Geoscience and Remote Sensing (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/geors.2022.050104
ISSN 2523-2592 Vol. 5 Num. 1

Reviewing Methods for Controlling Spatial Data Quality from Multiple Perspectives

Danling Chen

Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China

Keywords: Spatial Data, Geographic Information System, Uncertainty, Error, GIS

Abstract: Spatial data is the core and operation object of geographic information system (GIS). The quality of spatial data determines the application of GIS and the effectiveness of decision-making to a great extent. This article introduces two important types of spatial data, vector data and raster data. Then, this paper discusses the uncertainty and sources of errors in spatial data, and discusses the methods of checking and preventing uncertainty and errors from the aspects and processes of digitization, so as to ensure the quality of spatial data. Finally, this paper explores cutting-edge approaches to improving spatial data quality, including the Area preserving method for improved categorical raster resampling, and using hierarchical grid index to detect and correct errors in vector elevation data. By studying effective data quality control methods, the quality of spatial data in GIS can be guaranteed, and the basic guarantee for the wide application and development of geographic information science can be provided.

1. Types of Spatial Data

1.1. Vector Data

Vector data represents the location of a map graphic or geographic entity in terms of X, Y, and Z coordinates. Vector data generally expresses the spatial location of geographic entities as accurately as possible by recording coordinates [1].

Points, lines, and areas are treated differently in vector data. In addition to the x, y coordinates, the vector data structure stores information about the point entity's type, drawing symbols, and display requirements. Line entities are linear features made up of line elements. When exported, line entities can be solid or dashed. In general, arcs and chains describe any continuous and complex curve [2]. Polygon vector coding expresses not only location and attributes, but also shape, neighbourhood, and hierarchical structure of the area [3].

1.2. Raster Data

Raster data is a data form in which space is divided into regular grids, each grid is called a unit, and corresponding attribute values are assigned to each unit to represent an entity [4].

The work area is divided into rows and columns by a decomposition force to form many grids, each grid unit is called a pixel, and the grid data structure is actually a pixel array, that is, a

collection of pixels in the form of a matrix. Because raster data is arranged by rules, the entity position relationship is implicit in the row and column numbers. Each grid element's code represents an entity's attribute or its encoding. Each pixel can have a different "gray value" depending on the representation information of the represented entity [5].

1.3. Comparison of Vector and Raster Data

The vector and grid structures of spatial data are entirely different methods for simulating GIS.

In general, the benefits of grid and vector structures are limited. Generally, the grid model is better suited to large-scale and small-scale regional problems like natural resources, environment, agriculture, forestry, geology, and strategic layout research in the city planning stage. On the other hand, the vector model is better suited for zoning, land management, and utility management applications. The two models can also be mixed and displayed on the same screen [6].

Table 1: Comparing vector data with raster data

Vector data	Raster data
Accurate spatial location	low positional precision
The network connection method fully describes the topological relationship.	Difficulty establishing network connections
Graphs can be retrieved, updated, and synthesized.	Facilitate planar data processing
Difficulty in mathematical simulation	Mathematical simulation is convenient
Difficulty in overlaying multiple maps	Multiple map overlay analysis is convenient
Cannot directly process digital image information	Can directly process digital image information
High cost of data output	Low cost of technology development

1.4. Conversion between Vector and Raster Data

Raster and vector data formats have their plusses and minuses, which are determined by the processing method of the geographic information system and the characteristics of these two data formats.

In general, raster data is used as a background layer for digitising vector features. An advanced computer program can also automatically extract vector features from an image. Examples of features are sudden changes in the colour of adjacent pixels in an image, which a computer program looks for to create a vector feature. This feature is usually only found in specialised GIS software.

Converting vector data to raster data can be helpful in some cases, but it loses attribute data. Converting vector data to raster allows non-GIS users to view it as an image on their computer without special GIS software.

2. Uncertainty and Error in Spatial Data

2.1. Uncertainty

Uncertainty refers to the fact that the objective world or the entity itself is subject to variation. Due to the limitations of human understanding of objective entities and phenomena and the ambiguity of expression, the original data is inherently uncertain, and the data is then used for GIS analysis and processing, inevitably resulting in uncertain analysis results [7]. Imprecision, ambiguity, and vagueness are all examples of uncertainty.

2.2. Error

.Error is defined as the accuracy between a recorded measurement, and its numerical value is

inaccurate for most purposes [8]. Errors can be classified as random, systematic, or gross errors.

2.3. Reasons for Uncertainty and Error

Uncertainties and errors in spatial data are inevitable. Various operations, transformations, and processing in the spatial database will introduce bias: the more conversions, the more errors and uncertainties [9]. Therefore, it is critical to understand their sources at each stage and link them to the GIS system to ensure data quality.

Table 2: Sources of Uncertainty and Error at Various Stages

Stage	Source of Uncertainty and Error
Measurement	People error (Alignment error, reading error, adjustment error), instrument error(Noperfect, lack of calibration, uncorrected), environmental impact(low air, air pressure, temperature, magnetic field, signal interference, wind, light source),GPSData error (signal accuracy, receiver accuracy, positioning method, processing algorithm, coordinate transformation, orbital signal, etc.), etc.
Mapping	Drawing control points, editing, clearing, synthesis, duplication, color registration, etc.
Input	Manuscript quality, operator error (experience skills, physiological factors, work attitude), paper deformation, mathematical instrument accuracy, digitization method, etc.
Process	Geometric correction, coordinate transformation, projection transformation, data offset, data format conversion, topology matching, map overlay, etc.
Output	Scale error, output device error, media instability, etc.

3. Methods to Improve the Quality of Spatial Data

3.1. Data Quality Inspection

Traditional manual method

Manual quality control compares digital data to the data source. The visual method includes drawing on the transparent map and superimposing the original image. Finally, the attribute part is inspected by comparison with the original one by one or other methods [10].

Metadata methods

The dataset's metadata contains a wealth of data quality information that can assess data quality and track quality changes over time [11].

Geographical Correlation Law

The correlation of geographical feature elements assesses the spatial data quality. For example, the position of the river does not have to be on the convex connecting line of the contour line when superimposing two layers of data. If a problematic data layer cannot be identified, it can be overlaid with other high-quality layers.

3.2. Prevention of uncertainty and error

Data quality control is a complex process. It must address each process and link that contributes to error generation and diffusion.

Digital equipment selection

The digitiser's resolution and accuracy determine the quality of the digitisation. So, when choosing digital equipment, keep in mind that parameters like resolution and accuracy must meet design accuracy requirements. Generally, the digitiser's resolution should be 0.025mm, the scanner's resolution should be 0.083mm [12].

Data preprocessing

The original maps, tables, and other documents are sorted and counted to minimise digitisation errors and maximise digitisation efficiency. The segment closed curves and extended linear features on the map, as most GIS software can only store a limited number of vertices for linear entities, and segmenting linear features can help reduce digitisation errors, thereby improving digitisation accuracy.

Map Orientation

The coordinates of the points on the map collected by the digitising tracking head are the digitiser's plane coordinates. The accuracy and configuration of the digitiser determine these coordinates, as do the points themselves. So, when digitising a map, the map orientation must be converted from the digitiser plane coordinates to the actual geographic coordinates.

Digital Accuracy

The point method produces less error than the switch flow method, which is frequently used in practice. Digitisation quality is affected by the width, density, and complexity of the geographic element graphics. For example, thick lines are more likely to cause errors than thin lines, complex curves than flat lines, and dense elements than sparse elements [13].

4. The Latest Progress and Future Trends in Spatial Data Quality

4.1. Area Preserving Method for Improved Categorical Raster Resampling

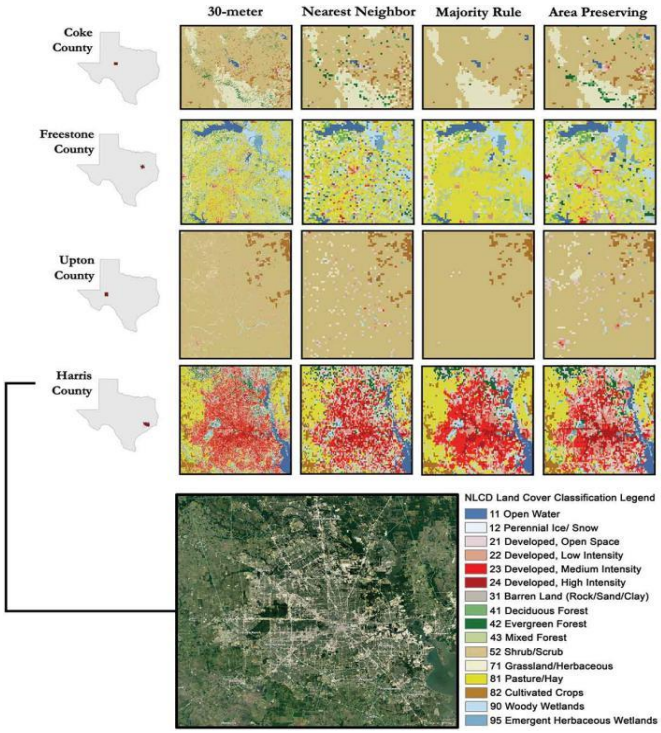

Figure 1: Four Texas counties resampled from the 30-meter National Land Cover Dataset using the nearest neighbor, majority rule and area preserving methods

Upscaling source data to a target map is frequently required to reconcile granularity differences between raw and "analysis-ready" data. Users are forced to choose between map structure and map diversity by default when using standard resampling methods (nearest neighbour and majority rule) [14]. This method generates more representative maps in terms of variety and structure, better retain minority classes, and generates maps that are more (or equally) accurate for both users and producers. The tool is scalable in performance and has a serviceable R-based implementation, as

shown in Figure 1.

The map was resampled using the AP, nearest neighbour, and majority rule methods. As a result, the resampled map is devoid of lesser-known classes.

Four Texas counties were chosen to highlight a range of patterns found in the visual evaluation of all resampled maps.

4.2. Using Hierarchical Grid Index to Detect and Correct Errors in Vector Elevation Data

To identify and correct elevation errors in vector elevation data, this method extracts, counts, visualises and analyses the elevation errors present in contour and elevation point data. The internal hierarchical grid model of the map with the wrong contour line elevation and the height conflict of the height point, dot line. The hierarchical grid model uses a dynamic operator as a floating template for data extraction, which improves operational efficiency and ensures recognition and correction accuracy [15]. The elevation error identification at the edge of the contour map frame and the internal elevation of the contour map frame are completed by summarising the vector elevation data elements' error types and spatial characteristics. Finally, the algorithm function is tested using verification data, and its accuracy and efficiency are compared to existing methods, as shown in Figure 2.

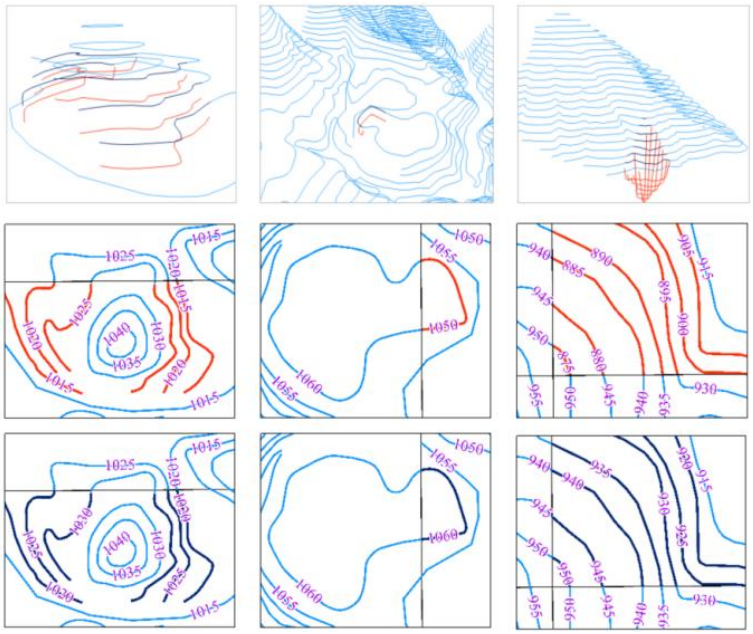

Figure 2: Comparison of contours before and after correction

Construction and application of Digital Topographic Map in Youyang County, Chongqing and Construction of High-precision DEM in Beibei District, Chongqing have used this method.

5. Conclusion

This article discusses two forms of spatial data that are vital to understand: vector data and raster data. As a result of these discussions, this article analyzes the uncertainty and sources of mistakes in spatial data, as well as the techniques for detecting and avoiding uncertainty and errors from the aspects and processes of digitization in order to assure the quality of geographical data. Finally, this paper examines cutting-edge approaches to improving spatial data quality, such as the Area preserving method for improved categorical raster resampling and the Using hierarchical grid index to detect and correct errors in vector elevation data, both of which are discussed in this paper. In the

course of researching appropriate data quality control techniques, it will be possible to ensure the quality of spatial data in GIS while also providing the fundamental assurance for the widespread use and growth of geographic information science.

References

[1] Carver, S. J., & Brunsdon, C. F. Vector to raster conversion error and feature complexity: an empirical study using simulated data. International Journal of Geographical Information Systems, 1994, 8(3), 261-270. https://doi.org/10.1080/02693799408901999

[2] Couclelis, H. The certainty of uncertainty: GIS and the limits of geographic knowledge. Transactions in GIS, 2003, 7(2), 165-175.

[3] Piwowar, J. M., Ledrew, E. F., & Dudycha, D. J. Integration of spatial data in vector and raster formats in a geographic information system environment. International Journal of Geographical Information Systems, 1990, 4(4), 429-444. https://doi.org/10.1080/02693799008941557

[4] Peuquet, D. Vector/raster options for digital cartographic data. (1983).

[5] Kwang-Soo, K., Min-Soo, K., & Kiwon, L. On integrated scheme for vector/raster-based GIS's utilization. IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing - A Scientific Vision for Sustainable Development, (3-8 Aug. 1997).

[6] Jian-Jun, C., Cheng-Hu, Z., & Wei-Ming, C. Area error analysis of vector to raster conversion of areal feature in GIS. (2007).

[7] Oberkampf, W. L., DeLand, S. M., Rutherford, B. M. Error and uncertainty in modeling and simulation. Reliability Engineering & System Safety, (2002). 75(3), 333-357. https://doi.org/https://doi.org/ 10.1016/ S0951-8320(01)00120-X

[8] Messina, J. P., Evans, T. P., Manson, S. M., Shortridge, A. M., Deadman, P. J., & Verburg, P. H. Complex systems models and the management of error and uncertainty. Journal of Land Use Science, (2008). 3(1), 11-25. https://doi.org/10.1080/17474230802047989

[9] Wang, G., Gertner, G. Z., Fang, S., & Anderson, A. B. A methodology for spatial uncertainty analysis of remote sensing and GIS products. Photogrammetric Engineering & Remote Sensing, (2005). 71(12), 1423-1432.

[10] Sun Qing-hui, C. T.-h., Zhao Jun-xi, Zhong Da-wei, Shao Shi-xin. Errors and Uncertainties Analysis of Spatial Data Processing Model. (2007).

[11] Li Nan, Z. J. Error Analysis and Quality Control of Spatial Data. (2017).

[12] Li, D., Zhang, J., & Wu, H. Spatial data quality and beyond. International Journal of Geographical Information Science, (2012). 26(12), 2277-2290.

[13] Yongling, W. (2001). Analysis of Error Sources in Spatial Data and Discussion on Quality Control.

[14] Johnson, J. M., & Clarke, K. C. An area preserving method for improved categorical raster resampling. Cartography and Geographic Information Science, (2021). 48(4), 292-304.

[15] Kangning, L. Research on error recognition and correction of vector elevation data based on hierarchical grid index. (2021).