RESEARCH ON CHINA WESTERN SPATIAL DATA QUALITY CHECK AND EVALUATION TECHNIQUE SYSTEM

Cao Yang [a, b,] *, Song Weidong [a], Lei Bing [b]

[a]College of Surveying and Geographic science, LiaoNing Technical University. No 47,Zhonghua Rd, Fuxin, China 123000 - caoyang1983628@163.com song_wd@163.net
[b]Key Laboratory of Geo-Informatics of State Bureau of Surveying and Mapping, CASM. No 16, Beitaiping Rd, Haidian District, Beijing, China 100039 - leibing@casm.ac.cn

Commission IV, SS-17

KEY WORDS: Spatial Data, Quality Control, Digital Linear Graphic, Cloud Model, Western Area

ABSTRACT:

The national western spatial data quality check method based on entity and evaluation method based on the cloud theory and rough set have been put forward according to the National Western Surveying and Mapping Project on 1:50000 Topological Maps Blank Area. In the first place, national western spatial data quality problem is analyzed and possible quality problems are described. Secondly, a digital linear graphic spatial data quality model is built and the quality elements and sub-elements are elaborately summarized. Next, spatial computing operators that the check process demands are given. Each concrete check rule is comprised of the union of different computing operators. And then, all check rules constitute spatial data quality check rules database. In the forth place, the weight of each index is calculated according to importance of attribute in rough set. Cloud decision generator transforms indexes value into qualitative evaluation. Finally, the homologous software of the spatial data quality check and evaluation is developed to control spatial data quality. Therefore a complete spatial data quality control and evaluation technique system is founded. It shows that the check and evaluation methods are feasible and software has higher automation from the experiment.

1. INTRODUCTION

Maximum Nowadays, with the rapid development of the information technique, the use of geographical information system is now being phased in. It is well known that the spatial data quality is the blood of GIS. Therefore it is absolutely essential and compulsory to control the spatial data quality.

The National Western Surveying and Mapping Project on 1:50000 Topological Maps Blank Area uses modern surveying method and high-tech equipments to collect more than 5000 sheets of 1:50000 topographic map which covers national western two million square kilometers land area. It also builds national western basic geographic information database, the related theme geographic information system and information-sharing service system to continuously monitor and effectively update the western geographic information changes. However if there is no effective spatial data quality control measure, the spatial database will exist in problems, errors, or even mistakes, which makes the related theme geographic information systems and information-sharing services not to support the decision. Therefore how to strictly check and scientific evaluate the spatial data quality in the process of building national western basic geographic information database is of great significance.

At present the GIS uncertainty studies mainly concentrate in the positional uncertainty and attribute uncertainty research. Moreover the theoretical basis of the studies originates from the traditional theory of probability and mathematical statistics, and the research is based on the random uncertainty of GIS quality control. Whereas the studies on how to control the spatial data quality during the data acquisition process to ensure quality requirements are more than few. The need for such kind of software that can check and evaluate the spatial data quality is more and more urgent. There is a bunch of domestic and foreign GIS data quality check software. But most of the software strongly depends on the specific GIS platform and has weak data exchange functions. Furthermore, its automation level is not high. The most important of all, it can never automatically describe the quality check results in quantity manner. Hence to develop a general use and high level automatic spatial data quality check and evaluation software is more important and valuable.

The paper is organized as the following: In the first place, China western spatial data quality problem is analyzed and possible quality problems are summarized. Secondly, a DLG spatial data quality model is built. Next, spatial data quality control based on geographical features is discussed. In the forth place, a new spatial data evaluation method based on cloud theory and rough set is put forward. Finally, the detailed introduction of the homologous check and evaluation spatial data quality software is given. Therefore a complete spatial data quality check and evaluation technique system is founded.

2. BUILDING THE SPATIAL DATA QUALITY MODEL BASED ON GEOGRAPHIC FEATURE

Spatial data storage format has two kinds: vector and raster. Different types of spatial data should have different model for

The research is financially supported by sub-project of Western China Surveying and Mapping: Comprehensive Database Construction and Data Quality Checking Software

data quality control. This paper is brought forward a referential template containing spatial data quality elements and sub-elements according to several ISO standards, the characteristics of the western region and the practical requirements, and what is more, builds the DLG quality model by means of principle of hierarchy and abstract reasoning.

2.1 The characteristics of China western spatial data

Because the western area is the barren or marginal region of China, the inspective content is different from the traditional. The inspective content primarily lies in river net, residential area and its facilities, traffic, pipeline, state boundary and administrative region, contour lines and earthiness, vegetation, place name and label.

Since the geographic environment of western region is special, it is impossible to adopt all the traditional spatial data quality control methods to check and evaluate the western spatial data. For example, the DLG are principally from satellite image, so it is infeasible to do the exploration and annotation on the spot.

2.2 The digital linear graphic spatial data quality model

Vector data is one kind of important spatial data, and what is more, DLG is the main source of vector data. To build DLG spatial data quality model should consider the factors in all directions, such as DLG product characteristics, the purpose of use and user requirements and the metadata. The terms “Space Reference”, “Positional accuracy”,” Attribute accuracy“,” Logical consistency“,” Completeness“,” Meta-quality“ are used for spatial data quality elements in the light of western data specificity. The DLG spatial data quality model is base on the quality elements (Table 1).

Element	Sub-element
Space Reference	Coordinate frame
Positional accuracy	Horizontal positional accuracy
Attribute accuracy	Nominal
	Attribute value
Logical consistency	Conceptual consistency
	Format consistency
	Geometric consistency
	Topological consistency
Completeness	Excess data
	Absent data
Meta-quality	Meta-quality

Table 1. The digital linear graphic spatial data quality model

3. THE CHINA WESTERN SPATIAL DATA QUALITY CHECK METHOD BASED ON ENTITY

The first thing to be done is to summarize the spatial computing operators that the check process demands. Next, each concrete check rule is comprised of the union of different computing operators. And then, all check rules constitute spatial data quality check rules database. Finally, the abstract point, line and area are construed as geographic entity, and a set of check proposals are formulated for each entity.

3.1 Spatial computing operator

Spatial computing operators are divided in three types: computing spatial relationships operators, computing topology operations operators and computing geometry operators. The computing spatial relationships operators include EqualsExact, Within, Contains, Crosses, Disjoint, Overlaps and Touches. The computing topology operations operators contain ConvexHull, Touches, Difference, SymDifference, Intersection and Union. The computing geometry operators consist of computing shape area operator and computing shape length operator.

3.2 Check rule

The check rules are classified by quality elements. Space reference check rules contain map border, kilometer grid and geographic system check rules. Positional accuracy check rule include geometric networks check rules, intersection check rules, overlap check rules, pseudo-node check rules, terminal vertex check rules, layer geometric validation check rules, feature validation check rules, geometric noise check rules, duplicate digitization check rules, geographical area check rules and sheet join check rules. Attribute accuracy check rule are mainly dataset layer name check rule, theme coding accuracy check rule, feature classification coding accuracy check rule, enumerated cell value accuracy check rule. Logical consistency check rules consist of spatial overlap, intersection, within, touches relationship check rules, place name and its Chinese phonetics consistency check rule, place name validation check rule, river flow direction check rule, river structure line attribute check rule, attribute table definition accuracy check rule, the elevation value consistency between elevation point and contour line. Completeness check rules are comprised of layer completeness check rule, feature completeness check rule, data format accuracy check rule, data file storage and absence check rules. Meta-quality check rule is used when checking the meta-quality.

3.3 Check proposal for western spatial entity

All the feature should be checked from the aspects of space reference, positional accuracy, attribute accuracy, completeness and meta-quality.

1) River net
River net should be checked by river flow direction check rule, river structure line attribute check rule and place name and its Chinese phonetics consistency check rule.

2) Residential area and its facilities
Residential area and its facilities should be checked by place name and its Chinese phonetics consistency check rule, spatial overlap, intersection, within, touches relationship check rules and attribute table definition accuracy check rule.

3) Traffic
Traffic should be checked by spatial overlap, intersection, within, touches relationship check rules and attribute table definition accuracy check rule.

4) Pipeline
Pipeline should be checked by spatial overlap, intersection, within, touches relationship check rules and attribute table definition accuracy check rule.

5) State boundary and administrative region

State boundary and administrative region should be checked by spatial overlap, intersection, within, touches relationship check rules, attribute table definition accuracy check rule, place name validation check rule.

6) Contour lines and earthiness
State boundary and administrative region should be checked by spatial overlap, intersection, within, touches relationship check rules, attribute table definition accuracy check rule, the elevation value consistency between elevation point and contour line.

7) Vegetation
State boundary and administrative region should be checked by spatial overlap, intersection, within, touches relationship check rules, attribute table definition accuracy check rule.

8) Place name and label
Place name and label should be checked by place name validation check rule and attribute table definition accuracy check rule.

4. THE SPATIAL DATA QUALITY EVALUATION METHOD BASED ON CLOUD THEORY AND ROUGH SET

A new spatial data quality evaluation method that combines cloud theory and rough set in the advantage of dealing with uncertainty information presents a new evolution way to assess spatial data quality. The weight of each index is calculated according to importance of attribute. Cloud decision generator transforms indexes value into qualitative evaluation.

4.1 Calculating index weights

4.1.1 Index: The index system is established on the grounds of the digital linear graphic spatial data quality model (see Figure 1). It has two levels, and the first level include six indexes, namely quality elements.

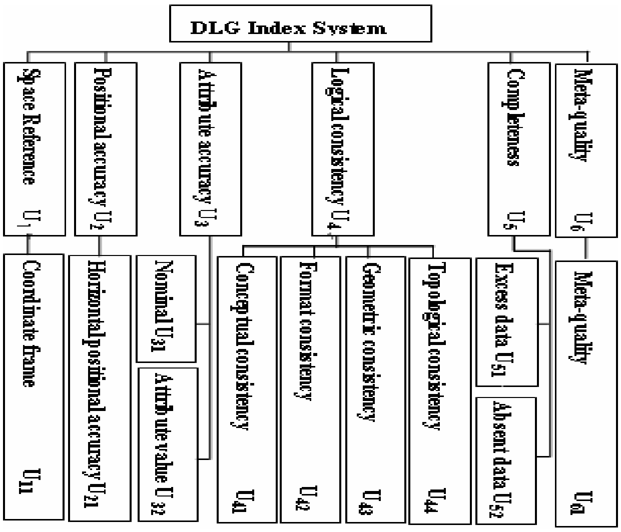

Figure 1. The digital linear graphic spatial data quality evaluation index system

4.1.2 Discretizing the indexes value: The 200 more sheets of DLG is evaluated through Defection Subtraction Score Method, and the results form the sample database. Figure 2 represents the index values of some sheets of DLG from the sample database.

DLG	**1**	**2**	**3**	**4**	**…**
U_1	91	82	63	59	…
U_{11}	95	82	62	59	…
U_2	97	91	78	30	…
U_{21}	97	90	77	31	…
U_3	92	85	66	62	…
U_{31}	88	84	69	64	…
U_{32}	97	90	63	63	…
U_4	96	82	72	61	…
U_{41}	94	76	73	59	…
U_{42}	91	80	70	60	…
U_{43}	100	84	69	62	…
U_{44}	97	88	61	60	…
U_5	91	83	70	65	…
U_{51}	89	85	74	70	…
U_{52}	93	81	60	60	…
U_6	93	81	67	60	…
U_{61}	92	81	68	60	…
Qualitative Evaluation	Excellent	Good	Qualified	Unqualified	…

Table 2. The original index values information

The statistical index values collected are all sequential real numbers, but the rough set can only deal with discrete attribute value, as a result, the indexes values are discretized firstly. All indexes value falls into four grades: excellent, good, qualified and unqualified according to Table 3. And for the sake of concision, 1 is used for representing excellent, 2 for good, 3 for qualified, 4 for unqualified. Table 4 is through data processing of Table 2.

Quality Grade	Index Value Interval
Excellent	90～100
Good	75～89
Qualified	60～74
Unqualified	0～59

Table 3. The relationship between index value and grade

DLG	**1**	**2**	**3**	**4**	**…**
U_1	1	2	3	4	…
U_{11}	1	2	3	4	…
U_2	1	1	2	4	…
U_{21}	1	1	2	4	…
U_3	1	2	3	3	…
U_{31}	2	2	3	3	…
U_{32}	1	1	3	3	…

U_4	1	2	3	3	…
U_{41}	1	2	3	4	…
U_{42}	1	2	3	3	…
U_{43}	1	2	3	3	…
U_{44}	1	2	3	3	…
U_5	1	2	3	3	…
U_{51}	2	2	3	3	…
U_{52}	1	2	3	3	…
U_6	1	2	3	3	…
U_{61}	1	2	3	3	…
Qualitative Evaluation	Excellent	Good	Qualified	Unqualified	…

Table 4. The discretized index values information

4.1.3 Calculating index weights: First of all，each index weight in the lowest level of index system should be calculated. Then each index weight in the second-lowest level is supposed to be calculated. Finally every index weight is calculated by parity of reasoning. Each and every attribute importance is different in term of certain specific problem. The importance of attribute a (a∈C) is estimated by decision table(U,C∪D) in rough set. U is a non-empty finite set of objects. C is a non-empty finite set of attributes. D is a non-empty finite set of decision attributes. And C∪D=Φ. Let D and C be subsets of A. D depends on C in a degree k denoted by $C \Rightarrow_k D,$ if

$$k = \gamma(C,D) = \frac{|POS_C(D)|}{|U|} \quad (1)$$

Where $POS_C(D) = \bigcup_{X \in U/D} \underline{C}(X),$ called *C*-positive region of *D*.
The common method to calculate the a importance is through the formula γ(C,D)-γ(C-{a},D)。Each sub-element weight is firstly calculated, and then is the element weight. For example when calculating the element weight, C={U_1,U_2,U_3,U_4,U_5,U_6},D={ Qualitative Evaluation }. Lastly all the weights in same level are processed so as to make their summation equal to 1. Table 5 lists each index weight.

4.2 Using the digital characteristics of clouds to represent qualitative evaluation

Cloud model is a model of the uncertain transition between a linguistic term of a qualitative concept and its numerical representation. In short, it is a model of the uncertain transition between the qualitative and the quantitative. The digital characteristics of clouds well integrate the fuzziness and randomness of linguistic terms in a unified way, which lays a foundation of knowledge representation. Cloud theory is made up of cloud model, uncertainty reasoning and clouds transformation. Cloud theory combines qualitative calculation with quantitative calculation, which can better resolve the spatial data evaluation.

Element	Element Weight	Sub-element	Sub-element Weight
Space Reference	0.1	Coordinate frame	1
Positional accuracy	0.2	Horizontal positional accuracy	1
Attribute accuracy	0.3	Nominal	0.6
		Attribute value	0.4
Logical consistency	0.2	Conceptual consistency	0.3
		Format consistency	0.2
		Geometric consistency	0.2
		Topological consistency	0.3
Completeness	0.15	Excess data	0.5
		Absent data	0.5
Meta-quality	0.05	Meta-quality	1

Table 5. The detailed index weights information

Cloud model has three digital characteristics: Ex (Expected Value), En (Entropy) and He (Hyper Entropy). Set U as a domain comprising accurate numbers. The expected value Ex is the position at U corresponding to the center of gravity of the cloud. In other words, the element Ex in the universe of discourse is fully compatible with the linguistic term. The entropy En is a measure of the coverage of the concept within the universe of discourse. It can be also considered as a measure of fuzziness of the concept. The hyperentropy He is the entropy of the entropy En. It is a measure of dispersion of the cloud drops.

Normal cloud is the most important kind of cloud model because the normal distribution has the application in almost every branch of social and natural scientific field. Using the digital characteristics of clouds to represent qualitative evaluation is to determine the three digital characteristics values according to the collected index value data, and then to generate the evaluation normal cloud. The main steps are as follows:

1) Classify the objects

Each object in sample dataset is classified according to qualitative evaluation type.

2) Calculating every object index value in each qualitative evaluation type

The index value vi of object xi in each qualitative evaluation type is calculated via the formula:

vi=xi(u1,u2, u3, u4, u5, u6)$\otimes$ W (2)

where ui is index value of certain quality elem$\otimes$t, means vector transvection, W is weight vector。

3) Using backward cloud generator to obtain the normal cloud of each qualitative evaluation

The index value vi of object xi has been inputted as a cloud drop to backward cloud generator ,and then backward cloud generator is used to get the normal cloud of each qualitative evaluation

Through the data processing in table 2 the normal cloud of each qualitative evaluation are obtained: Excellent =A1(100,6.0,0.01),Good=A2(82,5.0,0.005), Qualified =A3(67,5.0,0.005), Unqualified=A4(0,11.1,0.002)。

4.3 Transforms indexes value into qualitative evaluation by cloud decision generator

1) Calculating the indexes value of each quality element
The indexes value V is calculated by formula (2) too.

2) Calculating the membership degree of normal cloud
The index value V is regarded as the input of X-conditional cloud generator to get the membership degree of the object.

3) Deciding the evaluation grade via maximum membership principle

The qualitative evaluation of the object is gained according to the maximum membership principle.

For example, if the index values V of object X are (91, 82, 88, 85, 92, 90), hence the evaluation va⊗ble S=V W=91×0.1+82×0.2+88×0.3 +85×0.2+92 ×0.15 +90×0.05=87.2.Input S into the X-conditional cloud generator of normal cloud A1、A2、A3 and A4 to get the membership degree of A1,A2,A3 and A4. The results are 0.219、0.78、0.0008 and 0.0002. According to the maximum membership principle, the qualitative evaluation of object X should be good.

5. TECHNIQUE OF QUALITY CHECK AND EVALUATION SOFTWARE REALIZATION

5.1 System traits

The complicated spatial data quality characteristics are perceived as rule and model, which is used as the theoretic basis to check and evaluate the different data quality via program. Based on above research results, a kind of universal software is developed, which is named as Vector Data Quality Control and Evaluation Tool. It has the following features:

1) It has proprietary intellectual property rights. The software is developed totally based on C# and Microsoft Access,and can realize the vector spatial data input and visualization.

2) It is very flexible. Dynamically modifying the spatial control quality elements, weights and check proposals at any time is possible. In addition, the importance of geographic feature can be changed so as to better to classify the check error type.

3) It is the most convenient interactive software. Both the check and evaluation results can be visualized. The check result can be viewed on screen by means of detailed reports and highlight. The user is able to see the evaluation results though minute and detailed score table.

5.2 System overall structure

Figure 2 represents the software overall structure.

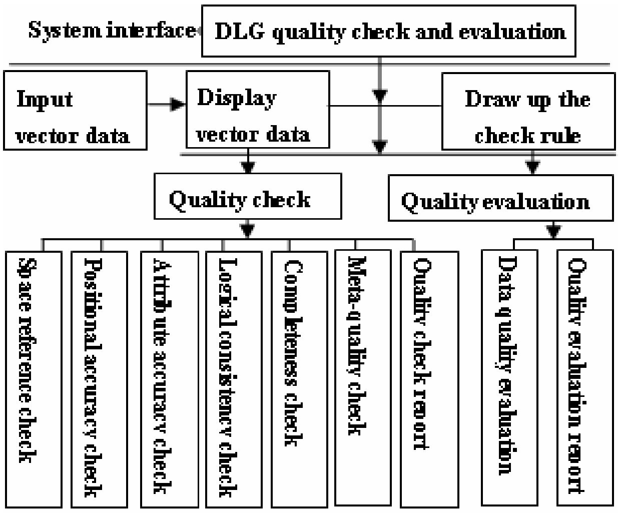

Figure 2 Software overall structure

5.3 System functions

1) Draw up check proposals
Because different productive sectors have different needs and requirements, there is need for formulate the check proposals interactively.

2) Input and display the vector spatial data
It calls for displaying the check error when checking the data, therefore the software should be able to input and display the vector spatial data.

3) Check data quality
DLG is checked in the aspect of the space reference, positional accuracy, attribute accuracy, logical consistency, completeness and meta-quality.

4) Show and confirm the check results
The check result can be viewed on screen by means of detailed reports and highlight. The user is able to zoom in on the error to confirm the check results.

5) Quality evaluation
DLG is evaluated based on the check results.

6) Output results
The detailed records of data quality check and evaluation results can be exported in Word document storage.

5.4 Quality check and evaluation work flow

Figure 3 represents the quality check and evaluation work flow.

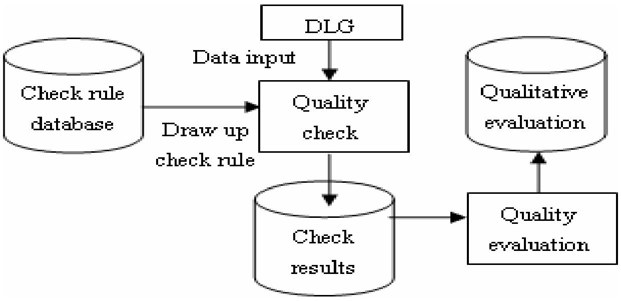

Figure 3 The quality check and evaluation work flow

6. CONCLUSIONS AND OUTLOOK

Firstly, the features need to be checked and data quality check rules are researched and concluded according to national area characteristics. A new spatial data quality evaluation method which applies cloud theory and rough set is put forward. The experiments shoes that this evaluation method is more objective and scientific. Each function in quality check and evaluation software has been tested by western spatial data. The outcome shows that this software can be used universally, the results of testing and evaluating are reliable and proper, and the test speed is fast. The research will continue with the completion of the check and evaluation software functions, such as pan, stop checking and so on.

ACKNOWLEDGMENT

The authors wish to thank sub-project of Western China Surveying and Mapping: Comprehensive Database Construction and Data Quality Checking Software for financial support.

REFERENCES

FAN Hong,ZHAN G Zuxun,DU Daosheng,2005. Quality Evaluation Model for Map Labeling. Geo-spatial Information Science, 8(1), pp.72~78.

GAO Jianxin, 2006. Research and Status on GIS Uncertainty. GEOSPATIAL INFORMATION, 4(5), pp.4~6.

Howard Veregin, David P. Lanter,1995. DATA-QUALITY ENHANCEMENT TECHNIQUES IN LAYER-BASED GEOGRAPHIC INFORMATION SYSTEMS. Comput Environ. and Urban System,19(1),pp.23~36.

HU Shiyuan,LI Deren,LI Deyi, 2007. Mining Weights of Land Evaluation Factors Based on Cloud Model and Correlation Analysis. Geo-spatial Information Science,10(3),pp.218~222.

Lars Harrie, 2003. Weight-Setting and Quality Assessment in Simultaneous Graphic Generalization. The Cartographic Journal, 40(3),pp.221~233.

LIU Dajie, LIU Chun,2001. The Status of Researching on Uncertainty of Spatial Data and Quality Control in GIS. ENGINEERING OF SURVEYING AND MAPPING, 10(1), pp.6~10.

Pepijn van Oort,2005. Spatial data quality: from description to application. Optima Grafische Communicatie, Optima Graphic Communication, Rotterdam.

Rui Li, Bir Bhanu_, Chinya Ravishankar, Michael Kurth, Jinfeng Ni, 2007. Uncertain spatial data handling: Modeling, indexing and query. Computers & Geosciences, 33 (1),pp.42~61.

Sylvain Bard,2004. Quality Assessment of Cartographic Generalisation. Transactions in GIS， 8(1),pp. 63~81.

ZHU Qin, CHEN Songli, HUANG Du,2004. Key Issues on Quality Standardization of Geospatial Data. Geomatics and Information Science of Wuhan Universit ,29(10),pp.863~866.