IEEE Access
Multidisciplinary Rapid Review Open Access Journal

Received December 7, 2017, accepted January 9, 2018, date of publication January 15, 2018, date of current version February 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2793302

A Matching Algorithm Based on Voronoi Diagram for Multi-Scale Polygonal Residential Areas

JIANHUA WU[1,2], YANGYANG WAN[2], YAO-YI CHIANG[3], ZHONGLIANG FU[4], AND MIN DENG[1]

[1]School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
[2]School of Geography and Environment, Jiangxi Normal University, Nanchang 330022, China
[3]Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90007 USA
[4]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

Corresponding author: Jianhua Wu (wjhgis@126.com)

This work was supported in part by the National Natural Science Foundation of China under Projects 41201409 and 41561084, and in part by the China Scholarship Council through a scholarship under Grant 201409470010.

ABSTRACT Matching spatial entities (e.g., polygonal residential areas) from sources of significantly different map scales is challenging. The reason is that the same entities in two map scales have significant variations in their positions, structure shapes and numbers, and topological relationships. Traditional matching methods based on minimum boundary rectangles (MBRs) or buffers usually lead to missed matches or mismatching. Furthermore, most of the previous approaches on entity similarity calculation are designed for datasets with specified map scales, which cannot directly apply to another set of dataset with a different scale. In this paper, we present a general approach using the Voronoi diagram for spatial entity matching on multi-scale datasets. Our approach first employs an efficient algorithm to construct the Voronoi diagram from the small-scale dataset. Next, the approach traverses each Voronoi polygon to find the corresponding large-scale features as the matching candidates (for each small-scale feature). Using the Voronoi diagram for identifying matching candidates does not require a manually determined search space (in contrast to the buffer-based approach). Also, our algorithm effectively uses the Voronoi diagram to prune the number of matching candidates even when the sources for matching contain large inconsistent position deviations. Finally, our approach utilizes three similarity indexes, namely, the convex hull shape similarity, convex hull area similarity, and overlapping area ratio to confirm the final matching results. We conducted experiments on two sets of datasets of two cities in China. The scales of the tested datasets were 1:10 000 and 1:50 000 and 1:1000 and 1:10 000. We compared our Voronoi-based method to both the MBR and buffer-based methods. The experiments showed that our method outperformed both the previous methods in generality and quality. Specifically, for the datasets where the inconsistent position deviations were large (i.e., the datasets of 1:1000 and 1:10 000 scales), the average F-measure of our results were 12.46%, 20.8%, and 64.45% higher than the MBR-based, 6-m buffer-based, and 3-m buffer-based methods, respectively.

INDEX TERMS Shape similarity, entity matching, Voronoi diagram, multi-scale, data conflation.

I. INTRODUCTION

Geographic data from different sources have their respective data qualities, and their geographic features have varying geometric shapes, topological structures, geometry accuracy, details of attributes, coding schemes, semantic representations, and spatial relationships [1]. A generate strategy for integrating multi-source geographic data is to adopt map conflation techniques to combine or update the geometry and attributes of the same entities from different sources [2]. Specifically, entity matching is a key technology that uses a series of similarity indexes to identify the features in multi-source, multi-scale, or multi-temporal map data that represent the same geographic phenomenon [3]–[5]. The majority of theories and methods of entity matching originated from a map conflation project of the United States Census Bureau between 1983 to 1985 [6]. After thirty years of development, researchers have achieved significant progress and plenty of research results. It includes a variety of similarity indexes of spatial entities [2], [7]–[10] and matching strategies for datasets with the same or multiple scales [11]–[15]. As well as matching accuracy [9], [15], [16]. Nevertheless, there still exists challenges to implement a generic matching method of polygonal residential area datasets with different scales.

First, the existing methods such as buffers or minimum bounding rectangles (MBR) based approaches for computing matching candidates have poor adaptability because of the uncertainties in producing multi-scale spatial data and changes in the ground truth. Figure 1 shows examples of multi-scale polygonal residential areas covering the same area where geographic features significantly vary in geometric structure and shape, topology, the number of geometries, spatial position, size, etc. In Figure 1, the red wireframes represent geographic features of a small-scale source (1: 10,000), the yellow areas represent features of a large-scale source (1: 1,000), the hatched areas represent user-specified buffers, and the green lines represent the MBRs. For matching entities from the two datasets, the user manually sets a buffer radius according to the map scales for computing the matching candidates. A small buffer size can lead to a missed match. For example, in Fig. 1(a), the large-scale features 715 and 716 (f_L715 and f_L716) are not entirely covered in the six-meter buffer so they could be discarded during the matching process. If a large buffer size is used, undesired features could be included in the matching process (e.g., f_L1339 and f_L1361 in Fig. 1 (b)). Acquiring matching candidates based on MBRs can also result in missed matches if the positional variation is significant. In Fig. 1(c), the small-scale f_S62 matches with all large-scale features in the figure, but the MBR of f_S62 only covers a fraction of the features (e.g., f_L1252 and f_L1242 do not intersect with the MBR).

Second, many existing similarity indexes (e.g., [17]–[19]) do not handle multi-scale polygonal residential areas matching. The reason is that the existing similarity indexes are sensitive to the positional uncertainties of the matching datasets (especially for trans-scale, for which the denominator of the small scale is five times more than that of the large scale). For instance, the tangent space-based shape similarity proposed in [19] depends on a manually specified buffer for identifying matching points in two match candidates. Figure 1(d) shows that a six-meter buffer cannot help to find the corresponding point P' from P (i.e., f_S122 fails to match f_L907). Also, the similarity indexes based on feature area ratio are usually used to match datasets with the same or similar map scales where most of the entity matching is a one-to-one relationship (e.g., [20]). When the difference in map scales between the matching datasets is large, the feature representations (e.g., shapes and area sizes) could vary significantly, which is hard for an area-ratio based approach to handle. For instance, a universal threshold for the area ratio cannot handle all the matching cases in Figure 1(e).

In this paper, we present a Voronoi diagram-based matching algorithm using geometric similarity indexes. Our matching algorithm handles geographical datasets without attributes or have significant attribute differences (e.g., the difference in schemas, naming, or coding conventions). Residential area Matching Method Based on Voronoi

FIGURE 1. Examples of the buffer and MBR based entity matching: (a) Acquiring matching candidates using a small buffer size, (b) Acquiring matching candidates using a large buffer size, (c) Acquiring matching candidates based on MBRs, (d) Examples of identifying matching points using a six-meter buffer and (e) The same entities from different map scales vary significantly.

diagram, with robust, adaptive similarity indexes and efficient matching strategies (hereafter the matching method is termed RMMBV). Using the Voronoi diagram, RMMBV can efficiently prune the matching space without manually setting the distance thresholds (e.g., the buffer size) while limiting the number of missed matches even when the two sources have significant or inconsistent position deviations.

RMMBV focuses on matching multi-scale (trans-scale) polygonal residential areas and aims to enhance the matching quality and generality from the previous work. The first step of RMMBV is an efficient algorithm for constructing the Voronoi diagram for identifying a set of matching candidates in the large-scale dataset for each small-scale feature. This step does not require manually determined search space (i.e., a distance threshold). Next, RMMBV utilizes a feature combination strategy based on a generalized nearest distance between a matching candidate and the target feature to identify one-to-one and one-to-many matches. The feature combination strategy employs adaptive similarity indexes, which are robust to uncertainties in the data sources and handle one-to-many matches in multi-scale residential area datasets.

The remainder of this paper is organized as follows. Section II presents the related work. Section III explains our RMMBV method. Section IV describes our experiments and results. Section V concludes our work and discusses future directions.

II. RELATED WORK

In general, the process of entity matching involves two main stages: identifying matching candidates and determining the final one-to-one or one-to-many matches. For searching matching candidates, most of the existing methods rely on a manually specified buffer (e.g., [12], [15], [16], [20]) or the MBR (e.g., [5], [14], [18], [21], [22]) to define a search space. However, as discussed in Section 1, these methods are not robust to handle multi-scale polygonal residential areas with large or inconsistent positional offset. Huang and Jiang [23] presented preliminary work using the Voronoi diagram for entity matching in trans-scale polygonal residential areas. Their work did not require a manually setting buffer or MBR, but they did not provide a detailed algorithm and experiment results. Yan and Wang [24] used a method based on the Voronoi polygons for cartographic generalization. Their algorithm for creating Voronoi polygons required lots of inter-visibility computation. In contrast, this paper presents a complete algorithm for finding matching candidates using the Voronoi diagram for polygonal residential area dataset. Our algorithm focuses on matching multi-scale (trans-scale) polygonal datasets of the residential area. Our method of creating the Voronoi diagram is based on interpolated points in the area boundaries, and it does not require expensive inter-visibility computation.

Once a set of matching candidates are identified, the next step in entity matching is to compare the source feature to be matched with its matching candidates to establish a one-to-one or one-to-many match. Hao *et al.* [18] presented a comparison method using similarity indexes based on all vertexes extracted from the matching candidates. Their method cannot handle one-to-many matches due to the difficulty in obtaining outline vertexes from the composite polygon of the matching candidates. Fang *et al.* [26] used mathematical morphology to describe and calculate similarities of individual buildings at the aspects of shape, construction, and interior extending direction. Their algorithm strengthened the identification ability of similarity index, while it was difficult to solve similarity calculation for the one-to-many case due to the difficulty in obtaining the outline of several polygons. Fan *et al.* [5] and Fu *et al.* [19] used a tangent-space-based shape similarity index. This similarity index depended on a predefined buffer to find pairs of matching vertexes for similarity calculation. The algorithm does not work if the predefined buffer fails to locate matching vertexes (e.g., when the positional discrepancies are significant). Therefore, it is difficult to be used for matching multi-scale polygonal residential areas with positional uncertainties. Tong *et al.* [12] determined the feature with maximum total probability as the corresponding feature. In fact, their judgment is not necessarily correct when spatial entities changed (e.g., fL1438 in Fig. 1(e) has the maximum probability of matching f_S65 than other features in the small-scale dataset, yet it is not the corresponding feature of f_S65).

Zhao [21] handled one-to-many cases using convex hulls to compare their similarity. This algorithm first divided the convex-hull into many sectors, then calculated and compared the similarities in the aspects of direction, distance, and area for the counterpart sectors. The algorithm is inefficient and time-consuming. Zhao *et al.* [25] proposed an algorithm for multi-scale polygonal feature matching based on geometry moments and overlay analysis. Their algorithm used overlapping area ratio for similarity calculation after the centroids of the source and target features coincided and selected the feature combination with maximum similarity value as the matching feature combination. Their method is time-consuming due to calculating candidates many times, and realistically the matching features with the greatest similarity value are not necessarily same entities, for instance, the maximum similarity value is 0.5, it means the entities could have changed, thus the related features cannot be confirmed as identical entities. Huang and Jiang [23] used the Voronoi diagram to address trans-scale residential area data matching. Their method used the overlapping area ratio between the target feature and its corresponding Voronoi polygon to determine whether they were same entities. However, the similarity index is so weak that it can result in mismatching.

In this paper, we strengthened the identification ability of similarities by incorporating the information including the shape of convex-hull, convex hull area, and overlapping area. Our method addresses the one-to-many correspondence using the combination strategy based on the generalized nearest distance.

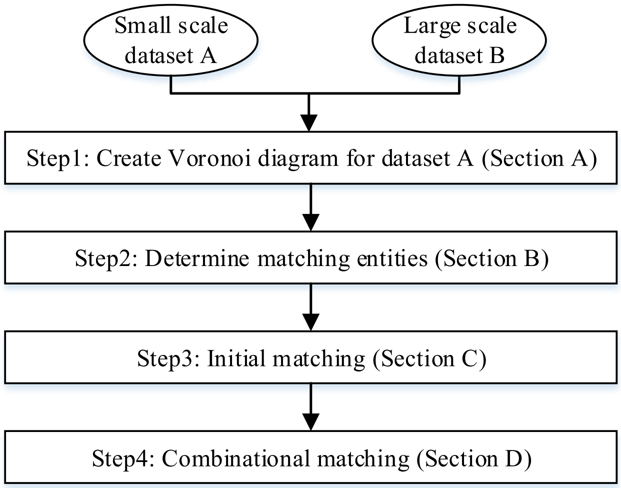


FIGURE 2. Matching workflow based on Voronoi diagram.

III. VORONOI DIAGRAMS BASED RESIDENTIAL AREAS MATCHING METHOD

Fig. 2 shows the RMMBV workflow for matching multi-scale polygonal residential areas. In Fig. 2, "dataset A" stands for the small-scale dataset or the source dataset, and "dataset B" represents the large-scale dataset or the target dataset. Herein the small-scale (large-scale) dataset is the dataset with a relatively smaller (larger) map scale of the two source datasets in the matching process. The overall matching process includes four stages explained in detail in the following subsections.

A. CREATE THE VORONOI DIAGRAM

RMMBV creates a Voronoi diagram from the small-scale dataset and then uses the Voronoi polygons of individual small-scale features to identify a set of matching candidates in the large-scale datasets. We propose an efficient algorithm for creating the Voronoi diagram from the small-scale residential area dataset. The idea is as follows.

Given the small scale dataset $A = \{A_1, A_2, A_3, \ldots, A_n\}$ where A_i is a feature and $i = 1, 2, \ldots, n$ (a total of n features). The geometry of A_i is a polygon $P_i = \{P_{i0}, P_{i1}, P_{i2}, P_{i3}, \ldots, P_{im}\}$ where P_{ij} is a vertex of P_i and $j = 1, 2, \ldots, m$ (a total of m vertexes). To create the Voronoi diagram, RMMBV first inserts evenly distributed points (interpolation points) on each of the polygon boundaries. RMMBV determines the number of interpolation points based on the map scales. This ensures RMMBV to insert only a small amount of points in the polygon boundary and improve the efficiency for computing the Voronoi diagram. If the map scale of dataset A is known, RMMBV inserts k (see the formula (1)) points in the boundary of Pi at an equal distance interval. In formula (1), *MapScale(A)* is the denominator value of map scale of dataset A. *MapScale(A)* is divided by 1,000 to represent the actual ground length corresponding to 1 mm (millimeter) in a paper map. The term 1/10 represents the minimal proportion of 1 mm distance

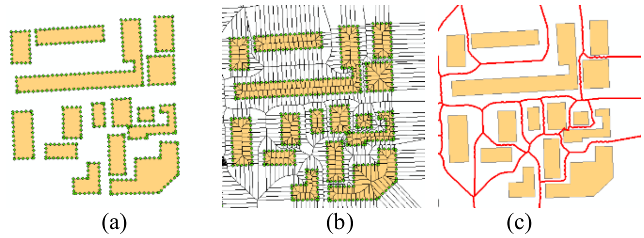


FIGURE 3. The process of creating Voronoi diagrams for polygonal residential areas: (a) Discrete points, (b) Voronoi polygons created by discrete points and (c) The merged Voronoi polygons.

on a paper map that human eyes can distinguish (namely human eye resolution) [27]. The term $\lambda(\leq \lambda \leq 10)$ is the distance tolerance coefficient. Considering data error and computation efficiency, here we use λ equal to 4. The function $Perimeter(P_i)$ represents the length of the perimeter of P_i. The function *Int []* is a function that returns the integer part of the input value. If the map scale of dataset A is unknown, RMMBV inserts a predefined k (we suggest $k \geq 4$) points at an equal distance interval on each side of P_i.

$$k = Int\left[\frac{Perimeter\ (P_i)}{\lambda * \frac{MapScale(\mathrm{A})}{1000} * \frac{1}{10}}\right] \tag{1}$$

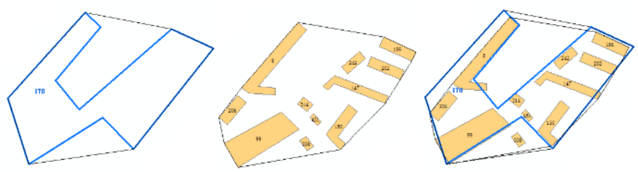
Finally, RMMBV uses the inserted points and the vertexes of every polygon (Fig. 3(a)) to construct the Voronoi diagram (Fig. 3(b). For each polygon P_i, RMMBV merges the Voronoi polygons overlapping P_i to obtain the final Voronoi polygons of the polygonal residential areas (Fig. 3(c)).

B. DETERMINE MATCHING ENTITIES

Once we have the Voronoi polygons for each feature in the small-scale datasets, RMMBV starts to find matching candidates in the large-scale datasets for evaluating possible matches by enumerating the compositions of the matching candidates. In this section, we describe three similarity indexes and their rules for evaluating a match.

The shape similarity index of convex hulls. The size, internal structure, and even the overall position of the features that represent the same ground object could be different greatly in two data sources, but the shape and size of their convex hulls maintain a relatively high similarity. Therefore, we use the shape similarity of the convex hulls as one of the RMMBV similarity indexes. Given a small-scale feature A_i and a set of composite features from the larger scale datasets. RMMBV establishes the convex hulls for A_i and the composite polygon of the target features CA_i, denoted by *ConvexHull* (A_i) and *ConvexHull*(CA_i), respectively. Figure 4 shows the convex hulls of the small-scale polygon and the large-scale combination polygon.

When the edges of the convex hull are arranged clockwise, the azimuth of each edge increases monotonically (starting from the edge with the smallest azimuth). This property guarantees that the shapes of the two convex hulls are similar if each edge (or several edges with a similar azimuth) can find



(a) (b) (c)

FIGURE 4. Convex hulls of polygon and group of polygons: (a) ConvexHull(A_i), (b) ConVexHull(CA_i) and (c) The overlay effect of convex hulls.

one or a set of corresponding edges with a similar azimuth and similar length. Our algorithm computes the shape similarity of convex hulls based on the azimuth and length of each edge of the two convex hulls. This approach does not require inserting many points nor needs a buffer to find initial point like the tangent-space-based algorithm (Fu and Shao 2010) and the shape-descriptor-based algorithm (Hao et al. 2008). RMMBV computes the shape similarity of convex hulls based on the above idea as follows.

First, for each edge azimuth E of $ConVexHull(A_i)$, RMMBV finds azimuth set E'_{Set} of $ConVexHull(CA_i)$ where the difference between the E and E' (E' indicates each azimuth in E'_{Set}) is smaller than an angle threshold T_{Angle}. The formula (2) shows the condition for matching the similar azimuths.

$$\left(\left|E-E'\right| \leq T_{Angle}\right)\ or\ \left(\left|360-E+E'\right| \leq T_{Angle}\right)\ or\ \left(\left|360+E-E'\right| \leq T_{Angle}\right) \tag{2}$$

where 360 is the maximum of azimuth. Using the 360 in formula (2) aims to deal with the case that E (E') is near 360 and $E'(E)$ is near 0. The matching result E'_{Set} of E may include zero to more azimuths. We build the matching pair relationship for E and E'_{Set}. Processing each azimuth by this way until azimuths of $ConVexHull\ (A_i)$ are exhausted. After that, our algorithm conducts backward matching for the azimuths that are unmatched in azimuths of $ConVexHull(CA_i)$ and builds the matching pair relationship for E_{set} and E'.

Second, RMMBV conflates the matching pairs if there is one same azimuth $E(/E')$ of different matching pairs in source azimuth data (/target azimuth data). Last the elements in the conflating results and other matching results to be separately written into array *Alist[]* (stores source azimuth data) and *Blist[]* (stores target azimuth data). Supposing the corresponding edge length sets of the source data and the target data of an azimuth matching pair are *Alen[i]* and *Blen[i]*, the formula of calculating the shape similarity of two convex hulls is (3), as shown at the bottom of this page. In formula (3), *simShape* represents the shape similarity index of convex hulls. *Max()* represents the maximum function and *Min()* represents the minimum function. *Perimeter()* represents the function of obtaining the length of the perimeter of a polygon. The parameter n represents the amount of the new matching pairs. The parameter p represents the amount of the elements in *Alen[i]*. The parameter q represents the amount of the elements in *Blen[i]*. A matched pair should have the shape similarity index of convex hulls larger than a user defined threshold.

The area similarity index of convex hull. Because the convex hulls of matching features can maintain high size similarity even when map scale is different, RMMBV uses the areas of convex hulls to design a similarity index for matching the same entities as follows. Assuming the two convex hulls are $ConvexHull(A_i)$ and $ConvexHull(CA_i)$, the calculation formula for the area similarity index is as follows (4), shown at the bottom of this page, where *simConvexHullArea* represents the area similarity index of the convex hulls, the *Min()* is the function of computing the minimum value, the *Max()* is the function of computing the maximum value, the *Area()* is a function of computing polygon area. A matched pair should have the convex hull area similarity larger than a user defined threshold.

Overlap area ratio of the target features. Matched features should have a high degree of overlapping in their positions after positional rectification. In matching multi-scale polygonal residential areas, the matching candidates should have a high degree of overlapping with the convex hull of the source feature after eliminating positional inconsistency of the convex hulls. Therefore, we design the overlap area ratio. Its calculation process as follows. First, computing the coordinate difference between two centroids of $ConvexHull\ (A_i)$ and $ConvexHull(CA_i)$. Then moving CA_i towards to centroid of A_i according to the coordinate difference to eliminate positional inconsistency. Here the translation result of CA_i is denoted as ${CA_i}^T$. Finally, the formula for calculating overlap area ratio as follows.

$$simOverlapArea = \frac{Area({CA_i}^T \cap ConvexHull(A_i))}{Area({CA_i}^T)} \tag{5}$$

A matched pair should have overlap area ratio similarity larger than a user defined threshold.

$$simShape = \frac{\sum_{i=0}^{n-1} Min\left(\sum_{j=0}^{p-1} Alen\,[i]\,[j]\,,\ \sum_{j=0}^{q-1} Blen[i]\,[j]\right)}{Max(Perimeter(ConVexHull\ (A_i)), Perimeter(ConVexHull(CA_i)))} \tag{3}$$

$$simConvexHullArea = \frac{Min(Area(ConvexHull(A_i)), Area(ConvexHull(CA_i)))}{Max(Area(ConvexHull(A_i)), Area(ConvexHull(CA_i)))} \tag{4}$$

FIGURE 5. The process of the initial matching.

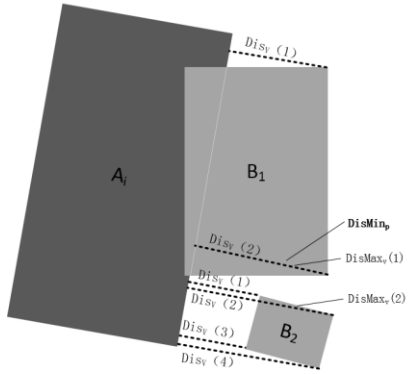


FIGURE 6. The generalized nearest distance.

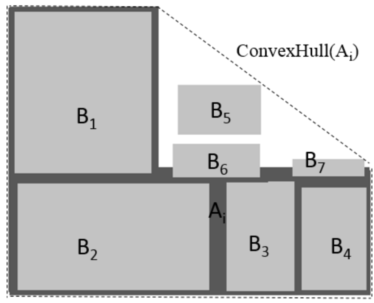


FIGURE 7. The diagram of matching result after eliminating the coordinate difference between centroids of two convex hulls.

C. INITIAL MATCHING

RMMBV conducts entity matching after generating Voronoi diagram for the small-scale dataset. To improve matching efficiency and quality, we divide entity matching into two stages called initial matching and combinational matching.

Assuming the features in dataset B that have at least 50% overlapping area with the Voronoi polygon V_i of a source feature A_i constitute a matching candidate set CA_i. Initial matching means matching the source feature A_i with CA_i' (CA_i' is the set of the features in CA_i that have at least 50% overlapping area with A_i) by calculating the similarity indexes (section II B). If the calculation results meet the similarity criteria, recording the matching result. Figure 5 shows the features in CA_i' (highlighted in blue).

D. COMBINATIONAL MATCHING

Combinational matching aims to find the feature in $CA_i''(CA_i'' = CA_i - CA_i')$ belongs to the matching result set of A_i so that RMMBV could improve matching precision. The process of combinational matching as follows: each time adding one feature B_i in CA_i'' to CA_i' according to the generalized nearest distance (the next paragraph demonstrates its definition by Figure 6). *Bi* is the feature with minimum generalized nearest distance. Then matching A_i with the new CA_i' and calculating their three similarity indexes. If the calculation results meet the similarity criteria, recording the matching result. Executing the above steps until all features in CA_i'' exhausted. Finally, RMMBV establishes the matching relationship of those features that meet similarity criteria in the last matching.

Figure 6 demonstrates the definition of the generalized nearest distance. The purpose of using the generalized nearest distance is to keep the area of the new convex hull as small as possible after combining one candidate feature so that maintaining the convex-hull similarity of the matching pair.

As shown in Figure 6, first, we suppose A_i is a feature of dataset A, $B_1, B_2, B_3, \ldots, B_n$ are the features of CA_i'' and suppose B_j has m vertexes. RMMBV uses the vertexes of B_j that are outside of A_i if B_j intersects with A_i to calculate the nearest distance $Dis_V(k)(k = 1, 2, 3, \ldots, m)$ from each vertex of B_j to A_i. Next, the maximum of the nearest distance of B_j and A_i can be calculated and denoted by $DisMax_V(j) = \mathrm{Max}(Dis_V(1), Dis_V(2), \ldots, Dis_V(m))$, where *Max()* is the maximum function. Finally, RMMBV can use formula (6) to calculate the generalized nearest distance.

$$DisMin_P = Min(DisMax_V(1), DisMaxV(2), DisMaxV(3), \ldots, DisMax_V(n)) \tag{6}$$

where $DisMin_P$ indicates the generalized nearest distance, *Min ()* is the minimum function.

After that, To guarantee the accuracy of matching result, RMMBV needs further to validate the correctness of the feature in matching result after eliminating the coordinate difference between centroids of the two matched convex hulls. As shown in Figure 7, in the matching result of RMMBV, A_i matches with B_1 trough B_7 because they meet the similarity criteria, visually A_i matches with B_1, B_2, B_3, B_4 and B_7 (or only B_1, B_2, B_3, B_4). To remove the features like B_5, B_6 and make sure B_7, we design the following strategy to address this problem.

RMMBV determines the feature belongs to the correct matching result according to the overlap area ratio of the target feature after eliminating the coordinate difference between centroids of the matched convex hulls. If the overlap area ratio of the target feature is low to a certain extent, it is unsuitable for a correct candidate. Considering the target feature with a smaller area size has less impact on the shape change of combination of candidates and vice versa, for example, B_7 has less impact on the shape change of combination of candidates than B_6. Based on our mapping experiences, we design three levels of thresholds for the overlap area ratio of the target feature based on area ratio k of the target feature area to the source feature area. If k of the target feature that interests with the source feature less than or equal to 0.2%, we confirm the target feature is a correct matching feature if its overlap area ratio is greater than zero (threshold 1). If k of the target feature that interests with the source feature greater than 0.2% and less than 5%, we confirm the target feature is a correct matching feature if its overlap area ratio is greater than 40% (threshold 2). If k of the target feature that interests with the source feature greater than or equal to 5%, we confirm the target feature is a correct matching feature if its overlap area ratio is greater than 80% (threshold 3). RMMBV removes the target feature that does not interests with the source feature.

IV. EXPERIMENT AND RESULT ANALYSIS

This section first introduces two test datasets of residential areas used in our experiment. Then we describe the evaluation metrics for assessing the performance of RMMBV for entity matching. Finally, we report and compare the entity matching results from RMMBV and both the MBR and buffer-based methods.

A. EXPERIMENTAL DATA

We tested two groups of polygonal residential areas of different scales to validate the feasibility and performance of RMMBV. Fig. 8 shows the first group of datasets, which include 1: 10,000 and 1: 50,000 polygonal residential areas (named 1Res1W and 1Res5W, respectively) that represent the same district of the Zhejiang Province, China (approximately 10.9 km^2). 1Res5W is the source data (small scale), which contains 174 features, and 1Res1W is the target data (large scale), which contains 543 features. We manually identified 102 matching pairs in this group of datasets as the ground truth. The second group of datasets (Fig. 9) include 1: 1,000 and 1: 10,000 polygonal residential areas (named 2Res1K and 2Res1W, respectively) covering a suburban district of Beijing, China (approximately 0.98 km^2). 2Res1W is the source data (small scale), which contains 199 features, and 2Res1K is the target data (larger scale), which contains 2,434 features. The number of manually identified matching pairs in this group of datasets was 151. In general, through observations of the two groups of datasets, the second group of datasets have a larger positional inconsistency and more aggressive generalization entity representations (in the small-scale dataset) than the first group.

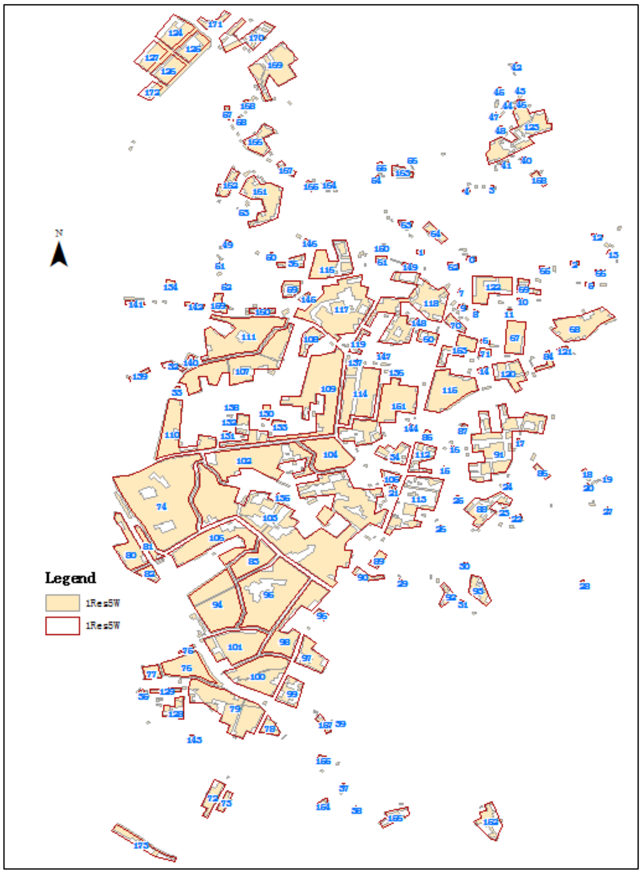


FIGURE 8. The experimental residential area data of map scale1:10,000 and 1: 50,000.

FIGURE 9. The experimental residential area data of map scale1:1,000 and 1: 10,000.

B. EXPERIMENTAL RESULT AND ANALYSIS

We implemented RMMBV in a program developed with the Microsoft Visual Studio .Net C# and Esri ArcGIS Engine 10.2. The experimental computer configuration environment is of Windows 7 64-bit operating system, Intel Core2 Duo CPU processor, and 4GB memory. In our

TABLE 1. Similarity criteria for determining the same entities.

Parameter set	Similarity criteria
1	TAngle=3，TShape =0.75，TArea =0.75，TOverlap=0.75
2	TAngle=3，TShape =0.75，TArea =0.6，TOverlap=0.75
3	TAngle=3，TShape =0.72，TArea =0.6，TOverlap=0.75
4	TAngle=3，TShape =0.7，TArea =0.6，TOverlap=0.75
5	TAngle=6，TShape =0.75，TArea =0.75，TOverlap=0.75
6	TAngle=6，TShape =0.75，TArea =0.6，TOverlap=0.75
7	TAngle=6，TShape =0.72，TArea =0.6，TOverlap=0.75
8	TAngle=6，TShape =0.7，TArea =0.6，TOverlap=0.75
9	TAngle=10，TShape =0.75，TArea =0.75，TOverlap=0.75
10	TAngle=10，TShape =0.75，TArea =0.6，TOverlap=0.75
11	TAngle=10，TShape =0.72，TArea =0.6，TOverlap=0.75
12	TAngle=10，TShape =0.7，TArea =0.6，TOverlap=0.75

experiments, we designed twelve sets of similarity criteria based on prior knowledge to test the robustness of RMMBV (Table 1). In Table 2, *TAngle* is the azimuth similarity threshold of convex hull edges, *TShape* is the shape similarity threshold of the convex hull, *TArea* is the area similarity threshold of the convex hull, and *TOverlap* is the threshold of overlap area ratio.

We used *Recall*, *Precision*, and *F-Measure* to assess the matching quality. The *F-Measure* is comprised of *recall* and *precision* for an integral evaluation for the matching quality:

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{7}$$

Where *Recall* equals to C/E, C represents the number of correct matching pairs, E represents the number of actual matching pairs (ground truth). And *Precision* equals to C/R, R represents the number of matching pairs in matching results.

We conducted experiments on data matching of our two groups of experimental data based on three matching methods: RMMBV presented in this paper, the MBR-based method, and the buffer-based method. For the buffer-based method, we determined the buffer radius as follows. According to the principle that the survey error is not beyond the triple value of standard deviation, here we replaced standard deviation with the ground distance that corresponds to the paper map distance 0.1 millimeters that human eye can distinguish. Therefore, the buffer radius of 1Res5W can be denoted by $0.1*3*50 = 15$, where 50 (meter) is ground distance represented by 1 millimeter on the paper map. Likewise, we calculated the buffer radius 3 meters for 2Res1W. Considering the great data differences of the second group of datasets, we also designed another buffer with a radius of 6 meters to investigate the matching quality. For both the MBR-based method and the buffer-based method, we used the MBR (used the spatial relation of ''intersects'') or the buffer (used the spatial relation of ''contains'') for searching candidates while still used our proposed similarity indexes to judge the same entities. The matching result of the first group of datasets shown in Table 2. Where PS value represents Parameter set shown in Table 1, T is time in seconds,

TABLE 2. The matching result of 1Res5W and 1Res1W.

PS	C	E	R	Recall	Precision	F-Measure	Method	T (s)
1	38	102	38	37.25%	100.00%	54.28%	RMMBV	4
	35	102	35	34.31%	100.00%	51.09%	MBR	4
	32	102	32	31.37%	100.00%	47.76%	Buffer	3
2	50	102	50	49.02%	100.00%	65.79%	RMMBV	5
	46	102	46	45.10%	100.00%	62.16%	MBR	4
	43	102	43	42.16%	100.00%	59.31%	Buffer	3
3	56	102	57	54.90%	98.25%	70.44%	RMMBV	4
	52	102	53	50.98%	98.11%	67.10%	MBR	6
	48	102	49	47.06%	97.96%	63.58%	Buffer	4
4	58	102	62	56.86%	93.55%	70.73%	RMMBV	5
	54	102	58	52.94%	93.10%	67.50%	MBR	5
	50	102	54	49.02%	92.59%	64.10%	Buffer	3
5	60	102	61	58.82%	98.36%	73.62%	RMMBV	4
	57	102	58	55.88%	98.28%	71.25%	MBR	5
	51	102	52	50%	98.08%	66.23%	Buffer	3
6	73	102	75	71.57%	97.33%	82.49%	RMMBV	4
	69	102	71	67.64%	97.18%	79.76%	MBR	5
	63	102	65	61.76%	96.92%	75.45%	Buffer	3
7	80	102	83	78.43%	96.39%	86.49%	RMMBV	6
	76	102	79	74.51%	96.20%	83.98%	MBR	5
	69	102	72	67.65%	95.83%	79.31%	Buffer	4
8	84	102	91	82.35%	92.31%	87.04%	RMMBV	5
	80	102	87	78.43%	91.95%	84.66%	MBR	5
	73	102	80	71.57%	91.25%	80.22%	Buffer	4
9	79	102	80	77.45%	98.75%	86.81%	RMMBV	4
	76	102	77	74.51%	98.70%	84.92%	MBR	4
	68	102	69	66.67%	98.55%	79.53%	Buffer	4
10	90	102	94	88.23%	95.74%	91.83%	RMMBV	5
	86	102	90	84.31%	95.56%	89.58%	MBR	5
	78	102	82	76.47%	95.12%	84.78%	Buffer	4
11	97	102	101	95.10%	96.04%	95.57%	RMMBV	5
	93	102	97	91.18%	95.88%	93.47%	MBR	5
	85	102	89	83.33%	95.51%	89.00%	Buffer	4
12	99	102	106	97.06%	93.40%	95.19%	RMMBV	5
	95	102	102	93.14%	93.14%	93.14%	MBR	5
	87	102	94	85.29%	92.55%	88.77%	Buffer	4

Fig. 10 shows the graphical representation of the Recall (Fig. 10(a)), Precision (Fig. 10(b)), and F-Measure (Fig. 10(c)) from the matching results at different similarity criteria. RMMBV, the MBR-based method and buffer-based method had the maximum F-Measure with the parameter set 11, respectively. All of the three methods achieved similar results (i.e., their recalls were 95.57%, 93.47%, and 89.00% respectively when the F-Measures were at the maximum), which was because the first test datasets had small differences in geometric position and morphological structure.

Table 3 shows the matching results of the second group of experimental data. Where Buffer (6m) represents using a buffer with a radius of 6 meters and Buffer (3m) represents using a buffer with a radius of 3 meters.

Figure 11 shows the graphical representation of the Recall (Fig. 11(a)), Precision (Fig. 11(b)), and F-Measure

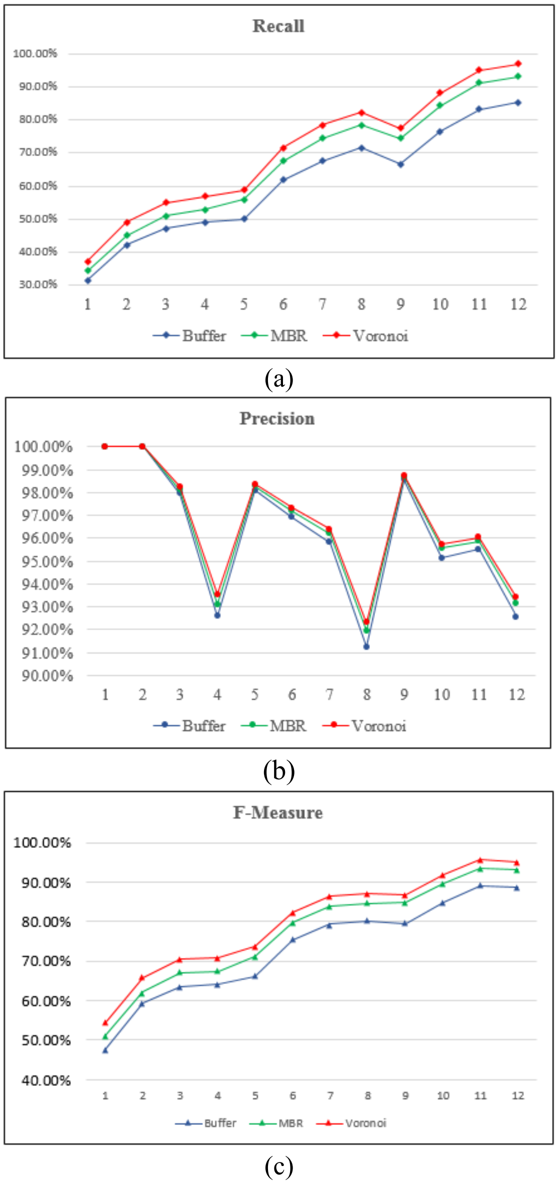


FIGURE 10. Matching results of 1Res5W and 1Res1W: (a) Comparison of recalls of three matching methods, (b) Comparison of precisions of three matching methods, (c) Comparison of F-measures of three matching methods.

(Fig. 11(c)) from the matching results at different similarity criteria. RMMBV had the maximum F-Measure with the sixth parameter set, the MBR-based method had the maximum F-Measure with the 11th parameter set, the 6 meters buffer-based had the maximum F-Measure with 7th parameter set, the 3 meters buffer-based had the maximum F-Measure with 4th parameter set respectively. Figure 10 shows RMMBV outperformed the other two methods when the best set of parameters were used for individual methods (i.e., the recalls were 74.17%, 68.21%, 51.66% (6 meters buffer) and 9.27% (3 meters buffer) when F-Measures were at the maximum). In addition, the average F-Measures of RMMBV were 12.46%, 20.8% and 64.45% higher than the MBR-based, 6 meters buffer-based, 3 meters buffer-based methods, respectively. This experiment also showed that the

TABLE 3. The matching result of 2Res1W and 2Res1K.

PS	C	E	R	Recall	Precision	F-Measure	Method	T (s)
1	89	151	101	58.94%	88.12%	70.63%	RMMBV	23
	70	151	82	46.36%	85.37%	60.09%	MBR	27
	58	151	73	38.41%	79.45%	51.79%	Buffer (6m)	12
	9	151	18	5.96%	50.00%	10.65%	Buffer (3m)	6
2	96	151	107	63.58%	89.72%	74.42%	RMMBV	20
	75	151	88	49.67%	85.23%	62.76%	MBR	25
	63	151	78	41.72%	80.77%	55.02%	Buffer (6m)	15
	10	151	21	6.62%	47.62%	11.63%	Buffer (3m)	6
3	99	151	110	65.56%	90.00%	75.86%	RMMBV	20
	75	151	93	49.67%	80.65%	61.48%	MBR	25
	66	151	81	43.71%	81.48%	56.90%	Buffer (6m)	17
	12	151	22	7.95%	54.55%	13.87%	Buffer (3m)	6
4	100	151	111	66.23%	90.09%	76.34%	RMMBV	27
	77	151	93	50.99%	82.80%	63.11%	MBR	28
	65	151	83	43.05%	78.31%	55.56%	Buffer (6m)	17
	14	151	26	9.27%	53.85%	15.82%	Buffer (3m)	6
5	107	151	121	70.86%	88.43%	78.68%	RMMBV	28
	85	151	104	56.29%	81.73%	66.67%	MBR	33
	67	151	92	44.37%	72.83%	55.14%	Buffer (6m)	19
	9	151	29	5.96%	31.03%	10.00%	Buffer (3m)	6
6	112	151	129	74.17%	86.82%	80.00%	RMMBV	22
	90	151	111	59.60%	81.08%	68.70%	MBR	30
	74	151	99	49.01%	74.75%	59.20%	Buffer (6m)	17
	13	151	33	8.61%	39.39%	14.13%	Buffer (3m)	6
7	113	151	133	74.83%	84.96%	79.58%	RMMBV	28
	88	151	116	58.28%	75.86%	65.92%	MBR	34
	76	151	102	50.33%	74.51%	60.08%	Buffer (6m)	19
	12	151	37	7.95%	32.43%	12.77%	Buffer (3m)	6
8	114	151	136	75.50%	83.82%	79.44%	RMMBV	21
	88	151	121	58.28%	72.73%	64.71%	MBR	27
	73	151	106	48.34%	68.87%	56.81%	Buffer (6m)	18
	14	151	41	9.27%	34.15%	14.58%	Buffer (3m)	6
9	108	151	131	71.52%	82.44%	76.60%	RMMBV	19
	82	151	116	54.30%	70.69%	61.42%	MBR	20
	70	151	100	46.36%	70.00%	55.78%	Buffer (6m)	12
	9	151	35	5.96%	25.71%	9.68%	Buffer (3m)	6
10	114	151	139	75.50%	82.01%	78.62%	RMMBV	20
	88	151	124	58.28%	70.97%	64.00%	MBR	20
	75	151	107	49.67%	70.09%	58.14%	Buffer (6m)	13
	13	151	40	8.61%	32.50%	13.61%	Buffer (3m)	6
11	116	151	142	76.82%	81.69%	79.18%	RMMBV	19
	103	151	127	68.21%	81.10%	74.10%	MBR	20
	73	151	109	48.34%	66.97%	56.15%	Buffer (6m)	13
	13	151	42	8.61%	30.95%	13.47%	Buffer (3m)	6
12	116	151	143	76.82%	81.12%	78.91%	RMMBV	21
	92	151	129	60.93%	71.32%	65.71%	MBR	20
	74	151	111	49.01%	66.67%	56.49%	Buffer (6m)	12
	15	151	48	9.93%	31.25%	15.08%	Buffer (3m)	6

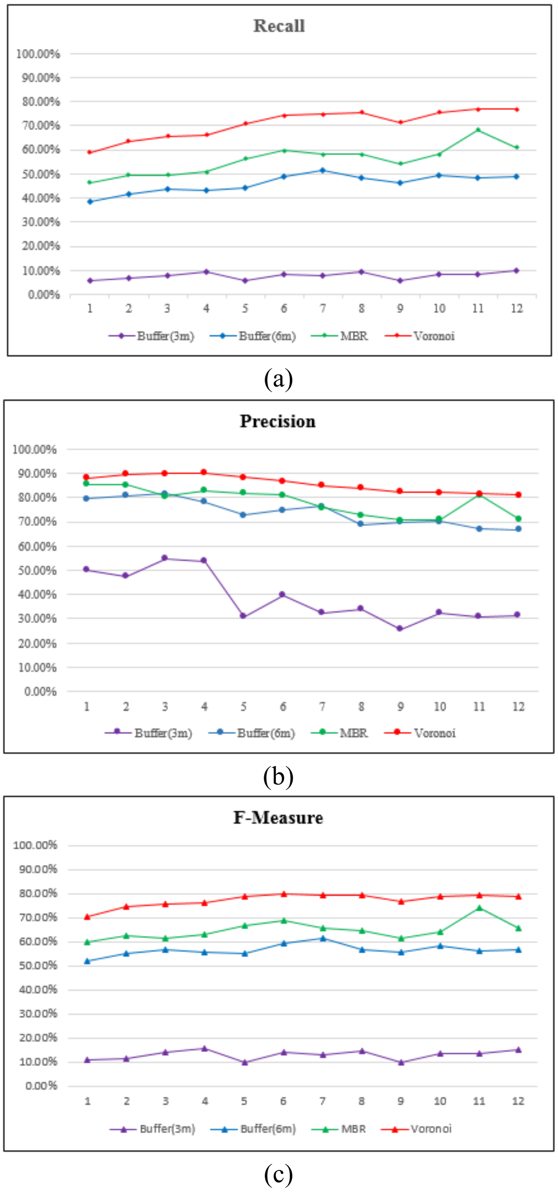


FIGURE 11. Matching results of 2Res1W and 2Res1W: (a) Comparison of recalls of different matching methods, (b) Comparison of precisions of different matching methods, (c) Comparison of F-measures of different matching methods.

matching result can be affected greatly by the buffer size when adopting the buffer-based method.

In summary, RMMBV outperformed the MBR and buffer-based methods. The experiment shows that RMMBV has a great advantage over the other two methods when the position difference was large (the second group of test datasets). The reason is that when there were significant differences in the spatial positions and (matching) feature numbers, the MBR-based and buffer-based method could cause missed matches while it has less impact on RMMBV. Regarding the matching time, the MBR-based method required more computation time because the larger number of matching candidates. For instance, for matching the second group of data using the 12 sets of similarity criteria, the average processing time

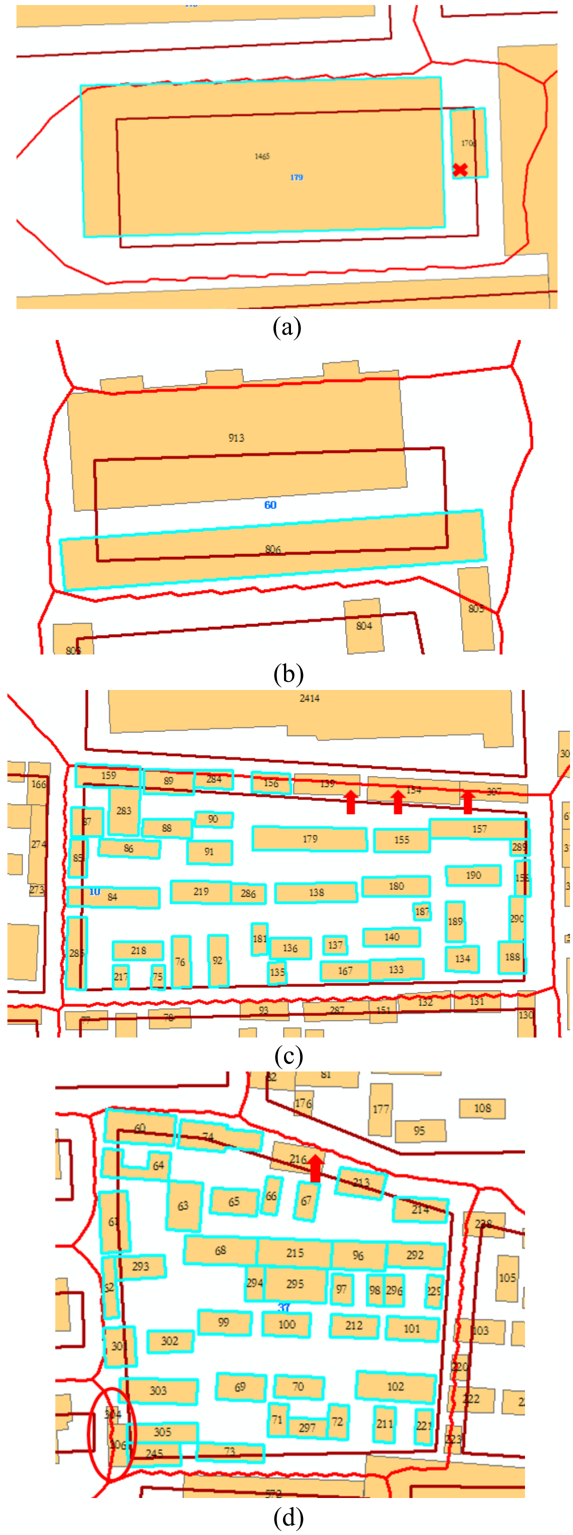


FIGURE 12. Cases of matching error: (a) f_S179 wrongly matched fL1706, (b) f_S60 wrongly matched the f_L806, (c) Three target features were missed and (d) One feature was missed.

of the RMMBV, MBR-based, 6 meters buffer-based and 3 meters buffer-based methods were 22.58, 25.75, 15.33, and 6 seconds. The buffer-based method required less time than the other two methods because the buffer-based method only used a smaller of matching candidates (which lead to missed matches).

The errors in the matching results of RMMBV were mainly from two categories. The first type of error was caused by the change of geographic objects. Fig. 12 (a) and Fig. 12 (b) show an example in which entities in the same place from different map scales have changed. When RMMBV adopted the 11th parameter set, the initial matching algorithm of RMMBV incorrectly matched f_S179 with f_L1706 because f_S1706 was a new geographic object near f_S1465. For the same reason, f_S60 was incorrectly matched f_L806 in the initial matching because they met the thresholds of the 11th parameter set in the initial matching. The second type of error was caused by potential mapping errors or inconsistent data quality. As shown in Fig. 12 (c), in the ground truth, f_S10 should match with all the highlighted large-scale features and the features indicated by red arrow. Because the size difference between f_S10 and the combination of the target features was large (probably from mapping errors), f_S10 missed to match three features using RMMBV (the features indicated by red arrow). In Fig. 12 (d), compared to the manual matching result, f_S37 missed to match f_L216. Without comparing feature attributes, it was not clear that the features in the red circle were new features due to the change of geographic objects or missed matches.

The matching quality depended on the selection of a good set of similarity criteria. In practice, the user should test the matching framework on a small sample set and compare the matching results to the ground truth to determine the best parameters. Through testing RMMBV using the two groups of experimental datasets with different map scales and positional differences, we found the matching results were relatively good when using the 11th set of similarity criteria. In the case of not knowing the ground truth of sample data, we recommend the users to use the parameter set like the 11th set of similarity criteria for polygonal residential area matching.

V. CONCLUSION AND FUTURE WORK

This paper presented a generic entity-matching framework for multi-scale polygonal residential areas. The matching framework is based on the Voronoi diagram and a novel combination matching strategy with similarity calculation models. In comparison to the traditional MBR-based and buffer-based methods, our method can improve the matching recall and precision, especially for multi-scale datasets with inconsistent positional deviations from different sources. The algorithm we designed can apply to matching multi-scale (or trans-scale) polygonal residential area datasets, which improves the generality of the existing matching methods.

We plan to test our approach on more varieties of polygonal residential area datasets with various map scales, improve the computational efficiency of our algorithm, and further explore the possibility of using the Voronoi diagram in multi-scale linear road matching.

REFERENCES

[1] M. F. Goodchild, "Attribute accuracy," in *Elements of Spatial Data Quality*, S. C. Guptill and J. L. Morrison, Eds. New York, NY, USA: Elsevier, 1995, pp. 59–80.

[2] A. Samal, S. Seth, and K. Cueto, "A feature-based approach to conflation of geospatial sources," *Int. J. Geogr. Inf. Sci.*, vol. 18, no. 5, pp. 459–489, 2004.

[3] D.-R. Li, J.-Y. Gong, and Q.-P. Zhang, "On the conflation of geographic databases," *Sci. Surv. Mapping*, vol. 29, no. 1, pp. 1–4, 2004.

[4] B. S. Yang, Y. F. Zhang, and X. C. Luan, "A probabilistic relaxation approach for matching road networks," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 2, pp. 319–338, 2013.

[5] H. Fan, A. Zipf, Q. Fu, and P. Neis, "Quality assessment for building footprints data on OpenStreetMap," *Int. J. Geogr. Inf. Sci.*, vol. 8, no. 4, pp. 700–719, 2014.

[6] A. Saalfeld, "Conflation automated map compilation," *Int. J. Geogr. Inf. Syst.*, vol. 2, no. 3, pp. 217–228, 1988.

[7] M. A. Cobb *et al.*, "A rule-based approach for the conflation of attributed vector data," *GeoInformatica*, vol. 2, no. 1, pp. 7–35, 1998.

[8] Q. Zhang, D. Li, and J. Gong, "Shape similarity measures of linear entities," *Geo-Spatial Inf. Sci.*, vol. 5, no. 2, pp. 62–67, 2002.

[9] C. Beeri, Y. Kanza, E. Safra, and Y. Sagiv, "Object fusion in geographic information systems," in *Proc. 30th VLDB Conf.*, Toronto, ON, Canada, 2004, pp. 816–827.

[10] M. Zhang, W. Shi, and L. Meng, "A generic matching algorithm for line networks of different resolutions," in *Proc. 8th ICA Workgroup Gen. Multiple Represent.*, Coruña, Spain, 2005, pp. 1–8.

[11] V. Walter and D. Fritsch, "Matching spatial data sets: A statistical approach," *Int. J. Geogr. Inf. Sci.*, vol. 13, no. 5, pp. 445–473, 1999.

[12] X. Tong, S. Deng, and W. Shi, "A probability-based multi-measure feature matching method in map conflation," *Int. J. Remote Sens.*, vol. 30, no. 20, pp. 5453–5472, 2009.

[13] J. O. Kim, K. Yu, J. Heo, and W. H. Lee, "A new method for matching objects in two different geospatial datasets based on the geographic context," *Comput. Geosci.*, vol. 36, no. 9, pp. 1115–1122, 2010.

[14] B. Zhao, M. Deng, Z. Xu, and H. Liu, "Development of general rules for matching multi-scale area objects," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 36, no. 8, pp. 991–994, 2011.

[15] X. Tong, D. Liang, and Y. Jin, "A linear road object matching method for conflation based on optimization and logistic regression," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 4, pp. 824–846, 2014.

[16] L. Li and M. F. Goodchild, "An optimisation model for linear feature matching in geographical data conflation," *Int. J. Image Data Fusion*, vol. 2, no. 4, pp. 309–328, 2011.

[17] E. A. Wentz, "A shape definition for geographic applications based on edge, elongation, and perforation," *Geogr. Anal.*, vol. 32, no. 1, pp. 95–112, 2000.

[18] Y. Hao, W. Tang, Y. Zhao, and N. Li, "Areal feature matching algorithm based on spatial similarity," *Acta Geodaetica Cartogr. Sinica*, vol. 37, no. 2, pp. 204–209, 2008.

[19] Z.-L. Fu, S.-W. Shao, and C.-Y. Tong, "Multi-scale area entity shape matching based on tangent space," *Comput. Eng.*, vol. 36, no. 17, pp. 216–218, 2010.

[20] Z. L. Fu and J. H. Wu, "Entity matching in vector data," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 37, no. B4, pp. 1467–1472, 2008.

[21] B. B. Zhao, "A study on multi-scale vector map objects matching method and its application," Ph.D. dissertation, Central South Univ., Hunan, China, 2011.

[22] J. F. Luo, "Automatic matching of multi-scale polygon features constrained by road network," *Appl. Res. Comput.*, vol. 31, no. 11, pp. 3247–3249, 2014.

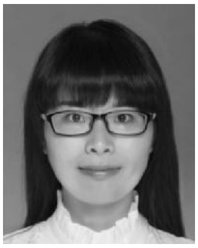
[23] W. Huang and J. Jiang, "Simple geometry matching of multi-scales spatial data," *Remote Sens. Inf.*, no. 1, pp. 27–31, 2011.

[24] H. W. Yan and J. Y. Wang, "Map group object description and automatic generalization," Beijing, China: Science Press, 2009.

[25] D. B. Zhao, Y. H. Sheng, and K. Zhang, "An algorithm for multi-scale one-to-many areal feature matching based on geometry moments and overly analysis," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 36, no. 11, pp. 1371–1375, 2011.

[26] A.-L. Fang, H.-W. Yan, and L.-P. Zhang, "Description and calculation of similarity degrees between individual buildings in multi-scale map space," *Sci. Surv. Mapping*, vol. 37, no. 1, pp. 98–100, 2012.
[27] Z. F. Pan, *Principles and Methods of Digital Mapping*. Wuhan, China: Wuhan Univ. Press, 2004.

JIANHUA WU received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University in 2008. He was a Visiting Scholar with the Spatial Sciences Institute, University of Southern California from 2015 to 2016. He is currently an Associate Professor with the School of Geography and Environment, Jiangxi Normal University. His current research interests include intelligent perception and service of geospatial information, geospatial data matching and integration, geographic information systems, and graphics recognition. He ever participated in and hosted numerous GIS projects (two projects granted by the National Natural Science Foundation of China).



YANGYANG WAN received the master's degree in cartography and geographical information system from the School of Geography and Environment, Jiangxi Normal University, in 2016. Her current research interests include geospatial data matching and integration and geographic information engineering.

YAO-YI CHIANG received the Ph.D. degree in computer science from the University of Southern California in 2010. He is currently an Associate Professor with the Spatial Sciences Institute, University of Southern California. He develops computer algorithms and applications that discover, collect, fuse, and analyze data from heterogeneous sources to solve real-world problems. He is also an expert on digital map processing and geospatial information system. He has authored a number of articles on automatic techniques for geospatial data extraction and integration. His general area of research is artificial intelligence and data science, with a focus on information integration and spatial data analytics.

ZHONGLIANG FU received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University in 1996. He is currently a Professor and the Doctoral Supervisor with the School of Remote Sensing and Information Engineering, Wuhan University. His current research interests include geospatial data management and update, geospatial data matching and integration, remote sensing image processing, and geographic big data mining.

MIN DENG received the Ph.D. degrees from Wuhan University in 2003 and the Asian Institute of Technology in 2004. He is currently a Professor and the Doctoral Supervisor with the School of Geosciences and Info-Physics, Central South University. His current research interests include geospatial data update, spatio-temporal data mining, and spatio-temporal analysis and modeling.