

Article

# A Multi-Scale Residential Areas Matching Method Considering Spatial Neighborhood Features

Jingzhen Ma <sup>1,2,3</sup> , Qun Sun <sup>1</sup>, Zhao Zhou <sup>1</sup>, Bowei Wen <sup>1</sup> and Shaomei Li <sup>1,\*</sup>

<sup>1</sup> Institute of Geospatial Information, Information Engineering University, Zhengzhou 450000, China; zb50mjz@163.com (J.M.); sunqun@371.net (Q.S.); zhouzhaochxy@163.com (Z.Z.); gis0803@163.com (B.W.)

<sup>2</sup> Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Zhengzhou 450000, China

<sup>3</sup> Key Laboratory of Spatiotemporal Perception and Intelligent Processing, Ministry of Natural Resources, Zhengzhou 450000, China

\* Correspondence: lsalsm@sina.com; Tel.: +86-0371-8163-6102

**Abstract:** Residential areas is one of the basic geographical elements on the map and an important content of the map representation. Multi-scale residential areas matching refers to the process of identifying and associating entities with the same name in different data sources, which can be widely used in map compilation, data fusion, change detection and update. A matching method considering spatial neighborhood features is proposed to solve the complex matching problem of multi-scale residential areas. The method uses Delaunay triangulation to divide complex matching entities in different scales into closed domains through spatial neighborhood clusters, which can obtain many-to-many matching candidate feature sets. At the same time, the geometric features and topological features of the residential areas are fully considered, and the Relief-F algorithm is used to determine the weight values of different similarity features. Then the similarity and spatial neighborhood similarity of the polygon residential areas are calculated, after which the final matching results are obtained. The experimental results show that the accuracy rate, recall rate and F value of the matching method are all above 90%, which has a high matching accuracy. It can identify a variety of matching relationships and overcome the influence of certain positional deviations on matching results. The proposed method can not only take account of the spatial neighborhood characteristics of residential areas, but also identify complex matching relationships in multi-scale residential areas accurately with a good matching effect.

**Keywords:** residential areas matching; spatial neighborhood; Delaunay triangulation; similarity; Relief-F algorithm



**Citation:** Ma, J.; Sun, Q.; Zhou, Z.; Wen, B.; Li, S. A Multi-Scale Residential Areas Matching Method Considering Spatial Neighborhood Features. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 331. <https://doi.org/10.3390/ijgi11060331>

Academic Editors: Sisi Zlatanova and Wolfgang Kainz

Received: 6 April 2022

Accepted: 30 May 2022

Published: 31 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Spatial object matching is an important prerequisite to realize multi-source spatial information fusion, spatial object change detection and dynamic update, and its purpose is to establish the corresponding relationship between the same object with the same name in different spatial data sources. Residential area matching is an important foundation of spatial data processing and applications. However, data from various sources differs in many ways, including geometry, attributes, and spatial relationships. This poses significant challenges for spatial data matching, especially at different spatial scales. Under the effects of cartographic generalization, the spatial distribution of residential areas is more diverse, and matching relationships between data are more complex [1–6]. As such, research on matching residential areas, especially multi-scale residential areas, is of great significance.

Many scholars have conducted studies on areal entity matching that have focused on two main aspects: measures of similarity and matching strategies. Studies on measures of similarity mainly use geometric, topological, and semantic features of areal elements as similarity features, among which geometric features are most used. In the calculation of

spatial similarity, the location, area, direction and shape of entities are generally used as indicators to calculate the similarity of elements from different sources, and each indicator is assigned a weight according to the features of different entities, and then the total similarity is obtained. Masuyama et al. [7] calculated the possibility of matching according to the degree of overlap of areas. Hao et al. [8] generalized similarities in the shape, position, and size of areal entities and identified matching entities based on their degree of total similarity. An et al. [9] proposed a multilevel description method for measuring the shape similarity of multi-scale areal entities. Luo et al. [10] combined spatial and semantic features of areal entities to identify the best matching objects. Some studies [11,12] have also calculated the similarities of contextual environmental features to match polygon residential areas, and mainly uses the area and perimeter of the triangle around the target building to calculate the contextual similarity.

Among the studies that have been conducted on matching strategies, Wang et al. [4] proposed a new method to match residential areas using skeleton line mesh of black region, and the matching object is transformed into skeleton line mesh matching. Tong et al. [13] proposed a matching model based on probability theory that matches targets by calculating the probability of candidate targets. Wu et al. [14] constructed a Voronoi diagram of polygon residential areas and used convex hull similarities to match areal entities. Wang et al. [15] used geometric features of sample data to train a neural network model and determine final matching results. Other studies [16,17] have used the probabilistic relaxation method to solve the matching matrix and obtain globally optimal matching results.

Although the above studies have addressed issues with areal entity matching to some extent, most involved matching entities at the same or similar scales, and their methods do not apply to complex matching at multiple scales. This study was conducted to develop a multi-scale residential area matching method that takes into account spatial neighborhood features. This method divides complex matching entities at multiple scales into closed domains using spatial neighborhood clusters. It also considers the influence of the features of residential areas themselves and spatial neighborhood features on the matching results, matching the different matching relationships to obtain the final matching results.

## 2. Complex Matching Candidate Determination Method for Multi-Scale Polygon Residential Areas

### 2.1. Complex Matching Relationship Analysis of Multi-Scale Polygon Residential Areas

Because of factors such as data sources, spatial accuracy, and cartographic generalization, there are significant differences in geometric, attribute, and topological features of residential elements in multi-scale spatial data. According to the entity correspondence of residential areas in different scale data and the number of entities included, the matching mode of multi-scale residential areas can be subdivided into 1: 0, 0: 1, 1: 1, 1: N, M: 1 and M: N, as shown in Table 1.

**Table 1.** Matching Mode of Multi-scale Residential Areas.

Matching Mode	Large Scale Data	Small Scale Data
1: 0		
0: 1		

Table 1. Cont.

Matching Mode	Large Scale Data	Small Scale Data
1: 1		
1: N		
M: 1		
M: N		

(1) 1: 0 mode, the residential areas existing in the large scale data have no corresponding targets in the small scale data. This matching relation may be caused by the selection algorithm. When large scale data is reduced to small scale data, it is necessary to make trade-offs due to the reduction of map size, resulting in the abandonment of corresponding targets in small scale data.

(2) 0: 1 mode, the residential areas existing in the small scale data have no corresponding target in the larger scale data. The matching relationship may be caused by the multi-source of spatial data or the temporal difference of data at different scales. If the large scale data is relatively new, the residents corresponding to the small scale data are deleted, while the small scale data is not updated.

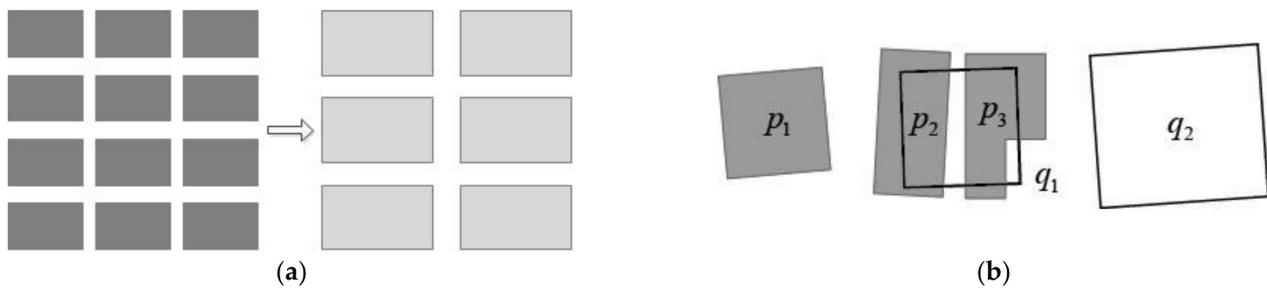
(3) 1: 1 mode, a residential area in the large scale data corresponds to a polygon in the small scale data. This situation involves map simplification, where the geometry (size, shape, etc.) may change despite the existence of one-to-one polygon object at different scales.

(4) 1: N mode, one residential area in the large scale data corresponds to N polygon residential areas in the small scale data. It may be caused by multi-source, multi-temporal or data error of spatial data of different scales.

(5) M: 1 mode, M polygon residential areas in the large scale data correspond to one settlement in the small scale data. This situation is mainly caused by the merge operator. In the scale transformation, multiple residential areas are merged into one, so there will be a many-to-one matching relationship.

(6) M: N mode, multiple polygon residential areas in large scale data correspond to multiple polygon residential areas in small scale data. This situation is mainly caused by typification operators, which can maintain the consistency of spatial features despite the inconsistency of the number of targets at different scales.

Compared with the same scale spatial data matching, multi-scale residential areas have more complex matchings, and the determination of matching target is more complicated. First, most of the data at smaller scales is generalized from data at larger scales; therefore, there are corresponding relationships between the data in terms of cartographic generalization. For example, the merge operator merges multiple objects into one object, resulting in a one-to-many matching relationship. Many-to-many matching relationships are mainly derived from the typification operator in cartographic generalization [18]. Although the quantities before and after generalization are inconsistent, the typical features of elements can be maintained, as shown in Figure 1a. In addition, combinations of different cartographic generalization operators (including selection, simplification, and displacement) are usually used. The combined use of simplification, merge, and displacement operators changes the shapes of elements and causes certain displacements, so it becomes more difficult to identify matching relationships between elements, as shown in Figure 1b.



**Figure 1.** Effects of Cartographic Generalization on Multi-scale Spatial Data. (a) Polygon Residential Areas Many-to-many Corresponding Relationships; (b) Displaced Data.

In present study, a two-way matching strategy was used to solve complex matching relationships. The main idea of two-way matching strategy is to discover the possible matching relationship between two kinds of data by exchanging the identity of the matched object and the object to be matched. Traditional methods with two-way matching strategy mainly uses the buffer zone or Minimum Bounding Rectangle (MBR) candidate matching methods, which exchange matching data and reference data and find matching relationships in multi-scale data. Traditional two-way matching is effective for one-to-one and one-to-many matching, but it often leads to false and missing matches in many-to-many matching generated by a typification operator. With incremental convex hull, group object detection, target clustering, and other methods, the efficacy of many-to-many matching is also unremarkable [19,20].

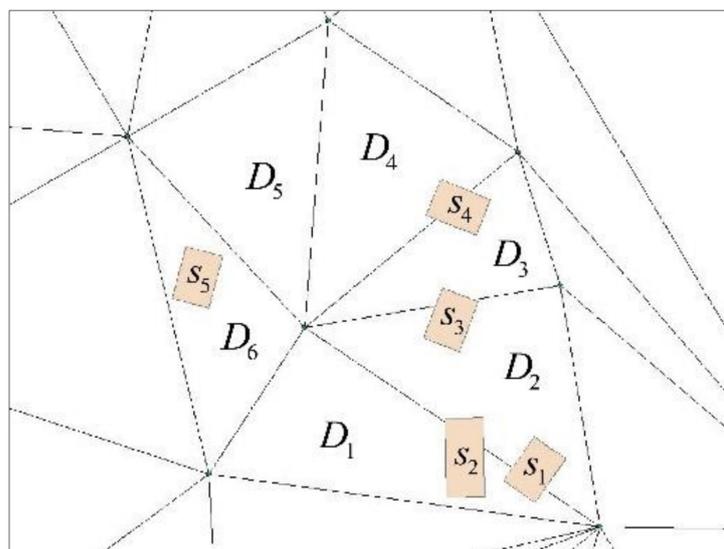
## 2.2. Determining Many-to-Many Matching Candidates Based on Spatial Neighborhood Clusters

For complex matching of multi-scale polygon residential areas, especially many-to-many matching relationships, it is necessary to obtain corresponding candidate elements in the data at different scales, that is, to determine candidate elements with many-to-many matching relationships. To this end, this study introduces the concept of spatial neighborhood clusters, which adheres to the basic principles of gestalt. It can aggregate elements with consistent spatial distribution patterns and adjacent elements together to ensure that the integrity of the aggregated element set is not undermined [21]. In many-to-many matching, spatial elements are aggregated elements with neighboring relationships, so they can be divided by spatial neighborhood clusters. In this method, Delaunay triangulation network was used to construct the spatial adjacency relations of residential areas, and the candidate elements were divided into closed spatial neighborhoods. Then, the spatial neighborhood cluster composed of several elements was obtained by further screening through the spatial distance, which was called the matching aggregation factor set.

Let two types of data to be matched be small-scale data  $S = \{s_1, s_2, \dots, s_m\}$  and large-scale data  $C = \{c_1, c_2, \dots, c_n\}$ , among which  $s_i$  and  $c_i$  are polygons to be matched in data  $S$  and  $C$ , and  $m$  and  $n$  are corresponding polygon numbers respectively. First, based on the initial matching results, dividing  $S$  and  $C$  into two types, the matched entities were labeled  $S^1$  and  $C^1$ , and the unmatched entities were labeled  $S^2$  and  $C^2$ , wherein  $S^1$  and  $C^1$  are a 1 : 1 and 1 :  $N$  match, and  $S^2$  and  $C^2$  are a 1 : 0 and  $M$  :  $N$  match, respectively. Then, obtain the geometric center of the first type of entity ( $S^1$  and  $C^1$ ), and use Delaunay triangulation to construct the spatial neighborhood relationship of  $S^1$  and  $C^1$  based on the geometric center point. Finally, merge the triangular spaces that cover the second type of entity ( $S^2$  and  $C^2$ ). Entities within these triangulated spaces are spatial neighborhood clusters with  $M$  :  $N$  matched features.

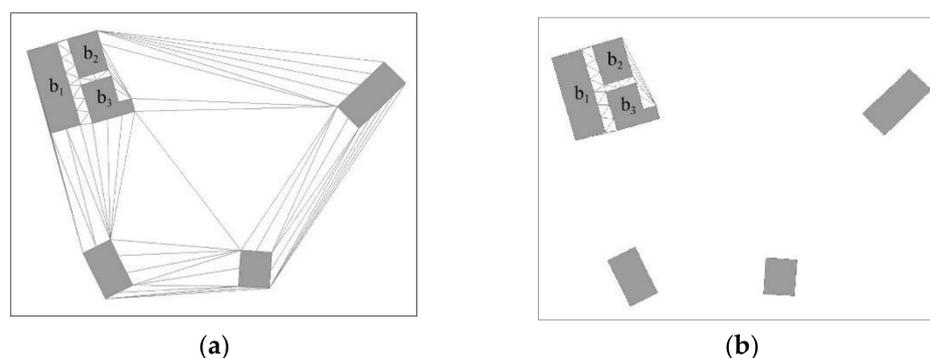
Because the first type of entity does not contain an  $M$  :  $N$  match, a Delaunay triangulation is constructed, and the  $M$  :  $N$  match is distributed in the triangular space constructed by the geometric center points of the first type of entity ( $S^1$  and  $C^1$ ). Then, the triangular spaces of the covered  $S^2$  and  $C^2$  are merged, which can prevent the aggregate element set of the  $M$  :  $N$  match from being undermined, thus ensuring its continuity and edges.

As can be seen from Figure 2,  $D$  means triangle, if  $D_1$ ,  $D_2$ , and  $D_3$  are intersected with  $s_1$ ,  $s_2$ , and  $s_3$  from the second type of entity, then the triangles are merged,  $D_3$  and  $D_4$  are intersected with  $s_4$ , and  $D_3$ ,  $D_4$  are then merged. Finally,  $D_1$ ,  $D_2$ ,  $D_3$  and  $D_4$  are merged into a triangular neighborhood space. In addition,  $s_5$  only intersects with  $D_6$ , so  $D_6$  is a single triangular space.



**Figure 2.** Adjacent Triangles Merging Graph.

After the neighborhood spaces of the triangles are determined, the areal entities ( $S^2$  and  $C^2$ ) located in the triangular spaces can be found by spatial analysis. There may be multiple areal entities in the same triangular space, but the spatial neighborhood clusters at this point are not necessarily the final many-to-many matching candidate element set. Some areal elements are far apart, so they do not meet the  $M : N$  match spatial proximity rule. Therefore, it is necessary to further narrow the scope to eliminate spatial neighborhood clusters and determine the final  $M : N$  matching element set. Drawing on the existing constrained Delaunay triangulation method, the method employed was as follows: (1) The edge nodes of polygon residential areas were extracted. Given that there are few residential area nodes, long, narrow triangles are likely to affect the construction of the Delaunay triangulation. Based on the literature [4], the residential nodes were densified. (2) A Delaunay triangulation was constructed that was constrained by the edges of the residential area, as shown in Figure 3a. (3) Based on the constrained Delaunay triangulation, the triangles of length smaller than  $d_t$  that connect different polygons were retained ( $d_t = G\mu$ , where  $G$  is the denominator of the scale and  $\mu$  is 0.4 mm [22]). Based on the connections of the triangle edges, the spatial neighborhood clusters of neighboring residential areas, that is, the final  $M : N$  matching candidate element set, were obtained, as shown in Figure 3b. Finally, the spatial domain needs to be calculated as a whole, so the convex hull method is introduced. A convex hull is the minimal convex polygon that contains all points of a cluster and represents the basic structure for describing the shape of a spatial object with small variable values. The convex hull processing is carried out for this spatial domain, and the similarity calculation is carried out to determine the matching relation of  $M : N$ .



**Figure 3.** Obtaining Spatial Neighborhood Clusters. (a) Constrained Delaunay Triangulation; (b) Spatial Neighborhood Clusters.

According to the above steps, the pseudo-code of the many-to-many matching candidate determination method based on spatial neighborhood cluster is as Algorithm 1:

---

**Algorithm 1** Spatial Neighborhood Clusters Algorithm

---

Input: small scale residential data  $S$ , large scale residential data  $C$ , divide  $S^1, S^2$  and  $C^1, C^2$ ;

Output: candidate matching cluster

/\*start\*/

/\*1. Construct Delaunay triangulation \*/

GetCentroid ( $S^1$  and  $C^1$ )

Delaunay ( $S^1$  and  $C^1$ )

/\*2. Merge triangulation space \*/

For  $i$  in  $S^2$

    If  $S^2(i)$  or  $C^2(i)$ .Intersection (Delaunay\_triangulation):

        Cluster\_tri.append (Delaunay\_triangulation)

Spatial\_Analysis (Cluster\_tri and  $C^2(i)$  or  $S^2(i)$ )

Get (Cluster\_origin)

/\*3. The final spatial cluster \*/

Delaunay\_constraint (Cluster\_origin)

Calculate( $d_t$ )

Get (Cluster\_final)

Return Cluster\_final

/\*end\*/

---

### 3. Calculating Similarity Taking into Account Spatial Neighborhood Features

In the real world, geographic elements have their own spatial features, and they are related to and mutually influence their surrounding neighborhood elements. The spatial features of residential entities that can be matched with each other have certain similarities, but their neighborhood features should also be similar. Sometimes, relying on their own features alone will not lead to accurate matching results. In Figure 1b, for example,  $p_2$  and  $p_3$  overlap spatially with  $q_1$ , but if neighborhood features are considered, it is relatively obvious that  $p_2$  and  $p_3$  are matching entities with  $q_2$ . Therefore, in the process of residential matching, it is necessary to consider the spatial similarities of the entities themselves, as well as to consider the similarities of spatial neighborhood features, and fully combine geometric and topological features of elements to obtain accurate matching results.

#### 3.1. Similarities in Features of Residential Areas

Traditional similarity measures mainly include location, direction, area and shape, etc. The following is a brief introduction.

### 3.1.1. Location Similarity

Spatial distance reflects the locational proximity of geographic entities, and it is an important measure of geometric similarity. The location of residential area is mainly reflected by its centroid, so the location similarity can be reflected by calculating the distance of centroid. If  $P_a(x_1, y_1)$  and  $P_b(x_2, y_2)$  are the geometric centroids of two polygon residential areas, the Euclidean distance is used to calculate position similarity as follows:

$$sim_{self\_pos} = 1 - \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{U} \quad (1)$$

where  $U$  is the maximum distance between the edge vertices of two polygon residential areas.

### 3.1.2. Direction Similarity

Direction is an important distinguishing feature of polygon residential areas, and it is generally the long axis of the Minimum Bounding Rectangle. The similarity of the directions of the two residential areas can be reflected by comparing the difference of the long axis direction of the MBR. If  $\theta_1$  and  $\theta_2$  are the direction values of two polygon settlements, then their direction similarity is calculated as follows:

$$sim_{self\_dir} = 1 - \frac{|\theta_1 - \theta_2|}{\theta_\tau} \quad (2)$$

where  $\theta_\tau$  is the direction threshold value, which is usually  $\pi/2$ .

### 3.1.3. Area Similarity

Area size is an important feature of a residential area. The smaller the difference is in the area of residential areas being matched, the greater the possibility is of them being similar entities. If  $A_1$  and  $A_2$  are the areas of two polygon residential areas, their area similarity is defined as follows:

$$sim_{self\_area} = 1 - \frac{|A_1 - A_2|}{Max(A_1, A_2)} \quad (3)$$

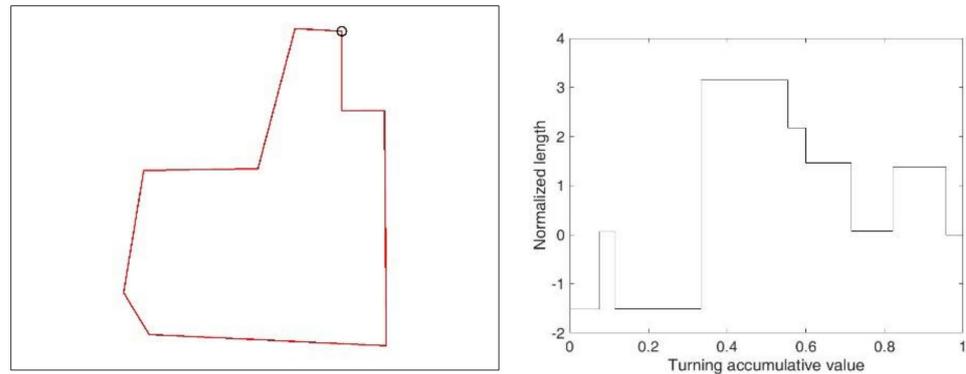
where  $Max(A_1, A_2)$  is the maximum value of  $A_1$  and  $A_2$ .

### 3.1.4. Shape Similarities

Shape is another important visual feature of polygon settlements, which can be widely used in detection, recognition, and matching of polygon objects. In this study, the turning function is used to describe the shape features of polygon residential areas. A polygon can be represented using a list of angle-length pairs, whereby the angle at a vertex is the accumulated tangent angle at this point, while the corresponding length is the normalized accumulated length of the polygon at this point [23]. Figure 4 shows the change of tangent angles ( $y$ -axis) along the normalized accumulated length of the polygon sides ( $x$ -axis) in the clockwise direction from the starting point (black hollow circle), respectively. From that point of view, the tangent angle can be regarded as a function of the normalized accumulated length. This kind of representation is invariant to rotation because it contains no orientation information. Furthermore, it is invariant to scaling, since the normalized length makes it independent of the polygon size. The shape similarity of the turning function is calculated as follows:

$$sim_{self\_shape} = 1 - \frac{\int_0^1 |e_a(l) - e_b(l)| dl}{\max(\int_0^1 e_a(l) dl, \int_0^1 e_b(l) dl)} \quad (4)$$

where  $e_a(l)$  and  $e_b(l)$  are the accumulated values of the corners of polygon residential areas.



**Figure 4.** Turning Function Representation.

In summary, the total similarity of features of the residential area itself is calculated as follows:

$$sim_{self} = \sum_{i=1}^4 w_i sim_{self\_X} \tag{5}$$

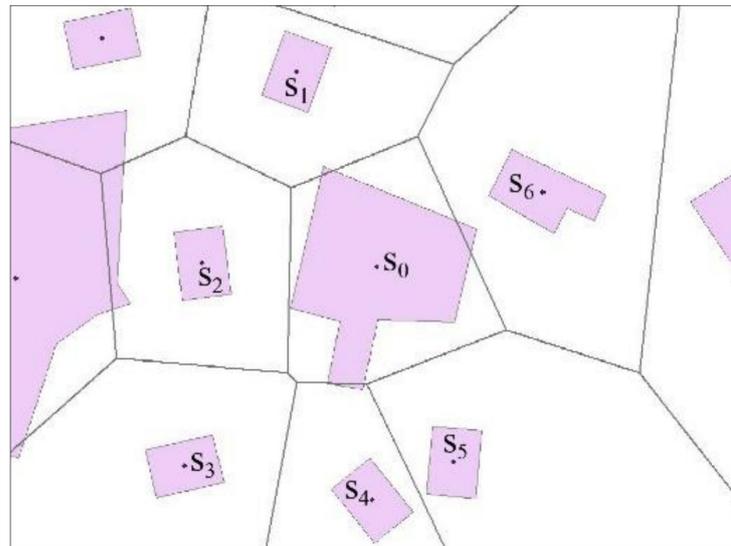
where  $X$  denotes *pos, dir, area, shape* i.e., position, direction, area, and shape features, and  $\sum_{i=1}^4 w_i = 1$ .

### 3.2. Similarity of Spatial Neighborhood Features of Residential Areas

#### 3.2.1. Identifying Spatial Neighbors

In order to calculate the similarity of spatial neighborhood features on the basis of the similarity of residents' own features, it is necessary to determine the spatial proximity target of residents' elements first. In this study, we used a Voronoi diagram to determine the space neighbors of small-scale polygon residential data  $S$ . A Voronoi diagram is a spatial neighborhood analysis tool that effectively conveys the spatial proximity of geographic entities. It is divided according to the nearest neighbor principle, and each point is associated with its nearest neighbor. It can ensure that each element has a certain number of space neighbors. First, the geometric centers of polygon residential areas in  $S$  were obtained, and then the Voronoi diagram was constructed. Space neighbors of polygonal elements can be determined according to distribution relationships. Figure 5 shows that, according to the spatial distribution of the Voronoi diagram, the space neighbors of residential area  $s_0$  are  $s_1$ – $s_6$ . For polygon residential area  $s_i$  with space neighbor  $s_h$ ,  $Vor$  represents the Voronoi diagram spatial distribution, and its space neighbor set is defined as  $N_i = \{s_h : (s_i, s_h) \in Vor\}$ .

Since the large-scale data set  $C$  contains more detailed and more geographic elements, the spatial distribution is quite different from that of the small-scale data set  $S$ , and the same method does not guarantee consistent space neighbors. Therefore, for the large-scale residential area data set  $C$ , defining the set of other areal targets in data set  $C$  with a distance from  $c_j$  that is smaller than  $d_\tau$  as space neighbors can be expressed as follows:  $N_j = \{c_k : d(c_j, c_k) \leq d_\tau\}$ , where  $c_j$  is the candidate matching entity of  $s_i$ ,  $c_k$  is the space neighbor of  $c_j$ , and  $d_\tau$  is the distance threshold, which is determined according to the accuracy of the data.

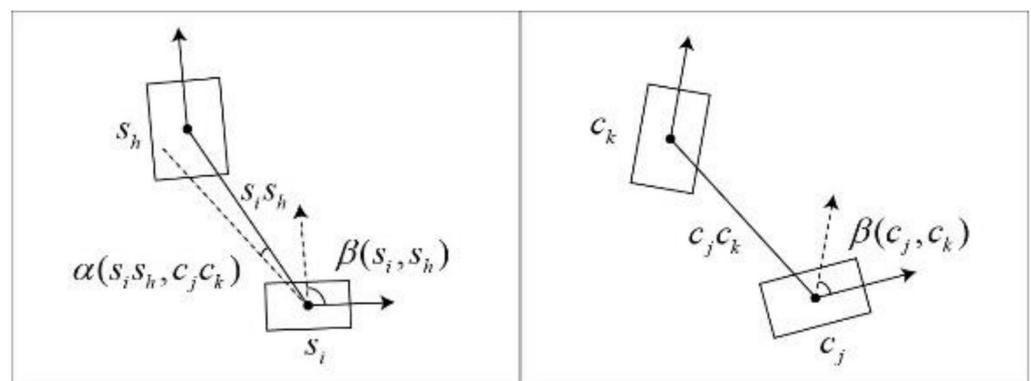


**Figure 5.** Spatial Neighbors of Residential Areas.

Using the Voronoi diagram and the distance threshold method, it is possible to determine the space neighbors of the multi-scale polygon residential area data and then calculate the space neighbor similarity of polygon residential areas at multiple scales.

### 3.2.2. Spatial Neighborhood Similarity

Once spatial neighbors are determined, it is necessary to determine specific matching relationships by calculating the spatial neighborhood similarity of polygon residential areas. If  $s_h$  and  $c_k$  are the space neighbors of  $s_i$  and  $c_j$ , then  $(s_h, c_k)$  is the neighbor candidate match of  $(s_i, c_j)$ , the relative position, relative direction, relative area, and relative shape similarities of their spatial neighborhood environments can be calculated as Figure 6 shows.



**Figure 6.** Relative Relationships of Polygon Residential Areas.

#### (1) Relative position similarity

Relative position is determined primarily from the distance and direction features of  $(s_i, c_j)$  and  $(s_h, c_k)$ , as shown in Figure 6. The calculation formula is as follows:

$$\begin{cases} r_{dis} = 1 - \frac{|d(s_i, s_h) - d(c_j, c_k)|}{\max_{s_m \in N_i, c_n \in N_j} (d(s_i, s_m), d(c_j, c_n))} \\ r_{dir} = 1 - \frac{\alpha(s_i, s_h, c_j, c_k)}{\pi/2} \\ sim_{rel\_pos} = r_{dis} \times r_{dir} \end{cases} \quad (6)$$

where  $d(s_i, s_h)$  and  $d(c_j, c_k)$  are the geometric center distances of areal targets;  $N_i$  and  $N_j$  are the space neighbor sets of  $s_i$  and  $c_j$ ; and  $r_{dis}$  is the relative distance relationship of the neighbor candidate match.  $\alpha(s_i, s_h, c_j, c_k)$  represents the relative direction of the centroid connection of  $(s_i, s_h)$  and  $(c_j, c_k)$ , with a threshold value of  $\pi/2$ . The relative position similarity  $sim_{rel\_pos}$  is obtained by multiplying  $r_{dis}$  and  $r_{dir}$ ,  $sim_{rel\_pos} \in [0, 1]$ , and the greater the relative distance  $r_{dis}$  and the relative direction  $r_{dir}$  are, the more similar the relative position values of  $(s_i, s_h)$  and  $(c_j, c_k)$  are and vice versa.

### (2) Relative direction similarity

The main directions of the areal targets are shown in Figure 6. Relative direction similarity is the difference between the main directions of  $(s_i, s_h)$  and  $(c_j, c_k)$ :

$$sim_{rel\_dir} = 1 - \frac{|\beta(s_i, s_h) - \beta(c_j, c_k)|}{\pi/2} \quad (7)$$

where  $\beta$  is the difference between the main directions of the two areal targets,  $\beta \in [0, \pi/2]$ , and  $sim_{rel\_dir}$  represents the direction similarity of  $(s_i, s_h)$  and  $(c_j, c_k)$ , where the greater the value is, the more similar the relative directions of the target pair are and vice versa.

### (3) Relative area similarity

Relative area similarity is determined mainly by the sizes of the areal targets and is calculated as follows:

$$\begin{cases} sim_{rel\_area} = \frac{1}{1 + (\phi(s_i, s_h) - \phi(c_j, c_k))^2} \\ \phi(s_i, s_h) = 1/\phi(s_i, s_h), \phi(c_j, c_k) = 1/\phi(c_j, c_k), \text{ if } \phi(s_i, s_h) > 1 \end{cases} \quad (8)$$

where  $\phi(s_i, s_h) = Area(s_i)/Area(s_h)$ ,  $\phi(c_j, c_k) = Area(c_j)/Area(c_k)$ , and  $Area$  are the areas of corresponding targets and  $sim_{rel\_area}$  represents the relative area similarity between  $(s_i, s_h)$  and  $(c_j, c_k)$ . To ensure that  $\phi(s_i, s_h) \in [0, 1]$ , if  $\phi(s_i, s_h) > 1$ , both  $\phi(s_i, s_h)$  and  $\phi(c_j, c_k)$  are counted in reverse.

### (4) Relative shape similarity

$$\begin{cases} sim_{rel\_shape} = \frac{1}{1 + (\delta(s_i, s_h) - \delta(c_j, c_k))^2} \\ \delta(s_i, s_h) = 1/\delta(s_i, s_h), \delta(c_j, c_k) = 1/\delta(c_j, c_k), \text{ if } \delta(s_i, s_h) > 1 \end{cases} \quad (9)$$

where  $\delta(s_i, s_h) = e(s_i)/e(s_h)$ ,  $\delta(c_j, c_k) = e(c_j)/e(c_k)$ , and  $e(\cdot)$  are areal target shape values calculated using the target function, and  $sim_{rel\_shape}$  represents the relative shape similarity between  $(s_i, s_h)$  and  $(c_j, c_k)$ . To ensure that  $\delta(s_i, s_h) \in [0, 1]$ , if  $\delta(s_i, s_h) > 1$ , both  $\delta(s_i, s_h)$  and  $\delta(c_j, c_k)$  are counted in reverse.

Based on Equations (6)–(9), the overall neighborhood similarity of  $(s_i, s_h)$  and  $(c_j, c_k)$  is expressed as follows:

$$sim_{rel}(s_i, c_j; s_h, c_k) = \sum_{i=1}^4 w_i sim_{rel\_Y} \quad (10)$$

where  $Y$  represents  $pos, dir, area, shape$ , that is, the position, direction, area, and shape features, and  $\sum_{i=1}^4 w_i = 1$ .

Because there is more than one neighboring matching candidate of  $s_i$  and  $c_j$ , it is necessary to generalize all the neighbor matching targets of  $s_i$  and  $c_j$  and calculate the overall neighborhood similarity as follows:

$$sim_{rel} = \sum_{s_h \in N_i} \max_{k \in N_j} (sim_{rel}(s_i, c_j; s_h, c_k)) / Num \quad (11)$$

where  $Num$  represents the number of space neighbors of  $s_i$  in  $N_i$ .

### 3.3. Using the Relief-F Algorithm to Determine Similarity Weight Values

In previous studies, weight values have been determined largely using empirical values, which is subjective and does not guarantee optimal weight values. The Relief-F algorithm is a multi-class feature selection algorithm that is an update of the Relief algorithm. The Relief-F algorithm is a method for calculating the weights of features based on sample learning. It has been used in many fields because of its simplicity, intuitiveness, and computational efficiency [24,25]. The algorithm assigns initial weights to each feature and then refines the weights by updating the formula, finally obtaining the weights of the various features. The Relief-F algorithm was used in this study to determine the weights of the various similarity features.

If the matching entity feature set is  $R = \{R_1, R_2, \dots, R_N\}$ ,  $R_i = \{R_{i1}, R_{i2}, \dots, R_{iM}\}$  represents the  $i^{\text{th}}$  matching entity feature (the features in this study are distance, area, direction, and shape, so  $M = 4$ ), and  $Class = \{Class_1, Class_2, \dots, Class_p\}$  represents the matching entity features of different groups. The  $k$  nearest neighbor samples were obtained from the sample sets of the same and different groups as  $R_i$ , and the following formula was used to update the weight values of the  $j^{\text{th}}$  feature of  $R_i$ .

$$\begin{cases} w_j^i = w_j^{i-1} + \frac{1}{k \times N} [diff_{miss}(R_i, k) - diff_{hit}(R_i, k)] \\ diff_{miss}(R_i, k) = \sum_{m=1}^k \frac{|R_{ij} - R_{hitmj}|}{\max(R_{*j}) - \min(R_{*j})} \\ diff_{hit}(R_i, k) = \sum_{C \notin Class(R_i)} \left[ \frac{P(C)}{1 - P(Class(R_i))} \sum_{m=1}^k \frac{|R_{ij} - R_{missmj}|}{\max(R_{*j}) - \min(R_{*j})} \right] \end{cases} \quad (12)$$

where  $diff_{miss}(R_i, k)$  and  $diff_{hit}(R_i, k)$  represent the differences between  $R_i$  and the same group and the different group, respectively;  $\max(R_{*j})$  and  $\min(R_{*j})$  are the largest and smallest values of the  $j^{\text{th}}$  feature, respectively;  $P(C)$  is the probability of class  $C$ , that is, the number of samples in class  $C$  as a proportion of the total samples; and  $R_{hitmj}$  and  $R_{missmj}$  represent the  $j^{\text{th}}$  eigenvalues of samples in the same group and different group as  $R_i$ .

Equation (12) was used to determine the weight values in Equations (5) and (10). According to the calculated feature similarity  $sim_{self}$  of the residential area itself and the spatial neighborhood similarity  $sim_{rel}$ , the average of the two was taken as  $sim = (sim_{self} + sim_{rel})/2$  to obtain the final similarity value of the entities to be matched,  $s_i$  and  $c_j$ , with the final matching relationships determined according to the set threshold.

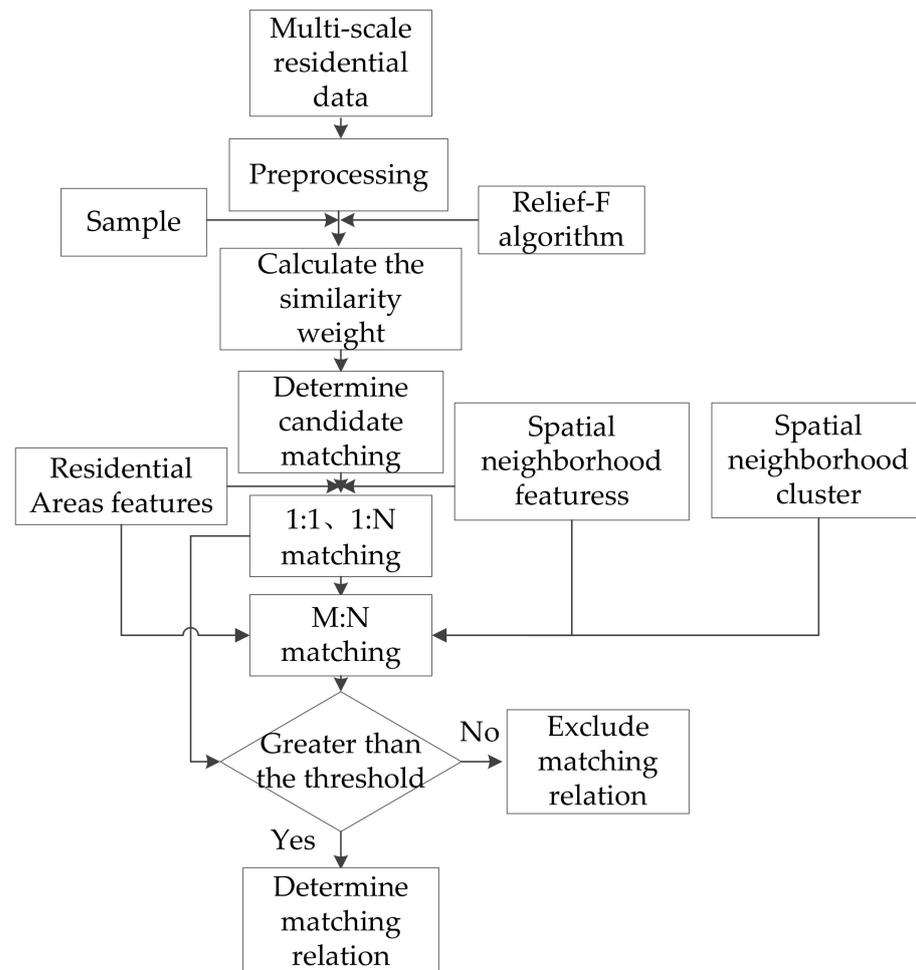
## 4. Test and Analysis

### 4.1. Matching Process and Test

The following is the testing process for the multi-scale residential area matching method proposed in this study (Figure 7):

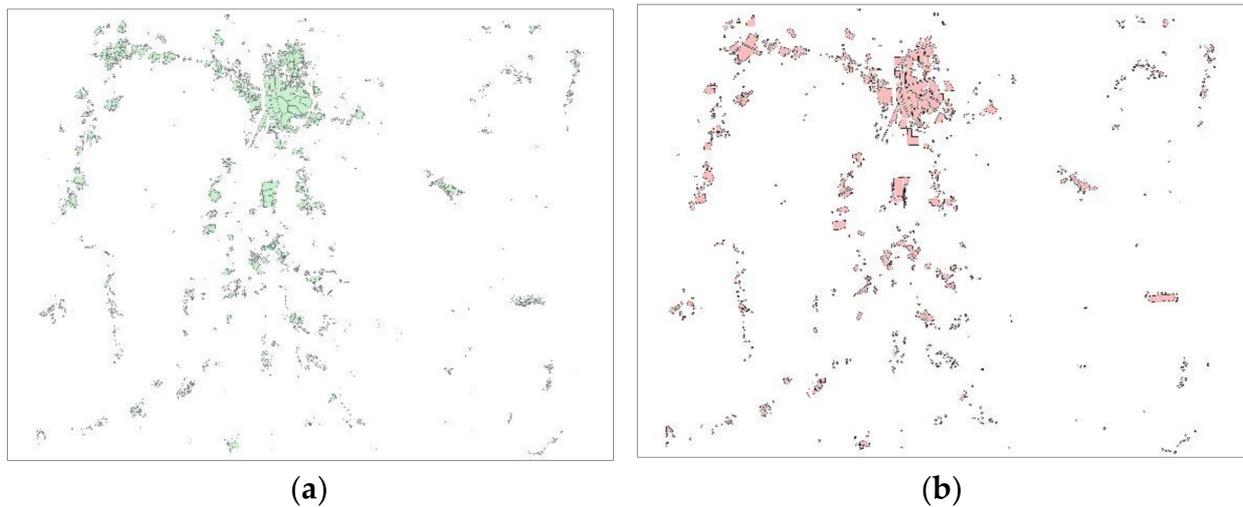
1. Data preprocessing is conducted, which mainly includes data format conversion, coordinate system conversion, projection alteration, and topology checking;
2. Weight values are determined by selecting a certain number of positive samples and calculating the sub-features of similarity of a residential area and spatial neighborhood similarity, using the Relief-F algorithm to determine the weight values of the various features;
3. Initial matching is undertaken using Minimum Bounding Rectangle to screen candidate matching entities and then by calculating spatial similarity values and determining 1 : 1 and 1 :  $N$  candidate matches;
4. Spatial neighborhood clusters are determined using the method described in Section 2.2, based on the initial matching. The matched entities were labeled  $C^1$  and  $S^1$ , and the unmatched entities were labeled  $C^2$  and  $S^2$ , which mainly included  $M$ :  $N$  matching and 1: 0 matching. Finally, the Delaunay triangulation network is created to divide the many-to-many matching spatial domain;

5.  $M : N$  matching is conducted by performing convex hull processing on the obtained aggregated element set of the spatial neighborhood clusters and converting them into single entities for matching, and spatial similarity is calculated to determine the matching relationship  $M : N$ ;
6. All matching results are obtained and evaluated, and the matching relation is mainly determined by the spatial similarity value.



**Figure 7.** The Matching Process of Multi-scale Residential Areas.

To verify the effectiveness of the algorithm proposed in this study for matching multi-scale residential areas, the ArcGIS Engine 10.2 software was used to design and implement the algorithm using the Python programming language. We used a computer with the Microsoft Windows 10 operating system, an Intel i7 processor, 8 GB of memory, and a 512-GB hard drive. The test used residential data at scales of 1:50,000 and 1:10,000 for an area of the city of Ningbo, Zhejiang Province in eastern China, which included 689 and 2016 polygon entities, respectively. The selected area has both densely populated urban areas and more sparsely populated suburban areas, which are suitable for matching. The test data are shown in Figure 8.



**Figure 8.** Test Data. (a) 1:10,000 Residential Area Data; (b) 1:50,000 Residential Area Data.

#### 4.2. Test Results and Analysis

Five groups of positive samples were selected from typical areas, such as urban, rural, and suburban areas, of the experimental data, with 20 samples in each group. The method proposed in this paper was used to calculate the similarity values for location, direction, area, and shape features of the residential areas themselves and their spatial neighborhoods. The Relief-F algorithm was used to refine the weight values of the features. During the training process, a value of 8 for  $k$  was obtained after running the algorithm 20 times. The similarity feature weight values in Equations (5) and (10) were obtained and are shown in Table 2. In addition, in accordance with the experience and knowledge of experts [14,17,26], and after much feedback, the distance threshold was set to  $d_\tau = 50m$ , and the similarity threshold was set to  $sim_0 = 0.6$ .

**Table 2.** Weight Values of Similarity Features.

Weight Value	Position	Direction	Area	Shape
Equation (5)	0.396	0.238	0.176	0.190
Equation (10)	0.317	0.262	0.193	0.228

To evaluate the matching efficacy of the algorithm proposed in this study, values of the matching accuracy  $P$ , recall  $R$ , and  $F$  were calculated as follows by comparison with manual matching results (through visual observation):

$$\begin{cases} P = \frac{f(C)}{f(C)+f(W)+f(M)} \\ R = \frac{f(C)}{f(C)+f(U)} \\ F = \frac{2PR}{P+R} \end{cases} \quad (13)$$

where  $f(C)$  is the number of correct matches,  $f(W)$  is the number of mismatches,  $f(M)$  is the number of matches that could not be determined manually, and  $f(U)$  is the number of missed matches.

We used the test data to compare the method employed in this study with the area overlap rate [7] and own feature similarity [8] matching methods from other studies, with the area overlap rate threshold set to 0.5 and the similarity threshold set to 0.6. The match results statistics are shown in Table 3. It can be seen that the accuracy, recall, and  $F$  value of the method employed in this study were all above 90%, but the accuracy and recall of the matching methods employed in the other studies [7,8] were in the range of 70 to 80%. Since the experimental data used in this paper were preprocessed, that is, there were no

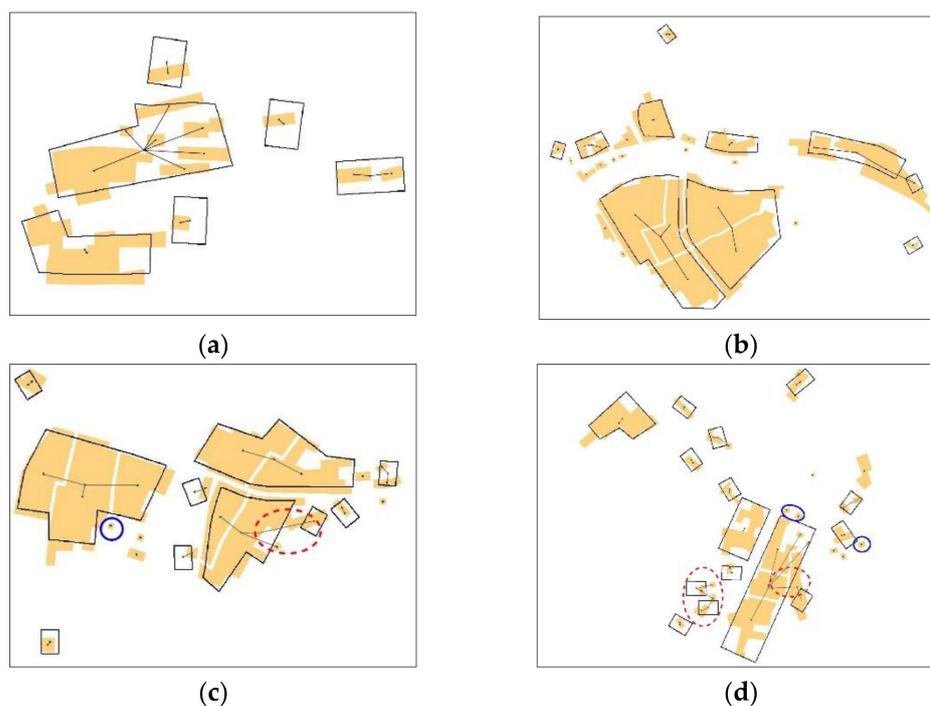
large geometric deviations, the area overlap ratio matching method was relatively good, but some 1 :  $N$  matches could not be fully matched and were prone to mismatches, and some entities with individual geometric deviations were prone to missed matches. In terms of multi-scale matching, because of significant differences in spatial features between some matching entities, there were many 1 :  $N$  matches. The matching method based on own feature similarity is prone to mismatches and missed matches. Furthermore, the methods used in the other studies [7,8] could not effectively identify  $M$  :  $N$  matching relationships.

**Table 3.** Match Results Statistics.

Matching Method	$f(C)$	$f(W)$	$f(M)$	$f(U)$	$P/\%$	$R/\%$	$F/\%$	Running Time/s
This study	579	40	18	52	90.9	91.8	91.3	85
Previous study [7]	465	89	18	117	81.3	79.9	80.6	13
Previous study [8]	402	105	18	164	76.6	71.0	73.7	21

However, the two methods used in the other studies are less complex, so they are quicker than the method used in this study. The method in this paper needs to determine candidate targets for many times, which is time-consuming in the traverse process. In addition, Voronoi diagram and Delaunay triangulation network are also used, so the overall speed is slightly lower.

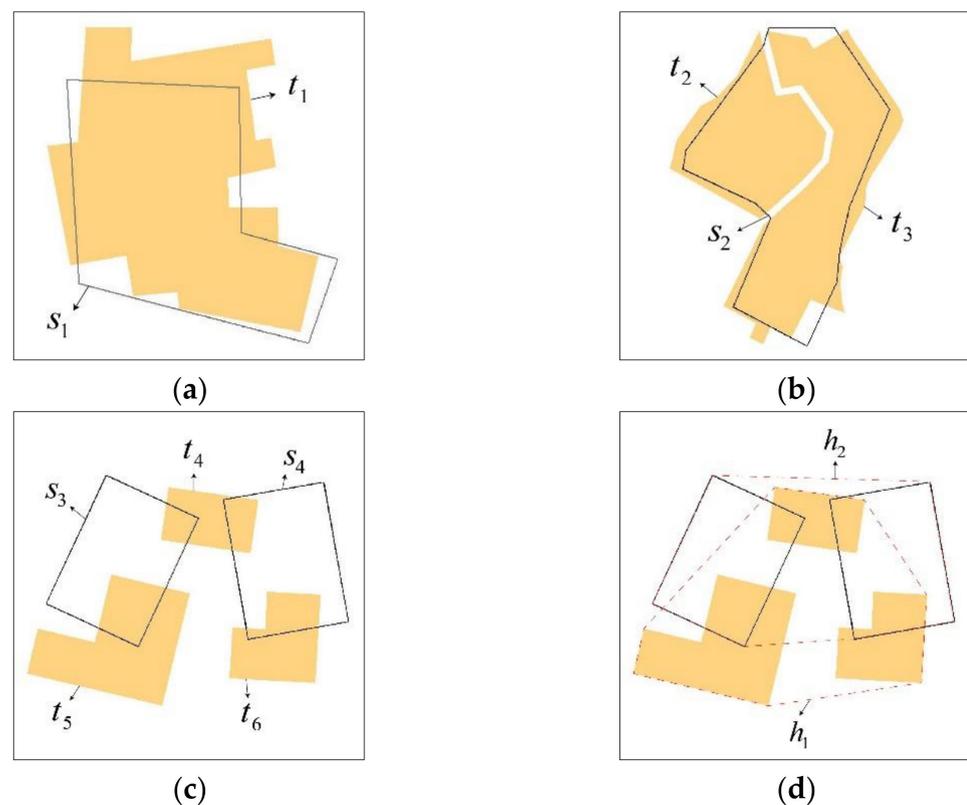
Figure 9 shows some of the detailed matching results for the data from the test area, and the drawing specification mainly expresses the matching relation by connecting the centroid of gravity of different residential areas. It can be seen that the method employed in this study can effectively handle multi-scale polygon residential matching, can identify various types of matching relationships, and has good matching efficacy. It can be seen from Figure 9a,b that the distribution of polygon residential areas is relatively regular, with no significant positional deviations and mostly 1 : 1 matching and simpler 1 :  $N$  matching, so these polygon residential areas are easier to match, and the matching results are better.



**Figure 9.** Matching Results of the Proposed Method.

If there is a complex spatial distribution, the spatial similarity features of the residential areas themselves are not sufficiently obvious, so matching these entities depends more on spatial neighborhood features. The method used in this study takes this into account, thereby improving matching efficiency somewhat. Figure 9c,d show that the overall matching results of polygon residential areas were good, and the method proposed in this study, which is based on spatial neighborhood clusters, can accurately identify complex  $M : N$  matches, as indicated by the areas in red dotted lines in the figure. However, because of differences in the features of individual elements being too large or matching relationships being relatively fuzzy, mismatches and missed matches also occurred, as indicated by areas in blue solid lines in the figure.

Figure 10 shows a matching sample selected from the results set. The similarities of the different entities were obtained from the calculations, as shown in Table 4. In Figure 10a,  $s_1$  and  $t_1$  are a  $1 : 1$  matching relationship. The own feature similarity and spatial neighborhood similarity of the two areal entities were both very high, so a matching relationship could be accurately determined between the two. In Figure 10b,  $s_2$  and  $t_2, t_3$  are a  $1 : N$  match. The similarity between the two entity types is relatively high, and the matching relationship between the three can also be accurately determined through two-way matching. In Figure 10c,  $s_3$  and  $s_4$ , together with  $t_4, t_5$ , and  $t_6$ , are an  $M : N$  match caused by typification. After initial matching, a specific matching relationship could not be determined, but an  $M : N$  candidate matching relationship could be determined using the method based on spatial neighborhood clusters. Convex hull processing was then performed. As Figure 10d shows, this matching relationship was converted into a  $1 : 1$  simple matching, and the similarity value was calculated to determine the matching relationship between the elements.



**Figure 10.** Matching Samples Selected from the Results.

**Table 4.** Matching Results for Figure 9.

Matching Sample	Entity Pair	Similarity	Match
Figure 10a	$s_1 : t_1$	0.883	Y
Figure 10b	$s_2 : t_2$	0.676	Y
	$s_2 : t_3$	0.724	Y
Figure 10c	$s_3 : t_4$	0.504	N
	$s_3 : t_5$	0.541	N
	$s_4 : t_4$	0.492	N
	$s_4 : t_6$	0.518	N
Figure 10d	$h_1 : h_2$	0.795	Y

## 5. Conclusions

This paper introduced a multi-scale residential area matching method that takes into account spatial neighborhood features. This method analyzes the complex matching relationships of multi-scale polygon residential areas and proposes a many-to-many candidate matching determination method that is based on spatial neighborhood clusters to achieve complex many-to-many matching relationships. It also combines similarities of polygon residential areas themselves and spatial neighborhood similarities, taking into account geometric and topological features of residential areas, to obtain accurate matching results. The experimental results show that this method can identify complex one-to-many and many-to-many matching relationships in multi-scale residential areas and overcome the influence of positional deviations on matching results, with a high degree of matching accuracy.

Although the method is believed to be very useful, there are issues for further investigation concerning its practical applications. One is that the primary match is found correctly, but other matches can be lost in special cases, which usually happens if several objects have been aggregated, but one of them is more similar to the target object. The other is that when the scale difference between the data to be matched is great, the residents are greatly affected by cartographic generalization factors, and their own features and spatial neighborhood features will change, which is easy to cause mismatching. Future studies should focus on adapting this method to matching polygon residential areas with scales of different spans and extending the method to matching other areal elements.

**Author Contributions:** Conceptualization, Jingzhen Ma; methodology, Jingzhen Ma, Qun Sun and Shaomei Li; software, Jingzhen Ma; validation, Bowei Wen; formal analysis, Jingzhen Ma; investigation, Zhao Zhou; resources, Qun Sun; data curation, Bowei Wen; writing—original draft preparation, Jingzhen Ma; writing—review and editing, Jingzhen Ma, Qun Sun and Shaomei Li; visualization, Jingzhen Ma; supervision, Shaomei Li; project administration, Zhao Zhou; funding acquisition, Jingzhen Ma and Qun Sun. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was supported by the National Natural Science Foundation of China (42101454, 41901397), the Fund Project of Zhongyuan Scholar of Henan Province (202101510001) and the Joint Fund of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province and Key Laboratory of Spatiotemporal Perception and Intelligent processing, Ministry of Natural Resources (212102).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data and codes presented in this study are available on request from the authors.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ruiz, J.; Ariza, F.; Ureña, M.; Elidia, B. Digital Map Conflation: A Review of the Process and a Proposal for Classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [\[CrossRef\]](#)
2. Xavier, E.; Ariza, F.; Ureña, M. A Survey of Measures and Methods for Matching Geospatial Vector Datasets. *ACM Comput. Surv.* **2016**, *49*, 1–34. [\[CrossRef\]](#)
3. Fu, Z.; Yang, Y.; Gao, X.; Zhao, X.; Fan, L. An Optimization Algorithm for Multi-characteristics Road Network Matching. *Acta Geod. Cartogr. Sin.* **2016**, *45*, 608–615.
4. Wang, X.; Qian, H.; He, H.; Chen, J.; Hu, H. Matching Multi-source Areal Habitations with Skeleton Line Mesh of Blank Region. *Acta Geod. Cartogr. Sin.* **2015**, *44*, 927–935.
5. Guo, Q.; Xie, Y.; Liu, J.; Wang, L.; Zhou, L. Algorithms for Road Networks Matching Considering Scale Variation and Data Update. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 381–388.
6. Kim, J.; Vasardani, M.; Winter, S. Similarity Matching for Integrating Spatial Information Extracted from Place Descriptions. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 56–80. [\[CrossRef\]](#)
7. Masuyama, A. Methods for Detecting Apparent Differences Between Spatial Tessellations at Different Time Points. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 633–648. [\[CrossRef\]](#)
8. Hao, Y.; Tang, W.; Zhao, Y.; Li, N. Areal Feature Matching Algorithm Based on Spatial Similarity. *Acta Geod. Cartogr. Sin.* **2008**, *37*, 501–506.
9. An, X.; Sun, Q.; Xiao, Q.; Yan, W. A shape multilevel description method and application in measuring geometry similarity of multi-scale spatial data. *Acta Geod. Cartogr. Sin.* **2011**, *40*, 495–501.
10. Luo, G.; Zhang, X.; Qi, L.; Guo, T. The Fast Positioning and Optimal Combination Method of Change Vector Object. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 1285–1292.
11. Samal, A.; Seth, S.; Cueto, K. A Feature-based Approach to Conflation of Geospatial Sources. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 459–489. [\[CrossRef\]](#)
12. Kim, J.; Yu, K.; Heo, J.; Lee, H. A New Method for Matching Objects in Two Different Geospatial Datasets Based on the Geographic Context. *Comput. Geosci.* **2010**, *36*, 1115–1122. [\[CrossRef\]](#)
13. Tong, X.; Deng, S.; Shi, W. A Probabilistic Theory-based Matching Method. *Acta Geod. Cartogr. Sin.* **2007**, *36*, 210–217.
14. Wu, J.; Wan, Y.; Chiang, Y.; Fu, Z.; Deng, M. A Matching Algorithm Based on Voronoi Diagram for Multi-Scale Polygonal Residential Areas. *IEEE Access* **2018**, *6*, 4904–4915. [\[CrossRef\]](#)
15. Wang, Y.; Chen, D.; Zhao, Z.; Ren, F.; Du, Q. A Back-Propagation Neural Network-Based Approach for Multi-Represented Feature Matching in Update Propagation. *Trans. GIS* **2015**, *19*, 964–993. [\[CrossRef\]](#)
16. Zhang, Y.; Huang, J.; Deng, M.; Fang, X.; Hu, J. Relaxation Labelling Matching for Multi-scale Residential Datasets Based on Neighboring Patterns. *Geomat. Inf. Sci. Wuhan Univ.* **2018**, *43*, 1098–1105.
17. Zhang, X.; Ai, T.; Stoter, J.; Zhao, X. Data Matching of Building Polygons at Multiple Map Scales Improved by Contextual Information and Relaxation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 147–163. [\[CrossRef\]](#)
18. Xu, J.; Wu, F.; Zhu, J.; Qian, H. A Multi-to-Multi Matching Algorithm Between Neighborhood Scale Settlement Data. *Geomat. Inf. Sci. Wuhan Univ.* **2014**, *39*, 340–345.
19. Huh, Y.; Kim, J.; Lee, J.; Yu, K.; Shi, W. Identification of Multi-scale Corresponding Object-set Pairs Between Two Polygon Datasets with Hierarchical Co-clustering. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 60–68. [\[CrossRef\]](#)
20. Zhang, L.; Guo, Q.; Sun, Y. The Method of Matching Residential Features in Topographic Maps at Neighboring Scales. *Geomat. Inf. Sci. Wuhan Univ.* **2008**, *33*, 604–607.
21. Liu, L.; Zhu, D.; Zhu, X.; Ding, X.; Guo, W. A Multi-scale Polygonal Object Matching Method Based on MBR Combinatorial Optimization Algorithm. *Acta Geod. Cartogr. Sin.* **2018**, *47*, 652–662.
22. Yang, M.; Ai, T.; Yan, X.; Chen, Y.; Zhang, X. A Map-algebra-based Method for Automatic Change Detection and Spatial Data Updating across Multiple Scales. *Trans. GIS* **2018**, *22*, 435–454. [\[CrossRef\]](#)
23. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [\[CrossRef\]](#)
24. Kononenko, I. Estimating attributes: Analysis and extension of RELIEF. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.
25. Du, S.; Luo, L.; Cao, K.; Shu, M. Extracting Building Patterns with Multilevel Graph Partition and Building Grouping. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 81–96. [\[CrossRef\]](#)
26. Chen, L.; Zhang, X.; Lin, H.; Yang, M. Change Analysis and Decision Tree Based Detection Model for Residential Objects across Multiple Scales. *Acta Geod. Cartogr. Sin.* **2018**, *47*, 403–412.