STUDY ON BIG DATABASE CONSTRUCTION AND ITS APPLICATION OF SAMPLE DATA COLLECTED IN CHINA'S FIRST NATIONAL GEOGRAPHIC CONDITIONS CENSUS BASED ON REMOTE SENSING IMAGES

Tao CHENG [a, *], Xu ZHOU [a], Yunpeng JIA [a], Gang YANG [a], Ju BAI [a]

[a] National Geomatics Center of China, 28 Lianhuachi West Road, Haidian District, Beijing, China - chengtao@ngcc.cn

KEY WORDS: National Geographic Conditions Census, Sample Data, Big Data, Database, Distributed File System

ABSTRACT:

In the project of China's First National Geographic Conditions Census, millions of sample data have been collected all over the country for interpreting land cover based on remote sensing images, the quantity of data files reaches more than 12,000,000 and has grown in the following project of National Geographic Conditions Monitoring. By now, using database such as Oracle for storing the big data is the most effective method. However, applicable method is more significant for sample data’s management and application. This paper studies a database construction method which is based on relational database with distributed file system. The vector data and file data are saved in different physical location. The key issues and solution method are discussed. Based on this, it studies the application method of sample data and analyzes some kinds of using cases, which could lay the foundation for sample data’s application. Particularly, sample data locating in Shaanxi province are selected for verifying the method. At the same time, it takes 10 first-level classes which defined in the land cover classification system for example, and analyzes the spatial distribution and density characteristics of all kinds of sample data. The results verify that the method of database construction which is based on relational database with distributed file system is very useful and applicative for sample data’s searching, analyzing and promoted application. Furthermore, sample data collected in the project of China's First National Geographic Conditions Census could be useful in the earth observation and land cover’s quality assessment.

1. INTRODUCTION

The collecting work of sample data for interpreting land cover based on remote sensing images is important in the project of China's First National Geographic Conditions Census (Leading Group Office of China's first National Geographic Conditions Census of the State Council, 2013). Until June 30th 2015, 3 million of sample data have been collected all over the country and the quantity of data files reaches more than 12 million and has grown in the following project of National Geographic Conditions Monitoring. In order to facilitate the following application, how to store and manage these massive sample data scientifically is becoming an important issue in front of the data managers.

With the development of computer technology and information systems, using database for storing, managing and analyzing the massive sample data has become a mainstream method currently (Liu L, 2007; Han J, 2013). In the project of National Geographic Conditions Census, it also use database to manage the sample data to improve data retrieval efficiency. Mature relational database technology uses structured query language to store data in two-dimension table. The technique was pioneered by Oracle Company in the 1970s and is still used most widely currently (Jason, 2014).

However, the quantity of sample data files is very large; besides, each sample data contains varied format data, such as ACCESS, JPG, TIFF, TFW, and XML and so on. Dividing from the view of data model, ACCESS belongs to structured data, JPG, TIFF and TFW belong to unstructured data, and XML belongs to semi-structured data. The typical characteristic of unstructured data is that the field length is unequal, and each field maybe consists of sub fields which are repeated or not repeatable data form. The semi-structured data situates between fully structured data (such as relational data) and completely unstructured data (such as voice, image files, etc.). The typical characteristic of the semi-structured data is that the structure and content of the data mix together, which have no obvious distinction. It is not convenient to use relational database table to store this type of data. So, at the same time of the method for data management and computing ability has improved in a very big promotion, it also faces some challenges (Meng X F, et al. 2013; Liu Z H, et al. 2014).

Therefore, according to the characteristics of sample data, a database construction method which is based on integration of relational database with distributed file system is researched. On the basis of built sample big database, it studies application mode. And it selects large sample data set to validate the research methods combining with spatial analysis.

* Corresponding author

2. COLLECTING METHOD OF SAMPLE DATA

2.1 What is sample data

The project of China's First National Geographic Conditions Census has defined what sample data is. The sample data consists of real picture photographed in field ground and corresponding image example clipped from remote sensing images, including ZY-3, TH-1, WorldView-1, WorldView-2, GeoEye-1, QUICKBIRD, IKONOS, etc., which can be used for interpreting land cover based on remote sensing images. Among these data, the picture is photographed by using a digital camera shooting in the field, which can clearly show object’s morphological features in a certain range; the image example is clipped from aerospace remote sensing images, whose shooting range and content are consistent with the picture.

Furthermore, the image example is clipped in a size of 511 pixel * 511 pixel, and makes the main body locating in the middle part of the image example as far as possible, also ensures shooting point is also within the range of image example. If this size is not suitable for the above conditions, it also can expand to a larger range appropriately.

The real picture and the image example have four kinds of relationships, which are one to one, one to many, many to one, and many to many. The two data reflect the morphological features of object in image from different sides respectively, which can confirm each other and help interpreting land cover for interpreter persons.

2.2 Rule of sample data collecting

The project of National Geographic Conditions Census has developed the technology rule for sample data collecting, which stipulates the collecting content, collecting requirement, and data storage requirement, and so on.

2.3 Product form of sample data

Products of sample data include real picture, remote sensing image example, and meta-database describing the real picture and remote sensing image example.

Each data has respective type, different format, and different standard. Real picture uses JPG format which is according to EXIF standard; remote sensing image example uses uncompressed TIFF format; coordinate information file uses TIFF WORLD document format; projection information file uses XML format which is according to Open Geospatial Consortium (OGC) standard; image metadata uses XML format; and meta-database uses ACCESS database format.

The meta-database includes three tables, which are PHOTO table, SMPIMG table, and PHOTO_IMG table. 19 attributes are recorded in PHOTO table for describing real picture, which are PHID, FILE, PHTM, LONG, LAT, DOP, ALT, MMODE, SAT, AZIM, AZIMR, AZIMP, DIST, TILT, ROLL, CC, REMARK, CREATOR, FOCAL; 14 attributes are recorded in SMPIMG table for describing remote sensing image example, which are IMGID, IMGFILE, SRCTYPE, SRCRES, SRCTIME, SRCBAND, LULONG, LULAT, RULONG, RULAT, LBLONG, LBLAT, RBLONG, RBLAT; and 5 attributes are recorded in PHOTO_IMG table for describing the relationship between real picture and remote sensing image example, which are PHID, IMGID, OPERATOR, EXAMINER, FDATE.

3. DATABASE CONSTRUCTION

3.1 General process

Before loading in the database, the original sample data has to be underwent a series of processing, including pre-processing, quality checks (Cheng T, 2015), problem solving, and file rearranging, and so on.

Also, the point vector data is generated by using the spatial position information recorded in the PHOTO table (Leading Group Office of China's first National Geographic Conditions Census of the State Council, 2015), such as point longitude (LONG attribute) and latitude (LAT attribute). The spatial processing is important for sample data’s application.

In the data warehousing, the technology of oracle database is used in the project. Structured data is stored directly in the oracle database table; point vector data is stored in oracle spatial space in a field defined SDO_Geometry style; unstructured data and semi-structured document data are stored in the distributed file system (Zhou J, et al. 2014).

3.2 Key technology

(1) Attribute information acquisition

The acquisition of any attribute information of sample data which can help users to understand sample data’s detail is a complex technical difficulty in the data collecting.

For the PHOTO table, in the stage of original real picture photographing, if the photographic equipment supports automatic recording camera pose parameters and camera parameter information, much attribute information can be obtained by reading the EXIF metadata labelled of GPS sign, including PHTM labeled of DateTimeOriginal, LONG labeled of GPSLongitude, LAT labeled of GPSLatitude, DOP labeled of GPSDOP, ALT labeled of GPSAltitude, MMODE labeled of GPSMeasureMode, SAT labeled of GPSSatellites, AZIM labeled of GPSImgDirection, AZIMR labeled of GPSImgDirectionRef, CREATOR labeled of Artist, FOCAL labeled of FocalLengthIn35mmFilm, and so on.

For the SMPIMG table, attribute information mainly be obtained from remote sensing image source, SRCTYPE, SRCTIME could be obtained by reading metadata of image source; SRCRES, SRCBAND could be obtained by reading image source directly; LULONG, LULAT, RULONG, RULAT, LBLONG, LBLAT, RBLONG, RBLAT could be obtained by theoretical calculating.

Other attribute information should be obtained by human-computer interaction method.

(2) Graphics drawing

In the process of remote sensing image example acquisition, the photographing position and scope of the real picture should be drawn in the image example, which can directly reflect real picture’s photographing range and main body.

The photographing position is decided according to LONG and LAT attributes, and the photographing scope is calculated by FOCAL attribute as formula (1):

$$v = 2 * atan(18 / FOCAL) \quad (1)$$

where v is photographing scope

FOCAL is the value of FOCAL attribute

The photographing position and scope are both drawn as straight linear graphic in raster image. Straight linear graphic is always composed of pixels in a given matrix which are optimal approximating to a line. There are usually three algorithms for drawing a pixel wide line, which are numerical differential method (DDA), midpoint method and Bresenham algorithm. Among them, Bresenham algorithm is used most widely for its advantage of only needing integer addition and multiplication calculation which avoids stepping in floating calculation (Sun J G, 1998; Li Y R, et al. 2011). Thus, this paper selects Bresenham algorithm for straight linear graphic drawing.

(3) Database construction

In order to update and maintain the massive sample data collected all over the country conveniently, the sample data should be neatened and stored in some dividing units according to national administrative divisions, generally in county administrative divisions or municipal administrative divisions.

In the distributed file system, all sample data in an administrative division maintain a fixed coupling storage structure; each administrative division is arranged in parallel folder with others, and is gathered to the superior administrative divisions step by step, which are municipal, provincial, and national. This method is advantageous for fast retrieval, editing, moving, delete sample data, and so on.

In the relational database, the physical locations of sample data in the distributed file system can be obtained by reading the real picture’ attribute of filename (FILE) stored in the PHOTO table and the image example’ attribute of filename (IMGFILE) stored in the SMPIMG table, which record the relative file path.

4. APPLICATION METHOD OF SAMPLE DATA

The purpose of database construction for massive sampled data is to provide application services, a large amount of valuable information can be mined by big data calculation and analysis (Li Q Q, et al. 2014). Two aspects of application are explored in this paper, which are direct application and derivative application. The direct application provides basic sample data information retrieved from database; and the derivative application provides regular feature information by spatial analyzing on the basis of basic information.

(1) The application for reflecting land covers type and ground surface feature

In some interpreting work, the land cover type cannot be interpreted accurately based on remote sensing image only, and there is no field work result for referencing. In this circumstance, interpreters could retrieve sample data from database to assist interpreting work by using the basic sample data information in research area.

(2) The application for reflecting land covers type in the area of similar geographical environment

The interpreting work based on remote sensing image could be developed in the area of similar geographical environment with sample data. Although there is no field work or it is difficult to reach for field work or interpreters don’t plan to carry out field work in this area. The land cover type could be decided by using correlation analysis method through similar spectral and texture contrasting. Also, the spectral, texture, shape and other features exhibiting in sample data can be used as a priori knowledge for supervised classification.

5. TEST REGION

Shaanxi province locates in northwest of China, links east and west regions. As the original place of the ancient Silk Road, Shaanxi province is the front position opening to the west. Five centers of transportation logistics, technology innovation, industry cooperation, cultural tourism and financial cooperation will be built in Shaanxi province. It plays an important role in the Silk Road Economic Belt construction.

Therefore, this paper selects Shaanxi province as the study area, and takes the sample data in this area for example to validate the efficiency of database construction in key processing process. Meanwhile, some application results are analyzed and discussed.

5.1 General situation of study area

According to the administrative divisions of the People's Republic of China 2015 (Ministry of Civil Affairs of the People’s Republic of China, 2015), Shaanxi Province’s area is about 210 thousand km^2, and has 107 county administrative divisions within the scope. The total population is 39 million 260 thousand.

Shaanxi province locates between the geographical location of 105.483 degrees and 111.250 degrees E longitude, and 31.700 degrees and 39.583 degrees N latitude. The north-south length is about 880km; east-west width is from 160km to 490km. Shaanxi province crosses two basins of the Yellow River and Yangtze River, the average elevation is 1127m, the overall topography exhibits characters of north-south high and middle low. The northern part is the northern Shaanxi plateau, middle part is the Guanzhong plain, and the south part is the Qinba mountainous area. The largest area of these topography regions is Loess Plateau area, accounts for 40%. The domestic climate difference is very big. Due to the complex topography and climate, natural resources and natural resources are extremely rich, and land cover’s diversity is obvious.

5.2 Data processing

The quantity of sample data collected in Shaanxi province reaches to 132 thousand, in the distributed file system, sample data are neatened and stored in 107 dividing units according to county administrative divisions. In order to improve the efficiency of data inspection, 107 units are divided into 5 groups. The specific efficiency is shown in Table 1; computer’s configuration is Windows 7 operating system of 64 bit, 32GB of memory.

Quantity of sample data	132 thousand
Quantity of data files	566 thousand
Time of data inspection	107 minutes
Computer memory occupied	50～2000 MB

Table 1. Efficiency of sample data inspection

5.3 Result analysis

The land cover classification system defines 10 first-level classes, which are cultivated land, garden plot, forest land, grassland, building construction, road, artificial structure, artificial stack land, bare land and water; also it defines 87 third- level classes (Leading Group Office of China's first National Geographic Conditions Census of the State Council, 2013). Some application achievements are obtained by using the method based on sample data collected in Shaanxi province.

(1) Characteristics of spatial distribution and density of the sample data

The spatial distribution of sample data is shown in Figure 1. 10 first-level classes and 82 third- level classes are classified in Shaanxi province.

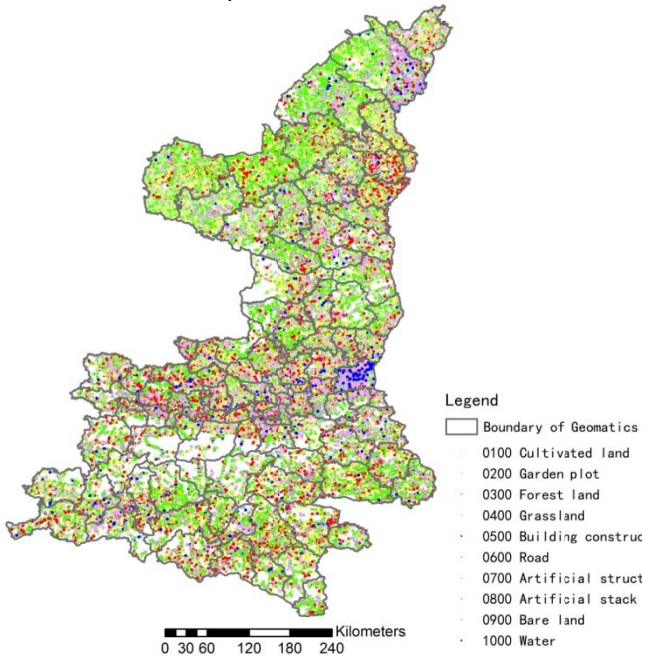

Figure 1. Thematic map of sample data’s spatial distribution

It takes 10 first-level classes for example; the percentage of each class’s quantity is shown in Figure 2.

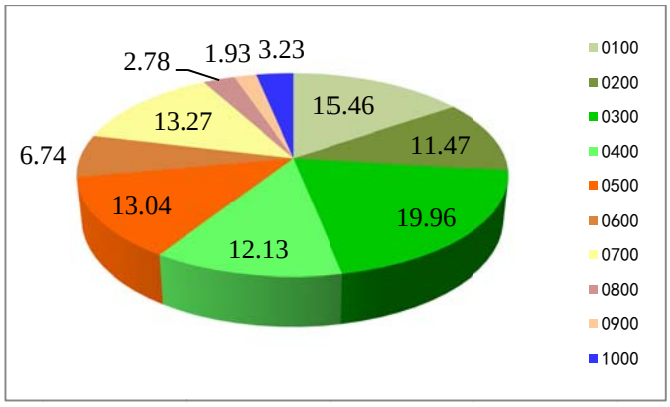

Figure 2. Percentage of each class’s quantity

As shown in Figure 2, the quantity of forest land sample data is the most, amounts to 19.96% of the whole study area. They distribute in all the county administrative divisions. The second most is cultivated land, amounts to 15.46% of the whole study area.

Sample data’s density of each county administrative division could be calculated by the quantity data and area data. The results are shown in Figure 3.

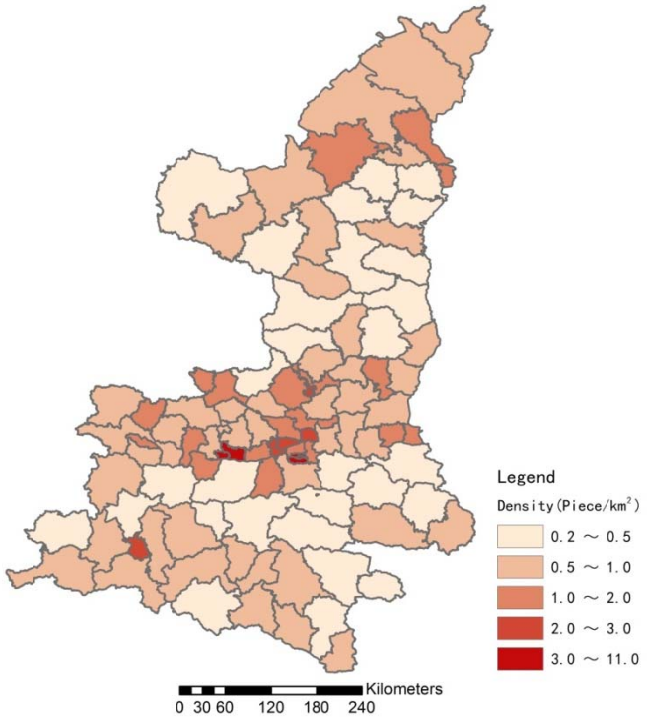

Figure 3. Thematic map of sample data’s density for each county administrative division

The value range of density is from 0.2 to 11.0, which could indirectly reflect the diversity of land cover in different regions.

(2) Morphologic feature in real picture and spectra feature in image example in the same season

Although the times of real picture photographing and image example obtaining are always different, but still a lot of sample data can be selected whose real picture photographing season is same as image example obtaining season by data analyzing. This feature is important to reflect the contrasting relationship of land cover’s morphologic feature and spectra feature, which can be used as priori knowledge for automatic

classification.

Figure 4, Figure 5, and Figure 6 show a sample data of coniferous forest. The real picture (Figure 4) was photographed in July 2014, and the image example (Figure 5) was obtained in July 2011, they are both obtained in July, which is the same season. Figure 6 is the spectral curve map of 4 bands (blue, green, red, and near infrared) on 16 bit Quickbird image. This feature can help users to interpret land cover’s characteristics more accurate.

Figure 4. Real picture of coniferous forest

Figure 5. Image example of coniferous forest

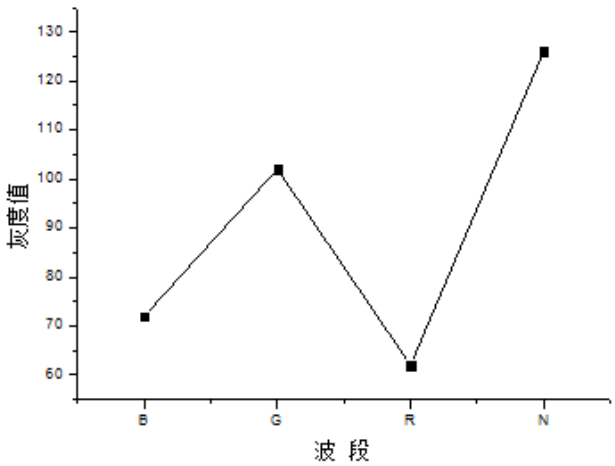

Figure 6. Spectral curve of image band

6. CONCLUSIONS

(1) Sample data accumulated in the project of China's First National Geographic Conditions Census are valuable resources for users. The remote sensing source images where the image samples collected from are very widely, including WorldView-1/2, GeoEye-1, QUICKBIRD, IKONOS, pléiade-1A/1B, ZY-3, Aerial photograph, and so on. For the basic level application, they can provide abundant interpretation symbol information for interpreters, so as to improve accuracy of land cover and land use classification. For the deep level application, some regular characteristic information can be mined through spatial analysis of massive sample data, which can reflect the spatial distribution characteristics, regional characteristics and the diversity of land cover.

(2) Sample data are produced according to the requirements of technology rule developed in the project of China's First National Geographic Conditions Census. They are all standardized products through standardized processing, which would promote the application of them.

(3) Using sample data in the adjacent area or similar geographical region, remote sensing image interpretation can be developed through similar spectral, texture contrasting and geographical correlation analysis methods. In some areas where is difficult to reach or limit to reach, sample data can be collected and used for interpretation, which has practical significance for earth observation and land cover’s quality assessment.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Innovation Development Foundation of National Geomatics Center of China (No.2018-KJ-G01), and National Engineering Project of Geographic National Conditions Monitoring (No.2016-GQ-03-8).

REFERENCES

Cheng T., 2015a. Study on the Method of Sample Data’s Quality Checking during Database Construction in Geographic National Conditions Investigation. *Bulletin of Surveying and Mapping*, 0(10), pp. 103-106.

Han J., 2013. *Research on some key technologies of big data-as-a-service*. Beijing, China: Beijing University of Posts and Telecommunications, pp. 12-32.

Jason P., 2014. *Oracle Database 12c SQL & PL/SQL (3)*. Beijing, China: Tsinghua University Press, pp. 1-22.

Leading Group Office of China's first National Geographic Conditions Census of the State Council., 2013. *Contents and Indexes of National Geographic Conditions Census*. Beijing, China: Surveying and Mapping Press, pp. 1-5.

Leading Group Office of China's first National Geographic Conditions Census of the State Council., 2013. *Technical Method for Data Collection in National Geographic Conditions Census*. Beijing, China: Surveying and Mapping Press, pp. 138-154.

Leading Group Office of China's first National Geographic Conditions Census of the State Council., 2015. *Technical Method for Database Construction in National Geographic Conditions Census*. Beijing, China: Surveying and Mapping Press, pp. 24-38.

Liu L., 2007. *Research on global massive remote sensing image data distributed management technology*. Changsha, China: National University of Defense Technology, pp. 1-20.

Li Q Q., Li D R., 2014a. Big Data GIS. *Geomatics and Information Science of Wuhan University*, 39(6), pp. 641-644+646.

Li Y R., Dong F G., Liu Y., et al. 2011a. Straight line generation algorithm based on pixel line. *Journal of Image and Graphics*, 16(10), pp. 1896-1899.

Liu Z H., Zhang Q L,. 2014a. Research overview of big data technology. *Journal of Zhejiang University (Engineering Science)*, 48(6), pp. 957-972.

Meng X F., Ci X., 2013a. Big Data Management: Concepts, Techniques and Challenges. *Journal of Computer Research and Development*, 50(1), pp. 146-169.

Ministry of Civil Affairs of the People’s Republic of China., 2015. *Administrative division jane book of the People’s Republic of China 2015*. Beijing, China: Sinomap press, pp. 1-7.

Sun J G., 1998. *Computer Graphics*. Beijing: Tsinghua University Press, pp. 165-170.

Zhou J., Wang W P., Meng D., et al. 2014a. Key technology in distributed file system towards big data analyses. *Journal of Computer Research and Development*, 51(2), pp. 382-394.