

PRELIMINARY STUDY ON INTEGRATION OF STATISTICAL DATA AND SPATIAL DATA IN GIS

Zhao rong Dong chun

(Chinese Academy of Surveying and Mapping, Beijing, 100039,P.R. China)

Abstract: Statistical data is import information resource of sustainable development of territories, including data of population, agriculture, industry, and economy, etc. It is a foundation for sustainable development of territories to integrate statistical data and spatial data adopting GIS technology. With analyzing characteristics and types of geographic information for sustainable development of territories, this paper analyzes common methods of integrating statistical data and spatial data in GIS, then puts forward a method of integration by “spatializing technology”. As a case study, through exploring the correlation of geographical factors and the distribution of people in the same region, theory and methods of establishing spatial distribution of population data are introduced. Through this integration, we can efficiently manage, display and analyze statistical data with existing capabilities of GIS.

1 GEOGRAPHIC INFORMATION FOR SUSTAINABLE DEVELOPMENT

According to incompletely statistics, about eighty percent of fulfillment activities in national economy developments and society developments have relation with geographic information, and about forty percent of the technology troops are engaged in keeping on developing work about resources environment. These activities and people are at various degrees of making use of geographic information. Research on the classification, characteristics of the geographic information, can provide the foundation for the sharing and further using of geographic information.

The geographic information for sustainable development can be defined as follows: It is the information about resources and environment of the Earth where the mankind exists.

1.1 The characteristics of the geographic information for sustainable development

The characteristics of the geographic information for sustainable development:(1) it is the foundation of the information for sustainable development, and is the same important to the material resources and the energy resources, and is productivity element and wealth; (2) it is geographic data which is collected primitively or has been handled through the experiment; (3) it has characteristics including the long-term, system, primitivity ,and commonweal; (4) it is the foundation of basic and applied research with multi-level and multi-discipline , and has the extensive and applied background with applied purpose.

1.2 The classification of geographic information for sustainable development

1.2.1 According to its contents

- (1) resources type(population, land, water, forest, mineral, ocean, species)
- (2) environment type(environment monitor, environment protection, environment management, disaster prevention, disaster reduction, disaster relief)
- (3) economy type(industry, agriculture, energy, transportation, communication, business, finance...etc.)

(4) society type(party, government , army, law, science, education, culture, hygiene...etc.)

1.2.2 According to its source

- (1) investigation data(population, soil, land, resources, administration...etc.)
- (2) statistical data(regional economy, the annual statistics report, yearbook...etc.)
- (3) observation(survey and map, remote sensing, environment and weather monitor ...etc.)
- (4) graph and text data(historical file, simulating production data, design map)

2 COMMON METHODS OF INTEGRATION OF STATISTICAL DATA AND SPATIAL DATA

From the characteristics and classification of the geographical information, a conclusion can be drawn that geographical information include a lot of statistical information, which is one of important information resources. Along with the development of GIS, GIS can provide technical support for integrating various statistical data with spatial data.

With the development of geography information system, the storage of attribute or statistical data in GIS mainly have two kinds of methods. The first method is that attribute or statistical data and spatial data are respectively stored in different files, having independent management. Attribute or statistical data is managed by common used RDBMS, can be organized,managed,queried independently, can be integrated with spatial data through public field of tables when there is needs to operate them together(see the figure 1).Another method is that attribute or statistical data and spatial data are stored in the same file on the platform of GIS system, and unify the maintenance and management(see the figure 2).

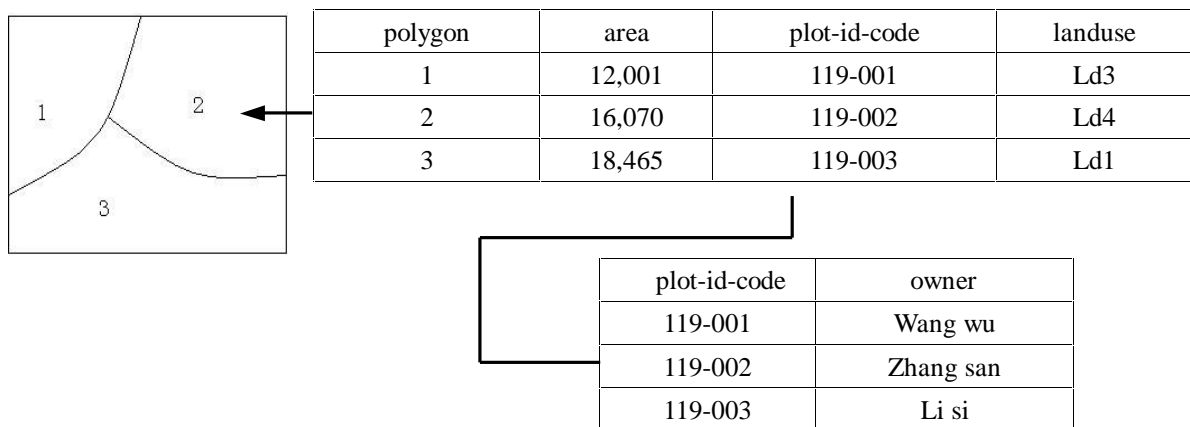


Figure 1: Attribute or statistical data integrate with spatial data through public field of tables

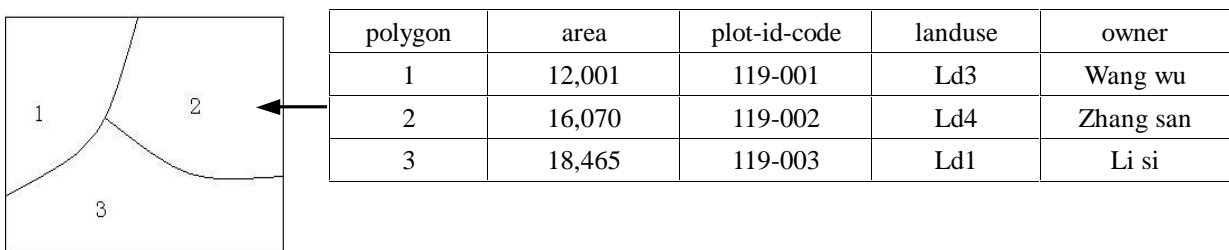


Figure 2: Attribute or statistical data integrate with spatial data through unified mangement

The first method can efficiently manage attribute or statistical data with the use of RDBMS such as Microsoft Access, the consistency maintenance of public field of tables is very important, otherwise the conjunction of this two kinds of data will result in disorder easily. Although the second method can avoid blemish of the first method, easy to carry out

the consistency maintenance, querying of spatial data and statistical data, but because the function of management of statistical data is weak in GIS, the workload of changing statistical data into tables increase consumedly. The precondition of these two methods is that the statistical unit of statistical data must be in accordance with the unit of spatial data, otherwise their integration can't be carried out effectively. For example, the statistical data with the province unit can be integrated with the coverage of province boundaries, can't be integrated with the coverage of county boundary.

3 SPATIALIZING OF STATISTICAL INFORMATION

As comprehensive reflection of the society economy phenomenon, statistical information usually has space attribute. In recent years, there is more studies on spatializing of statistical information, particularly the study on population's spatial distribution. At present because spatial unit is different between statistical data and spatial data, a series of problems have been produced. (1)Spatial difference of statistical data can not be revealed because of it's low spatial resolution. (2)Having hindered the share and integration between statistical data and spatial data. (3)Having prevented from the application of new technology such as GIS in statistics. Spatializing of statistical information is one of the best methods to solve these problems. That is to distribute statistical data within administration region unit on geographic grid with a definite size, and to create spatial distribute database of statistical data, which can be easily integrated with geo-spatial data such as land use, land cover, environment background data, and making this database become one part of national network for sustainable development, providing the service for researching sustainable development, providing the basis for science research and government decision-making.

Based on GIS technology, by creating geographical parameter database including topographical factor(inhabitant sites, water system, highway line, and railway line etc.), DEM factor, slope factor, aspect factor, and land cover factor etc., by exploring the correlation of geographical parameters and the distribution of statistical data in the same region by means of quantitative analysis and qualitative analysis, the distribution model of statistical data is established. Weight coefficient of each factor is gained. And then counting the values of each factor in the standard grid(1km*1km), considering the restricted condition of he distribution of statistical data(such as elevation, slope, water, etc.), the quantitative distribution of statistical data in spatial grid is established, creating the data foundation for the integration with geo-spatial data.

4 THE TEST ON POPULATION'S SPATIAL DISTRIBUTION

4.1 Original data

The original data include national topographical database at the scale of 1:250000(covering administrative boundaries,highway line, railway line, inhabitant sites), DEM database at the scale of 1:250000, land cover database at the scale of 1:100000(covering farmland, grassland, dry land, water system, inhabitant sites, and un-utilized land etc.), the fifth census database at the county level in 2000 year. Unit population(1km*1km) is established based on above-mentioned database.

4.2 The flow of spatializing population

Through spatial analysis, statistical analysis, and data filter for original database based on GIS, geographical parameter database and population grid database are created. Detailed processing flows as follows(see figure 3).

4.2.1 Creation of geographical parameter database

Geographical parameter database is a kind of database that stores the quantitative values of various geographical factors in each geographical cell. Inhabitant sites, water system, highway line, railway line etc. are most important

landform elements, which describe the particular landscape in a geographic region. They are the important parameters because they have tight correlation with population spatial distribution. Therefore they are contents in geographical parameter database. The characteristics of population distribution in vertical direction show that population distribution has intensive tendency to the low and flat topography. Some researches have proved that different types of land cover can affect weights of population distribution model.

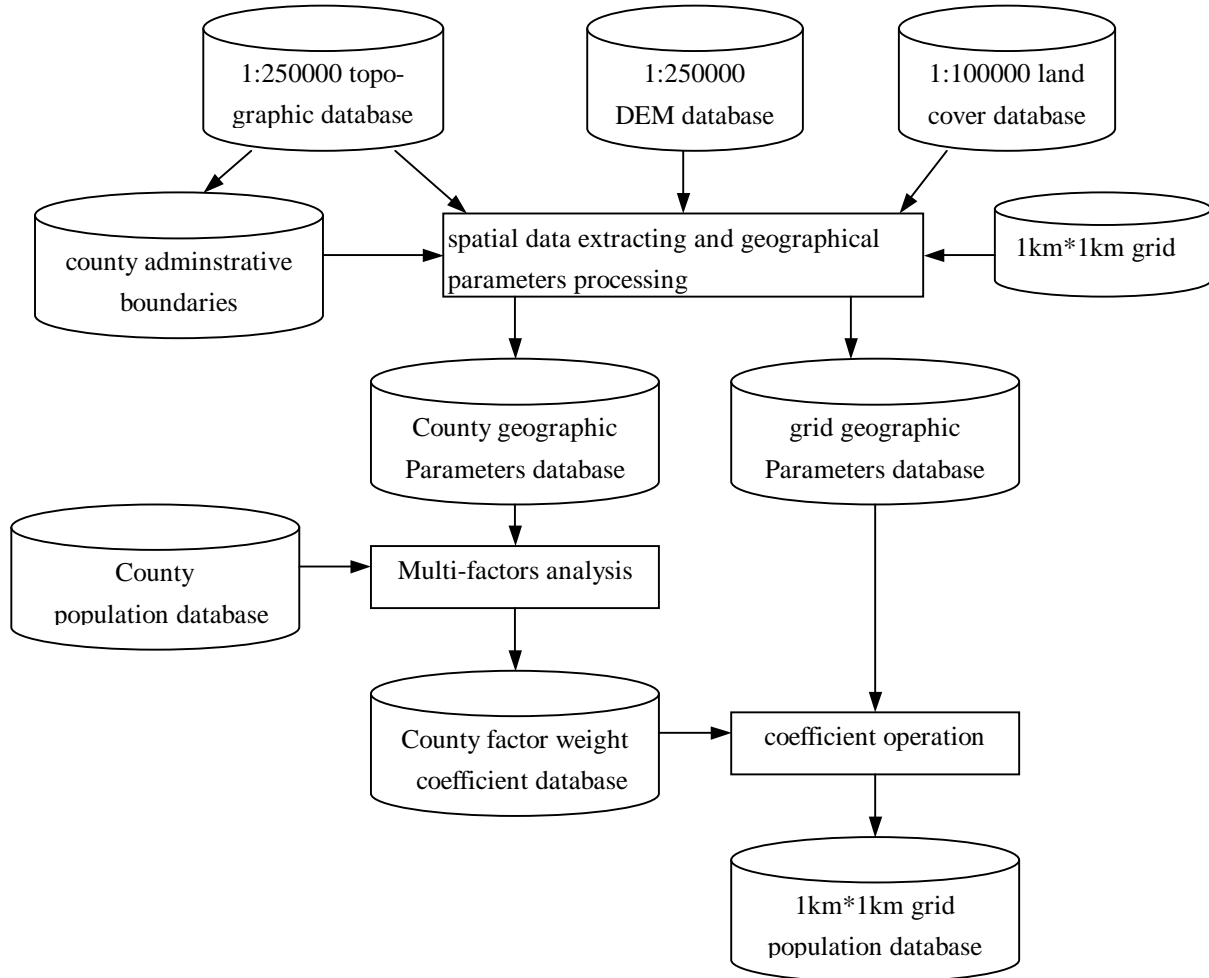


Figure 3: The flow of establishing spatial distribution of population data

- (1) Topographic factors such as highway line, railway line, inhabitant sites are obtained from topographical database at the scale of 1:250000 using vector overlay tools in GIS.
- (2) Altitude belt factor(divided into nine levels) is obtained from DEM database at the scale of 1:250000 using raster overlay tools in GIS. To represent altitude difference of each province in China, adopting different level standard according to altitude range of each province.
- (3) Land cover factors such as farmland, grassland, dry land, water system, inhabitant sites, and unutilized land etc. are obtained from land cover database at the scale of 1:100000.

4.2.2 Establishment and standardization of weight factor database

Here simple correlative coefficient is regarded as factor weight. Negative correlative coefficient and it's factor weight are defined as zero.

Each factor is standardized in county unit. For example railway factor, the number of 1km*1km grids in certain county is n, then standardization of railway factor is as follows:

$$\text{Standardized length of railway in number } i \text{ grid} = (\text{length of railway in number } i \text{ grid}) / (\text{total length of } n \text{ grids})$$

Standardized factor weight of of railway in number i grid = weight coefficient of railway * Standardized length of railway in number i grid

But calculating weight of the same category factor is different, for example, land cover is classified into six categories, i.e land1,land2,land3,land4,land5,land6, then it's weight is defined as follows:

Weight of land cover in number i grid = weight coefficient of land1 * area of land1 in number i grid + ... + weight coefficient of land6 * area of land6 in number i grid

Standardized weight of land cover in number i grid = (weight of land cover in number i grid) / (total weight of n grids)

4.2.3 calculating compositive weight and population of grid

Compositive weight of grid is defined as the sum of each factor's weight. Standardization of compositive weight is:

Standardization of compositive weight in number i grid = (compositive weight in number i grid) / (total compositive weight of n grids)

Finally, calcuting population of each grid:

Population in number i grid = total population of county * Standardization of compositive weight in number i grid

Through above-mentioned flows, national raster population database is established with 1km * 1km cell size(see figure 4, one part of database – population distribution of guizhou province). To query and analyze population information in GIS is very simple through this kind of integration(see figure 5).

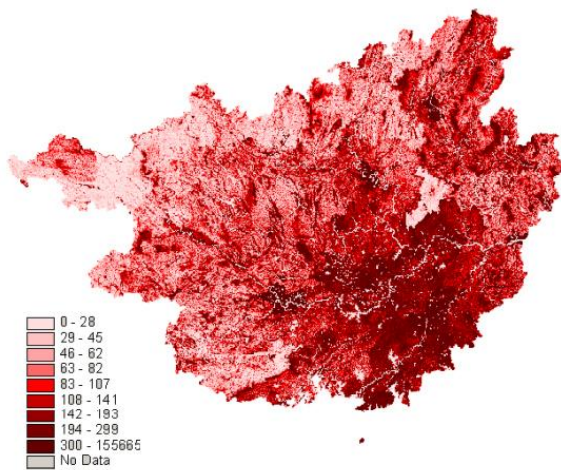


Figure 4: 2000 year population distribution of guizhou province, china



Figure 5: Query population information in discretional polygon region with GIS

5 CONCLUSIONS

This paper analyzes common methods of integrating statistical data and spatial data in GIS, and puts forward a method of integration by “spatializing technology”. As a case study, through exploring the correlation of geographical factors and the distribution of people in the same region, theory and methods of establishing spatial distribution of population data are introduced. Through this integration, we can efficiently manage, display and analyze statistical data with existing capabilities of GIS. Because the factor that influences population distribute is diverse and complicated, the model of population distribution needs further study.

Based on GIS technology, establishing GIS systems which cover spatial data, social, economical and scientific information, and realizing spatializing of statistical data and integration of all kinds of thematic data with geographic infrastructure data, will provide new technical methods for integrating,analyzing,representing,issuing of statistical data, will provide better service for government decision-making.

References

- Dong chun, Zhao rong, Liu ji-ping et al.,2003.Application of geographical parameter database to establishment of unit population databse. Chinese Geographical Science,13(1),34-38
- Kang-tsung Chang,2003.Introduction to geographic information systems. Science Press. Beijing,103-119
- Dong chun, Liu ji-ping, Zhao rong, Wang guixin,2002. An discussion on correlation of geographical parameter with spatial population distribution. Remote Sensing Information.Beijing,68(4) ,61-64
- Li pei,2003.Study on sharing mechanism of geographic information for sustainable development. Proceedings of the seventh China Association for Geographic Information System symposium.46-50
- An kai et al., 2003. Study on constructing statistical information system in Gansu province. Proceedings of the seventh China Association for Geographic Information System symposium.598-602

Zhao rong

Professor

male

was born in Aug. 1966

graduated from Dept. of Geography, Nanjing University, Nanjing and received the Master of Science degree at Cartography and Remote Sensing in 1992.

I am now working at Chinese Academy of Surveying and Mapping, Beijing ,P.R. China. My research interest is in the fields of GIS Project, application of spatial data etc.

My e-mail: zhaorong@casm.ac.cn

Postal adress: 16 Beitaiping road, Beijing, P.R.China

Postcode: 100039