 0.2, and the weight of general names is 0.1. So we evaluate the comprehensive quality of the geographical names data using Equation 4.

$$S = S_{im} \times 0.7 + S_{le} \times 0.2 + S_{ge} \times 0.1 \quad (4)$$

where S = the comprehensive score of whole geographical names
S_{im} = the score of important geographical names
S_{le} = the score of less important geographical names
S_{ge} = the score of general geographical names
0.7 is the weight of important geographical names
0.2 is the weight of less important geographical names
0.1 is the weight of general geographical names

3.2.5 Quality classification: Based on the overall evaluation score, we can classify the overall quality level of the geographical names. The division method is shown in Table 4. At the same time, we can also calculate the quality score of a single geographical name according to formula 1. According to the quality score of a single geographical name, the quality level of each geographical name can be divided according to Table 4, and then the quality level of each geographical name can be divided according to the quality level of each geographical name. In this way, we can know which place names are included in each quality class, providing a more specific reference for producers to improve quality.

Quality Score	Quality Level
90≤S≤100	excellent
75≤S＜90	good
60≤S＜75	pass
S＜60	fail

Table 4. Settings for quality level classification.

4. TEST

Based on the method proposed in this paper, 6064 geographical names were selected as the geographical names to be evaluated (Figure 1), and 5 types of reference data including GD, SW, SG, BD, and CD were selected as the evaluation reference data. 6064 geographical names are located in Fuzhou, Fujian Province. Fuzhou is located in the southeast coast of China. It is a city with medium economic development. The geographical names to be evaluated include the administrative division geographical names, government unit names, landmark building geographical names, school geographical names, catering and entertainment and other important, secondary and general geographical names. Using the method proposed in this paper, a confusion matrix was established for the geographical names and reference data in this area, and the complete consistency and fuzzy consistency between geographical names and various reference data were counted. Finally, the comprehensive quality of the data of 6064 geographical names is evaluated and analysed, and some geographical names are verified on the spot. The test results show that the overall evaluation of geographical names data using this method is basically consistent with the evaluation of the field inspection, and has a strong reference.

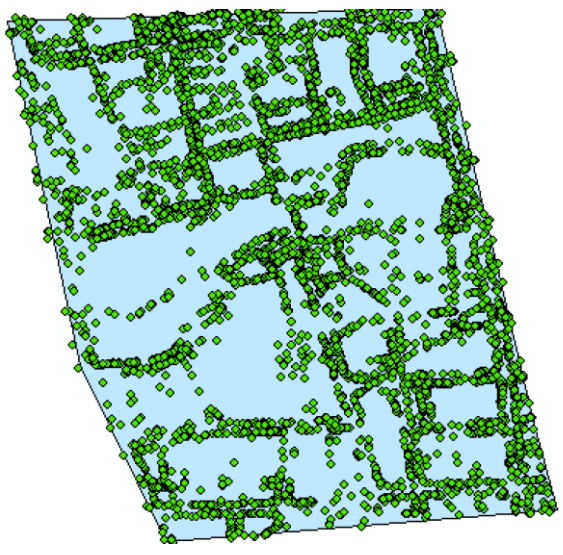

Figure 1. Figure test geographical names.

According to the evaluation process proposed in this paper, we first performed data cleaning and data processing on 5 types of reference data including GD, SW, SG, BD, and CD. We chose Euclidean distance as the spatial matching factor, and the search radius was 10 meters. The name of the attribute data is used as the first attribute matching factor, and two methods of exact matching and fuzzy matching are selected for matching. After the geographical names to be evaluated are completely matched with each type of reference data, the calculated complete consistence rate is shown in Figure 2.

Figure 2. Figure complete consistence rate.

According to the inverse ratio of complete consistence rate, we can determine the weight of each type of reference data in the evaluation. In this test the weight distribution of each type of reference data in the evaluation shown in Figure 3.

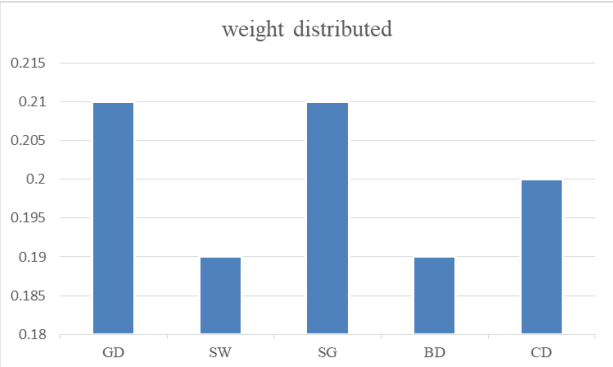

Figure 3. Reference data weight distribution diagram

Through machine text fuzzy matching, we get preliminary fuzzy matching results (partly see Figure 4, 5). For more than 1 fuzzy match, we performed manual analysis and finally confirmed the matching result. The fuzzy matching results of important geographical names, less important geographical names and general geographical names were classified and counted, and the fuzzy coincidence rates of the three types of geographical names and the five reference source data were obtained respectively. The details are shown in Table 5, and the distribution diagram is shown in Figure 6 shown.

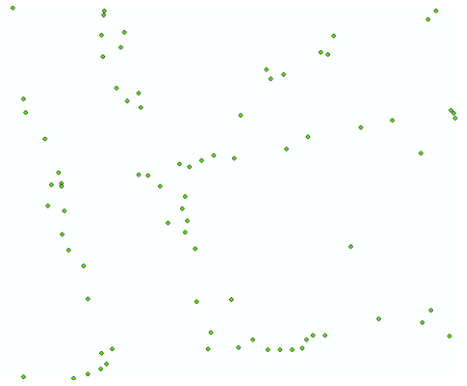

Figure 4. Geographical names to be evaluated(partly)

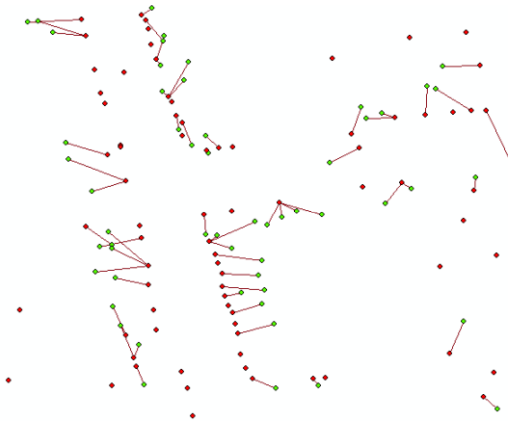

Figure 5. Fuzzy Consistency Matching Results(partly)

	important	Less important	general
GD	98	87	78
SW	96	89	80
SG	95	86	84
BD	99	90	82
CD	97	88	79

Table 5. fuzzy consistency of geographical names.

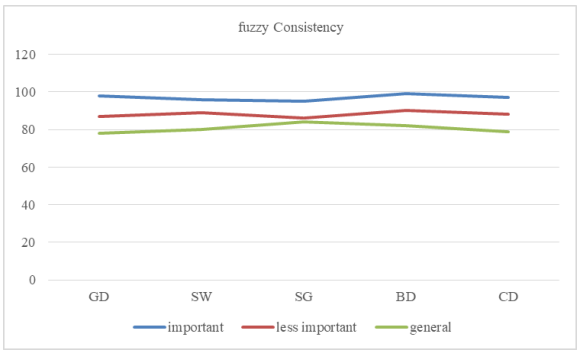

Figure 6. Fuzzy Consistency Distribution Diagram

Using the fuzzy coincidence rate and weight of important geographical names, less important geographical names and general geographical names, according to formula 1 and formula 3, the geographical names quality score of the test area

can be calculated. The final score for geographical names in this experimental area was 93 points.

40 geographical names were randomly selected in the sample area for field verification. The verification results found that 33 geographical names were consistent with the field geographical names, 3 geographical names had been abandoned, 1 geographical name had been changed, and 3 geographical names were wrong.

5. CONCLUSION

This paper proposes a geographical names data quality evaluation method. It used confusion matrix and weight. Different reference data have different weight. Different kind of geographical names to be evaluated have different weight. It has been verified by experiments. This method has certain reliability in evaluating the quality of geographical names. Evaluation results can provide a basis for users to better use data. It can also provide a more accurate reference for producers to improve the quality.

REFERENCES

Qin Xuexiu. J., 2016. Three Forms of Placename Data and Their Demand. Bulletin of Surveying and Mapping, 2011(10), 68-69

Niu Ruchen. J., 2016.What is the use of toponymy. China Geographical names, 2016(4),10-13

Fu Haojun, Fan Chengxiao, Zhang Haibo, J., 2019. Method of Geo-Name Automatic Acquisition and Information FusionBased on Vector Map Data. Journal of Geomatics, 44(2)，53-56

Zeng Yanwei, Tan Mingjian, etc, 2009. GB/T24356-2009, Specitications for Quality Inspection and Acceptance of Surveying and Mapping Products, China National Standard Management committee, Beijing, China

Kim J, Vasardani M, Winter S, J., 2017. Similarity matching for integrating spatial information extracted from place descriptions. International Journal of Geographical Information Science, 31(1), 56-80

Ji Xiaoyan, Zhou Min. J., 2006. A study of processing of global basic geographic base map database. Bulletin of Surveying and Mapping, 2006(7), 45-48

Cheng Gang, Lu Xiaoping. Matching algorithm for Chinese geographical names by similarity in consideration of semantics of general names for places. Acta Geodaetica et Cartographica Sinica, 43(4),404-410,418

Cao cunxiang, Huo liang, etc, J., 2019. Research on global name translation and matching method based on multiple data sources. Science of Surveying and Mapping. 44(7), 171-175

Du Haoqiang. J., 2014. A Discussion on Evaluation of the Data Quality for Geographic Name and Address. Geomatics & Spatial Information Tecchnology. 37(10), 209-211