Quality Assessment of Attribute Data in GIS Based on Simple Random Sampling

LIU Chun  SHI Wenzhong  LIU Dajie

1 Introduction

Recently, some new methods have been involved in the research on the accuracy of attribute data\(^{[1-2]}\), for example, the fuzzy set theory is used to measure the uncertainty of attribute classification\(^{[3-4]}\), rough set\(^{[5]}\) to discuss the description and measurement of spatial data classification accuracy\(^{[6-8]}\).

The existing research of attribute data accuracy assessment pays more attention to the raster data than the attribute data in vector GIS. The analysis technologies focusing on the misclassification of remote sensing data may be insufficient when dealing with attribute data accuracy in vector GIS because they are complex and without quantitative value. So it is important to find a method for assessing the attribute data accuracy in vector GIS during the data collecting and data application.

2 Rate of Disfigurement Model

Most attribute data is stored with data record in database and lack of logical relation between each other so that it is difficult to quantify its quality. The sampling method is effective to obtain the comprehensive assessment of the attribute data quality. So we discuss here:

1. With an attribute data set \( X \), sampling data unit with the size of \( n \) from its total data set size \( N \);  
2. After quality sampling inspection with the sampling size \( n \), a valid quality sampling inspection with the sampling size \( n \), a valid statistic index \( a \) should be found according to the result of the quality sampling inspection, and the statistic index...
is expected to represent the quality of the attribute data set effectively, so the \( a \) is the mathematic model used to assess the attribute data accuracy.

The rate of disfigurement (RD) index is a mathematic model that is be brought forward to measure the quality of the attribute data. The RD model means the number of defects in certain number of attribute data briefly. So a valid RD index can be used to measure the quality of the attribute data for it includes the number of defects from the quality sampling inspection.

Each attribute data set has a total data size, denoted by \( N \). It is used to represent the data size of the whole attribute data set. That is to say, if a data unit is taken as an inspection unit, \( N \) is the total of all attribute data sets. In the meantime, the set also has the sampling size denoted by \( n \), which is used to represent the data size of the inspected attribute data. That is to say, if a data unit is taken as an inspection unit, \( n \) is the sum of the whole inspected data sets.

In mathematics, an attribute data set is termed as \( X \), its total is \( N \) (the total size of \( X \) is \( N \)), and the simple random sampling without replacement is used to inspect the data sets. If the sampling size is \( n \), the number of the obtained defects for sampling is \( y \).

The sampling size is \( n \), and the sampling result always has two statuses as,

\[
y_j = \begin{cases} 1, & \text{if } j \text{ unit is a defects} \\ 0, & \text{if } j \text{ unit is not a defects} \end{cases}
\]

where \( j \) represents the \( j \)th sampling unit for the sampling procedure. So the number of the defects \( y \) is,

\[
y = \sum_{j=1}^{n} y_j
\]

The ratio of \( y \) to \( n \) is termed as the estimation of RD for sampling, and represented by

\[
a = \frac{y}{n}
\]

If the total unit size is \( N \), and the number of its defects is \( Y \), \( a \) is regarded as the estimation of RD of the total data set \( \left( U = \frac{Y}{N} \right) \).

The number of defects and RD of attribute data can be obtained by the method of sampling inspection, and the defects can be judged according to the digital data standard or the known value. So it is simple and practical to use the mean and variance of the RD to assess the quality of attribute data.

### 3 Mean and variance representation of statistical RD model

The estimated value of RD has a connection with the sampling size, sampling schemes and the result of sampling inspection because it is a statistical model. So the estimation of RD will be obtained by a certain sampling method and then be used for quality assessment of attribute data. To a statistical index, mean and variance are generally used to describe statistical characteristic. So the mean and variance of statistical RD model must be obtained before the attribute quality assessment in order to master the statistical features, and further prove its application feasibility. Here, the mean and variance of statistical RD model is conducted under a general definition of sampling.

For an attribute data set \( X \) whose total data size is \( N \), considering the general circumstance, sampling arbitrary data unit in the total attribute data set and Eq. (1), let

\[
a_j = \begin{cases} 1, & \text{if } j \text{ attribute data unit is sampled} \\ 0, & \text{if } j = 1,2,\ldots,N \text{ other} \end{cases}
\]

then the sum of the defects in the attribute data set will be

\[
y = \sum_{j=1}^{n} y_j = \sum_{j=1}^{N} a_j y_j
\]

Simple random sampling without replacement is adopted, so each data unit has the same sampling probability, and the probability distribution for \( a_j \) is

\[
p = \begin{array}{c|cc}
  a_j & 1 & 0 \\
  \hline
  p & \frac{n}{N} & 1 - \frac{n}{N} \\
\end{array}
\]
Thus the mean of $a_j(E(a_j))$ is

$$E(a_j) = P(a_j = 1) = \frac{n}{N} \quad (6)$$

According to Eq. (3), the mean of the estimation of RD $a$ ($E(a)$) is

$$E(a) = E(\frac{Y}{N}) = \frac{1}{n} \sum_{j=1}^{N} E(a_j) y_j$$

$$= \frac{1}{n} \cdot \frac{N}{N} \sum_{j=1}^{N} y_j = \frac{Y}{N} = u$$

$$\quad (7)$$

where $Y$ is the sum of defects in total attribute data set; $u$ is RD of total attribute data set.

Here $a_j$ has only two different values of 0 or 1, so $a_j$ obeys the binomial distribution according to the principle of statistic. As a result,

$$E(a_j) = \frac{n}{N}, V(a_j) = \frac{n}{N} - \frac{1}{n} \quad (8)$$

$$\text{cov}(a_j, a_k) = -\frac{n}{N(N-1)}(1 - \frac{n}{N}) \quad (9)$$

where $E(a_j)$ and $V(a_j)$ are the mean and variance of $a_j$, and $\text{cov}(a_j, a_k)$ is the correlation coefficient between $a_j$ and $a_k$.

Denoting $f = \frac{n}{N}$, we can obtain the Variance of the estimation of RD $a$ ($V(a)$):

$$V(a) = V(\frac{1}{n} \sum_{j=1}^{N} y_j)$$

$$= \frac{1 - f}{nN} \left( \sum_{j=1}^{N} y_j^2 - \frac{2}{N-1} \sum_{j=1}^{N} \sum_{k>j}^{N} y_j y_k \right)$$

$$\quad (10)$$

For $(\sum_{j=1}^{N} y_j)^2 = \sum_{j=1}^{N} y_j^2 + 2 \sum_{j=1}^{N} \sum_{k>j}^{N} y_j y_k$, Eq. (10) can be altered to

$$V(a) = \frac{1}{n} - f \left[ \frac{1}{N-1} \sum_{j=1}^{N} y_j^2 - \frac{1}{N} (\sum_{j=1}^{N} y_j)^2 \right]$$

$$= \frac{1 - f}{nS^2} \quad (11)$$

where

$$S^2 = \frac{1}{N-1} \sum_{j=1}^{N} (y_j - Y)^2$$

$$= \frac{1}{N-1} \sum_{j=1}^{N} (y_j^2 - NY)^2 \quad (12)$$

Because $y_j$ has two different values of 0 or 1, and $\sum_{j=1}^{N} y_j = \sum_{j=1}^{N} y_j$, Eq. (12) can be altered to

$$S^2 = \frac{1}{N-1} (Nu - Nu^2) = \frac{N}{N-1}u(1-u) \quad (13)$$

Substituting Eq. (13) into Eq. (11), we have

$$V(a) = \frac{N-n}{n(N-1)}u(1-u) \quad (14)$$

The sampling variance $S^2$ should be estimated by the sampled variance $s^2$, and RD $u$ is also unknown and to be estimated, so

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y}) = \frac{na(1-a)}{n-1} \quad (15)$$

From Eq. (11), the estimation of variance of RD $a$ ($V(a)$) is

$$V(a) = \frac{1-f}{n} \cdot \frac{n}{n-1} a(1-a) \quad (16)$$

Because the estimation of $a$ is mainly influenced by $f$ but $f$, if $f = n/N$ is small enough ($f < 0.05$ or $f < 0.1$), then $1-f \approx 1$ can be considered. In the meantime, if RD $a$ is also small enough, Eq. (16) can be written as

$$V(a) \approx \frac{a(1-a)}{n-1} \quad (17)$$

then the mean and variance representation of the statistical RD model is gotten by the simple random sampling with the sampling size of $n$.

$$E(a) = u$$

$$V(a) \approx \frac{a(1-a)}{n-1} \quad (18)$$

When $n/N$ is small, Eq. (18) is the estimation of the mean and variance of RD $a$, which are the two main statistical characteristic values of the RD model. As a result, no bias is included when the RD model is used to assess the attribute data, so the mean of the RD model is the estimation of the attribute data while the variance is the confidence estimation of the quality of the attribute data. The generalization of the two statistical characteristic values can realize the assessment of the attribute data quality.

### 4 Assessment of attribute data quality for single digital map

As the quality of a single data set needs to be inspected (e.g. the attribute data of a single digital map), the whole attribute data can be regarded as a sampling inspection data set, so the inspection schemes can be described as follows: obtain the sample data whose data size is $n$ from the total data set
N, then the sum of the defects y can be gotten by quality inspection, and the RD value a can also be computed. The total data set size \( N \) can be explained by example of attribute data of digital road map. Total 22 data fields have been listed in Table 1 as partial attribute data description for digital road. So the total attribute data set size \( N \) can be \( 1000 \times 22 = 22000 \) if each single digital map includes 1000 digital roads.

| ID | Attribute field         | ID | Attribute field                   |
|----|-------------------------|----|-----------------------------------|
| 1  | Serial number of road   | 12 | Limited length of vehicle         |
| 2  | Average speed           | 13 | Limited weight of vehicle         |
| 3  | Traffic direction       | 14 | Limited width of vehicle          |
| 4  | Length                  | 15 | Carriageway number                |
| 5  | Width                   | 16 | Most carriageway number           |
| 6  | Road degree             | 17 | Least carriageway number          |
| 7  | Road form               | 18 | Limited traffic capacity          |
| 8  | Road function           | 19 | Pay road                          |
| 9  | Status of road          | 20 | Limited number of vehicle         |
| 10 | Limited height of vehicle | 21 | Ownership                         |
| 11 | Restriction of passing | 22 | Open time                         |

According to the principle of statistic, when the total attribute data size \( N \), sample size \( n \) and the difference \( (N-n) \) are big, the sum of defects y will approximately obey the normal distribution, so the assessment of the quality of the attribute data for a single digital map can be described as follows.

1) The sample estimation of RD \( \hat{a} \) can be used to assess the quality of the attribute data for a single digital map. The smaller \( \hat{a} \) is, the better the quality of the attribute data is. Conversely, the larger \( \hat{a} \) is, the worse the quality will be.

2) The standard variance of the sample estimation of RD is \( \sqrt{\text{Var}(\hat{a})} = \sqrt{\frac{\hat{a}(1-\hat{a})}{n-1}} \), which can represent the sampling accuracy and confidence of the estimation of RD.

3) The condition equation \( \Pr(u-a<d) = 1-\alpha \) can be explained that the absolute bias of the estimation value of RD from the real value will be smaller than a constant under a certain confidence. When the sample size is large enough, \( a \) can be considered to obey the normal distribution \( N(a, \text{Var}(a)) \), so

\[
\frac{u-a}{\sqrt{\text{Var}(a)}}
\]

obeys the normal distribution of \( N(0,1) \). Then the upper limit of \( a \) can be expressed as

\[
u_\alpha = a + \mu_{1-\alpha} \sqrt{\frac{a(1-a)}{n-1}}
\]

where \( \mu_{1-\alpha} \) is the division point of the normal distribution under the confidence of \( 1-\alpha \).

As \( \hat{a} \approx \frac{\hat{a}}{n} \), then

\[
u_\alpha = \hat{a} + \mu_{1-\alpha} \sqrt{\frac{\hat{a}}{n}}
\]

When \( \hat{a} < \nu_\alpha \), the RD value \( a \) can have the confidence of \( 1-\alpha \) in assessing the attribute data quality.

5 Assessment of attribute data quality for a batch of digital maps

The quality of a batch of data sets \( (K \) data sets, e.g. attribute data sets of \( K \) digital maps, needs to be inspected when \( K \) data sets conduct the quality assessment. The inspection schemes can be described as follows: get sample data sets \( n_k (k=1,2,...,K) \) from the \( k(k=1,2,...,K) \) data sets whose total sizes are \( N_k (k=1,2,...,K) \), respectively. For each data set, the sum of the defects \( y_k \) can be gotten by quality inspection, and the RD value \( a_k \) can also be calculated.

Thus, the total estimation of RD for the \( K \) data sets can be calculated by

\[
\hat{a} = \frac{1}{k} \sum_{k=1}^{K} a_k
\]

So the assessment of the quality of the attribute data for a batch of digital maps can be described as follows.
1) The sample estimation of RD $\tilde{u}$ is used to assess the quality of the attribute data for a batch of digital maps. The quality of the attribute data of a batch of digital maps is better when the value $\tilde{u}$ is smaller.

2) The standard variance of the sample estimation of RD for a batch of digital maps is

$$\sqrt{\text{Var}(\tilde{u})} \approx \sqrt{\frac{\tilde{u}(1-\tilde{u})}{n-1}},$$

which can represent the sampling accuracy and confidence of the estimation of RD for a batch of digital maps.

3) Same to the single digital map, the upper limit of the quality of attribute data for such a batch of digital maps can be given as

$$u_c = \tilde{u} + \mu_{(1-\alpha)} \sqrt{\frac{\tilde{u}(1-\tilde{u})}{n-1}}$$

When the estimation of RD of attribute data $\tilde{u} < u_c$, the quality of the attribute data of a single digital map is under control and has the confidence of $1-\alpha$.

As an example, the attribute data of a batch of digital maps need to be inspected for quality. The size of the attribute data for each map is about 4,000, and about 350 attribute records from each map are sampled. The confidence level is $\alpha=0.05$, hence the RD value of attribute data for such a batch of digital maps is given in Table 2.

When $n=350$, the mean of RD for a batch of digital maps is:

$$\tilde{u} = \frac{1}{25} \sum_{k=1}^{25} u_k = 0.01872$$

Here, $\tilde{u} = 0.01872$ is the quality assessment value for attribute data of such a batch of digital maps.

Without considering $f$, the upper limit of attribute data quality RD value for such a batch of digital maps is

$$u_c = 0.01872 + \mu_{0.05} \sqrt{\frac{0.01872(1-0.01872)}{350-1}} = 0.030645$$

With considering $f$, the upper limit of attribute data quality RD value for such a batch of digital maps is

$$u_c = 0.01872 + \mu_{0.05} \sqrt{\frac{0.01872(1-0.01872)(1-f)}{350-1}}$$

Table 2: RD values for a batch of digital maps

| Map_Num | Total data size | Sample data size | RD value |
|---------|----------------|-----------------|----------|
| k       | for each map   | for each map    | $u_k$    |
| 1       | 4,019          | 351             | 0.012    |
| 2       | 4,032          | 352             | 0.017    |
| 3       | 4,023          | 352             | 0.023    |
| 4       | 3,987          | 348             | 0.022    |
| 5       | 4,203          | 367             | 0.014    |
| 6       | 3,976          | 347             | 0.03     |
| 7       | 4,095          | 358             | 0.009    |
| 8       | 4,100          | 358             | 0.015    |
| 9       | 3,900          | 341             | 0.023    |
| 10      | 3,987          | 348             | 0.021    |
| 11      | 4,023          | 352             | 0.019    |
| 12      | 3,926          | 343             | 0.019    |
| 13      | 4,093          | 358             | 0.02     |
| 14      | 4,033          | 352             | 0.008    |
| 15      | 4,000          | 350             | 0.028    |
| 16      | 4,024          | 352             | 0.022    |
| 17      | 3,987          | 348             | 0.021    |
| 18      | 3,872          | 338             | 0.019    |
| 19      | 3,986          | 348             | 0.019    |
| 20      | 4,023          | 352             | 0.02     |
| 21      | 3,893          | 340             | 0.015    |
| 22      | 4,026          | 352             | 0.023    |
| 23      | 4,082          | 357             | 0.022    |
| 24      | 4,100          | 358             | 0.019    |
| 25      | 3,860          | 340             | 0.014    |

The upper limit of attribute data quality RD value becomes a little bigger without considering $f$, but in the same numerical degree. So the sample ratio of $f$ can be neglected for simple equation in practical application while estimating the variance of the RD value.

Fig. 1 is the RD spread of the attribute data for each map in the batch of digital maps. The dashed line in the figure is the upper limit (without considering $f$) of the quality of the attribute data under the confidence of 95%. The dot dashed line is the quality assessment of attribute data quality for such a batch of digital maps. From Fig. 1, the quality distribution of attribute data for the batch of digital maps can be easily found, and total attribute
quality can also be visually obtained.

Besides, figures similar to Fig. 1 can illustrate the attribute data status visually. The upper line in Fig. 1 is the quality limit for the batch of digital maps under certain confidence, so the quality of the attribute data may vary if their qualities extend to the upper limit. As Fig. 1 shows, the quality of the whole attribute data in this batch are stable and under control for no point extending through the upper line. In the meantime, from Fig. 1 it can also be deduced that the attribute data quality has no systematic distribution and moves up and down in a certain range without a trend change. So the figure of quality spread can be used for monitoring and controlling the quality of the attribute data quality.

6 Conclusions

This paper is focused on the point of assessing the quality of the attribute data for vector data. The statistical rate of disfigurement model is put forward under the simple random sampling firstly. On the basis of characteristics of the statistical model, the two main statistical indexes, the mean and variance of the RD model, are deduced mathematically. As the result, the non-bias estimation of the RD resulting from the defects inspection reveal the suitable for assessing the quality of the attribute data. Furthermore, the quality of the attribute data assessment schemes based on the mean and variance of RD for single and a batch of digital maps are discussed. The method is proved that it is of easy operation and has the tight theory accordance with applications and thus has the practical merit in the assessment of attribute data quality. At the same time, the method is a new one for the research of the uncertainty in GIS, and can spread in the data collecting procedure.

References

1 Gopal S, and Woodcock C (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogrammetric Engineering and Remote Sensing, 60(2):181-188
2 Zadeh L A (1965) Fuzzy sets. Information and Control, 8:338-353
3 Wang F, Hall G B (1996) Fuzzy representation of geographical boundaries in GIS. International Journal of Geographical Information Systems, 10: 573-590
4 Brown D G (1998) Classification and boundary vagueness in mapping pre-settlement forest types. International Journal of Geographical Information Science, 12(2): 105-129
5 Pawlak Z (1982) Rough sets. International Journal of Computer and Information Sciences, 11:341-356
6 Worboys M F (1998a) Imprecision in finite resolution spatial data. Geoinformatica, 2:257-280
7 Ahlqvist O, Keukelaar J, Oukbir K (2000) Rough classification and accuracy assessment. International Journal of Geographical Information Science, 14:475-496
8 Worboys M F (1998b) Computation with imprecise geospatial data. Computers, Environment and Urban Systems, 22:85-106
9 Caspary W, Joos G (1999) Statistical quality control of geodata. International Symposium on Spatial Data Quality, Hong Kong Polytechnic University.
10 Liu, Liu D J (2001) Study on sampling inspection schemes to digital products in GIS. Geo-Spatial Information Science, 4(1):62-67
Joos G (1999) Assessing the quality of geodata by testing consistency with respect to the conceptual data schema. In: Craglia M, Onsrud H eds. Geographic Information Research: Trans-Atlantic Perspectives, London; Taylor & Francis.

Hay A M (1988) The derivation of global estimates from a confusion matrix, International Journal and Remote Sensing, 9:1 395-1 398.

Card D H (1982) Using know map categorical marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering and Remote Sensing, 48(3), 431-439.

Schilling E G (1982) Acceptance Sampling in quality control. New York, Marcel Dekker.

Stephen, Stehman V (1996) Estimating the Kappa coefficient and its variance under stratified random sampling. Photogrammetric Engineering & Remote Sensing, 62(4), 401-402.

UCGIS (1996) Research priorities for geographic information science. Cartography and Geographic Information Systems, 23(3), 115-127.

Bishop Y, Fienberg S, Holland P (1975) Discrete multivariate analysis: theory and practice. Cambridge; MIT Press.

Zadeh L A (1965) Fuzzy sets. Information and Control, 8, 338-353.

Congalton R (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environment, 37, 35-46.

Congalton R G, Green K (1999) Assessing the accuracy of remotely sensed data. principles and practices. London, Lewis Publishers.

Chrisman N (1982) A theory of cartographic error and its measurement in digital databases. In: Proceeding of Auto-Carto 5, 159-168.

Liang X J, Zhu D P (1998) The method and principle of the sampling inspection. Shanghai: Huadong Normal University Press. (in Chinese).

Liu D J, Shi W Z (1991) Accuracy analysis and quality control of spatial data in GIS. Shanghai: Shanghai Press of Science and Technology. (in Chinese).

Ma Z, Redmond R L (1995) Tau coefficients for accuracy assessment of classification of remote sensing data. Engineering & Remote Sensing, 61(4), 435-439.