Construction of multi-scale grid for massive land survey data

Jia Zhang[1], Xiulian Wang[1], Xiaotong Zhang[1], Xiaofei Bai[1,*], and Qiang Chen[2]

[1]China Land Survey and Planning Institute, Beijing, China
[2]Data Intelligence Information Technology Co,.Ltd, Beijing, China

Abstract. In the face of ever-growing and complex massive multi-source spatiotemporal data, the traditional vector data model is increasingly difficult to meet the needs of efficient data organization, management, calculation and analysis. Based on the simple and widely used geographic grid data organization model, this paper designs a technical method to convert vector data into multi-scale grid data, establishes a unified, standardized and seamless land spatial grid data model, and analyses the area accuracy of multi-scale grid data. Practice shows that the model can better meet the needs of multi-scale geospatial information integration and analysis, and it is easy to carry out distributed data processing, which provides technical support for the efficient organization, fusion and analysis of spatiotemporal data.

1 Introduction

Globally, using big data to improve social governance and enhance government services and regulatory capabilities is becoming a trend [1]. In recent years, with the deep integration of spatial information technology and resource management, spatial data is growing in variety and complexity. At present, the national land basic database has accumulated more than 20 categories of geospatial data involving mountains, rivers, fields, lakes and grasses, with a total amount of more than 300TB, and the data scale is still growing rapidly. For a long time, these data are managed by different departments, and the standards of data scale, coordinate system and data format are different, which causes the barriers of data analysis and application [2]. At the same time, China is a country with vast territory, diverse types of resources, complex land use types and land use structure [3], so the spatial and temporal scales of data required by various geoscience applications are different [4]. In order to meet the needs of fine and diversified multi-source and multi-scale spatial data integration analysis of natural resources, it is necessary to establish a unified data organization model supporting multi-dimensional information interoperability. As a concise data model with a long history, grid is easy to integrate spatial data of different scales and uneven distribution, which is an important tool for spatial analysis and spatial data mining [5]. From digital map to spatial information grid, in order to store, manage and express spatial information, relevant scholars have carried out many beneficial researches on grid data models such as latitude-longitude grid, triangle grid, quadrilateral grid and Voronoi-based grid [6-8]. However, most of the related researches focus on grid generation algorithm, and there are few reports on constructing multi-scale grid for large-scale vector data and its engineering application. Based on the standardized geographic graticule, this paper constructs a multi-scale land spatial grid system on the land survey data, to provide a solution and base map for multi-source spatial data fusion and calculation of natural resources.

2 Grid Construction

2.1 General ideas

The massive vector data of land survey cover the national survey area seamlessly, which is the geospatial base map of all kinds of natural resources data. As a kind of widely used, concise and scientific geographic grid, geographic graticule is suitable for the integration and analysis of continuous, contiguous and large-scale geospatial information, and it is an effective carrier for building land survey grid data. The grid cells of geographic graticule are non-uniform, its area and shape change with latitude. For a better control of space and area, and facilitate distributed data processing, a grid system is constructed by the idea of divide and conquer, the basic technical ideas are as follows:

- In the survey area of nationwide, constructing a control grid system, define the control grid scale, and grid encoding.
- Calculate the area of control grid unit one by one.
- Acquiring vector data of land survey within all control grids scope.
- Unify the grid cell starting point and define the cell size, then feature to raster.
- Data accuracy analysis.

The technology roadmap is shown in Fig. 1.

* Corresponding author: bxflxq@126.com

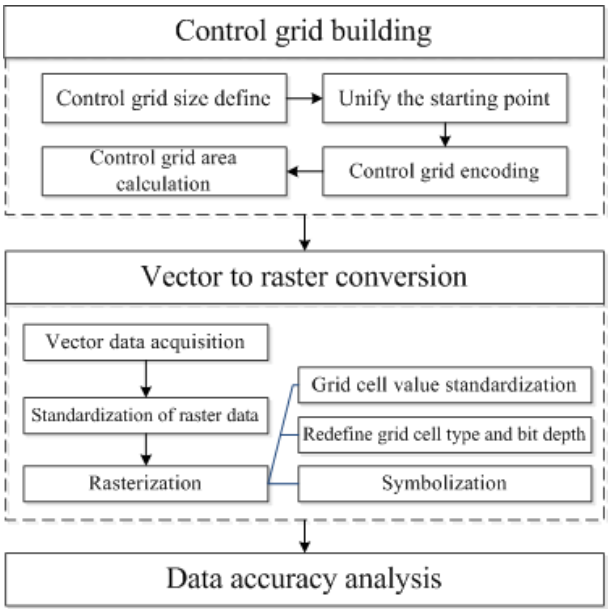

Fig. 1. Technology roadmap.

2.2 Control Grid Building

The control grid is a geographic graticule covering all data space; it is the benchmark grid for constructing multi-scale grid, so that the grid scope, grid size, grid code and grid area need to be defined first.

2.2.1 Control grid size and scope

National land survey data covers extra-large scale spatial scope, the northernmost end is 53°33′ N, the southernmost end is 3°52′ N, the westernmost end is 73°40′ E, the easternmost end is 135°2′ E. The control grid should cover the above space and select the appropriate grid scale. Too large a control grid unit will reduce the accuracy of area calculation and is not conducive to distributed computing and processing, too fine grids will increase data complexity. Considering the needs of data calculation and analysis, a 30 minutes (30') grid cell is selected as the basic unit of the control grid, that is, the longitude and latitude differences between adjacent control grid units are both 30'. There are 12423 control grid units nationwide, from west to east, starting at 73°30′ E, ending at 135°30′ E, involving 123 units; from south to north, starting at 3°30′ N, ending at 54° N, involving 101 units.

2.2.2 Control grid encoding

For better data integration and sharing, the control grid is coded according to the geographic graticule coding rules defined in the national standard of geographic grid (GB / T 12409-2009), shown as Fig. 2. Each code consists of five elements: quadrant code, grid interval code, interval unit code, latitude and longitude code and grid number. Quadrant code includes values of 'NW, NE, SW and SE', which represent the four regions of the world: northwest, northeast, southwest and southeast, with equator and prime meridian as intersection points; The grid interval code uses the integer value of the control grid size; The interval unit code represents the unit of grid scale, with 'D' as degree, 'M' as minute, and 'S' as second; The longitude and latitude code is generated by calculating the integer value of latitude and longitude; Grid number is generated by non-integer numerical calculation of latitude and longitude.

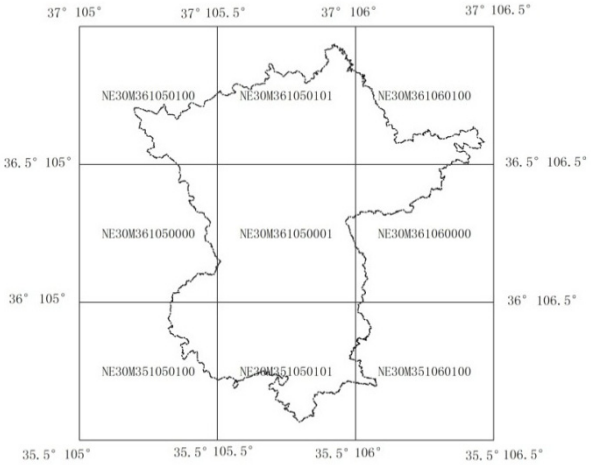

Fig. 2. Control grid unit code.

2.2.3 Control grid area

The vector features of land survey are calculated and summarized on the ellipsoid area (i.e. surface area of the earth ellipsoid), using the spheroid constants and related parameters of the 2000 national geodetic coordinate system (CGCS2000) [9]. In order to keep consistent with the land survey area measurement standard, the ellipsoid area is used as the area of the control grid, and the summary area of the vector features within the grid are used to adjust and correct the control grid area. The length of the global meridians is equal, so the length of the east and west sides of each grid unit is equal, but because the length of the latitude decreases from low latitude to high latitude, the southern side of each control grid unit is longer than the northern side. Obviously, the ellipsoid area of control grid units decreases from low latitude to high latitude. For the 30' control grid constructed in this paper, the ellipsoid area of the lowest latitude (3.5° N) grid unit is about 3072km2, and the ellipsoid area of the highest latitude (53.5° N) grid unit is about 1836km2. Area variation of control grid units is shown as Fig. 3.

2.3 Vector to raster conversion

Based on the control grid, the vector data of land survey and related natural resources within the grid scope can be quickly obtained. With the control grid units as index, the vector data can be converted into raster (grid cell) data, so that more finer-scale grids can be further defined.

2.3.1 Vector data acquisition

The national land survey data is huge, with more than 2TB of data and 300 million features in a single year,

massive spatial vector data must be partitioned for processing, the basic steps are:

- Overlay the control grid with the vector data of the county-level administrative divisions, record the spatial mapping relationship between them, and build the index of the control grid units and the counties.
- Extracting vector features of land survey within control grid unit scope.
- Split the vector features according to the control grid unit extent to establish the spatial relationship between the features and the grid units.
- Iterate the above operations for each control grid unit to obtain vector data corresponding to it.

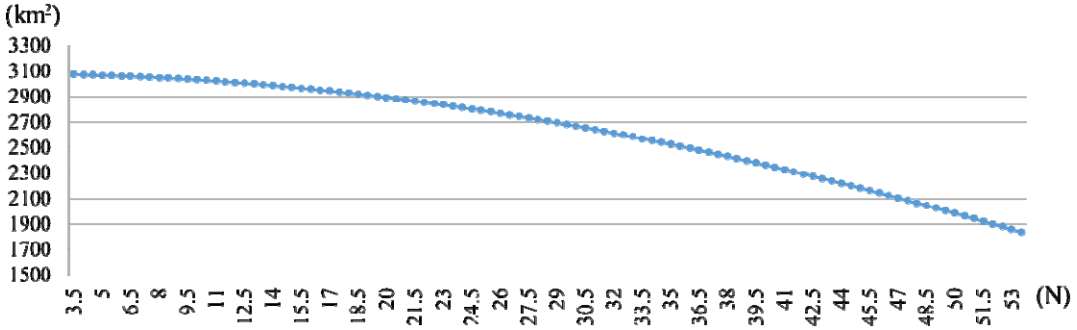

Fig. 3. Area variation of control grid units.

2.3.2 Standardization of raster data

To ensure data uniformity and specification, the starting point and cell size should be defined before rasterization.

(1) Unify the starting point of rasterization. Taking the lower left corner of the control grid as the starting point of the vector data rasterization, the size of the control grid unit is 30', so the longitude and latitude of the starting point are both integer times of 30'.

(2) Grid cell size and area defines. To ensure the seamless connection of raster data, the grid cell size should be an integral multiple of 30', which selected in this study are 0.00004° (1/25000), 0.0004° (1/2500), 0.004° (1/250) and 0.001° (1/100), shown as Fig. 4. The cell area is the mean value that the area of the control grid unit divided by the number of cells within it.

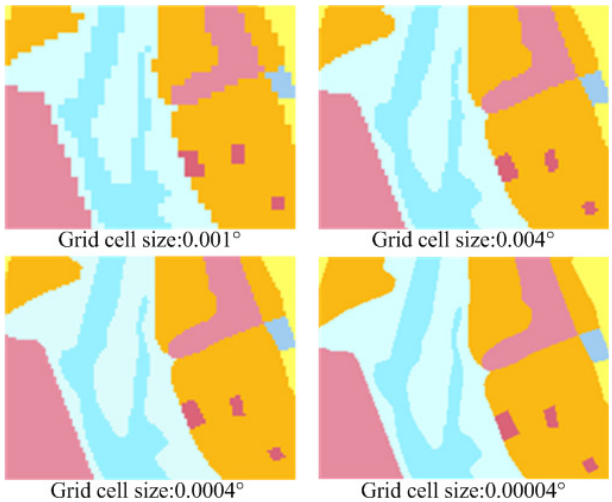

Fig. 4. Raster data of different cell sizes.

2.3.3 Rasterization

The general process of vector rasterization can be summarized as follows: Traverse vector features one by one, determine the grid cells within and on the boundary of each graphic, and assign vector features attribute values to these grid cells. In this paper, a distributed rasterization tool is developed based on ArcEngine to process vector data in each of the control grid unit in parallel. The Rule of Maximum Area (RMA) is used for the conversion, it can be interpreted as a dominance rule, which is based on the principle that if there are multiple types of a grid cell, the area-dominant type value is assigned to the grid cell [10].

(1) Grid cell value standardization. To improve the efficiency of rasterization, ArcEngine vector rasterization algorithm uses random values to assign attributes to grid cells, so that vector features with the same attributes have different type values after rasterization, which cannot meet the needs of data analysis. Therefore, we build a cell dictionary and design a standardized raster type value domain, which establishes the relationship between vector feature attribute values and raster type values.

(2) Redefine grid cell type and bit depth. After rasterization of vector features within different control grid unit, the grid cell types and bit depths may be inconsistent, making it impossible to perform uniform raster operations. To solve this problem, the grid cell type and bit depth are redefined by image reconstruction, so as to carry out efficient matrix calculations between grid cells.

(3) Symbolization. Based on the specification for cartographic symbols, the symbolic file of raster data is compiled, and the grid cell is rendered to support the data visualization and mapping.

3 Data accuracy analysis

Rasterization is a lossy conversion process, the loss of information varies with the size of the grid in the conversion [11, 12]. To ensure the correctness of data statistics, the accuracy of multi-scale grid data should be analyzed to select the grid scale that can meet the application requirements. In this paper, based on the vector data of land use in the Ningxia Hui Autonomous Region, four cell-size grid data (0.00004°, 0.0004°,

0.004° and 0.001°) were constructed. Taking arable land, garden, woodland, the grass, urban villages, industrial and mining land, transportation land, waters and water conservancy facilities land and other land as statistical indicators, land use classification area of four scales grid data were statistics and compared with the result of vector summarization, then the area deviation of different grid scales were obtained, shown as Table1 (Sv represents vector data statistical value, Sr represents raster data statistical value, the unit is hectare).

Table1. Statistical difference of classified area between grid data and vector data.

Grid size		arable land	garden	woodland	grass	urban & villages	transportation	waters	other land
S_r - S_v	0.00004°	-18.29	33.24	66.42	-161.83	37.51	13.75	10.26	-34.16
	0.0004°	14.75	58.77	-34.64	-235.61	157.09	-92.29	94.21	-19.84
	0.001°	-56.37	105.73	-131.24	197.11	-179.86	617.6	-425.56	-172.26
	0.004°	-210.5	1077.82	-3021.78	3018.39	1411.64	-1157.73	17.83	-1648.24

Obviously, with the increase of grid cell size, the difference between the statistical value of grid data and vector data increases gradually, and the data accuracy decreases. With (Sr-Sv)/Sv expresses the area difference ratio, when the grid cell size is 0.00004°, it shows a very good accuracy for the sum-up data deviation of each land use type within 0.04%, which can fully meet the needs of meso and macro analysis of natural resources management, shown as Fig. 5.

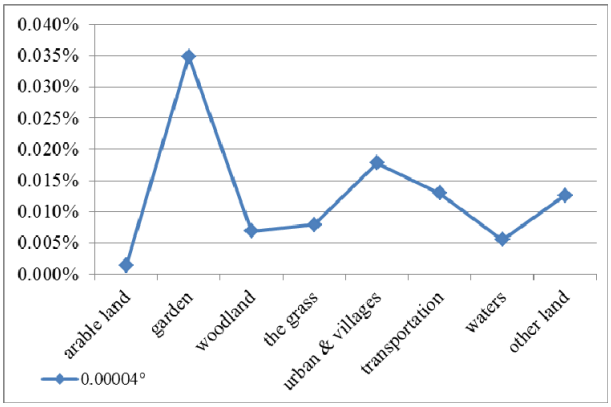

Fig. 5. Data deviation of land use types.

4 Conclusions

The analysis and application of massive vector data is the focus and difficulty of natural resource management. Because of the large sizes and complex types, massive land survey data are generally divided into administrative divisions or geographic units, which make it difficult to ensure a balanced computing and storage resources. The geographic grid data model built in this paper enables seamless coverage and efficient calculation of multi-scale grid data over a very large spatial range. It has been fully applied in Evaluation of National Resources and Environment Carrying Capacity and Territorial Spatial Development Suitability Evaluation, and has achieved good application results, which provides an effective implementation path for the integration and analysis of multi-source, heterogeneous and massive spatial data of natural resources.

Data accuracy is positively correlated with grid scale. The smaller the grid cell size, the higher the accuracy, and the higher consumption of computing and storage resources. Therefore, the focus of the follow-up work is to study the optimal grid scale according to the characteristics and application needs of land survey data, so as to further improve the efficiency of data production and analysis.

Acknowledgments

The authors gratefully acknowledge the support to this research from Annual Land Change Survey and Dynamic Monitoring Program and the Third Nationwide Land Survey Program (Ministry of Natural Resources of the People's Republic of China).

References

1. The State Council of China. (2015) Action Outline for Promoting the Big Data Development. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm
2. YAN J.M., WANG X.L., XIA F.Z. (2018) Remold New Pattern of Natural Resource Management: Target Orientations, Value Guidelines and Strategic Choices. China Land Science, 32: 1–7.
3. Research Group—"The Application of Spatial Statistics in A Schema of Multi-layer Sampling Frame of Spatial Information in the Resources Remote Sensing Investigation". (1999) A Schema of Multi-layer Sampling Framer of Spatial Information in the Resources Remote Sensing Investigation and Application of GIS Techniques. Statistical Research, 16: 37-41.
4. LIU M.L., TANG X.M., LIU J.Y., ZHUANG D.F. (2001) Research on Scaling Effect Based on 1km Grid Cell Data. JOURNAL OF REMOTE SENSING, 5:183-192.
5. CHEN S.P., CHEN Q.X., ZHOU C.H. (2002) Grid Mapping and Grid Computing, Science of Surveying and Mapping, 27: 1-6+2.
6. LI D.R., ZHU X.Y., GONG J.Y. (2003) From Digital Map to Spatial Information Multi-grid—A Thought of Spatial Information Multi-grid Theory. Geomatics and Information Science of Wuhan University, 28: 642-650.

7. BAI J.J., SUN W.B. (2011) Character Analysis and Comparison of Global Grid System. Geography and Geo-Information Science, 27: 1-5.
8. LI D.R. (2005) On Generalized and Specialized Spatial Information Grid. JOURNAL OF REMOTE SENSING, 9: 513-520.
9. Ministry of Natural Resources of the People's Republic of China. (2019) Technical regulation of the third nationwide land survey. Geological Publishing House, Beijing.
10. BAI Y., LIAO S.B., SUN J.L. (2011) Evaluating Methods and Scale Effects of Attribute Information Loss in Rasterization: A Case Study of 1:250000 Land Cover Data of Sichuan. Acta Geographica Sinica, 66: 709-717.
11. YANG C.J., LIU J.Y., ZHANG Z.X., WANG C.Y. (2001) Analysis of Accuracy Loss During Rasterizing Vector Data with Different Grid Size. JOURNAL OF MOUNTAIN SCIENCE. 19: 258-264.
12. CHEN J.J., ZHOU C.H., CHENG W.M. (2007) Area Error Analysis of Vector to Raster Conversion of Areal Feature in GIS. ACTA GEODAETICA et CARTOGRAPHICA S INICA. 36: 344-350.