THE STUDY ON THE QUALITY OF DATA SOURCE IN THE PROCESS OF FOUNDING GIS DATABASE

Zheng Yidong　Wudi　Yu caixia　Chen Huirong　HeYangqing
(Dept. of Hydrography and Cartography, Dalian Naval Academy, Dalian 116018, China)

Abstract: For GIS, data is very important. GIS spacial data can be collected by newest ways, for example, GPS, satellite images. Meanwhile there are many analogue data in the form of paper, such as chart, survey original and topographic map, which all can be digitized as GIS data source, what after all, is one economic and shortcut way. In the paper, quality probelms of data source in the process of founding GIS database are discussed, including preprocessing before digitalization, precision of hand followed digitizing or resolution and digital precision of scanning, as well as the error in the process of chart matching.
Key words: GIS,　data sources,　data quality,　digitalization

As we all know, data are geographic information systems (GIS) of blood, the quality of the data quality of GIS projects are the key to success. The impact of data quality factors have the quality of data sources, including a variety of measurement data, charts and remote sensing data, such as errors caused by survey and statistical property data error, as well as documents and other data errors. Library in the process of data quality issues, including the figure of the pre-treatment, hand tracking precision of digital or scanned digital vector resolution and accuracy, charts Edge errors in the process. This article will analyze data quality problem in the process of building GIS database.

1. Pretreatment before digitizing

Operations for the digital charts (job base chart), when the use of polyester film map, the deformation of generally less than 0.2%. The use of paper drawings map, the drawings with the size of humidity and temperature change, temperature change, the humidity from 0% to 25%, then the paper size may change by 1.6%. Paper because of swelling and

shrinkage is not the same, even if the humidity back to the original size, the drawings can not be reverted to the original scale. Therefore in the digital to determine the appropriate scale factor, through the affine transformation for geometric correction in order to reduce the job base deformation resulting position error to the corresponding accuracy. Charts job base map accuracy should meet the following requirements: chart profile point error $\leq$ 0.1mm, chart profile side error $\leq$ 0.1mm, chart profile diagonal error $\leq$ 0.15mm. For different job types and the proportion of digitized base map, we should pay attention to whether their projection line, the ratio match. For different projection mode, should be timely in the digital transformation for the system requirements for the projection mode. Ratio for different scales and accuracy should be recorded in the original data in order to estimate this possible error.

2. The process of digitizing orientation error analysis

Both hand followed digitizing and scanned digitizing, drawing charts because there are varying degrees of deformation, using a scanner to scan the course will also have a random error, the impact of these errors in the data editor can not deal with the process of elimination (This process can only eliminate or reduce the figure of the process of handling errors generated by partial or obvious error), so the need for geometric correction. Achieve a lot of geometric correction methods are generally carried out first to the map orientation (calibration). Directional error of the determination process is as follows:

Directional transformation formula is:

$$x_D = a_{11} + a_{12}x + a_{13}y + a_{14}xy \qquad 1$$

$$y_D = a_{21} + a_{22}x + a_{23}y + a_{24}xy \qquad 2$$

In these formulas, a_{11} a_{12} a_{13} a_{14} a_{21} a_{22} a_{23} a_{24} are the transformation parameters, for the determination of these parameters, based on four-point profiles of chart to find the coordinates of points, if the four-point measurement profile map coordinates is$(x_1 \quad y_1)$ $(x_2 \quad y_2)$ $(x_3 \quad y_3)$ $(x_4 \quad y_4)$ while the actual coordinates is $(x_{d1} \quad y_{d1})$ $(x_{d2} \quad y_{d2})$ $(x_{d3} \quad y_{d3})$ $(x_{d4} \quad y_{d4})$ will have

$$A_1 = B^{-1}X \qquad A_2 = B^{-1}Y \qquad 3$$

In these formulas,

$$A_1 = (a_{11}a_{12}a_{13}a_{14})^T \qquad A_2 = (a_{21}a_{22}a_{23}a_{24})^T$$

$$X = (x_{d1}x_{d2}x_{d3}x_{d4})^T \qquad Y = (y_{d1}y_{d2}y_{d3}y_{d4})^T$$

$$B = \begin{bmatrix} 1 & x_1 & y_1 & x_1y_1 \\ 1 & x_2 & y_2 & x_2y_2 \\ 1 & x_3 & y_3 & x_3y_3 \\ 1 & x_4 & y_4 & x_4y_4 \end{bmatrix}$$

Charts at the time of scanning because of the rotation angle smaller, at the same time chart diagram can be approximated as a rectangular profile, for the sake of convenience, may wish to set $x_1 = 0$ $y_1 = 0$ $x_2 = s_x$ $y_2 = 0$ $x_3 = s_x$ $y_3 = s_y$ $x_4 = 0$ $y_4 = s_y$ in these formulas, s_x s_y are respectively measuring east-west and the North-South map profile length, and thus can be drawn by the orientation error caused by estimating equation (a detailed proof see [4]):

$$M_{YD} = 25.4 / D[(1 - x' - y' + x'y')dx_1 + (x' - x'y')dx_2 + x'y'dx_3 + (y' - x'y')dx_4]$$

$$M_{XD} = 25.4 / D[(1 - x' - y' + x'y')dy_1 + (x' - x'y')dy_2 + x'y'dy_3 + (y' - x'y')dy_4]$$

In these formulas, $x' = x/s_x$ $y' = y/s_y$ D is the resolution of the scanner.

3. Digital tracking

Hand followed digitizing is a digital way of a lower degree of automation, the digital way, and its accuracy is also a result of the digital operator and its work without the same degree of fatigue, the operator of the higher labor intensity. With the large format scanner to continuously reduce the cost of scanning and vectorization technology continues to improve, this figure means of automatic scanning may become a complement to digital. Impact on the quality of digital data to track a number of factors, in these factors, the following factors influence:

(1) Digital base map of the geographical elements of the width, density and complexity of the digital quality of the results have a significant impact.

(2) Digitizer resolution and accuracy of digital data quality has a

direct and decisive impact.

(3) Digital operator skills and experience the difference the introduction of human factors errors are different, because of operator vision, operating habits, and fatigue degree of proficiency in different locations to determine the best sampling point, cross wire and the target point to determine the degree of coincidence there will be a certain degree of difference, the impact of digital quality.

(4) Mode of operation (such as mining point curve the number of ways and mining) will also affect the quality of digital data.

Assumed that the impact of various errors in line with the law of error propagation, hand followed digitizing integrated precision should be obtained under the style:

$$M = \pm\sqrt{M_1^2 + M_2^2 + M_3^2} \tag{4}$$

In this formula, M is hand express tracking digital integrated precision; M1 is express base job orientation error; M2 is express precision digitizers; M3 is express human factors error.

4. Digital scanning

Scanning with high-precision digital scanner to scan images and graphics to form raster data files, re-use scanning software vector raster data files on the deal, it will not be converted to vector graphics data. Standardize the electronic chart provides: graphics positioning control point scanning error is not more than 0.1mm, compared with the job base maps, vector after scanning point position error of not more than 0.15mm, Line error no greater than 0.2mm. Impact on the quality of digital scanning in addition to image quality factors, will also include scanning accuracy, directional accuracy, accuracy of vector loss.

(1) Scanner resolution and accuracy

For scanning resolution and accuracy of digital quality, Scanner is important. Therefore, It Is expected to select the appropriate scanner in accordance with the specific circumstances.

(2)Raster data vectorization precision loss

Vectorization of raster data in the process of refinement,tracking may introduce some error. Based on actual work experience, the complexity of automatic vectorization effect graphics very poorly, it will produce a large number of cross-lines, resulting in polygon tracking error. In this point, interactive vectorization should be the most

appropriate method. Therefore, selecting vectorization software should be not only concerned about the degree of automation (automatic vectorization software prices are often high), but also the following features: Intelligent speckle reducing, cutting, distortion correction, proportional control, the level of correction, grating Edit and interactive vector and so on. At present, common vectors such as software is GeoScan.

(3) errors resulting from the Ways of digitizing scanning

Scanner resolution is the scanning of geometric figure method error in the main error sources, reducing this error is the only way to improve the geometric resolution scanner. However, with the resolution improving, the volume of raster data increases to the square-class pace. This often resulted in depletion of computer storage resources, data processing time to extend the square-class. Integrated precision digital scans can be determined with the following equation:

$$M_s = \pm\sqrt{M_a^2 + M_b^2 + M_c^2} \tag{5}$$

In this formula, Ms is express scan digital integrated precision; Ma is express base job orientation error; Mb is express scanner accuracy; Mc is express vector of error.

5. Error analysis in the process of charts edge

In most cases, with the process of digitizing carried out by sub-sites, it is inevitable that digital maps adjacent to the border sometimes do not coincide after the geometric location. Therefore a graphics splicing problem come out. Currently, many digital systems and GIS mapping system access function are weak or do not have the ability to Edge Accuracy Assessment, concerning the database splicing process, property and the overall structure of analysis, docking edge elements of classification, this paper charts put out a further discuss Edge accuracy analysis. Edge is completed, will automatically generate a precision Edge document, to record each side of the accuracy of information received as a measure of positional accuracy of the figure of an indicator, then the accuracy in margin of error to measure. Under normal circumstances, the observed values are the true value X do not know, I really can not find the error Δ. In fact, because Edge is a series of the same feature point pairs were observed, arising from inconsistent methods used in treatment, so the calculation error can be used in

dual-observation method:

Assumptions on the X_i for two observation results are L_i' and L_i'', then the difference between dual observation are:

$$d_i = L_i' - L_i'' \qquad (\ =1 \quad 2 \quad \ldots\ldots \quad) \qquad (6)$$

because d_i the true value is zero, therefore, d_i is the real error.

The assumption that the observations are of the right to p_i, the unit Medium error (valuation) as follows:

$$m_0 = \pm\sqrt{\sum_{i=1}^{n} p_i d_i^2 / 2n} \qquad (7)$$

In this formula, $\sum_{i=1}^{n} p_i d_i^2 / 2n$ is sum of squares that weighted Double difference observations.

Medium error of the observation value L_i' and L_i'' is

$$m_{L_i}'' = m_{L_i}' = m_0\sqrt{1/p_i} \qquad (8)$$

The first i times the average of the observed value of the error (valuation) as follows:

$$m_{x_i} = m_{L_i}' / \sqrt{2} = m_0\sqrt{1/2p_i} \qquad (9)$$

If all the observed values are the same precision, make them the right 1, by the formula (7), (8) were the observed values of the error as follows:

$$m_L = \pm\sqrt{\sum_{i=1}^{n} d_i^2 / 2n} \qquad (10)$$

While the average error as follows:

$$m_{X_i} = m_L\sqrt{1/2} = \pm\frac{1}{2}\sqrt{\sum_{i=1}^{n} d_i^2 / 2n} \qquad (11)$$

In this formula, $\sum_{i=1}^{n} d_i^2$ is the all square of Double observed values

of the difference.

Through the practice shows that such a dual-observation method can better reflect the splicing Edge precision graphics.

6 Concluding remarks

GIS data are the most fundamental and important component of GIS is also a project in the largest part of investment. Data quality is good or bad, will directly affect the function of GIS systems and applications. Factors that affect the quality of the data are multi-faceted, built library in the process of data-quality problems have a considerable portion of the process from the figure, but the quality of source data can not be neglected. Introduced by the source data error on the impact of GIS projects are difficult to immeasurable. Based on the database in the process of data quality problem to do a simple analysis of GIS data quality control to provide some basis.

References

[1] Wuhan University of Surveying and Mapping Science and Technology, Department of Surveying Adjustment, Survey Adjustment basic 1996

[2] Fu-Ling Bian. Geographic Information Systems mapping principle and method, Beijing Press, 1996

[3] Zhao Jun. some discussion on the accuracy of the digital data process, Sichuan, Surveying and Mapping 2000,23 (1)

[4] Hao Xiangyang. precision analysis on map scanning digital point [J]. Journal of Surveying and Mapping 1996,25 (1)