Image Data Stream Organization and Online Analysis Application Based on Data Cube Technology

Huang Xinyuan[1], Gao Xiaoming[2*], Ouyang Sida[2], Fu Zhengbo[2]

[1]School of Geomatics，Liaoning Technical University，Fuxin 123000，China
[2]Land Satellite Remote Sensing Application Center, MNR, Beijing 100048,China

Keywords: pseudo-data cube, remote sensing data processing, ArcPy, spatial data analysis, Python multiprocessing, multi-scale spatial data cube, grid cell, large-scale remote sensing data processing.

Abstract

This study aims to explore the important role of data-like cube structures in modern remote sensing data processing and data analysis through ArcPy and Python multiprocessing techniques. A multi-scale spatial data cube is innovatively developed to improve the efficiency of remote sensing data management and optimize data analysis. The core of this study is to define and implement grid cells of different sizes that form the basis of data cube, and to quantify the efficient coverage of specific areas using Python multiprocessing techniques. Experiments were conducted in Hainan Province, and efficient data coverage of the whole Hainan Province was realized using the grid data method, which significantly reduced the amount of remote sensing data and processing time required. This shows that the method has successfully improving data coverage capacity and utilization efficiency. The results of this study not only demonstrate the effective application of data-like cubes in remote sensing data processing and analysis, but also provide new perspectives and methods for future complex spatial data analysis and large-scale remote sensing data processing.

1. Introduction

Remote sensing image information technology has been developing rapidly for a long time. A large number of remote sensing satellites have been successively launched, gradually forming various capabilities for earth observation under different conditions(CHEN Li, 2023). As the availability of data increases, it is particularly important to understand the methods of further analyzing and processing the remote sensing images and then refining this data to extract valuable knowledge to guide the scientific decision-making in the fields of agriculture, environment, military, transportation and so on(Zhou Hui, 2010)and(TIAN Hao，2012). However, the large amount of remote sensing data is not effectively utilized and occupies storage space for a long time, resulting in significant waste of resources(LI Deren et al., 2014). Therefore, the effective management and processing of these massive remote sensing data is a major challenge.

[1] mpared with the traditional scene-based processing method, the grid data is used to assess the validity of the image by whether it is cloud-free or not, which greatly improves the efficiency of data utilization. The storage and application of grid data is mainly based on ground positions, which allows different images of the same location to be managed in a unified way and is more conducive to providing standardized data interfaces. The consistency of its spatial dimension simultaneously facilitates temporal data updating and correlation analysis.

Currently, remote sensing images are usually processed on a scene basis, i.e., they are used and analyzed according to specific image scenes captured by remote sensing satellites. Due to the large spatial coverage of remote sensing images, scene based processing and analysis often leads to a considerable computational burden and is not conducive to large-scale spatial analysis.

In this context, this paper proposes a new remote sensing image processing method, which organizes remote sensing images in the form of "grid" data. Co

In addition, from the perspective of processing efficiency, standardized grid data processing is more conducive to computer language organization and parallel computing.

In 1996, Gray et al. (J. Gray et al., 1996) and（Han Jia-Wei et al., 2011)defined the data cube as a multidimensional data organization approach, which provides convenient means for data query and analysis by organizing the data in multiple dimensions such as time and location, thus greatly improving the efficiency of data utilization. We adopt this method to organize

Fund Projects: National Key R & D Plan Key Special Project (2022YFB3903601).
first author: Huang Xinyuan, E-mail:15643842098@163.com.
* corresponding author: Gao Xiaoming, E-mail:gaoxm@lasac.cn.

remote sensing images by multi-dimensions such as space and time to form a remote sensing image data cube, which improves scene-based processing of remote sensing images and facilitates large-scale spatial and temporal analysis.

This paper therefore proposes a new remote sensing image processing method that draws on the concept of the data cube, which significantly improves the efficiency and flexibility of data processing by organizing remote sensing images in the form of grid data. The design principle, implementation steps of this method, and its application to cases in Hainan Province are introduced in this paper, with the aim of providing new ideas and methods for the efficient management and application of remote sensing images.

2. Data Structure Model

In order to realize efficient and accurate processing of remote sensing images, a mathematical model was constructed with grid data as the core. The model is centered on four key components: grid, grid entity, index, and XML attributes, and these components and their roles in remote sensing image processing are described in detail below.

2.1 Grid

The grid created by ArcPy's CreateFishnet_management function forms the infrastructure of this model, dividing the vast geographic range of remote sensing images into standardized and manageable grids. Specifically, remote sensing images are segmented into equal-sized grid cells, each of which covers a specific geospatial region and contains remote sensing images for that region. This partitioning method transforms complex, large remote sensing images into a series of smaller, more manageable grids, paving the way for efficient management and parallel processing.

2.2 Grid entity

Grid entities are specific remote sensing image data within a grid cell. Each grid entity contains all the remote sensing data of the corresponding region, such as geographic information and spectral information. The remote sensing image is divided into several grid entities by the ExtractByMask function in ArcPy, which are then processed independently. Parallel processing of remote sensing images is thereby realized through multiprocessing in a pool library, greatly improving the data processing efficiency.

2.3 Index

To quickly locate each grid entity, an indexing system was designed for each grid entity. The index contains key information such as the position, size, and data type of the grid entity in the grid. This indexing mechanism improves the efficiency and convenience of accessing and processing specific grid entities.

2.4 XML attributes

In addition to the raw data of the remote sensing image, each grid entity is accompanied by associated metadata stored in XML attributes. This metadata may include detailed information such as geographic coordinates, timestamps, and image format, allowing for accurate data query and analysis.

In summary, through the four components of grid, grid entity, index, and XML attributes, this model provides a rigid and flexible framework for efficient processing of remote sensing images. In the next sections, the efficient use of this model for processing and analysis of remote sensing images is discussed.

3. Implementation Method

3.1 Processing flow

This study aims to realize effective processing and indexing of remote sensing data to support applications such as environmental monitoring and geographic information system (GIS) analysis. The main process is indicated in Fig. 1.

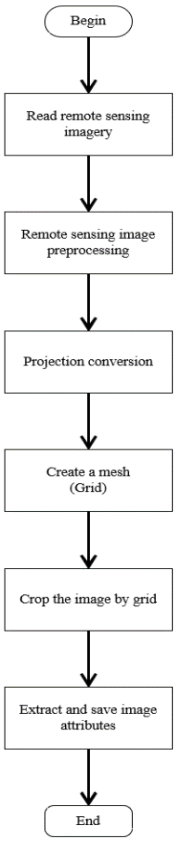

Figure 1. Flow Chart for remote sensing image data processing

Environment setup and data preparation: First, initialize the ArcPy environment and set up the workspace. Then, define the paths for inputting remote sensing image data and provincial boundary vector data.

Image preprocessing: For image data in JPEG format, use the ArcPy tool to convert to TIFF format

and make sure all data has the correct coordinate system.

Data projection and range determination: Project the provincial boundary data and remote sensing images into a unified planar coordinate system to calculate the extent of the provincial boundary.

Grid generation and adjustment: Create grids of different sizes (5 km, 10 km, 20 km, 40 km) based on the extent of the provincial boundaries. Adjust the grid to ensure that it fully covers the provincial boundaries.

Parallel processing: Use Python's multi-process feature to generate grids of different sizes in parallel for efficiency.

Image cropping and extraction: Crop the remote sensing image using the ExtractbyMask function according to the generated grid to extract the image portion corresponding to the grid.

Index file generation: Generate an index file in XML format for each cropped image to record the basic information and geographic location data of the image. Finally, a total tree-structured XML file is generated, containing all the data and locations.

Data storage and cleanup: Store cropped images and index files according to a defined directory structure, and clean up temporary files generated during the process.

3.2 Storage structure

The image data is stored in a hierarchical folder structure, with each folder corresponding to a grid of a particular size. Images for each grid are stored in subfolders named after the center of the grid. Each subfolder contains cropped remote sensing images and XML index files. The cropped images and index files are stored according to the structure indicated in Fig. 2.

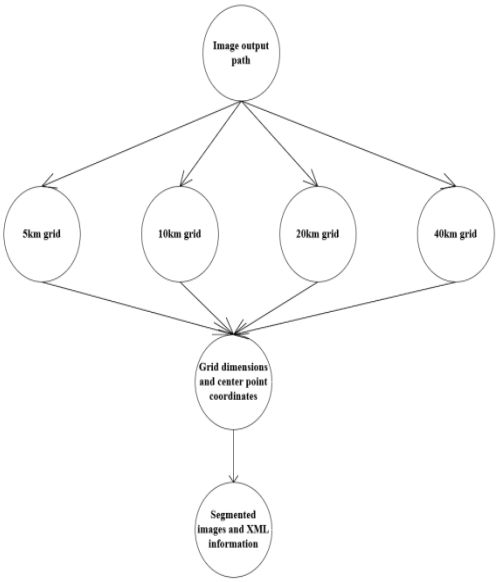

Figure 2. Hierarchical folder structure for storage remote sensing images and XML index files.

3.3 Indexing method for each cropped image using XML format

In remote sensing data processing and GIS analysis, an efficient data indexing method is essential to improve data retrieval speed and processing efficiency. In this study, an XML-based indexing method is adopted to organize and manage remote sensing image data and its metadata. The method aims to provide a flexible and easily extensible data indexing mechanism to support complex remote sensing data processing and analysis.

3.3.1 Index structure design

The structure of the index file is based on the XML format, which uses its hierarchical and self-describing characteristics to define a set of tagging systems for describing remote sensing images and grid information. Each index file corresponds to multiple remote sensing image data under one grid cell and contains the following main information:

GridID: A unique identifier for each grid cell that ensures the consistency of the index file.

Details: Each Details element has a unique ID attribute that identifies a specific remote sensing image instance, ensuring accurate referencing of the data.

GridName: The name of the grid that uniquely identifies the grid cell.

ImagePath: The path that records the location of the cropped remote sensing image file.

Rows and Cols: The image size, specifying the number of rows and columns of the image.

LongitudeMin, LatitudeMax, LongitudeMax, LatitudeMin: Geographic ranges that describe the minimum longitude, maximum latitude, maximum longitude and minimum latitude covered by the image.

SatelliteSource: Satellite data source that records the name of the satellite that produced the remote sensing image.

Through this structural design, the index file not only provides basic information about remote sensing images, but also supports efficient spatial queries and image retrieval by providing information about geographic ranges and grid cells.

3.3.2 Index generation

The generation of index files is done automatically in the last step of the remote sensing image data processing flow. Whenever the remote sensing image cropping and processing of a grid cell is completed, the corresponding index file will be generated based on the metadata information of the image and stored in the same directory as the image file. After the XML files of all grid cells have been generated, they are merged into one overall XML file. This automated index file generation mechanism ensures timely updating and accuracy of data indexing.

In summary, through this automated index generation process, the data index is updated in real time and the accuracy and efficiency of data retrieval and management is improved, laying a solid foundation for subsequent analysis and application.

4. Application Evaluation

Users can select different grid sizes according to their needs, quickly locate and access related remote sensing images through index files, and apply them to GIS, remote sensing data analysis, regional planning and other fields.

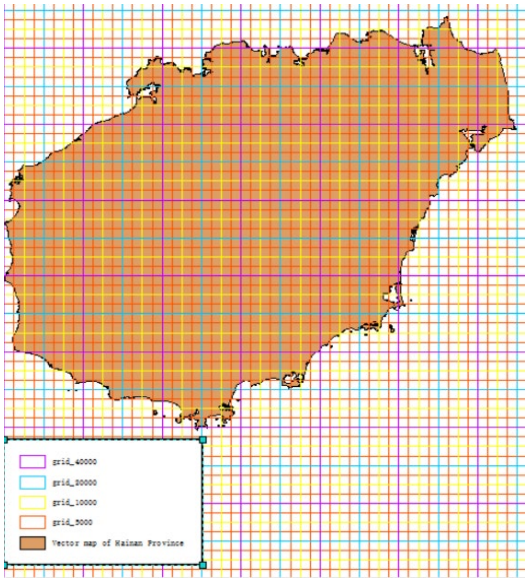

Figure 3. Grid Map of Hainan Province Vector Map

4.1 Advantages and disadvantages of grids and scenes

In the field of remote sensing image processing, the grid and scene processing methods each have their own characteristics and applicable scenarios. The grid processing method, such as the grid-based image segmentation used in this project, is advantageous because of its efficient data processing capability and ease of large-scale analysis. By dividing large-area images into regular small grids, the data volume for each processing task can be reduced, thereby increasing the processing speed and reducing the demand for computational resources. In addition, grid processing methods facilitate parallel computing, which further improves data processing efficiency. However, this method can lead to edge effects, where discontinuities or overlaps may occur at the edges of the grids, potentially causing errors in subsequent image analysis. In contrast, the scene processing method uniformly processes the entire landscape area of the image. This method can preserve the integrity of the image and avoid the edge effect problems encountered with grid processing. When dealing with geophysical phenomena with strong continuity, such as land cover change monitoring, scene processing can provide more coherent and consistent results. However, scene processing methods face challenges in terms of computational resources and processing efficiency when dealing with large-scale or high-resolution images.

4.2 Feasibility analysis of data production and industry applications

In the processing and analysis of remote sensing data, the feasibility of data production is closely related to the practicality of industry applications. Automated processing tools and methods can not only improve the efficiency and accuracy of data processing, but also greatly enhance the application value of remote sensing data in various fields.

4.2.1 Optimization of data production

Using the remote sensing data processing in Hainan Province as an example, the traditional method of covering the entire province's data may require the processing of over 100 scenes over a period of four months. This approach is not only time-consuming, but also inefficient. In contrast, by segmenting the data into grids, the number of images to be processed can be significantly reduced to 50 scenes, and the time required can be reduced to 1 month. This optimization significantly improves the feasibility of data production, including data coverage capability and utilization efficiency.The coverage method for Hainan Province, as shown in Figure 4 and Figure 5.

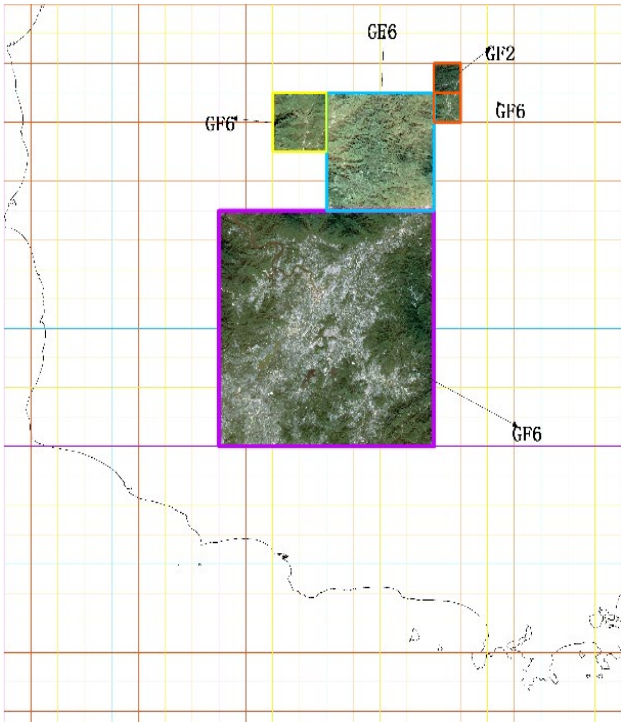

Figure 4. The application of different satellites in coverage

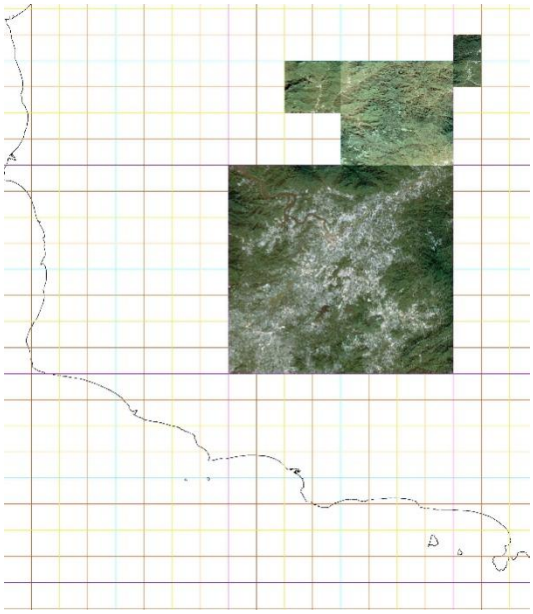

Figure 5. Image Overlay Applications

Automated processing processes such as spatial analysis, image cropping, and cloud detection using the ArcPy library in Python scripts further improve the efficiency of data processing. The application of multi-process parallel processing techniques and the introduction of data quality control tools, such as cloud cover detection, ensure efficient and reliable production of quality data.

4.2.2 Expansion of industry applications

The optimization of remote sensing data processing has not only improved the feasibility of data production, but also greatly expanded its application in various industries. In fields such as agriculture, environmental monitoring, urban planning, and disaster management, the rapid acquisition and processing of high-quality remote sensing data has become critical to supporting decision making, monitoring, and evaluation.

For example, in agriculture, optimized data processing methods can support real-time monitoring of crop health and improve the implementation efficiency of precision agriculture. In environmental protection, it can effectively monitor changes in forest cover or water pollution, and provide a scientific basis for the formulation and implementation of environmental policies. In disaster management, fast and accurate remote sensing data processing capability can provide timely disaster impact assessment and support the efficient implementation of rescue and recovery work.

Overall, through automated and optimized data processing workflows, the feasibility of remote sensing data production, including data coverage capability and utilization efficiency, and its practicality in various industrial applications can be greatly enhanced. With continuous technological advancement and expanding application scope, the processing and utilization of remote sensing data will become more flexible and efficient, thus providing greater support for socio-economic development and environmental protection.

5. Conclusions

5.1 Feasibility

Based on modern remote sensing techniques and GIS tools, the implementation of this gridded method is highly feasible. In terms of technical support, the advanced remote sensing data processing techniques and computational capabilities make this method applicable in a wide range of scenarios.

5.2 Practicability and integration

In scenarios where large-scale remote sensing data need to be efficiently processed and analyzed, such as urban planning, environmental monitoring, and disaster response, this method provides powerful support and is therefore considered practical.

This method can be easily integrated into existing GIS and remote sensing data processing flows.

5.3 Shortcomings and directions for improvement

5.3.1 Shortcomings

There are apparent boundary issues relating to this method. Grid edges may lead to data discontinuity and information loss. The existence of overlapping regions may lead to data duplication, redundancy, and reduced storage efficiency. When dealing with high resolution remote sensing data, there is a huge demand for storage space, which may be a limiting factor. Lastly, as the volume of data increases, maintaining and updating the data becomes more complex.

5.3.2 Directions for improvement

Boundaries and their impact can be reduced by using more flexible grid designs or boundary fusion techniques. Efficient data compression techniques can be utilized to reduce storage requirements. More advanced data indexing and retrieval systems can be developed to improve data management efficiency. Cloud storage techniques can be used to distribute storage and processing pressure and improve scalability.

In summary, the gridded processing method is feasible and practical in the field of remote sensing data management and analysis. Although there are some challenges, these shortcomings can be improved and solved through technological innovations and strategy adjustments.

Acknowledgements

The authors thank the editors and the reviewers for their constructive and helpful comments, which led to substantial improvement of this paper.This work was supported by National Key R & D Plan Key Special Project (No.2022YFB3903601).The authors are also

immensely grateful to the Land Satellite Remote Sensing Application Center, Ministry of Natural Resources, for their invaluable training and resources that greatly contributed to the completion of this research. The opportunities provided by the center have been pivotal in the successful development of this project.We would like to thank everyone who was involved in this study for their support and encouragement.

References

CHEN Li., 2023:Toward Trustworthy Intelligence for High-Resolution Remote Sensing Image Scene Classification. *Geomatics and Information Science of Wuhan University*, 48(12), 2104-2104.

Zhou Hui., 2010. Hierarchical Analysis Method for High Resolution Remote Sensing Images. Ph.D. dissertation,National University of Defense Technology, Changsha, Hunan, CHN.

TIAN Hao., 2012. Research on Image Understanding of Building Regions in Remote Sensing Images. Ph.D. dissertation,National University of Defense Technology, Changsha, Hunan, CHN.

LI Deren.,ZHANG Liangpei.,XIA Guisong.,2014:Automatic Analysis and Mining of Remote Sensing Big Data. *Acta Geodaetica et Cartographica Sinica* , 43(12), 1211-1216.

J. Gray., A. Bosworth., A. Lyaman and H. Pirahesh., 1996:Data cube: a relational aggregation operator generalizing GROUP-BY, CROSS-TAB, and SUB-TOTALS. *Proceedings of the Twelfth International Conference on Data Engineering*, New Orleans, LA, USA, pp. 152-159.

Han Jia-Wei., Kamber M., Pei Jian.,2011:*Data mining : concepts and techniques. Third Edition*. San Francisco : Morgan Kaufmann.