doi:10.3772/j.issn.1006-6748.2020.04.014

A spatial decomposition approach for accelerating buffer analysis of vector data①

Li Xiaohua(李晓华)[*,**], Guo Mingqiang②[***], Qi Xinhong[**]
([*] School of Safety Science and Engineering, Henan Polytechnic University, Jiaozuo 454003, P. R. China)
([**] Guizhou Coal Mine Design Research Institute Co., Ltd, Guiyang 550025, P. R. China)
([***] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, P. R. China)

Abstract

Parallel vector buffer analysis approaches can be classified into 2 types: algorithm-oriented parallel strategy and the data-oriented parallel strategy. These methods do not take its applicability on the existing geographic information systems (GIS) platforms into consideration. In order to address the problem, a spatial decomposition approach for accelerating buffer analysis of vector data is proposed. The relationship between the number of vertices of each feature and the buffer analysis computing time is analyzed to generate computational intensity transformation functions (CITFs). Then, computational intensity grids (CIGs) of polyline and polygon are constructed based on the relative CITFs. Using the corresponding CIGs, a spatial decomposition method for parallel buffer analysis is developed. Based on the computational intensity of the features and the sub-domains generated in the decomposition, the features are averagely assigned within the sub-domains into parallel buffer analysis tasks for load balance. Compared with typical regular domain decomposition methods, the new approach accomplishes greater balanced decomposition of computational intensity for parallel buffer analysis and achieves near-linear speedups.

Key words: high performance spatial computing, buffer analysis, parallel computing, load balancing, vector data

0 Introduction

A buffer in geographic information systems (GIS) is defined as the zone around a geometric geographic feature, measured in units of distance or time[1]. Buffer analysis plays an important role in many applications of GIS, such as environmental measurement and management[2,3], human health[4,5], landscape and urban planning[6,7], geographic data processing and representation.

According to the parallel strategies for spatial analysis, existing studies can be classified into 2 categories: the algorithm-oriented parallel strategy and the data-oriented parallel strategy.

The first is the algorithm-oriented parallel strategy, a strategy that generally changes the current spatial algorithms to make full use of the parallel computing framework to achieve better parallel performance[8-10]. Some researchers proposed a parallel buffer algorithm based on area merging and message passing interface (MPI) to improve the performance of buffer analysis on processing large datasets. A visualization-oriented buffer analysis method which was developed based on a fully optimized hybrid-parallel processing architecture was proposed by Ma et al.[11], they put forward an efficient spatial-index-based buffer generation method to generate the results.

The second category is the data-oriented parallel strategy, which mainly focuses on data partition and data organization to suit the corresponding parallel framework. Some researchers developed a distributed spatial index based on Apache Storm, which is an open-source distributed real-time computation system[12]. There are many great improvements in spatial index and data skew in Hadoop. A cluster-computing-oriented parallel vector buffer generating algorithm was proposed by Shen et al.[13], which contains a data partition method based on Hilbert space filling curve.

These parallel approaches mentioned above have obtained high performance of spatial operations. However, the improvement of each existing algorithm is a

① Supported by the National Natural Science Foundation of China (No. 41971356, 41701446) and National Key Research and Development Program of China (No. 2017YFB0503600, 2018YFB0505500, 2017YFC0602204).
② To whom correspondence should be addressed. E-mail: gmqandjxs@163.com
Received on Jan. 3, 2020

very complex work, and it requires vast redevelopment. In order to address the problem, a spatial decomposition approach for vector buffer analysis is proposed.

The rest of the paper is organized in the following. Section 1 articulates the spatial decomposition approach. Section 2 presents a series of experiments to demonstrate the effectiveness and performance of the new approach proposed by this paper. Conclusion and future work are given in Section 3.

1 Methods

1.1 Construction of computational intensity model

The relationship between the computing time and the number of a feature's vertices for the retrieve, buffer and write steps of polyline and polygon buffer analysis can be represented by linear model[14,15]. Thus, these models can be used to generate the computational intensity transform functions (CITFs), so as to estimate the computational intensity of generating a group of polyline and polygon buffer analysis results. First, the sub-CITFs of the single polygon or polyline feature can be built, as shown in Eq. (1) and Eq. (2).

$$CL(x) = (a_1 + a_2 + a_3)x + (b_1 + b_2 + b_3) \tag{1}$$

$$CP(x) = (a_4 + a_5 + a_6)x + (b_4 + b_5 + b_6) \tag{2}$$

where, CL, CP are the computing time of the polyline and polygon buffer analysis respectively; x is the number of vertices of a polyline or a polygon feature; a_1, a_2, a_3, a_4, a_5, a_6 are the slope of the functions of 3 steps respectively; and b_1, b_2, b_3, b_4, b_5, b_6 are the intercept respectively.

Then the overall CITFs can be constructed for a group of polylines and polygons.

$$WL = \sum_{i=1}^{n} CL(x_i) \tag{3}$$

$$WP = \sum_{i=1}^{n} CP(x_i) \tag{4}$$

where, WL is the overall computing time for a group of polylines, WP is the overall computing time for a group of polygons, n is the number of polylines or polygons, x_i is the number of vertices of the polyline or polygon feature i.

The CITFs can be used to estimate the computational intensity of generating buffers for a group of polylines or polygons. It is significant for the spatial representation of buffer analysis computational intensity.

1.2 Spatial representation of computational intensity

In order to ensure the effectiveness of the parallel scheduling method, the spatial distribution of computational intensity of buffer generation must be properly represented. In this research, the computational intensity surface (CIS) approach proposed by Wang et al.[15] is exploited to solve this issue. The spatial computational domain of a vector layer is divided into a group of regular lattices so that a computational intensity grid (CIG) for buffer analysis can be generated. A 4 × 4 CIG of polygon dataset is shown in Fig. 1, where W_{ij} is the computational intensity of the lattice at row i and column j in the grid. W_{ij} can be calculated by Eq. (3) and Eq. (4) for polylines and polygons respectively.

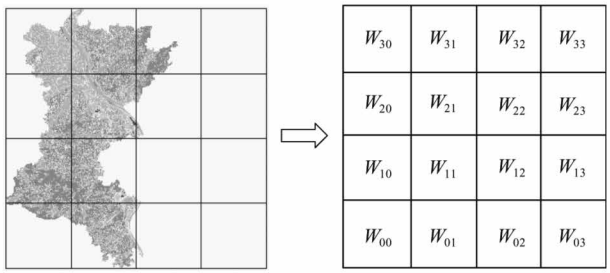

Fig. 1 A 4 × 4 CIG for the polygon buffer analysis

1.3 Spatial decomposition for parallel buffer analysis

The ideal decomposition for parallel buffer analysis is to ensure that every task has similar computational intensity, so that all the parallel tasks can be completed simultaneously. However, regular decomposition strategies only pay attention to decomposing sub-domains evenly in areas. Vertical decomposition (VD) method divides the whole domain to equally sized column-like sub-domains (Fig. 2(a)), while horizontal decomposition (HD) generates equally sized row-like sub-domains (Fig. 2(b)). And the vertical and horizontal decomposition (VHD) generates the block-like sub-domains by both columns and rows (Fig. 2(c)). These regular decomposition methods do not take the spatial distribution of the features in the dataset into consideration, and just decompose the computational domain to sub-domains evenly in areas. If the features are not homogeneous in space, the regular decomposition methods may result in great load imbalance among sub domains. Therefore, the spatial decomposition strategy based on computational intensity is proposed to address the problem.

The spatial decomposition (SD) method is based on HD or VD method. This approach can effectively divide spatial computational domain into sub-domains with same computational intensity. As shown in Fig. 3, after CIG is formed, the sum (W_0, W_1, W_2, W_3) of the computational intensity of the lattices is firstly computed for each row. W_{total} can be calculated as the overall computational intensity, and W_{task} can be calculated as the computational intensity of each sub task. Then,

the computational intensity grid needs to be scanned row by row, the computational intensity of each row should be compared with W_{task}. All of the rows will be scanned and all sub-domains with same computational intensity will be generated.

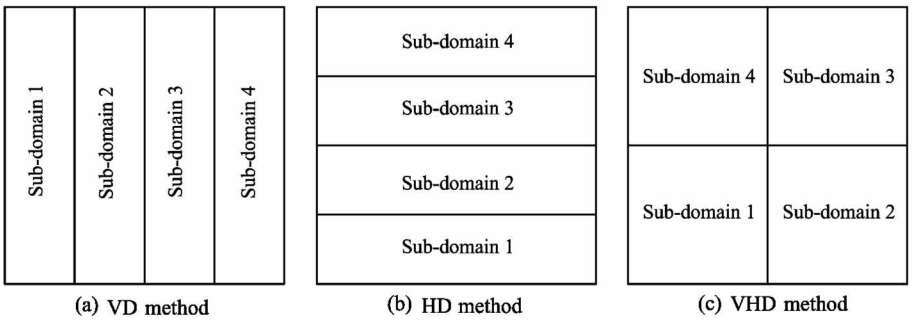

Fig. 2 Three regular decomposition methods

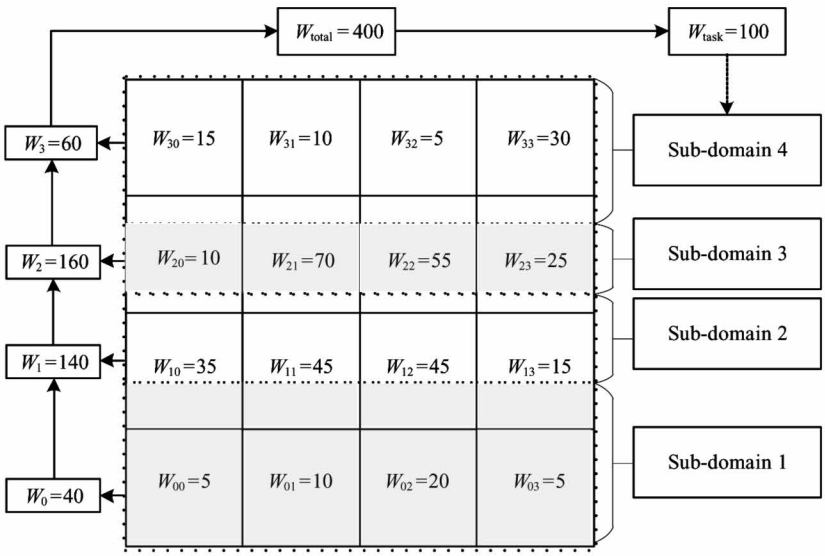

Fig. 3 The workflow of the SD method

2 Experiments

2.1 Experimental environment and dataset description

In order to evaluate the performance of the proposed method, a group of experiments are conducted in the parallel buffer analysis framework on QGIS platforms. SD is compared with regular decomposition methods. The computing nodes are composed of 2 Intel Xeon E5620 8-core CPUs at 2.4 GHz and 16 GB of memory. And the experiments are conducted by QGIS SDK 3.4.8.

Aiming to demonstrate the availability and efficiency of proposed SD methods, 2 real-world vector datasets are adopted in the experiments (Table 1). Dataset A and dataset B present the same geographic objects with different feature type.

Table 1 Description of vector dataset

Dataset name	Feature type	Number of features	Number of vertices	Size (MB)
Dataset A	polygon	93 368	3 994 495	127
Dataset B	polyline	100 574	3 994 495	134

2.2 Experiments and performance assessments

Firstly, the API of QGIS is selected to conduct the parallel buffer analysis task with varying numbers of threads. The sub-domains of parallel buffer analysis task are generated by VD, HD and SD respectively. The VD and HD methods divided the computational domain by area. These methods can be easily realized, but they will lead to great load imbalance. In this

work, 32 × 32 CIGs are used to conduct a group of experiments. Three various 8 sub-domains decomposition results of dataset A and dataset B are shown in Fig. 4 and Fig. 5. The decomposition results of VD and HD are uneven in computational intensity, while the computing load of sub-domains generated by SD is almost equal.

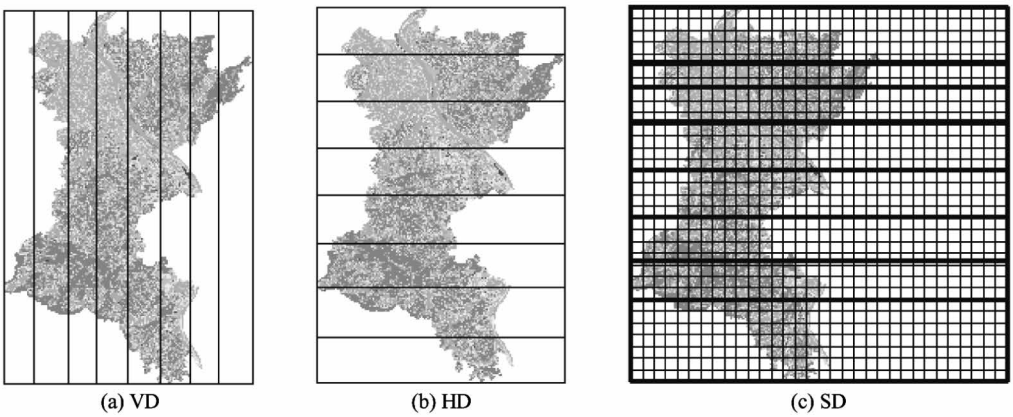

Fig. 4 Three types of decomposition for dataset A

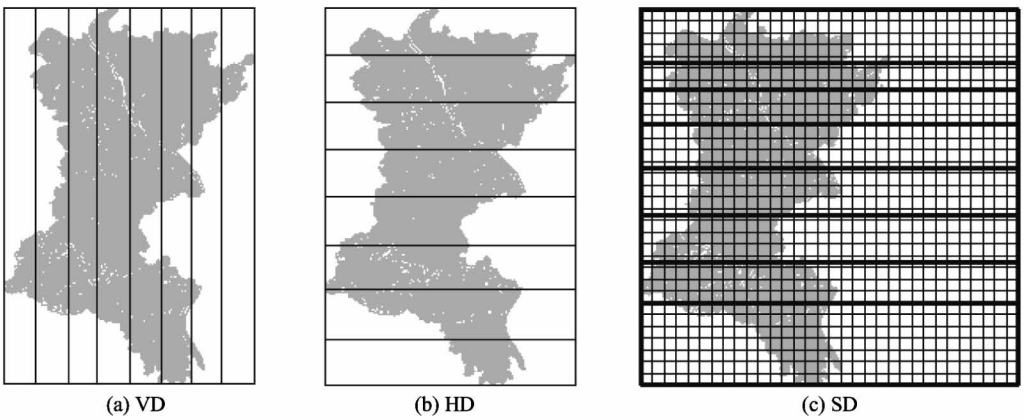

Fig. 5 Three types of decomposition for dataset B

A serial buffer analysis program using dataset A and dataset B is conducted to offer the benchmark for assessing the performance of the parallel program. The computing time of serial buffer analysis of dataset A is 23 076.335 ms, and that of dataset B is 50 272.347 ms. A set of experiments are carried out by using the parallel buffer analysis program with various numbers of threads (2 – 8). As shown in Fig. 6, the computing time decreases with the increasing number of threads, and SD achieves the best performance. Fig. 7 shows that SD achieves near-linear speedups on dataset A and dataset B, and the speedups are greatly higher than that

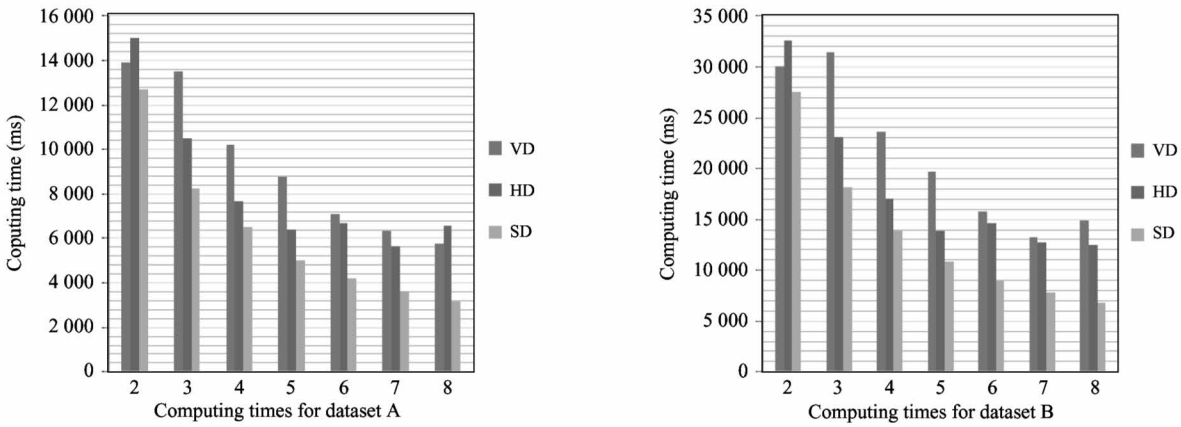

Fig. 6 Computing times of 3 decomposition methods

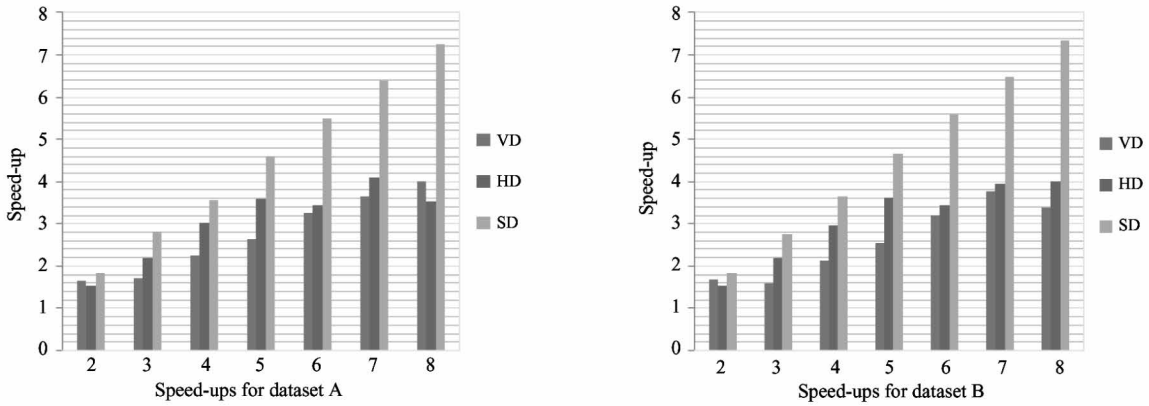

Fig. 7 Speed-ups of 3 decomposition methods

of VD and HD methods. The reason of near-linear speedups of SD is that the spatial distribution of computational intensity is taken into consideration. The new approach ensures that the workload is averagely assigned to parallel computing nodes.

3 Conclusion

With the growing volume of spatial data, existing vector buffer analysis algorithms cannot meet the demands of fast data processing. In this work, a spatial decomposition for vector buffer analysis based on spatial computational intensity is proposed, so as to generate balancing sub-domains in parallel environment.

With the relationship between the number of vertices and the buffer analysis computing time, the CITFs are generated to estimate the computational intensity. Based on the CITFs, CIGs of polyline and polygon are constructed to represent the spatial distribution of computational intensity for buffer analysis. The computational domain can be effectively divided by the spatial decomposition approach developed in this work.

Future work will focus on how to partition the vector features distributed in the adjacent area of 2 sub-domains, so as to further address the balance partition problem for vector data spatial analysis.

References

[1] Ramasubramanian L. A to Z GIS: an illustrated dictionary of geographic information systems[J]. *Journal of Planning Literature*, 2009,23(3): 263-264

[2] Saha A K, Gupta R P, Sarkar I, et al. An approach for GIS-based statistical landslide susceptibility zonation-with a case study in the Himalayas[J]. *Landslides*, 2005,2(1): 61-69

[3] Sliva L, Williams D D. Buffer zone versus whole catchment approaches to studying land use impact on river water quality[J]. *Water Research*, 2001,35(14): 3462-3472

[4] English P, Neutra R, Scalf R, et al. Examining associations between childhood asthma and traffic flow using a geographic information system[J]. *Environmental Health Perspectives*, 1999,107(9): 761-767

[5] Vine M F, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research[J]. *Environmental Health Perspectives*, 1997,105(6): 598-605

[6] Mhuireach G, Johnson B R, Altrichter A E, et al. Urban greenness influences airborne bacterial community composition[J]. *Science of the Total Environment*, 2016,571: 680-687

[7] Xiang W N. GIS-based riparian buffer analysis: injecting geographic information into landscape planning[J]. *Landscape and Urban Planning*, 1996,34(1): 1-10

[8] Liu C Y, Xiong L, Hu X Y, et al. A progressive buffering method for road map update using open street map data[J]. *ISPRS International Journal of Geo-Information*, 2015,4(3): 1246-1264

[9] Tveite H, Langaas S. An accuracy assessment method for geographical line data sets based on buffering[J]. *International Journal of Geographical Information Science*, 1999,13(1): 27-47

[10] Fan J, Ji M, Gu G, et al. Optimization approaches to MPI and area merging-based parallel buffer algorithm[J]. *Boletim de Ciencias Geodesicas*, 2014,20(2): 237-256

[11] Ma M Y, Wu Y, Chen L, et al. Interactive and online buffer-overlay analytics of large-scale spatial data[J]. *ISPRS International Journal of Geo-Information*, 2019,8(1): 1-14

[12] Zhang F, Zheng Y, Xu D P, et al. Real-time spatial queries for moving objects using storm topology[J]. *ISPRS International Journal of Geo-Information*, 2016, 5(10): 178

[13] Shen J X, Chen L, Wu Y, et al. Approach to accelerating dissolved vector buffer generation in distributed in-memory cluster architecture[J]. *ISPRS International Journal of Geo-Information*, 2018,7(1): 26

[14] Guo M, Guan Q, Xie Z, et al. A spatially adaptive decomposition approach for parallel vector data visualization of polylines and polygons[J]. *International Journal of Geographical Information Science*, 2015,29(8): 1419-1440

[15] Wang S, Armstrong M P. A theoretical approach to the use of cyberinfrastructure in geographical analysis[J]. *International Journal of Geographical Information Science*, 2009,23(2): 169-193

Li Xiaohua, born in 1983. He is currently a Ph. D candidate from Henan Polytechnic University. He received his B. S. and M. S. degrees from Henan Polytechnic University in 2007 and 2010 respectively. His research interests include GIS application, coal mine information, intelligence and gas disaster prevention.