Comparative Study on the Effect of Shape Complexity on the Efficiency of Different Overlay Analysis Algorithms

PENG ZHANG1,2, JUNFU FAN1, PANPAN ZHANG3, ZHIKUN ZHANG1, ZHENG CHEN1, AND LIUSHENG HAN1

1School of Civil and Architectural Engineering, Shandong University of Technology, Zibo, Shandong 255000, China
2State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
3National Ocean Technology Center, Tianjin 300112, China

Corresponding author: Junfu Fan (fanjf@sdut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 42171413, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2020MD015 and Grant ZR2020MD018, in part by the National Key Research and Development Program of China under Grant 2017YFB0503500, in part by the Major Science and Technology Innovation Project of Shandong Province under Grant 2019JZZY020103, and in part by the Young Teacher Development Support Program of Shandong University of Technology under Grant 4072-115016.

ABSTRACT Polygon shape is an important factor affecting the efficiency of vector polygon overlay analysis. Aiming at the problems that polygon shape is difficult to accurately measure and different algorithms have different sensitivities to the same polygon in vector polygon overlay analysis, 27 shape variables are selected from literature to establish the shape complexity models for Vatti, Greiner-Hormann and Martinez algorithms by multiple linear regression analysis in this paper. These three algorithms are suitable for arbitrary polygon clipping and include intersection, union, difference and symmetric difference operators. The experimental results show that the calculated performances of the intersection, union, difference and symmetric difference operators belonging to the same algorithm are similar to their sensitivity to the same polygon shape. The fitting correlation coefficient $R^2$ of the shape complexity model based on the Greiner-Hormann algorithm is the highest among the three algorithms, which can explain 99% of the performance of this algorithm.

INDEX TERMS Shape complexity, overlay analysis, multiple linear regression.

I. INTRODUCTION

Non-topological vector data have become the industry standard in Geographic Information System (GIS) due to their high rendering speed and small storage requirements [1]–[4]. As one of the core algorithms in the field of geocomputing [5], vector polygon overlay analysis have been widely used in geographical field.

Polygon overlay analysis refers to Boolean operations such as intersection, union, difference and symmetric difference of two feature layers in the same area [6]. The result is that new features are generated by dividing the original features, and the new feature layer integrates the attributes of the original two feature layers [7], [8]. The two feature layers that perform the Boolean operations are the subject layer and the clipping layer, which can be either a single polygon or a set of polygons [9], [10]. Polygon overlay analysis algorithms have been fully developed and improved in the past decades, and many algorithms have been proposed by relevant scholars to deal with various problems in polygon clipping. To deal with rectangular clipping polygons, scholars such as Sutherland-Hodgeman, Liang-Barsky, Foley and Maillot have proposed many solutions [11]–[15]. The polygons involved in the overlay analysis can also be circles, ellipses, arbitrary polygons and others [16]–[20]. However, in real applications, overlay analyses between arbitrary polygons are more common. Among them, Vatti algorithm [21], Greiner-Hormann algorithm [22] and Martinez algorithm [7], [9] are recognized as effective algorithms that can handle arbitrary polygon clipping problems in a limited time and obtain correct results [8], [23].
TABLE 1. Basic information statistics of the experimental data.

| Datasets                          | Number of polygons | Number of vertices | Number of holes | Data sources                  |
|-----------------------------------|--------------------|--------------------|-----------------|-------------------------------|
| World county-level vector boundaries | 255272             | 79522711           | 5192            | Database of Global Administrative Areas |
| Land patches in China             | 509219             | 12914282           | 30825           | OpenStreetMap                 |
| Water system in China             | 222411             | 12505545           | 28096           | OpenStreetMap                 |

FIGURE 1. Experimental data.

With the development of surveying and mapping technology and the progress of data acquisition technology, people have acquired a large amount of geospatial data, especially geographic polygons [24], [25]. Different from simple regular polygons, geographic polygons are extremely irregular and complex and may be concave, convex, self-intersecting or have many complex holes and island structures [26] containing dozens to tens of thousands of vertices. Polygonal shape is an important factor affecting the efficiency of overlay analysis [26]. The large differences among different various geographic polygons lead to significant efficiency differences in overlay analysis. Overlay analysis involving millions or even tens of millions of complex geographic polygons has become a common demand in GIS [27]. Therefore, the accurate measurement of polygon-shaped features is of utmost importance for quantitatively evaluating the performance of overlay analyses and optimizing the overlay analysis algorithm.

To describe the differences among shapes quantitatively, the concept of shape complexity was proposed [28], [29]. Attneave [30] first proposed to build a shape complexity model based on geometric features such as the number of vertices, symmetry, number of turns and angular variability, but this model only had a good effect on concave polygons. Global distance entropy, local angle entropy and polygon randomness were also used to measure the shape complexity of two-dimensional graphs [31]. Page [32] used the entropy of discrete curvature to calculate the polygon complexity, but their calculation results depended on the division of discrete sampling interval. Some scholars quantified the size of polygons by polygon features or the number of vertices, so as to achieve uniform division of data of different computing nodes [26], [33], [34]. However, the features used to consider
quantization of polygons are not comprehensive enough. In addition, some scholars used graphic boundaries, global structures, absolute curvature, symmetry, the area perimeter ratio of the graphics and the number of blocks to measure shape complexity [35]–[37], but do not consider the irregular and complex characteristics of geographical polygons. Based on the above literature, 27 shape features were selected to measure the complexity from four perspectives: spatial relationship, polygon attribute, spatial morphology and polygon auxiliary features.

In spatial analysis, we need to establish the relationship between the shape complexity of a polygon and the computational performance of the algorithm and express the shape complexity as a function of the problem size and the computational cost [28]. Initially, univariate models [38] were used to predict polygon complexity, but univariate models are not enough to accurately express polygon complexity. A weighted multivariate model [35], [39] is then introduced to assess complexity. The validity of the model is difficult to evaluate because the weighting coefficient is manually specified. Multiple linear regression method can express the relationship between shape complexity and algorithm performance more accurately and effectively. Therefore, this paper chooses this method to evaluate the complexity of polygons facing overlay analysis algorithms.

Some scholars have also made quantitative analyses of the influence of polygon-shaped features on the efficiency of the overlay analysis algorithm [30], but only a single Boolean operator of the Greiner-Hormann algorithm was
TABLE 2. Characteristics of three overlay analysis algorithms.

|                          | Vatti algorithm       | Greiner-Hormann algorithm | Martinez algorithm |
|--------------------------|-----------------------|---------------------------|--------------------|
| Data structure           | doubly linked list    | doubly linked list        | priority queue     |
| Algorithm thought        | scan line idea        | find parts of one polygon in another polygon | scan line idea |
| Application scene        | lower than Martinez algorithm when dealing with large scale data | small-scale data | large-scale data |
| Algorithm steps          | middle                | less                       | more               |

FIGURE 7. The flow chart of Vatti algorithm.

FIGURE 8. Types of intersections that lead to a subdivision (by Martinez, 2009).

models for the Vatti algorithm, Greiner-Hormann algorithm and Martinez algorithm by the multiple linear regression method to quantitatively evaluate the influence of polygon shape variables on the efficiency of these polygon clipping algorithms.

The rest of this paper is organized as follows. Section 2 introduces the data used in this paper, the overlay analysis algorithms, the selected polygon shape variables that may affect the efficiency of the overlay analysis algorithm, and the multiple linear regression method used to construct the model. In Section 3, shape complexity evaluation models are built for the algorithms, and then the effects of polygon shape variables on the efficiencies of different overlay analysis algorithms are analyzed quantitatively. Section 4 provides the conclusion and discussion drawn from this research.

II. DATA AND METHODOLOGY

A. DATA SOURCE

Planar geographic data generally include administrative divisions, traffic, water systems, buildings, vegetation, land use patches and other features, among which the administrative divisions, land use types and water systems contain large amounts of polygon data with varied complexities, such as simple polygons, concave polygons, polygons with holes and even self-intersecting polygons. Therefore, the above three data types used in this paper have the general characteristics and universality of all surface geographic data in an overlay analysis. The basic data information for this paper are shown in Table 1 and Figure 1 below.

adopted, and there is a lack of comprehensive comparisons of different algorithms. Therefore, in this paper, several shape variables were selected to build shape complexity evaluation
TABLE 3. Formulas and short descriptions of the spatial relationship variables.

| Shape variables          | Equation                | Description                                                                 |
|--------------------------|-------------------------|----------------------------------------------------------------------------|
| overlapping area (OA)    | $OA = A_{inter}$        | It measures the area of the overlap of two intersecting vector polygons.     |
| overlapping boundary (OB)| $OB = P_{inter}$        | It measures the perimeter of the overlap of two intersecting vector polygons.|
| overlapping points (OP)  | $OP = V_{inter}$        | It measures the number of vertices in the overlapping part of two intersecting vector polygons. |

TABLE 4. Formulas and short descriptions of the polygon attribute variables ($Apol$: area of the polygon).

| Shape variables          | Equation                | Description                                                                 |
|--------------------------|-------------------------|----------------------------------------------------------------------------|
| number of vertices (NOV)| $NOV = V_{pol}$         | It measures the number of vertices in the polygon.                          |
| number of concaves (NOC)| $NOC = V_{concave}$     | It measures the number of concave points in polygon vertices.               |
| ratio of concave points  | $RNOC = V_{concave} / (V_{pol} - 3)$ | It measures the proportion of concave points in the vertices of the polygon. |
| number of holes (NOH)    | $NOH = H_{hole}$        | It measures the number of inner rings in the polygon.                       |
| average nearest neighbor | $ANN = \frac{\sum d_i}{n} \cdot \frac{0.5}{\sqrt{n} / A_{pol}}$ | It measures the spatial clustering of polygon vertices. The polygon is highly complex when the ANN ratio of the vertices is small. $d_i$ is the nearest distance between each point to other points; $n$ is the number of vertices of the polygon. |
| frequency of vibration   | $Freq = 16(RNOC - 0.5)^8 - 8(RNOC - 0.5)^2 + 1$ | It represents the frequency of concave-convex variation of the polygon vertices. |

TABLE 5. Formulas and short descriptions of the spatial morphology variables ($Ach$: area of the convex hull of the polygon; $Apol$: area of the polygon; $Ambr$: area of the minimum bounding rectangle of the polygon; $Aepc$: area of the equal-perimeter circle of the polygon; $Aeac$: area of the equal-area circle of the polygon; $Ppol$: perimeter of the polygon; $Pch$: perimeter of the convex hull of the polygon; $Peer$: perimeter of the equal-area rectangle of the polygon; $Pmcr$: perimeter of the equal-area square of the polygon; $reac$: radius of the equal-area circle of the polygon; $rlmc$: radius of the minimum circumscribed circle of the polygon; $rlnc$: radius of the largest inscribed circle of the polygon.).

| Shape variables          | Equation                        | Description                                                                 |
|--------------------------|---------------------------------|----------------------------------------------------------------------------|
| concavity (Conv)         | $Conv = (A_{pol} - A_{min}) / A_{pol}$ | It measures the area difference between the polygon and its convex hull. |
| amplitude of the vibration (Ampl) | $Ampl = (P_{pol} - P_{ave}) / P_{pol}$ | It measures the perimeter difference between the polygon and its convex hull. |
| rectangularity (Rect)    | $Rect = A_{pol} / A_{ave}$      | It measures the area difference between the polygon and its minimum bounding rectangle. |
| circularity (Cir)        | $Cir = A_{pol} / A_{eqc} = 4\pi A_{pol} / P_{pol}^2$ | It measures the area difference between the polygon and its equal perimeter circle. |
| equivalent rectangular (ER)| $ER = P_{ave} / P_{pol} = \sqrt{A_{ave} / A_{pol}} \times P_{ave} / P_{pol}$ | It measures the perimeter difference between the polygon and its equal area rectangle. |
| fractality (Frac)        | $FR = 1 - \frac{\log(A_{pol})}{2\log(P_{pol})}$ | It measures the roughness and smoothness of the polygon boundary. |
| squareness (Squa)        | $Squa = P_{ave} / P_{pol} = 4\sqrt{A_{pol} / P_{pol}}$ | It measures the perimeter difference between the polygon and its square of equal area. |
| detour index (DI)        | $DI = P_{ave} / P_{ave} = 2\sqrt{A_{ave} \pi} / P_{ave}$ | It measures the perimeter difference between the convex hull of the polygon and that of its equal area circle. |
| exchange index (EI)      | $EI = A_{pol} / P_{ave}$       | It measures how much of the area inside a circle is exchanged with the area outside it to create the polygon. |
| perimeter index (PI)     | $PI = P_{ave} / P_{ave} = 2\sqrt{\pi A_{pol} / P_{pol}^2}$ | It focuses on the compactness of a polygon’s boundary. |
| girth index (GI)         | $GI = r_{ave} / r_{ave} = \sqrt{A_{pol} / \pi}$ | It is a measure of the thickness of the layer insulating its innermost core from its periphery. |
| range index (RI)         | $RI = r_{ave} / r_{ave} = \sqrt{A_{pol} / \pi}$ | It measures the difference between the radius of the circumscribed circle of the polygon and the radius of its equal area circle. |

B. POLYGON OVERLAY ANALYSIS ALGORITHMS

1) VATTI ALGORITHM

In 1992, Bala R. Vatti proposed a polygon-clipping algorithm based on sweep line algorithm that supports clipping operations between arbitrary polygons.

In Vatti algorithm, a polygon boundary is divided into a left boundary and a right boundary. The left and right boundaries are usually composed of groups of boundary lines, and each boundary line is composed of several edges. As shown in Figure 2, B2 and B3 are the right boundary
TABLE 6. Formulas and short descriptions of the polygon auxiliary variables (The polygon is meshed, and the mesh points inside the polygon are called inner grid points (IGPs); \( r_{igp} \): distance from the centroid to the IGP; \( \text{disp} \): distance between the inner grid points; \( \mu_{igp} \): distance from an IGP to the boundary; The polygon boundary is divided equally and new nodes are inserted at the bisection points, which are called interpolated boundary points (IBPs); \( r_{ibp} \): distance from the centroid to an IBP; \( \text{rdev} \): average distance between the centroid of the polygon and IBPs; \( \text{disp} \): inner shortest path length between IBPs; \( \text{rdev} \): distance between an IBP and the boundary of the ADC along a radial line.).

| Shape variables            | Equation                                                                 | Description                                                                 |
|----------------------------|--------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| normalized cohesion index  | \( n\text{Cohe} = \frac{0.9054 \times \sqrt{A_{pol}}}{p} \)             | It is a measure of overall accessibility from all points to others within a polygon. |
| normalized proximity index  | \( n\text{Prox} = \frac{2}{3} \times \frac{\mu_{igp}}{\mu_{ap}} \)         | It is a measure of overall accessibility from all inner points to the center of a polygon. |
| normalized spin (nSpin)     | \( n\text{Spin} = \frac{0.5 \times A_{pol}}{\mu_{ap} \pi} \)             | It measures the compactness of the polygon.                                  |
| normalized dispersion (nDisp) | \( n\text{Disp} = 1 - \frac{\mu_{ap}}{r_{adc}} = 1 - \frac{\mu_{ap}}{\mu_{bp}} \) | It measures the average distance between the centroid of the polygon and its boundary insertion point. |
| normalized depth (nDept)    | \( n\text{Dept} = \frac{\mu_{pol_{igp,s}}}{\mu_{ap_{igp,s}}} = \frac{3 \mu_{pol_{igp,s}}}{A_{pol} \pi} \) | It represents the irregular change degree of the boundary of the polygon.         |
| normalized traversal (nTrav) | \( n\text{Trav} = \frac{(\mu_{disp})_{rad}}{(\mu_{disp})_{pol}} = \frac{4 \sqrt{A_{pol}} / \pi}{(\mu_{ap})_{pol}} \) | It represents the average distance of the shortest internal path between the insertion points of the polygon boundary. |

and B1 and B4 are the left boundary; these boundaries are uniformly stored in a local minimum list (LML). The storage structure of the LML is shown in Figure 3. As shown in Figure 2 (a), each vertex of the polygon intersects a scan line, and the area between two adjacent scan lines is called a scan beam. By traversing all scan beams from bottom to top, the intersecting vertices of the polygon are found.

The execution process of Vatti algorithm is shown in Figure 4. Firstly, the polygon data is read and the vertex of the polygon is converted into LML, a unique data structure of Vatti algorithm. Then according to the idea of horizontal scanning, each scanning beam is processed in turn and the intersection point is calculated. Finally, the corresponding resulting polygon is constructed according to the different Boolean operations.

Murta [40] improved the Vatti algorithm to solve the problem that horizontal edges are not allowed in the input polygons, and opened the source of the algorithm, which is called the general polygon clipping library.

2) GREINER-HORMANN ALGORITHM

Greiner and Hormann proposed a more simple polygon-clipping algorithm based on previous studies in 1998 that also supports fast clipping between arbitrary polygons.

3) MARTINEZ ALGORITHM

Martinez proposed a simple and efficient polygon-clipping algorithm in 2009 that can be considered an extension of
TABLE 7. Correlations between the shape variables.

| OA   | OB  | OP | NOV | SOC | RNOC | CNV | NOH | Ampl | ANN | Freq | Rect | Cir | ER | Frac | Squa | DI | El | PI | GI | RI | nCoke | nProx | nSpin | nDisp | nDept | nTrav |
|------|-----|----|-----|-----|------|-----|-----|------|-----|-----|-----|-----|----|-----|------|-----|----|----|----|----|-------|-------|-------|-------|-------|-------|-------|
| 1    | 0.904 | 1  | 0.678 | 0.795 | 1  | 0.018 | 0.032 | 0.069 | 0.465 | 0.432 | 0.148 | 0.152 | 0.579 | 1  | 0.017 | 0.028 | 0.051 | 0.996 | 1  | 0.207 | 0.181 | 0.142 | 0.265 | 0.276 | 1  |

TABLE 8. Model summary.

|          | R   | R-square | Adjusted R-square |
|----------|-----|----------|-------------------|
| Vatti Intersection | 0.942 | 0.887 | 0.887 |
| Vatti Union | 0.944 | 0.891 | 0.891 |
| Vatti Difference | 0.945 | 0.892 | 0.892 |
| Greiner-Hormann Intersection | 0.994 | 0.989 | 0.989 |
| Greiner-Hormann Union | 0.994 | 0.989 | 0.989 |
| Greiner-Hormann Difference | 0.995 | 0.990 | 0.990 |
| Martinez Intersection | 0.947 | 0.897 | 0.897 |
| Martinez Union | 0.933 | 0.871 | 0.871 |
| Martinez Difference | 0.944 | 0.891 | 0.891 |
| Martinez Symmetric difference | 0.938 | 0.880 | 0.880 |

The classic plane scanning algorithm. He believed that the boundaries of the resulting polygon were made up of the parts of the boundaries of each polygon participating in a Boolean operation, so finding these parts and connecting them together made up the resulting polygon.

The execution process of Martinez algorithm is shown in Figure 7. The polygon vertex data is read to construct the corresponding data structure for each edge, and then all edges are inserted into the priority queue. Then, as shown in Figure 9, the idea of vertical scanning is used to scan each scanbeam in turn. The edges of the polygon are subdivided at the intersection of line segments, and the subdivided edges are inserted into the original priority queue, as shown in Figure 8. Next, the resulting edges are selected based on whether the current edge is inside another polygon. Finally, the edges selected in the previous step are connected to form the resulting polygon. Assuming that n is the number of vertices participating in Boolean operation and k is the number of intersections, the time complexity of the algorithm can reach $O((n + k) \log(n))$.

Martinez improved the algorithm in 2013 to make it easier to understand and implement and added a function for calculating the sub-contour of the polygon contour.

4) COMPARISON OF OVERLAY ANALYSIS ALGORITHMS

As shown in Table 2 below, although Vatti algorithm, Greiner-Hormann algorithm and Martinez algorithm are all suitable for clipping between arbitrary polygons, their applicable scenes and characteristics are different. The Martinez
algorithm is more efficient when dealing with polygon overlay analyses with millions of vertices or more, but its algorithm steps are more complex. When processing small-scale data, the Greiner-Hormann algorithm is suitable.

C. EXPLANATORY VARIABLES

This paper analyzes polygon shape variables that may affect the efficiency of an overlay analysis with extensive literature search and expert knowledge. Based on these studies, we selected 27 polygon shape variables related to the efficiency of polygon overlay analysis and divided these variables into four categories, including spatial relationship variables, polygon attribute variables, spatial morphology variables, and polygon auxiliary variables. The spatial relationship variables include the overlapping area (OA), overlapping boundary (OB) and overlapping points (OP). The polygon attribute variables include the number of vertices (NOV), number of concaves (NOC), ratio of concave points in vertices (RNOC), number of holes (NOH), average nearest neighbor (ANN), and frequency of vibration (Freq). The spatial morphology variables include the concavity (Conv), amplitude of vibration (Ampl), rectangularity (Rect), circularity (Cir), equivalent rectangularity (ER), fractality (Frac), squareness (Squa), detour index (DI), exchange index (EI), perimeter index (PI), girth index (GI), and range index (RI). The polygon auxiliary variables include the normalized cohesion index (nCohe), normalized proximity index (nProx), normalized spin (nSpin), normalized dispersion (nDisp), normalized depth (nDept), and normalized traversal (nTrav).

We used the c plus programming language to develop a program to calculate these variables as shown in Table 3,
FIGURE 10. Scatter graphs and linear fitting of polygon shape complexity and calculation time of 100,000 sets of training data based on the intersection, union, difference, and symmetric difference boolean operations of Greiner-Hormann algorithm, Martinez algorithm and Vatti algorithm.

Table 4 and Table 5. In addition, the variables in Table 6 were calculated with the Shape Metrics Tool. In Table 6, for boundary-based features, a set of equidistant interpolation points was used to represent the boundaries of the polygon. For this paper, we chose to insert 500 points evenly into the boundary. For grid-based features, the Shape Metrics Tool was used to divide the polygons into 20,000 grids equally by area to obtain the desired internal grid points.

D. REGRESSION MODEL

In early studies of shape complexity, the complexity was usually measured with the univariate method [32], [38]. Since the univariate method cannot accurately measure shape complexity, scholars gradually introduced a weighted linear combination of multiple variables to predict its complexity [31], [35], [39]. However, the weighting coefficients of each variable in the model were manually specified,
which made it difficult to evaluate the effectiveness of the model.

Regression analysis, as a more objective and effective method of developing models, is widely used in the fields of econometrics, econometrics and geography, statistics and so on. It mainly evaluates a model with the following parameters: 1) The determination coefficient $R^2$ (the closer $R^2$ is to 1, the better the model fitting effect is.) 2) F test (if $\text{Sig} > 0.05$, the model is not significant; If $0.01 < \text{Sig} < 0.05$, the model is significant; If $\text{Sig}$ value $< 0.01$, the model is extremely significant.) 3) T test, (the $\text{Sig}$ value can be used to judge the significance of characteristic variables, and then to delete redundant variables).

This paper used multiple linear stepwise regression analysis to construct the model. Compared with multiple linear regression analysis, this method has a more reasonable independent variable screening mechanism and can avoid the influence of non-statistically significant independent variables on the regression equation. In the process of analysis, every time a new variable is introduced, the substituted independent variable should be calculated again to check whether it has the value of remaining in the equation. Based on this process, the introduction and elimination of independent variables are carried out alternately until no new variable can be introduced or eliminated.

The regression method requires that there be no multiple collinearities between variables. Therefore, before regression analysis, we carry out correlation analyses between variables and eliminate variables with high correlation in advance to eliminate multiple collinearities between variables. The correlations between variables are measured with a correlation coefficient. The coefficient measures the correlation between two variables. Its value range is $-1$ to $1$. The number $1$ means that the two variables are completely correlated, $-1$ means that the two variables are completely negatively correlated, and $0$ means that they are completely uncorrelated.

### III. RESULT

In this paper, 100,000 polygons were randomly selected from the data in Table 1 above as the subject polygons, and a typical complex polygon was selected as the clipping polygon. Each subject polygon was subjected to overlay analysis.
with the clipping polygon, including the intersection, union, difference and symmetric difference operations of the Vatti, Greiner-Hormann, and Martinez algorithms.

Due to the subject polygon and the clipping polygon being in different positions, they were moved to the same position before clipping so the centroids of the two polygons coincided; then, the overlay analysis was carried out. According to the computational formulas of the 27 shape variables in Table 3 to Table 6, we calculated the variable values of each group of polygons. The SPSS tool was used to conduct the correlation analysis between these variables, and Table 7 was obtained. According to the data in the table, we excluded variables with high correlations and retained the following variables: overlapping area, overlapping points, number of vertices, number of concaves, number of holes, amplitude of the vibration, average nearest neighbor, frequency of vibration, and rectangularity.

Through the above correlation analysis, we excluded the variables that had high correlations with each other and eliminated multicollinearity among the variables. Then, we took the remaining 9 variables as independent variables and the calculation time of each Boolean operator as dependent variables to make multiple linear stepwise regression analyses one by one. The shape complexity evaluation models of intersection, union, difference and symmetric difference Boolean operators of the Vatti, Greiner-Hormann and Martinez algorithms are summarized in Table 8 below. We can see that the adjusted values of R^2 belonging to different Boolean operators of the same algorithm are close to each other. Therefore, when comparing goodness of fit, this paper discusses the algorithm as the basic unit rather than the operator. Table 9, Table 10 and Table 11 respectively show the ANOVA information of Vatti, Greiner-Hormann and Martinez algorithms. The values of sig. in Table 9, Table 10 and Table 11 are less than 0.05, thereby indicating that the models were statistically significant and well fitted. The R^2 of the Greiner-Hormann algorithm is significantly higher than that of the other two algorithms, and the evaluation model can explain the 99% calculation time performance of the Greiner-Hormann algorithm with high consistency. The adjusted R^2 of the Martinez
algorithm is close to that of Vatti algorithm and lower than that of Greiner-Hormann algorithm, between 0.87 and 0.89, which can explain 87% to 89% of the performance of the algorithms.

With these evaluation models constructed by the multiple linear regression method, we were able to calculate the shape complexity of each group of polygons participating in clipping. Therefore, based on each Boolean operator, the shape complexity of 100,000 polygons and their corresponding calculation time were plotted as scatter graphs with the polygon shape complexity as the horizontal axis and the calculation time as the vertical axis, and a linear regression operation was carried out. The results are shown in Figure 10.

There are very few outlier points in the scatter graph of the Greiner-Hormann algorithm, and more than 99% of them are concentrated near the regression line. With increasing complexity, these points gradually become scattered and the distance from the regression line becomes larger and larger. This indicates that with an increase of shape complexity, the fitting accuracy gradually decreases. In the scatter graph of the Martinez algorithm, most of the points are distributed near the regression line, and the distribution of points is relative uniform as the shape complexity increases. However, due to the existence of many abnormal points far from the regression line, the correlation coefficient of this algorithm is reduced. Compared with the other two algorithms, the distribution of the points in the scatter graph of the Vatti algorithm is more scattered. With increasing shape complexity, the points become more and more scattered and are located increasingly far from the regression line.

The above models were built with a stepwise multiple linear regression method. The probability of variable F entering the model was set to 0.05. If the probability of F was greater than 0.1, the variable was removed. The analysis results are shown in Figure 11, 12 and 13 below. In the process of step by-step regression, if the significance test result of the independent variable is greater than 0.05, it indicates that the independent variable has no statistical significance in the

![Figure 13](https://example.com/figure13.png)

**FIGURE 13.** Standardized coefficients of shape variables explaining Martinez algorithm efficiency using multiple linear stepwise regression.
model and should be deleted. The coefficient of the independent variable in Figure 10 is the standardized coefficient. A positive coefficient means that the influence of the variable on the dependent variable is positively correlated, while a negative coefficient means that the influence of the independent variable on the dependent variable is negatively correlated. In addition, the larger the coefficient, the greater the influence on the dependent variable.

In these evaluation models based on the Vatti, Greiner-Hormann and Martinez algorithms, the standardized coefficient of the number of polygon vertices is the highest and has the highest influence on the calculation time. The core step of the clipping algorithm is the intersection of the line segment, so the number of polygon vertices involved in clipping is the most important factor affecting the clipping algorithm. Each algorithm needs to construct its own data structure before clipping. The Greiner-Hormann algorithm adopts a doubly linked list structure, and the time complexity of its data structure construction process is $O(1)$. The GPC library uses a doubly linked list structure to store vertex data. However, a binary search tree is adopted in the construction of a scan line set, and its time complexity ranges from $O(\log n)$ to $O(n)$. As the number of vertices increases, the construction time also increases, resulting in a decrease in the fitting correlation coefficient for the overall calculation time and the shape complexity of the algorithm.

The standardized coefficients of the other explanatory variables are much lower than the number of vertices, which has a low impact on the algorithm performance. The standardized coefficients of different Boolean operators belonging to the same algorithm are relatively similar, and the standardized coefficients of different algorithms are different, and even have completely opposite effects on performance.

The average nearest neighbor variable has a positive effect on the performance of the Vatti and Martinez algorithms and has a higher effect on the performance of the Vatti algorithm but has a negative effect on the Greiner-Hormann algorithm, which may be because the Vatti and Martinez algorithms are based on the idea of scan line construction. The vertex density affects the performances of the algorithms to some extent. The amplitude of the vibration variable has a negative
FIGURE 15. Scatter diagram and linear fitting between polygon shape complexity and calculation time of 30,000 sets of test data based on the intersection, union, difference, and symmetric difference boolean operations of Greiner-Hormann algorithm.

In summary, based on the classification of polygon shape variables in this paper, polygon attribute variables number of vertices, number of holes, average nearest neighbor, and frequency of vibration have the greatest impact on the performance of an overlay analysis algorithm. The overall effect of spatial relation variables on the efficiency of overlay analysis algorithms is relatively weak. However, the standardized coefficient is 0.12 in the symmetric difference operation of overlapping points in the Martinez algorithm, indicating that this operation has a certain influence on the algorithm performance. For spatial morphological variables, amplitude of vibration and rectangularity have a weak influence on the algorithm efficiency. The polygon auxiliary variables have no effect on the performance of the algorithms.

Finally, to verify the validity of these models, we selected another 30,000 groups of new polygons from the dataset above as a test dataset. The shape complexity models obtained previously were used to calculate the shape complexity of each group of polygons, and the shape complexity was fitted with its corresponding calculation time, as shown in Figures 14, 15, and 16.

By comparing the regression model fitting correlation coefficients of the test dataset and the training dataset, we knew that the test data of the Vatti and Greiner-Hormann algorithms were basically consistent with the coefficient of the training data, and the fitting coefficient of the test data from the Martinez algorithm was higher than the training data. This is because there are fewer outliers in the training data.

After the above verification, the shape complexity evaluation models constructed for the overlay analysis algorithms in this paper have certain universality.
IV. CONCLUSION AND DISCUSSION

A. CONCLUSION
In the era of the rapid growth of the scale of geospatial data, in addition to the overlay analysis algorithm itself, the shape features of polygons involved in analysis are also important factors affecting overlay analysis efficiency. To quantitatively evaluate the influence of polygon shape features on the performance of overlay analysis algorithms, twenty-seven polygon shape variables were selected from the literature, and these variables were divided into four categories according to their characteristics: spatial relational variables, polygon attribute variables, polygon auxiliary variables, and spatial morphological features. By using the multiple linear regression method, shape complexity evaluation models for various overlay analysis algorithms were constructed, and quantitative measurements of the influence of shape complexity on the efficiency of overlay analysis algorithms were realized. Regression analysis can not only effectively reveal the internal relationships between complexity calculation models and computational efficiency but also reveal the relationship between explanatory variables in a model. By comparing the standardized coefficients of each shape variable in different evaluation models, we realized a quantitative analysis of the influence of shape variables on the efficiency of polygon overlay analysis algorithms, which expanded people’s cognizance of overlay analysis algorithms and were beneficial to deepening our understanding and optimization of the overlay analysis algorithms. In addition, since the efficiency of GIS spatial analysis algorithms is generally affected by the shape characteristics of spatial entities involved in the analysis, the method adopted in this paper has universality.

B. DISCUSSION
In this paper, the shape complexity evaluation models for the Vatti algorithm, Greiner-Hormann algorithm and Martinez algorithm were established by using the multiple linear step-wise regression method. The models can effectively measure the computational performance of these clipping algorithms. However, this method cannot completely explain the performance of an algorithm, especially the Vatti and Martinez algorithms. In addition to the shape characteristics of the polygon, the data structure adopted by the algorithm itself and the algorithm concept also affect efficiency of the algorithm.
These factors can be analyzed to further optimize the model. In addition, this paper only builds an evaluation model for the clipping between two polygons. Next, we can discuss the clipping between two polygon layers to build a shape complexity evaluation model suitable for coarse-grained overlay analysis.

REFERENCES

[1] R. McDonnell and K. Kemp, International GIS Dictionary. Hoboken, NJ, USA: Wiley, 1995.

[2] P. A. Burrough, R. McDonnell, R. A. McDonnell, and C. D. Lloyd, Principles of Geographical Information Systems. London, U.K.: Oxford Univ. Press, 2015.

[3] W. T. Reeves, “Particle systems—a technique for modeling a class of fuzzy objects,” ACM Trans. Graph., vol. 2, no. 2, pp. 91–108, 1983.

[4] M. N. DeMers, Fundamentals of Geographic Information Systems. Hoboken, NJ, USA: Wiley, 2008.

[5] S. Wang, E. Zhong, H. Lu, H. Guo, and L. Long, “An effective algorithm for lines and polygons overlay analysis using uniform spatial grid indexing,” in Proc. 2nd IEEE Int. Conf. Spatial Data Mining Geograph. Knowl. Services (ICSDM), Jul. 2015, pp. 175–179.

[6] J. Fan, T. Ma, M. Ji, Y. Zhou, and T. Xu, “Implementation and optimization of eight parallel polygon overlapping tools with OpenMP at the feature layer level in GIS,” Prog. Geogr., vol. 32, no. 12, pp. 1835–1844, 2013.

[7] F. Martínez, A. J. Rueda, and F. R. Feito, “A new algorithm for computing Boolean operations on polygons,” Comput. Geosci., vol. 35, no. 6, pp. 1177–1185, Jun. 2009.

[8] J. Fan, “Parallel algorithms for polygon overlapping,” Acta Geodaetica et Cartographica Sinica, vol. 45, no. 4, p. 502, 2016.

[9] F. Martínez, C. Ogayar, J. R. Jiménez, and A. J. Rueda, “A simple algorithm for Boolean operations on polygons,” Adv. Eng. Softw., vol. 64, pp. 11–19, Oct. 2013.

[10] S. Puri and S. K. Prasad, “Efficient parallel and distributed algorithms for GIS polygonal overlay processing,” in Proc. IEEE Int. Symp. Parallel Distrib. Process., Workshops PhD. Forum, 2013, pp. 2238–2241.

[11] I. E. Sutherland and G. W. Hodgman, “Reentrant polygon clipping,” Commun. ACM, vol. 17, no. 1, pp. 32–42, 1974.

[12] Y.-D. Liang and B. A. Barsky, “An analysis and algorithm for polygon clipping,” Commun. ACM, vol. 26, no. 11, pp. 868–877, Nov. 1983.

[13] T. M. Nicholl, D. T. Lee, and R. A. Nicholl, “An efficient new algorithm for 2-D line clipping: Its development and analysis,” in Proc. 14th Annua. Conf. Comput. Graph. Interact. Techn., 1987, pp. 253–262.

[14] P.-G. Maillot, “A new, fast method for 2D polygon clipping: Analysis and software implementation,” ACM Trans. Graph., vol. 11, no. 3, pp. 276–290, Jul. 1992.

[15] J. D. Foley, Computer Graphics: Principles and Practice. Reading, MA, USA: Addison-Wesley, 1996.

[16] H. Hang and L. Sun, “An algorithm for circle clipping against polygon window,” Comput. Technol. Develop., vol. 19, no. 5, pp. 235–237, 2009.

[17] M. Cao, H. Zhang, C. Zhou, Y. Sun, and H. Yu, “Vector circle clipping algorithm based on polygon window of hexagonal grid system,” J. Phys., Conf. Ser., vol. 1288, no. 1, Aug. 2019, Art. no. 012006.

[18] K. Weiler and P. Atherton, “Hidden surface removal using polygon area sorting,” ACM SIGGRAPH Comput. Graph., vol. 11, no. 2, pp. 214–222, 1977.

[19] L. Yong-Kui, G. Yun, H. You-Qun, and L. Yk, “An efficient algorithm for polygon clipping,” J. Softw., vol. 14, no. 4, pp. 845–856, 2003.

[20] F. Junfu, K. Weihua, M. Ting, Z. Chenghu, J. Min, and Z. Yuke, “RaPC: A rasterization-based polygon clipping algorithm and its error analysis,” Acta Geodaetica et Cartograph. Sinica, vol. 44, no. 3, p. 338, 2015.

[21] B. R. Vatti, “A generic solution to polygon clipping,” Commun. ACM, vol. 35, no. 7, pp. 56–63, 1992.

[22] G. Greiner and K. Hormann, “Efficient clipping of arbitrary polygons,” ACM Trans. Graph., vol. 17, no. 2, pp. 71–83, Apr. 1998.

[23] Y. K. Liu, X. Q. Wang, S. Z. Bao, M. Gombosi, and B. Žalik, “An algorithm for polygon clipping, and for determining polygon intersections and unions,” Comput. Geosci., vol. 33, no. 5, pp. 589–598, May 2007.

[24] S. Li, S. Dragicevic, F. A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, and T. Cheng, “Geospatial big data handling theory and methods: A review and research challenges,” ISPRS J. Photogramm. Remote Sens., vol. 115, pp. 119–133, May 2016.

[25] Z. Li, “Geospatial big data handling with high performance computing: Current approaches and future directions,” 2019, arXiv:1907.12182. [Online]. Available: http://arxiv.org/abs/1907.12182

[26] K. Zhao, B. Jin, H. Fan, W. Song, S. Zhou, and Y. Jiang, “High-performance overlay analysis of massive geographic polygons that considers shape complexity in a cloud environment,” ISPRS Int. J. Geo-Information, vol. 8, no. 7, p. 290, 2019.

[27] K. Zhao, B. Jin, H. Fan, and M. Yang, “A data allocation strategy for geocomputation based on shape complexity in a cloud environment using parallel overlay analysis of polygons as an example,” IEEE Access, vol. 8, pp. 185981–185991, 2020.

[28] R. B. Tilove, “Line/polygon classification: A study of the complexity of geometric computation,” IEEE Comput. Graph. Appl., vol. 1, no. 2, pp. 75–88, Apr. 1981.

[29] H. Zhang, J. Dong, S. He, and S. Lv, “Research on image complexity description method based on approximate entropy,” in Proc. Int. Conf. Intell. Transp., Big Data Smart City (ICITBS), Jan. 2020, pp. 918–920.

[30] F. Attneave, “Physical determinants of the judged complexity of shapes,” J. Exp. Psychol., vol. 53, no. 4, p. 221, 1957.

[31] Y. Chen and H. Sundaram, “Estimating complexity of 2D shapes,” in Proc. IEEE 7th Workshop Multimedia Signal Process., Oct. 2005, pp. 1–4.

[32] D. L. Page, A. F. Koschan, S. R. Sukumaran, B. Roui-Abidi, and M. A. Abidi, “Shape analysis algorithm based on information theory,” in Proc. Int. Conf. Image Process., vol. 1, 2003, p. 229.

[33] B. Chazelle and J. Incerpi, “Triangulation and shape-complexity,” ACM Trans. Graph., vol. 3, no. 2, pp. 135–152, Apr. 1984.

[34] M. Guo, Q. Guan, Z. Xic, L. Wu, X. Luo, and Y. Huang, “A spatially adaptive decomposition approach for parallel vector data visualization of polylines and polygons,” Int. J. Geograph. Inf. Sci., vol. 29, no. 8, pp. 1419–1440, Aug. 2015.

[35] H. Su, A. Bourdiane, and D. Crookes, “Scale adaptive complexity measure of 2D shapes,” in Proc. 18th Int. Conf. Pattern Recognit. (ICPR), vol. 2, 2006, pp. 134–137.

[36] X. Duan, X. Deng, and Q. Zuo, “The research on the complexity of progressive die edge,” J. Eng. Graph., no. 5, pp. 94–97, 2006.

[37] W. Saleem, A. Belayev, D. Wang, and H.-P. Seidel, “On visual complexity of 3D shapes,” Comput. Graph., vol. 35, no. 3, pp. 580–585, Jun. 2011.

[38] S. Piarra and T. Grajewski, “Describing shape and shape complexity using local properties,” in Proc. 3rd Int. Space Syntax Symp. Princeton, NJ, USA: Citeseer, 1995, p. 109.

[39] T. Brinkhoff, H.-P. Kriegel, R. Schneider, and A. Braun, “Measuring the complexity of polygonal objects,” in Proc. ACM-GIS. Princeton, NJ, USA: Citeseer, 1997, p. 9.

[40] A. Mutta. (1997). A General Polygon Clipping Library. [Online]. Available: http://www.cs.man.ac.uk/toby/alan/software/gpc.html