THE QUALITY ASSESSMENT AND SAMPLING MODEL FOR THE GEOLOGICAL SPATIAL DATA IN CHINA

Huan XIE[a, *], Xiao-hua TONG[a], Zuo-qin JIANG[b]

[a]Department of Surveying and Geo-informatics, Tongji University, No. 1239, Siping Road, Shanghai, 20092, P.R. China - xiehuantj@gmail.com, xhtong@mail.tongji.edu.cn
[b]Development and Research Center of China Geological Survey, No. 40, Xueyuan Road, Beijing, 100083, P.R. China - jzuoqin@cgs.gov.cn

Commission II, WG II/7

KEY WORDS: Sampling**,** Quality assessment**,** Geological data, Geographic information, ISO standards

ABSTRACT:

The sampling inspection model proposed by this paper is the key for the quality inspection and assessment of geological data, which is based on the present standards, research findings, and the quality characteristics of geological data. The sampling inspection model includes the geological data sampling schema, the sampling plans, the sampling method and the assessment method. This model solves the scientific sampling problems for geological data and has practical meanings on the data quality improvement. This model has been used in the Chinese standard of "Inspection and assessment for geological data". The main contexts of our research are: 1) Digital spatial geological data has biggish differences with traditional products on various characteristics, e.g. context, format, production process and quality characteristics. We can not simply apply the quality sampling methods which are commonly used for traditional industrial products. 2) The traditional percentage sampling method which is used in former data quality inspection has several disadvantages, so we choose a series of sampling plans to overcome them. Further, three types of sampling schema are proposed, i.e. sampling for one map sheet, sampling for multiple map sheets and sampling for attribute data. 3) Although differences existed in three sampling schemas, the sampling plans can be used for all sampling schemas are deduced in these paper. These sampling plans are based on the concept of accept quality limit (AQL). 4) Sampling methods, e.g. spatial sampling, judgmental sampling, stratified sampling, and probability sampling, are suggested according to the spatial characteristics and the uneven distribution of geological data. 5) Based on the thought that the data quality of geological spatial data is a comprehensive expression for several quality characteristics, an assessment method named the defect rate based data quality assessment method is suggested. Thus, the comprehensive data quality for all quality characteristics can be realized.

1. INTRODUCTION

1.1 General Instructions

Geology information (data) is a result of monitoring, researching, analysing and explaining the earth using many methods. The geology information is related to many specific fields, *e.g.* Geology, Geophysics, Geochemistry, Hydrology, Geological disaster and Drilling, and it is also the important basis for making decisions in exploring, developing, using, management and protection of resources. The collecting of geology data needs a lot of labours, resources and money. Moreover, it is impossible to re-collect part of the geological data. Therefore, guarantee the quality of geology data is of great significance in enhancing the scientific nature of the relevant decision-making.

Geological data has many characteristics *e.g.* complexity, huge data volume, multi-category, multi-source, multi-dimensional, multi-scale, therefore, sampling is an indispensable method in the in the assessment and evaluation of geological data. But the former sampling method used in the quality assessment of geological data is percentage sampling; this sampling method has the disadvantage that the sampling plan has different strictness on different population. The first sampling inspection standard is published by the Ministry of Electronics in 1978 and become national standard of China in 1987, *i.e.* GB/T 2828-87. The newest edition of this standard is GB/T 2828-2003(Yu 2004). In recent twenty years, the theories and applications of sampling got rapidly developed. Twenty two national sampling standards were published and different sampling inspection systems, *e.g.* lot by lot, by attributes, by variables, were set up. Therefore, based on the present standards and the research achievements, it is significant to solve the scientific sampling problems of geological data, which is also helpful to improve the quality of geological data.

1.2 Review on the present researches

The vast majority of geological data belong to spatial data; therefore, we can learn a lot from the present researches on the spatial data quality. The uncertainty and quality of spatial data is well-known a basis theory of geographic information science, and it's also an important issue to be resolved in the area of GIS theories and applications. Achievements were gained in related areas such as positional uncertainty, attribute uncertainty, spatial relationship uncertainty, spatial analysis uncertainty, data processing and quality control of spatial data, etc. These research results provide a theoretical basis for the spatial data quality control, sampling and testing theory.

* Corresponding author. This is useful to know for communication with the appropriate person in cases with more than one author.

The international standard on spatial data sampling is ISO 19114 published by the standard technical committee of digital geographic information and geomatics (ISO TC211). This standard, based on ISO 2859 and ISO 3951, suggested that the spatial sampling can be used for the sampling of spatial data and provide an example on how to define samples and how to consider the spatial characteristics during deciding the sampling method in the informative appendix of the standard.。The probability sampling methods and judgemental sampling which is non probability sampling method are recommended in this standard. In probability sampling methods, simple random sampling, stratified sampling and systematic sampling are most common used sampling methods.

In China, one of the earliest researches focusing on the quality and sampling methods of spatial data (Liu et al. 1999) analyzed the quality characteristics of GIS data and GIS products, discussed the process controlling and the sampling inspection of GIS data quality, and proposed the sampling schema which can be used in the quality assessment of GIS spatial data. This book suggests that the quality problems can be quantified by the number of defects, and define different levels of defects according to the degree that the problem may affect the quality of the data. We keep the definition of defects in our research.

"Specification for inspection, acceptance and quality assessment of digital surveying and mapping products" (GB/T 18316-2001) is a national standard which is published by State Bureau of Surveying and Mapping of P.R. China, and this standard is mainly used to assess the basic geographic spatial data. This standard has the same opinion with the above mentioned book in the definition of defects and the defect levels, but the sampling schema involved is only percentage sampling, and the sampling methods are simple random sampling and stratified sampling.
Above all, according to the present sampling standards and related researches, both probability sampling and judgemental sampling can be used in the sampling of geological data. The sampling schema can be either qualitative or quantitative. Related international standards are ISO19114, ISO2859-1 (corresponding national standard is GB/T2828.1-2003), ISO3951 (corresponding national standard is GB/T6378), ISO2859-3(corresponding national standard is GB/T13263).

Our paper will based on the present sampling standards and researches, considering the characteristics of geological data, studying the sampling model and quality assessment standard of geological spatial data.

2. THE QUALITY INSPECTION AND THE CHARACTERISTICS OF GEOLOGICAL DATA

Besides complex contents and the big amount of data volume, geological data has following quality related characteristics:

a) Major types of geological data

In geological data, both spatial data and non-spatial are included. According to previously practises on the assessment of geological data, the data are divided into three categories, *e.g.* spatial geological data; attribute geological data, and cartographical geological data. These three types cover almost all geological data except raster data.

b) Multi-datasets

The spatial geological data includes spatial dataset, cartographical dataset, metadata dataset and document dataset. The attribute geological data includes the above datasets except the cartographical dataset.

c) The characteristics of geological data production

The production of geological data in our country follows the same producing specification. The production departments will received the same training and produce the products separately. The production of one geological product always takes several years. And the assessment of the data is carried both during the producing process and the final inspection.

d) Full inspection and sampling are both suitable for the assessing of the quality characteristics of geological data

Considering the characteristics of geological data, the quality description frame is given in our paper. Data quality elements and data quality overview elements were defined. Those elements are the basis of quality inspection and assessment of geological data, they includes completeness, logical consistency, positional accuracy, thematic accuracy, decoration normalization and the corresponding quality sub-elements. Among this quality elements, completeness, logical consistency, and positional accuracy are suitable for being automatic inspected by the software, thus suitable for full inspection. While almost thematic accuracy and decoration normalization need to be inspected manually, thus sampling is more useful.

e) Individual varies in complexity and importance

According to the preliminary terms of sampling inspection, the individuals which composed the population should be consistency, which means the individual should be produced on the same condition at about the same period with the same sort, form, level, scale and ingredient. The geological data can almost meet these requirements, but several individuals have differences in the degree of complexity and significance.

Due to these characteristics, we got to know that sampling by attributes can be used to evaluate the quality elements such as the omitting, logical consistency, attribute and other un-continuous discrete variables. Although that some variables such as positional accuracy can be evaluated by continuous variables, we all transfer the continuous variables to discrete variables and all quality characteristics will be evaluated by sampling by attributes. The sampling by variables is not considered in our research now.

As we discussed, the production of geological data always last for several years. After the first period of production, especially after the first quality assessment, the quality of geological data becomes stable. Meanwhile, the inspection of geological data happens regularly, so this also meet the requirements of both continuous lots and stable quality, only the time between each lots is somehow to long. Sampling by attributes- lot by lot is one of the sampling models meet the requirements of the sampling of geological data. The sampling plan can be adjusted according to the quality of products. National standard GB/T2828.1-2003 is the corresponding sampling standard of the ISO2859.1. This standard has strict mathematic basis and broad application area. However, due to the special characteristics of geological data, how to use and improve this sampling standard and ISO19114 still need to be discussed and researched.

3. THE SAMPLING MODEL OF GEOLOGICAL DATA

3.1 Three types of sampling of geological data

Individual is the basic unit to make up the sample. In our research, any part of the dataset can be an individual. For geological data, map sheet, spatial feature, non-spatial objects and attribute items can all be considered as the basic unit of assessment. According different definitions of the individual, three types of sampling is suggested, *i.e.* sampling for one map sheet, sampling for multiple map sheets and sampling for attribute data. In these three types, the individual of sampling for the multiple map sheets is the map sheet, the individual and the population of the sampling for one map sheet are the spatial feature and the map sheet respectively, while the individual of the sampling for attribute data is the item.

3.2 The sampling schema using AQL

Adjusted sampling by attributes- lot by lot schema was used in our research and AQL (acceptable quality level) is used for the sampling of spatial geological data.

In one sampling plan, there are three parameters which are N, n and Ac (or Re), N means the number of population, n means the number of samples, and Ac (Re) determines whether these data is accepted or not according to the inspection result of samples. The plan is decided by the quality level- AQL and the population and can be directed found in the looking-up tables.

3.3 The same AQL for the sampling for one map sheet and the sampling for multiple map sheets

The AQL of a sampling plan is a level of quality routinely accepted by the sampling plan. It is generally defined as the percent defective (defectives per hundred units X 100%) that the sampling plan will accept 95% of the time. This means lots at or better than the AQL are accepted at least 95% of the time and rejected at most 5% of the time. Associated with the AQL is a confidence statement one can make. If the lot passes the sampling plan, one can state with 95% confidence that the quality level of the lot is equal to or better than the AQL (i.e., the defective rate of the lot < AQL). On the other hand, if the lot fails the sampling plan, one can state with 95% confidence that the quality level of the lot is worse than the AQL.

As we've discussed, in the quality assessment of geological data, when the individual is defined as the map sheet, the AQL means the percentage of unacceptable individuals. When the individual is defined as the spatial feature or the data item, the AQL means the percent defective. Thus we thought that the data range of AQL in GB/T2828.1-2003 was inconsistent with the practice situation of geological data. The AQL value of geological is concentrated in the range from 0 to 5 percents. We decreased the AQL spacing from 0 to 5.

Besides, the geological data has many quality characteristics and each quality characteristic might contain many inspection items. An item do not meet the quality requirements has only one defect. According to different influences to the quality of the data, the defects might have 2 to 6 levels. GB/T2828.1-2003 said that for each defect level, there should be a corresponding AQL. That will be somehow to complex to operate the sampling in practical operation. So we transferred the defects to the lowest level (that means the defect with the weakest influence to the data quality), and calculate the AQL after that.

3.4 The calculation of the AQL sampling plans

The AQL is a percent defective that is the base line requirement for the quality of the producer's product. The producer would like to design a sampling plan such that there is a high probability of accepting a lot that has a defect level less than or equal to the AQL. The probability can be calculated as

$$P_a(AQL) = \sum_{d=0}^{Ac} \frac{(n \times AQL)^d}{d!} e^{-(n \times AQL)} \tag{1}$$

Ideal operating characteristic curve would be perfectly perpendicular from 0 to 100% for a given fraction defective. Practical OC curves should be close to the ideal OC curve in which the lot with good quality should has high probability of acceptance while the lot with poor quality should has high probability of rejection (low probability of acceptance). In order to get a better operation characteristic curve, in the sampling plans of GB/T2828/1-2003 (ISO2859.1), with the increasing of the number of population-N, the number of samples- n increased, the probability of acceptance on the same AQL increased, while the ratio of n/N decreased.

Meanwhile, several inspection levels were given out in GB/T2828/1-2003 (ISO2859.1); different inspection level represents different sampling numbers. Decreasing the sampling number may increase the consumer's risk. The design of sampling plans should following these principles:

a) The lot with a quality of AQL should be accepted in a high probability, all probability of acceptance would be greater than 90% and less than 99%. One exception is that the sampling plans with acceptable number equals 0 have the probability of 88%.

b) The lot with a quality of AQL do not have a fixed probability of acceptance. On the same AQL level, the probability of acceptance increases with the increasing of the number of samples. With the same number of samples, the probability of acceptance increases with the increasing of the AQL.

c) In order to maintain the advantages of preferred numbers, the R5 and R10/2 preferred numbers were selected to decide the number of AQL and n in GB/T2828/1-2003 (ISO2859.1). Thus, in our research, base on the numbers of GB/T2828/1-2003 (ISO2859.1), the numbers of AQL and n were increased. R10 preferred numbers were used in both AQL and n, in which the common ration is $(10)^{1/10} \approx 1.259$, the number sequence is 10, 13, 15, 20, 25, 32, 40, 50, 65, 80, *etc.*, the number sequence of sample number is 2, 2.5(3), 3, 4, 5, 6, 8, 10, 13, 16, *etc.*. The using of preferred number in AQL and n makes a symmetry table of sampling plans. The preferred number also simplified the OC curves. If we list a table that the *AQL* as the row and the *n* as the columns and the value of the table is $n \times AQL$, then the $n \times AQL$s have the same value on one bevel.

According to these, our standard calculated the difference probability of acceptances, as we can see in Table 1. In this table Ac is the acceptable number, the numbers with parenthesis can not meet requirement b), thus in the sampling plan table, the number will be replaced by an arrow. Actually the numbers with underlines, e.g. 0.9583, 0.9786, 0.9834, 0.9857 do not meet the design requirements, either, but we can not overcome it. For instance, if $n{\times}AQL$ is 2.518, selecting Ac as 5, 6 or 7 can not meet the requiremens, but Ac is 6 will be better. In other circumstances, the selecting of Ac follows the same priciple.

Ac	*n×AQL*	P_{AQL}	*Ac*	*n×AQL*	P_{AQL}
0	0.1259	0.8817	6	2.518	0.9853
0/1	0.1589	(0.8531/ 0.9886)	7	3.170	0.9840
0/1	0.2000	(0.8187/ 0.9824)	8	4.000	0.9786
0/1	0.2518	(0.7774/ 0.9731)	10	5.024	0.9858
0/1	0.3170	(0.7283/ 0.9592)	12	6.325	0.9868
0/1	0.4000	(0.6703/ 0.9384)	14	7.962	0.9834
1	0.5024	0.9091	17	10.00	0.9857
1/2	0.6325	(0.8673/ 0.9736)	21	12.62	0.9896
2	0.7962	0.9531	25	15.89	0.9878
2/3	1.000	(0.9197/ 0.9810)	30	20.00	0.9865
3	1.262	0.9606	37	25.18	0.9898
4	1.589	0.9769	44	31.70	0.9850
5	2.000	0.983	55	40.00	0.990

Table 1. The probablity of acceptance (P_{AQL})of single normal sampling inspection

The sampling plans are easy to be obtained according to Table 1. Table 2 listed the designed value and practical value of P_{AQL}. We can found that the producer's risk varies from 1% to 12%. The sampling plans with *Ac* equals 0 have a relative lower P_{AQL}, if possible these plans should not be used.

n×AQL	*Ac*	P_{AQL} Designed	P_{AQL} Practical
0.1259	0	0.8817	0.8781-0.8869
0.5024	1	0.9091	0.9037-0.9245
0.7962	2	0.9531	0.9459-0.9595
1.262	3	0.9606	0.9569-0.9662
1.589	4	0.9769	0.9749-0.9814
2.000	5	0.983	0.9804-0.9875
2.518	6	0.9853	0.9828-0.9884
3.170	7	0.9840	0.9817-0.9881
4.000	8	0.9786	0.9734-0.9852
5.024	10	0.9858	0.9823-0.9896
6.325	12	0.9868	0.9839-0.9911
7.962	14	0.9834	0.9737-0.9897
10.00	17	0.9857	0.9799-0.9914
12.62	21	0.9896	0.9859-0.9939
15.89	25	0.9878	0.9844-0.9938
20.00	30	0.9865	0.9865-0.9920
25.18	37	0.9898	0.9840-0.9950
31.70	44	0.9850	0.9783-0.9937
40.00	55	0.990	0.9810-0.9939

Table 2. the probablity of acceptance (P_{AQL})of single normal sampling inspection - the designed and practical values

3.5 The adjustment of sampling plans

The adjustment of sampling plans mains emerged in two occasions, ***e.g.*** the adjustment during the inspection and the adjustment in re-inspection. In the data quality assessment of geological data, single normal inspection plan was first used, if the quality of some quality characteristics show a stable and improved quality, the sampling plan can be adjusted to reduced inspection. In re-inspection of geological data, producers should manager the problems found in the previous inspection. Reduced inspection is used for the products with an 'excellent' in the first inspection, normal inspection is used for the products with a 'pass', while tightened inspection is used for the products which are failed in the first inspection.

4. THE SAMPLING METHODS OF GEOLOGICAL DATA ASSESSMENT

Sampling method is the way to get samples. After the sampling plans of geological data is decided, this section will analyze the sampling methods may used.

4.1 Simple random sampling

Random sampling is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. When there are very large populations, it is often difficult or impossible to identify every member of the population, so the pool of available subjects becomes biased.

Simple random sampling is the basic sampling method to pick up samples.

4.2 Stratified sampling

Stratified sampling is commonly used probability method that is superior to random sampling because it reduces sampling error. A stratum is a subset of the population that shares at least one common characteristic. Examples of stratums might be males and females, or managers and non-managers. The researcher first identifies the relevant stratums and their actual representation in the population. Random sampling is then used to select a sufficient number of subjects from each stratum. "Sufficient" refers to a sample size large enough for us to be reasonably confident that the stratum represents the population. Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

Stratified sampling is in common use in geological data. When the map sheets in a lot is produced by different producers or the individuals differs in a comparatively large range, the stratified

4.3 Systematic sampling

Systematic sampling is often used instead of random sampling. It is also called an Nth name selection technique. After the required sample size has been calculated, every Nth record is selected from a list of population members. As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method. Its only advantage over the random sampling technique is simplicity. Systematic sampling is frequently used to select a specified number of records from a computer file.

For geological data, this sampling method is mainly used in the sampling of attributes.

4.4 Cluster sampling

Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondent within those areas is selected.

For geological data, this sampling method is mainly used in the sampling of attributes.

4.5 Spatial sampling

Considering the spatial distribution characteristics of geological data, the spatial sampling should not be simply sampled by areas. The spatial features should be considered in the spatial sampling of geological data. Judgemental by the experts, areas with sufficient spatial features (meet the requirements of the sampling plan decided by the population and AQL) would be selected for further inspection.

In our standard, different sampling methods were used in the three types of sampling. For the sampling for multiple map sheets, percentage sampling were changed, for instance, if the map amount is less than 8, judgemental sampling method will be implied. The stratified sampling method is used to solve the different complexities of individuals. Several principles were set up, e.g. the proportion of samples should be equal to the proportion of sub-population.

5. THE EVALUATION METHOD OF GEOLOGICAL DATA

In our standard, defects were used to describe the problems occurred in the data inspection. We classify the defects of geological data into 6 levels, i.e. fatal defect, very serious defect, serious defect, less-serious defect, light defect and less-light defect. A significant principle of the evaluation of the geological data quality is that the defects in the same level have approximately equal influences on the data quality.

Based on the defining and classification of defects, the evaluation and marking methods are suggested. The defective rate is used to evaluate the quality of geological data. The defects in variety levels are transformed to the lowest level of defect, i.e. less-light defect. Transformation indexes are given out. Integrated evaluation can be done for all datasets of geological data. This way of doing changes the traditional quality evaluation method which is evaluated by pass and fail.

The basic thoughts of defect based quality evaluation are:

a) Add up all defects in different levels.

b) Transform all defects into less-light defects and calculate the defective rate.

c) Taking AQL as the quality limit of the data using full inspection. Taking the average c/n in sampling plans as the quality limit of the data using sampling inspection. If V is the quality limit, the mark of the dataset can be calculated as

$$S_F=100-U_F\ (40/V_F)$$
$$S_S=100-U_S\times\ (40/V_S)$$
$$S=(S_FK+S_S\times(1-K))/2 \quad (2)$$

where S is the mark of dataset, U is the average defective rate, V is the quality limit, and the subscript F and S stand for full inspection and sampling inspection respectively. K is the proportion of full inspected items.

6. CONCLUSIONS

Sampling model is the key contents of the assessment and inspection of geological data. Considering the characteristics of geological data, base on the previous standards and researches, the sampling inspection model of geological data is proposed in our paper, including the sampling types, sampling plan, sampling methods and evaluation methods.

Above mentioned theories and results are already been used in the drafting of standard <The inspection and evaluation of geological data> and practical inspection proved that our methods can reduce sixty percents of labour. Our results are also been used to several project, ***e.g.*** Establishment of a Quality Model for the 1:1000 Topographic Data (2005-2006) of Hong Kong, and Data Quality Model for Macao 1:1000 Digital Data (2006). Corresponding specifications were also set up in related departments.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (Project No. 40771174 and 40301043), Program for New Century Excellent Talents in Universities (Project No. NCET-06-0381), Open Research Fund Program of the Geomatics and Applications Laboratory, Liaoning Technical University, China (Project No. 2006005), and grants from the Key Laboratory of Geo-informatics of State Bureau of Surveying and Mapping (Project No. 200618). We are also very grateful to persons and institutions, which provided accurate and reliable data for the original criteria.

REFRENCE

Cheng, J.C., Guo, H.D., Shi, W.Z., *et al.*, 2004. *Uncertainty of remote sensing data*, Publishing House of Science, Beijing.

China Geological Survey, 2001. *File format guide of the layers and attributes spatial database of regional hydrogeological map*. China geological Survey, Beijing.

China Geological Survey, 2002. *Guide of the building of geological map spatial database*. China geological Survey, Beijing.

Feng, S.Y., Shi, X.Q., 1996. *Sample survey - Theory, Methods and Practice*. Shanghai Science and Technology Publishing House, Shanghai.

Institute of Standardization for Surveying and Mapping, 2001. *GB/T 18316-2001 Specifications for inspection, acceptance and quality assessment of digital surveying and mapping products*. State Bureau of Quality and Technical Supervision, Beijing.

Hu, S.W., 2006. *GIS quality evaluation and reliability analysis*. Surveying and Mapping Press, Beijing.

ISO 19113:2002 *Geographic information -- Quality principles*

ISO 19114:2003 *Geographic information -- Quality evaluation procedures*

Liu, D.J., Shi, W.Z., Tong X.H., Sun, H.C., 1999. *The accuracy analysis and quality control of GIS spatial data*. Shanghai Scientific and Technological Education Publishing House, Shanghai.

Wang, J.Y., 2001. T*he principles of spatial information systems.* Publishing House of Science, Beijing.

Wang, X.Z., Shi, W.Z., Wang, S.L., 2003. *Fuzzy spatial information processing*. Publishing House of Wuhan University, Wuhan.

Shi, W.Z., 2005. *Principle of modelling uncertainties in spatial data and analysis*. Publishing House of Science, Beijing.

Tang, G.A., Liu, X.J., Lu, G.N., 2005. *The principles and methods of digital elevation model and geoanalysis*. Publishing House of Science, Beijing.

Xiao, H., Ma, Y.L., Zhang, Y.Z., *et al.*, 2003. *GB/T 2828.1-2003 Sampling procedures for inspection by attributes-Part1: Sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection*. General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Beijing.

Yu, S.Q., 1991. *Sampling Inspection and Quality Control*. Peking University Press, Beijing.

Yu, Z.F., 2004. *The latest practical guide of statistics sampling inspection and process control*. Standards Press of China, Beijing.

Zhu, C.Q., Shi, W.Z., 2006. *The principles and modelling of spatial analysis*.