

Article

Multiscale Spatial Polygonal Object Granularity Factor Matching Method Based on BPNN

Daoye Zhu ¹ , Chengqi Cheng ^{1,2}, Weixin Zhai ³, Yihang Li ⁴, Shizhong Li ⁵ and Bo Chen ^{6,*} 

¹ Center for Data Science, Peking University, Beijing 100871, China; zhudaoye@pku.edu.cn (D.Z.); ccq@pku.edu.cn (C.C.)

² College of Engineering, Peking University, Beijing 100871, China

³ College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China; zhaiweixin@cau.edu.cn

⁴ College of Urban and Environmental Sciences, Peking University, Beijing 100871, China; got@pku.edu.cn

⁵ TH Center of China, Beijing 100094, China; 1401110501@pku.edu.cn

⁶ Institute of Space Science and Applied Technology, Harbin Institute of Technology, Shenzhen 518055, China

* Correspondence: pkuchenbo@pku.edu.cn

Abstract: Spatial object matching is one of the fundamental technologies used for updating and merging spatial data. This study focused mainly on the matching optimization of multiscale spatial polygonal objects. We proposed a granularity factor evaluation index that was developed to promote the recognition ability of complex matches in multiscale spatial polygonal object matching. Moreover, we designed the granularity factor matching model based on a backpropagation neural network (BPNN) and designed a multistage matching workflow. Our approach was validated experimentally using two topographical datasets at two different scales: 1:2000 and 1:10,000. Our results indicate that the granularity factor is effective both in improving the matching score of complex matching and reducing the occurrence of missing matching, and our matching model is suitable for multiscale spatial polygonal object matching, with a high precision and recall reach of 97.2% and 90.6%.

Keywords: multi-scale; spatial polygonal object; match; granularity factor; BPNN



Citation: Zhu, D.; Cheng, C.; Zhai, W.; Li, Y.; Li, S.; Chen, B. Multiscale Spatial Polygonal Object Granularity Factor Matching Method Based on BPNN. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 75. <https://doi.org/10.3390/ijgi10020075>

Academic Editor: Wolfgang Kainz

Received: 19 November 2020

Accepted: 9 February 2021

Published: 13 February 2021

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



Copyright: © 2021 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 one of the most challenging and indispensable procedures in the processing and application of spatial data [1]. It is defined as the process of identifying the same real-world objects from different sources and establishing a corresponding relationship [2]. Spatial object matching has been applied widely in the updating, maintenance, and fusion of spatial data [3–5].

In recent years, following the rapid increase of volunteered geographic information, spatial object matching has received increasing attention regarding the improvement of the quality and evaluation of volunteered geographic information data [6,7]. However, geospatial data from different sources are often inconsistent in actuality, scale, geometry, and semantics [8–11]. These discrepancies can cause complications and low efficiency in spatial data matching, especially for data with different scales, because complex matching (1:N and M:N matching) is widespread in multiscale matching [12] and the characteristics between matching data are more blurred [13].

Traditional matching methods are based mainly on using the similarities of the geometric, attribute, and topological information of objects to obtain matching results [14–16]. In spatial polygonal object matching, the geometric measure is one of the most widely used information elements. Shape, length, size, and direction have all previously been used for the measurement of similarity [17–20]. Deng et al. [21] introduced the Hausdorff distance as a metric indicator for different types of spatial objects, which was then improved by Tong et al. [22] for two polylines in matching road networks. The ratio of

the area of overlap is considered an effective metric with which to measure polygonal object similarity [23,24] and it can be used to identify corresponding polygonal objects that are associated through shape, size, and direction [25–27]. Multifeature combination matching has received increasing attention because of its reliable matching performance [3]. However, despite its acceptable performance in terms of matching objects of the same scale, its performance is unsatisfactory in multiscale matching. This is because traditional matching methods have satisfactory efficiency and accuracy only in simple matching (1:1 matching) [28–30]. Therefore, Tong et al. [22] obtained complex matches via multiple iteration detection. Wang et al. [20] applied a traditional two-way matching strategy to obtain candidate matching pairs, but when the deviation of the positions of objects was large, many complex matches were missed. Moreover, a two-way matching strategy can tend to neglect smaller objects [31].

Because of the excellent performance of machine learning algorithms in classification problems, many spatial object matching models introduce machine learning algorithms to avoid the need of weighing feature similarity and select matching thresholds. Zhang et al. [18] proposed a one-to-one matching approach that aimed at the maintenance of a multiple representation database and tested the approach with four machine learning methods in building polygonal object matching: C4.5 algorithms proposed by Quinlan [32], the classification and regression tree proposed by Breiman et al. [33], the Naïve Bayes classifier proposed by Rish [34], and support vector machines proposed by Hsu and Lin [35]. McKenzie et al. [36] solved the weight assignment problem of multiple attributes in points of interest (POI) based on the regression model, and arrived at a weighted combination of all matchers; they also evaluated the results against an ordinal weight combination and an unweighted baseline. Ruiz-Lendínez et al. [37] proposed a methodology for matching the bidimensional objects of both area and point features extracted from geographical databases, and polygon-to-polygon matching is obtained by means of a genetic algorithm that allows the classification of area features from two geographical databases. Tong et al. [22] proposed an improved linear object matching approach, which combines the optimization model and logistic regression model to obtain a better matching result by detecting incorrect matches and missed matches that are included in the result obtained from using the optimization method for object matching in conflation. BPNN accumulates knowledge in weighted connections to mimic the function of a human brain [38]. Wang et al. [20] proposed an approach for multi-represented feature matching based on spatial similarity and BPNN. Although machine learning algorithms have a high level of automation and strong data adaptability, their ability to identify complex matching types is relatively weak, and it requires a machine learning feature factor to deal with complex matching.

To address the aforementioned problems, we propose a multiscale spatial polygonal object granularity factor matching method based on a backpropagation neural network (BPNN). We constructed a granularity factor evaluation index (GFEI) to promote the recognition ability of complex matches in multiscale spatial polygonal object matching (MSPOM), and we used BPNN to improve further the accuracy of MSPOM by our designed multistage matching workflow. This article is organized as follows. The proposed method is presented in detail in Section 2. The experimental design and the experimental results of the approach are presented in Section 3. Finally, Section 4 draws conclusions based on the findings of the research.

2. Methodology

2.1. MSPOM Similarity Evaluation Indices Combined with Minimum Bounding Rectangle (MBR)

To describe the geometric characteristics of MSPOM, a series of commonly used methods for evaluating similarity has been established through analyzing the spatial difference of MSPOM in terms of their geometric metrics such as distance, size, direction, and shape [16,20]. The minimum bounding rectangle (MBR) is generally used to represent the source objects and the target objects in MSPOM. For source objects

$a = \{a_1, a_2, \dots, a_n | n \geq 1\}$, and target objects $b = \{b_1, b_2, \dots, b_m | m \geq 1\}$ from a candidate matching pair, we combine geometric metrics with MBR. The MBR is the smallest rectangle parallel to the coordinate axes, which contains objects. It is simple, efficient, and is the most widely used approximate expression of geographical objects [39]. In our study, an MBR contains a set of spatial polygonal objects that contains one or more spatial polygonal objects. The maximum and minimum coordinates $\{c | x_{max}, x_{min}, y_{max}, y_{min}\}$ in the X and Y directions are calculated for the space coordinates of the vertices of all polygonal objects in this set, and the rectangle surrounded by the lines $\{L = c | c = x_{max}, x_{min}, y_{max}, y_{min}\}$ corresponding to c is used as MBR.

2.1.1. Distance Similarity

When describing the relative distance between a and b , we use the distance between the MBR centers of a and b , as shown in Figure 1. The smaller the relative distance between a and b , the higher their distance similarity. The formula for calculating the distance similarity between a and b is as follows:

$$S_{dis} = \frac{d_a + d_b}{d_a + d_b + d(P_a P_b)} \quad (1)$$

where P_a is the MBR center of a , P_b is the MBR center of b , $d(P_a P_b)$ represents the distance between the MBR centers of a and b , d_a and d_b represent half the diagonal lengths of the MBRs of a and b , respectively, and S_{dis} represents the distance similarity of a and b .

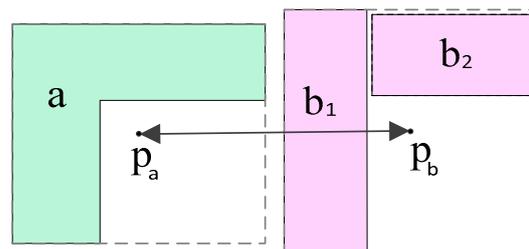


Figure 1. Distance between MBR centers of a and b .

2.1.2. Overlap Rate of Area

First, the MBR centers of a and b are obtained (i.e., P_a and P_b , respectively). Then, P_a and P_b are coincided. Finally, the shape and size similarities between a and b are reflected by the overlap area after the translation. As shown in Figure 2, the shaded part is the overlap area. In the figure, because b_3 does not overlap with a after the translation, b_3 is excluded in the calculation; thus, the final corresponding relationship is $(a : b_1, b_2)$. The formula for calculating the overlap rate of area of a and b is as follows:

$$S_{olp} = \frac{Area_{a \cap b}}{\sqrt{Area_a \times Area_b}} \quad (2)$$

where $Area_{a \cap b}$ denotes the overlapping area after the alignment of the MBR centers of a and b , $Area_a$ denotes the total area of the objects overlapping with b in a after the translation ($Area_b$ is the same as $Area_a$), and S_{olp} denotes the overlap rate of area of a and b .

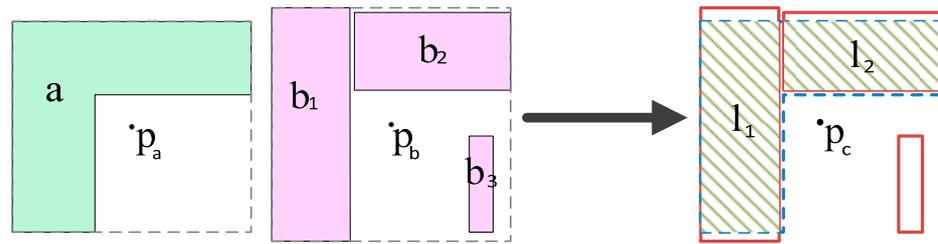


Figure 2. Overlapping area of a and b .

2.1.3. Direction Similarity

Directions of a and b are replaced by the diagonal directions of the MBRs. As shown in Figure 3, L_1 and L_2 are the directions of a and b , respectively.

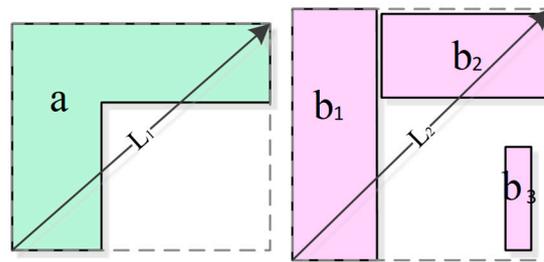


Figure 3. Directions of a and b .

Therefore, the direction similarity formula of a and b is as follows:

$$S_{dir} = 1 - \frac{|\theta(a) - \theta(b)|}{\theta_\tau} \quad (3)$$

where $\theta(a)$ and $\theta(b)$ are the directions of the polygonal objects a and b , $\theta(a) \in [0, \frac{\pi}{2}]$, $\theta(b) \in [0, \frac{\pi}{2}]$, and θ_τ is the direction threshold (generally, $\theta_\tau = \frac{\pi}{2}$).

2.1.4. Shape Similarity

The shape similarity of a and b is measured by the relationship between the lengths and widths of the MBRs of a and b . The closer a 's length and width are to b 's, the more similar a is to b . The MBR shape similarity formula of a and b is as follows:

$$S_{shape} = \frac{\min(H_{MBR(a)}, H_{MBR(b)})}{\max(H_{MBR(a)}, H_{MBR(b)})} \times \frac{\min(W_{MBR(a)}, W_{MBR(b)})}{\max(W_{MBR(a)}, W_{MBR(b)})} \quad (4)$$

where $H_{MBR(a)}$ and $H_{MBR(b)}$ represent the lengths of the MBRs of a and b , respectively, $W_{MBR(a)}$ and $W_{MBR(b)}$ represent the widths of the MBRs of a and b , respectively, and S_{shape} represents the MBR shape similarity of a and b .

2.2. Granularity Factor Evaluation Index (GFEI)

In the field of geographic information, granularity refers to the length, area, or volume of the smallest recognizable spatial object, namely, resolution [40,41]. The finer the granularity, the more accurate is the spatial attribute information expression. In contrast, the coarser the granularity, the coarser is the spatial attribute information expression [42]. In this study, we use the difference between the number of source objects and the number of target objects of the spatial polygonal object matching pair in spatial data as the granularity feature scale, so as to introduce the granularity factor. In geographic phenomena, a large proportion of statistical data follows a heavy-tailed distribution. The heavy-tailed distribution refers to the serious right deviation of the data distribution, i.e., a few high-value datasets are in

the “head” of the distribution, while most low-value datasets are in the “tail.” In MSPOM, there are mainly four matching types: 1:1, 1:N, M:N, and 1:0. Among them, 1:1 and 1:N matchings account for the majority, while M:N and 1:0 matching account for only a small proportion. Faced with this heavy-tailed distribution, it is usual for data distributions to be described by an exponential function, logarithmic normal distribution, or power law.

In this study, we constructed for the first time a unified evaluation index (i.e., GFEI) to measure complex matching and simple matching, which combines the distribution of matching types of MSPOM with the “head–tail segmentation function” [43]. This index is more advantageous and natural than the natural interval segmentation method.

The value of GFEI is between 0 and 1. When there are many m:n matchings ($m > 0$, $n > 0$), the mean value and the standard deviation of the value m/n are brought into the “head and tail segmentation function” and the value of GFEI is the normalized function value. GFEI is used to describe the matching scale, which refers to the ratio of the target object(s) and the source object(s) in a matching pair, which reflects the scale difference of two matching object sets in a matching pair. The larger the matching scale, the higher the value of GFEI of MSPOM, and vice versa. The formula of GFEI used in MSPOM can be expressed as follows:

$$GFEI_i = \frac{1}{\sqrt{2\pi S_{psd}^2}} \exp\left(\frac{-(\ln(R_{psdi}) - M_{psd})^2}{2S_{psd}^2}\right) \quad (5)$$

where R_{psdi} is the ratio of the fragmentation degree, M_{psd} is the average of R_{psdi} , and S_{psd} is the standard deviation of R_{psdi} . We use the following concrete formulas P to calculate R_{psdi} , M_{psd} , and S_{psd} :

$$P : \begin{cases} R_{psdi} = m_i / n_i \\ M_{psd} = \frac{1}{n} \sum_{i=1}^n R_{psdi} \\ S_{psd} = \frac{1}{n} \sum_{i=1}^n \sqrt{(R_{psdi} - M_{psd})^2} \end{cases} \quad (6)$$

where n is the number of R_{psdi} , i.e., the number of matching pairs.

When using GFEI to optimize the geometric similarity evaluation indices (Section 2.1), it is necessary to normalize GFEI. The normalization method used here can be defined as follows:

$$N_{GFEI} = 1 - \frac{GFEI_i - \min(GFEI_i)}{\max(GFEI_i) - \min(GFEI_i)} \times (1 - \varepsilon) \quad (7)$$

where ε is the threshold for determining correct matching pairs.

For M: N matching and partial 1: N matching, the low values of evaluation indices such as Overlap Rate of Area (Section 2.1.2) led to a low score of matching, which is lower than the evaluation threshold of correct matching, and therefore it is misjudged as incorrect matching. The head–tail segmentation classification further groups or stratifies the matching types in the heavy-tail distribution. First, the values of N_{GFEI} are calculated, which divide all candidate matching pairs into two parts. It then continues to divide the head values higher than the average until the head values are no longer in the heavy-tail distribution. After the weighted processing of N_{GFEI} , the scores of M: N matching and partial 1: N matching in Table 1 are improved, as shown in Table 2, such as $(a_{10}, a_{11} : b_1, b_2, b_3, b_5)$ and $(a_5 : b_{15}, b_{17}, b_{18}, b_{19}, b_{20})$. Thus, the score of the entire candidate matching pairs is no longer a heavy-tail distribution. Finally, the number and interval of the matching types are determined naturally. The correct matching candidate pairs will be identified for the corresponding matching types; for example, different types of matching pairs in Table 2 are identified correctly. N_{GFEI} produced using this procedure is one of the characteristics of training samples of “BPM training” in the data-processing workflow.

2.3. Matching Model Based on BPNN (BPM)

An artificial neural network (ANN) is an operation model consisting of a large number of interconnected neurons. Each neuron is a set output function. The connection between neurons is based on the weight of two neurons. ANNs mainly focus on neuronal characteristics, neural network topology, and learning rules. ANNs have reasonable intelligent characteristics, associative storage, self-learning, and fast search capabilities for determining optimal solutions. ANNs have been used successfully to solve many practical problems in various fields such as pattern recognition, automatic control, intelligent robots, and prediction. ANNs are used widely in scientific and engineering problems and attempts to simulate the recognition pattern of biological nervous systems [44].

In MSPOM, identical spatial polygonal objects in different maps have similarities in many feature descriptions. Therefore, the problem of MSPOM can be regarded as a problem of pattern recognition by an ANN. BPNN is a feed-forward learning model of backpropagation [45] and can achieve a global optimal approximation for the desired mapping and it has strong generalization ability [46]. We established a BPNN structure to judge whether candidate polygonal object matching pairs match or do not match.

The BPNN used here comprised three layers: an input layer, a hidden layer, and an output layer. The input layer refers to the characteristics described by similarity evaluation indices combined with MBR (Section 2.1) and GFEI (Section 2.2) of multiscale spatial polygonal objects. The hidden layer consists of the nodes set up in the BPNN. The output layer is the output of the matching result information of the model operation, i.e., 1 or 0, corresponding to “match” and “does not match,” respectively. At the same time, the output layer will derive the probability of a “match” (p) and the probability of a “does not match” ($1 - p$) outcomes. The criteria for judging whether to match are as follows:

1. When $p \geq 0.5$, the matching result is “match”; furthermore, the closer the value of p is to 1 or the value of $1 - p$ is to 0, the more similar the matching pair.
2. When $p < 0.5$, the matching result is “does not match”; furthermore, the closer the value of p is to 0 or the value of $1 - p$ is to 1, the more dissimilar the matching pair.

2.4. Matching Workflow

To efficiently apply BPM, a workflow was designed, as shown in Figure 4, comprising three processes: preparation, matching, and evaluation.

Step 1: In the preparation stage, the data are preprocessed, which includes format conversion, topology checking, and geometric coordinate transformation. The purpose of preprocessing is to resolve systematic errors between data from different sources [47].

Step 2: The matching stage is the focus of this study, which includes BPM training, first-matching, and last-matching. Model training is performed to acquire the geometric features and matching results of the training data. In first-matching, 1:1 and 1:N candidate matching pairs are detected using the MBR combinatorial optimization [48] (MBRCO) algorithm. MBRCO uses the objects' MBR to replace them so as to find their candidate matching objects, and adopts the combination threshold to avoid the excessive calculation of the exhaustive method, which can quickly and effectively screen the candidate matching pairs. Then, 1:1 and 1:N correspondences are obtained using the trained BPM. In final matching, spatial districts (SDs) are divided based on Delaunay triangulation, in which the remaining unmatched objects (M:N and 1:0) are distributed. The M:N candidate matching pairs are detected using the MBRCO algorithm and the M:N correspondences are obtained using the trained BPM. The remaining unmatched objects are 1:0 correspondences.

Step 3: In the evaluation stage, we evaluate the matching results by comparing them with the results of manual inspection.

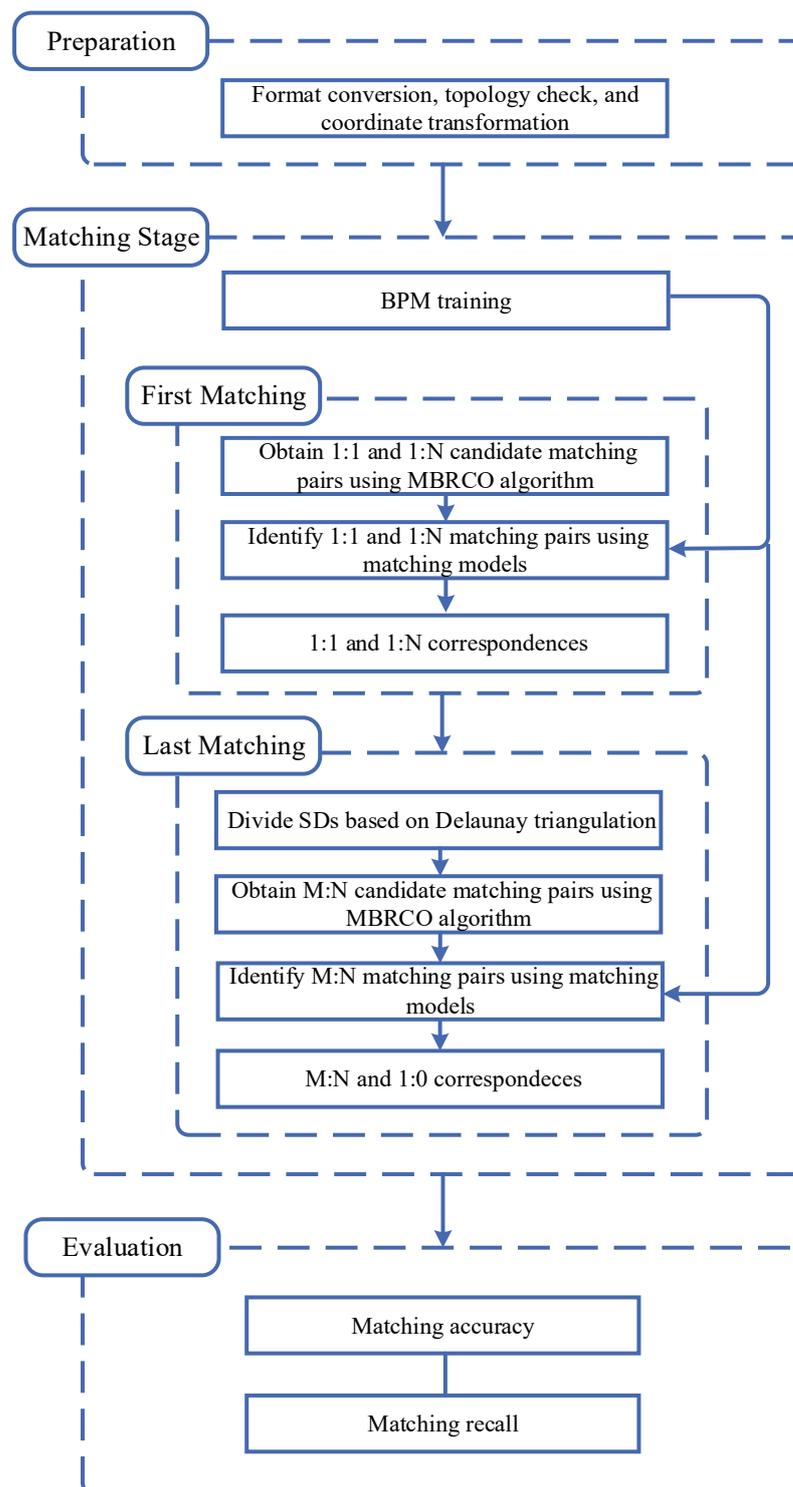


Figure 4. Multiscale Spatial Polygonal Object Matching workflow.

3. Experiments and Discussion

3.1. GFEI Index Analysis

Test data of GFEI index analysis is shown in Figure 5. The test data in Figure 5 include source data A and target data B. The source data A include 11 polygonal objects, namely, $\{a_1, a_2, \dots, a_{11}\}$ and the target data B include 26 polygonal objects, namely $\{b_1, b_2, \dots, b_{26}\}$ in Figure 5. There are six pairs of 1:1 matchings, three pairs of 1:N matchings, one pair of M:N matching, and zero pairs of 1:0 matching, and the 0:1 matching includes six polygonal

objects in B. The proportions of simple matching and complex matching in the experimental data are reasonable, making them suitable for the analysis test of GFEI.

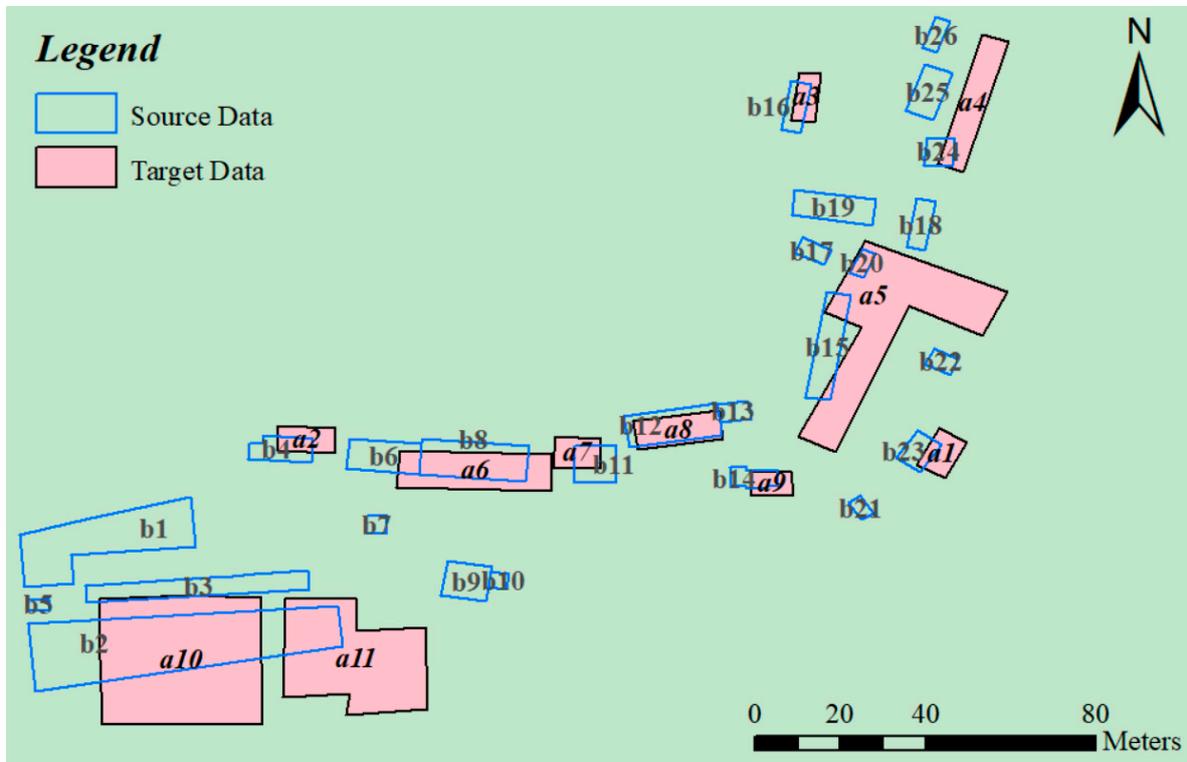


Figure 5. Test data of granularity factor evaluation index (GFEI) index analysis.

We define the formula of the combination of similarity evaluation indices as follows:

$$S_{all} = (S_{dis} \times w1 + S_{olp} \times w2 + S_{dir} \times w3 + S_{shape} \times w4) / (w1 + w2 + w3 + w4) \quad (8)$$

where $w1$, $w2$, $w3$, and $w4$ are weights of S_{dis} , S_{olp} , S_{dir} , and S_{shape} , respectively.

$$S_{all}' = (N_{GFEI} + S_{all}) / 2 \quad (9)$$

The weight ratio of the four similarity evaluation indices (Section 2.1) is set to 1:1:1:1 based on experience, which indicates that $w1 = 1$, $w2 = 1$, $w3 = 1$, and $w4 = 1$. The listed matching pairs are all correct matching pairs from manual recognition. The test results are shown in Table 1, where S_{all} represents the total similarity of the pairs to be matched and $R(S)$ represents the matching result based on the four similarity evaluation indices. We set the criterion of matching to $S_{all} \geq 0.75$ ($R(S)$: Match) and $S_{all} < 0.75$ ($R(S)$: Not recognized).

As indicated in Table 1, when the weighted similarity of the four indices ≥ 0.75 is taken as the criterion, all six matching pairs in 1:1 matching can be recognized as correct matching pairs. However, the values of S_{olp} and S_{shape} of the matching pairs are low in the complex matching. One matching pair is recognized in the 1:N matching, while two further matching pairs are not recognized as correct matching pairs. One matching pair is not recognized in the M:N matching. For example, the value of S_{olp} of matching pair ($a_5 : b_{15}, b_{17}, b_{18}, b_{19}, b_{20}$) is only 0.359 and that of S_{shape} is 0.684, leading to a value of the weighted similarity S_{all} of 0.685 (i.e., below the threshold value of 0.75), which means this matching pair cannot be recognized as a correct matching pair.

Table 1. Experimental results of the four similarity evaluation indices (Section 2.1).

Matching Type	Matching Pair	S_{dis}	S_{olp}	S_{dir}	S_{shape}	S_{all}	R(S)
1:1	$a_1 : b_{23}$	0.745	0.859	0.973	0.742	0.830	Match
	$a_2 : b_4$	0.709	0.909	0.971	0.882	0.868	Match
	$a_3 : b_{16}$	0.805	0.922	0.998	0.943	0.917	Match
	$a_7 : b_{11}$	0.721	0.865	0.922	0.778	0.822	Match
	$a_8 : b_{12}$	0.940	0.903	0.992	0.813	0.912	Match
	$a_9 : b_{14}$	0.739	0.848	0.943	0.803	0.833	Match
1:N	$a_4 : b_{24}, b_{25}, b_{26}$	0.779	0.496	0.902	0.643	0.705	Not recognized
	$a_5 : b_{15}, b_{17}, b_{18}, b_{19}, b_{20}$	0.811	0.359	0.885	0.684	0.685	Not recognized
	$a_6 : b_6, b_8$	0.822	0.874	0.988	0.784	0.867	Match
M:N	$a_{10}, a_{11} : b_1, b_2, b_3, b_5$	0.776	0.598	0.890	0.646	0.726	Not recognized

Now we introduce the proposed GFEI for further experiment. We use the normalized results of GFEI, N_{GFEI} , and S_{all} to do the 1:1 weighted processing. The test results are shown in Table 2.

Table 2. Test results of GFEI. (where S_{all}' represents the total similarity values of the matching pairs after the 1:1 weighted processing).

Matching Type	Matching Pair	R_{psd}	M_{psd}	S_{psd}	GFEI	N_{GFEI}	S_{all}'
1:1	$a_1 : b_{23}$	1	1.8	1.249	0.113	0.75	0.790
	$a_2 : b_4$	1			0.113	0.75	0.809
	$a_3 : b_{16}$	1			0.113	0.75	0.833
	$a_7 : b_{11}$	1			0.113	0.75	0.786
	$a_8 : b_{12}$	1			0.113	0.75	0.831
	$a_9 : b_{14}$	1			0.113	0.75	0.792
1:N	$a_4 : b_{24}, b_{25}, b_{26}$	3			0.091	0.86	0.783
	$a_5 : b_{15}, b_{17}, b_{18}, b_{19}, b_{20}$	5			0.063	1	0.842
	$a_6 : b_6, b_8$	2			0.108	0.775	0.821
M:N	$a_{10}, a_{11} : b_1, b_2, b_3, b_5$	2			0.108	0.775	0.751

We find the total similarity values of the matching pairs S_{all}' are all above the correct matching threshold of 0.75 after the weighted processing by N_{GFEI} , which avoids the occurrence of missing matching. For example, for matching pair ($a_5 : b_{15}, b_{17}, b_{18}, b_{19}, b_{20}$), R_{psd} is 5, and S_{olp} is only 0.359. Through the weighted processing of N_{GFEI} , S_{all}' effectively rises from 0.685 to 0.842, which takes this matching pair from being a not recognized matching pair to being a correct matching set whose threshold is ≥ 0.75 . We propose the correct matching recognition rate $P(T)$ to describe the overall degree of improvement. The formula can be expressed as follows:

$$P(T) = \frac{N(T)}{n} \times 100\% \quad (10)$$

where n is the number of correct matching pairs, and $N(T)$ is the number of matching pairs that can be recognized as the correct matching pairs in the candidate matching pairs based on the evaluation indices of MSPOM, which themselves are correct matching pairs.

The experimental results of the sample data in this section show that the recognition rate $P(T)$ is raised from 70% to 100% by the weighted processing of GFEI, which effectively reduces the occurrence of missing matching.

3.2. Experimental Data

BPM was compared and analyzed in the experiment of matching between 1:2000 and 1:10,000 polygonal objects of residential buildings and industrial facilities in the city of

Zhoushan, Zhejiang Province (China). The statistical description of the experimental data is shown in Table 3.

Table 3. Statistical description of experimental data.

	Dataset A	Dataset B
Place	Zhoushan, China	
Area	36 km ²	
Scale	1:2000	1:10,000
Production time	2012	2009
No. of polygons	6774	778

As shown in Figure 6, the relative deviation of the positions of larger polygonal objects is small, while that of smaller polygonal objects is comparatively larger. There are obvious differences in the data expression granularity, geometry, and location.



Figure 6. The superposition of dataset A and dataset B.

3.3. Model Training

3.3.1. Sample Selection

In the experiment, 950 candidate matching pairs, generated by about 20% of the polygonal objects in target dataset B, were selected at random as a sample for training BPM and comparison models. Half the sample data were selected as training data and the other half used as test data. The number of 1:1 candidate matching pairs was 98, which included 58 correct matching pairs and 40 incorrect matching pairs. The number of 1:N candidate matching pairs was 827, which included 356 correct matching pairs and 471 incorrect matching pairs. The number of M:N candidate matching pairs was 25, which included 5 correct matching pairs and 20 incorrect matching pairs. The purpose of incorrect matching pairs in the sample in Figure 7 is to allow the BPM to learn the mismatched features, so that BPM can better select the correct matching pairs from the candidate matching pairs. The matching results of the sample were artificially distinguished. The matching type distribution of the sample is shown in Figure 7.

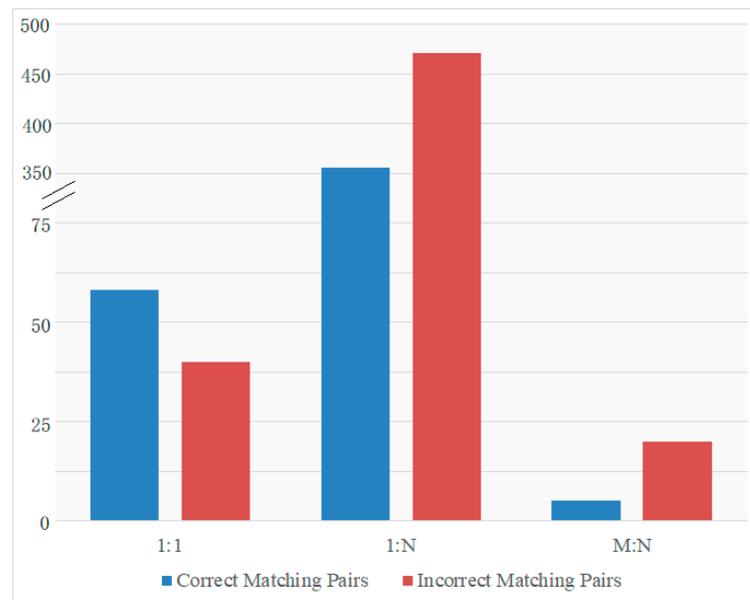


Figure 7. Matching type (1:1, 1:N, and M:N) distribution of the sample.

3.3.2. Models and Parameters

In this experiment, BPM was compared with multivariate logistic regression [22] matching model (MLRM) and SVM regression [49] matching model (SRM). Five indices proposed in Section 2.1 and 2.2 (distance similarity, overlap rate of area, direction similarity, shape similarity, and GFEI) were used as input factors in both BPM and comparison model settings. The traditional empirical weight matching model (EWM) was used as a control experiment, and the weights of the five factors were set to 1:1:1:1:1 in turn. Machine learning models can obtain weights by learning samples, whereas EWM is a model that allows us to assign weights to different factors according to experience.

The parameters of BPM, MLRM, and SRM were as follows:

1. BPM: We established a three-layer BPNN structure. The activation function was a logistic function, number of nodes in the hidden layer was 9, momentum factor was 0.9, learning efficiency was 0.01, and maximum number of iterations was 1000.
2. MLRM: The confidence interval of parameter estimation was 95%, maximum number of iterations was 1000, convergence value of parameters was 10^{-6} , and singularity tolerance was 10^{-8} .
3. SRM: The random number state was 1, kernel function was the radial basis function, and gamma value was 20.

3.4. Experimental Results and Discussion

To quantitatively evaluate the performances of MLRM, SRM, BPM, and EWM, this paper compares the matching results with the results of manual discrimination. To train the machine learning models better, matching was performed manually by cartographic engineers. We used the MBRCO method to obtain candidate matching pairs for cartographic engineers' matching discrimination. With the help of auxiliary information, the multi-phase remote-sensing image, according to cartographic engineers' matching experience, they made expert visual discrimination for complex matching such as M:N. Accuracy, recall, and F1-Measure evaluation methods are adopted. The calculation formulas M can be defined as follows:

$$M : \begin{cases} P = \frac{TP}{TP+FP+AM} \times 100\% \\ R = \frac{TP}{TP+FN} \times 100\% \\ F_1 = \frac{2*P*R}{P+R} \end{cases} \quad (11)$$

where P is the accuracy, R is the recall, TP is the number of correct matching pairs, FP is the number of incorrect matching pairs, AM indicates ambiguous cases that cannot be judged explicitly through manual inspection, and FN is the number of missing matching pairs. In the experiment, the buffer distance μ in obtaining candidate matching pairs was 30 m and the search threshold ε of MBRCO was 0.2. The matching performance of the four models is shown in Table 4.

Table 4. Matching results of four matching models.

Models	TP	FP	AM	FN	P	R	F_1
MLRM	631 81.1%	42 5.4%	14 1.8%	91 11.7%	91.8%	87.4%	89.5%
SRM	667 85.7%	15 1.9%	14 1.8%	82 10.5%	95.8%	89.1%	92.3%
BPM	687 88.3%	6 0.8%	14 1.8%	71 9.1%	97.2%	90.6%	93.8%
EWM	619 79.6%	51 6.6%	14 1.8%	94 12.1%	90.5%	86.8%	88.6%

To show the matching performance intuitively, a comparison histogram of accuracy, recall, and F1-Measure of the four matching models is shown in Figure 8.

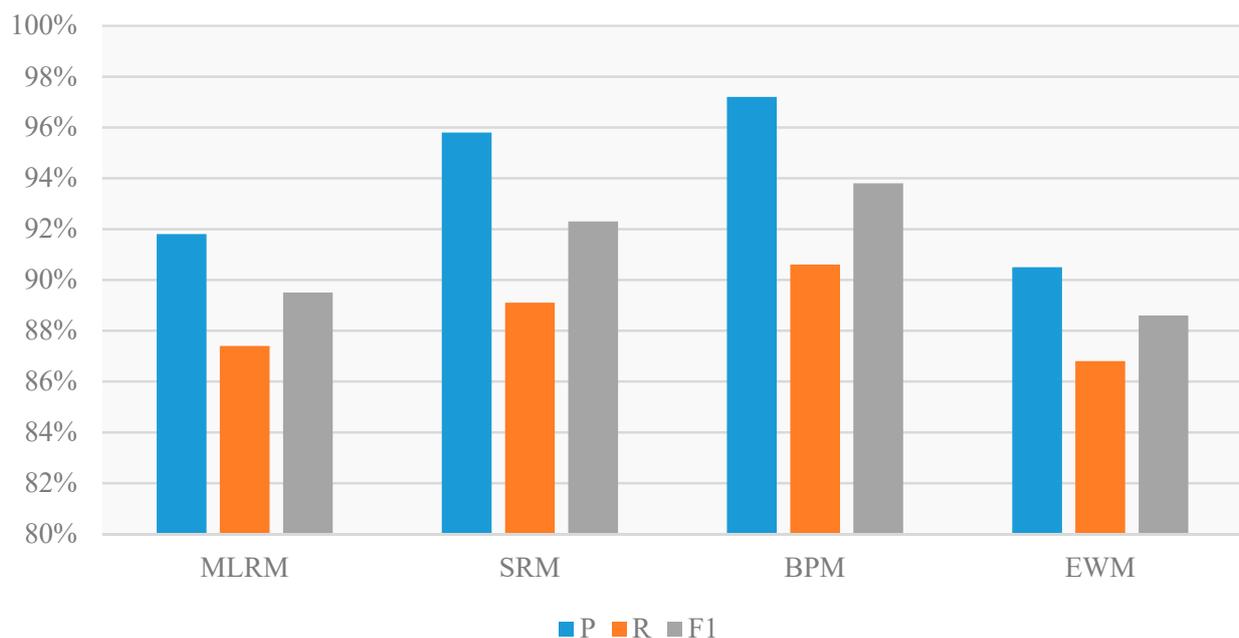


Figure 8. Comparison of P (Accuracy), R (Recall), and F_1 (F1-Measure evaluation) of the four matching models, multivariate logistic regression matching model (MLRM), SVM regression matching model (SRM), Matching Model Based on Backpropagation Neural Network (BPM), and empirical weight matching model (EWM).

The matching accuracies of the MLRM, SRM, and BPM are all higher than the EWM by 1.3%, 5.3%, and 6.7% respectively. As shown in Figure 9, the number of correct matching pairs of the MLRM, SRM and BPM is higher than the 619 pairs achieved by the EWM, which shows that the machine learning models avoid matching errors caused by inaccurate empirical factor weights. Through the training of the sample data, reasonable evaluation criteria are established. Among them, the matching accuracy of the BPM is highest, i.e., 6.7% higher than that of the EWM, which indicates that BPM is most suitable for improving the matching accuracy of multiscale spatial polygonal objects.

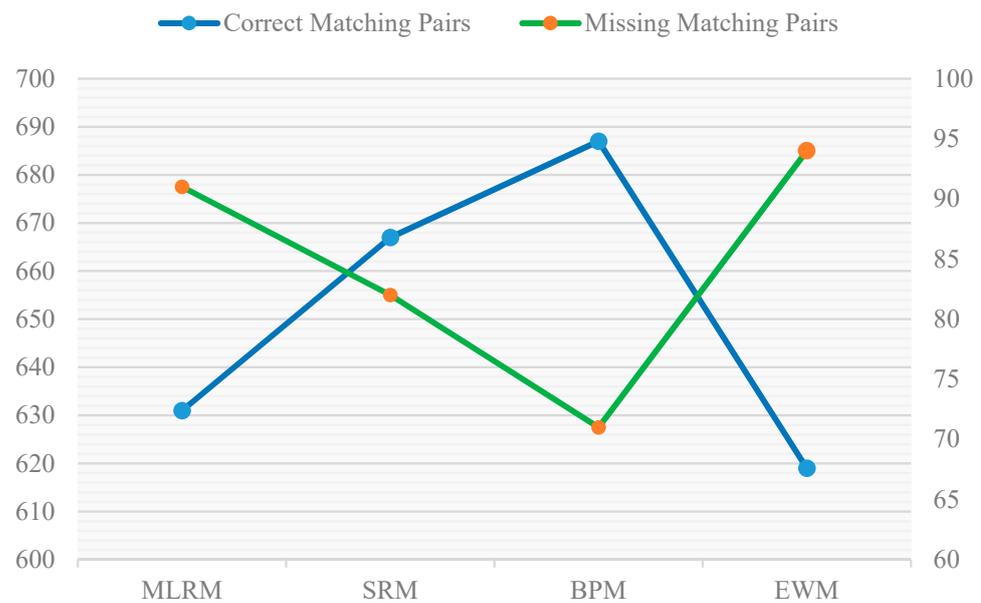


Figure 9. Comparison of the number of correct matching pairs and missing matching pairs of MLRM, SRM, BPM, and EWM.

The recalls of the MLRM, SRM, BPM and EWM are reasonably high with small differences because these four models adopt MBRCO to acquire candidate matching pairs. MBRCO can obtain more candidate matching pairs, which is conducive to selecting the best matching pairs in multiple candidate matching pairs. It also helps the missing matching situation and makes the recall satisfactory when the location deviation between the source data and the target data is large. As shown in Figure 9, the number of missing matching pairs of MLRM, SRM and BPM is slightly lower than the 94 pairs of the EWM, which shows that the machine learning models can also both reduce the number of missing matching pairs and improve the recall to a certain extent.

As shown in Table 4, the matching accuracies and recalls of MLRM, SRM, and BPM are all higher than the EWM and thus their overall performance (F_1) is also better. The F_1 index of the EWM is 88.6%, while that of the BPM is the highest (93.8%). This shows that the overall performance of BPNN is the best. The overall behavior of BPM matching results depends on the interconnection and interaction among the neurons that comprise the neural network, avoiding limitations. BPNN has a strong generalization ability and it can achieve global optimal approximation for the desired mapping. The F_1 index of the SRM is second best at 92.3%. This is because SVM follows the principle of seeking the optimal hyperplane and thus its classification accuracy is higher than that of either the MLRM or the EWM.

The similarity evaluation indices describing matching similarity, combined with GFEL, are not sufficiently accurate. MBR is an extent rectangle whose sides are parallel to the coordinate axes, and the extent is an approximation of the shape. In the future, we will construct a minimum area bounding rectangle, oriented toward the main direction of the object or the group of objects, that will better depict the shape, and substitute the direction similarity (Section 2.1.3) and shape similarity (Section 2.1.4) with one measure.

4. Conclusions

This study considered the matching optimization of multiscale spatial polygonal objects. The construction of matching similarity evaluation indices and matching optimization based on GFEL and BPNN were discussed. The proposed matching optimization method was demonstrated in an experiment of matching between 1:2000 and 1:10,000 polygonal features of residential buildings and industrial facilities in Zhoushan, Zhejiang Province (China).

Based on MSPOM similarity evaluation indices combined with MBR, GFEI was proposed to deal with complex matches in MSPOM. Experiments showed this index effective in improving the matching score of complex matching and in promoting the ability of matching recognition. Sample data were selected to train BPM, and BPM was then used to discriminate candidate matching pairs automatically. Experiments showed that combining machine learning models could effectively improve the accuracy of matching, and BPM was found most suitable for MSPOM among them.

The head–tail segmentation function was used to partition and quantitatively describe the complex matching and simple matching when constructing GFEI. In the future, we will consider looking for other effective functions with which to construct the factor model and we will compare their applicability.

Author Contributions: Daoye Zhu performed the research, analyzed the data and wrote the paper. Chengqi Cheng supervised the study. Bo Chen and Weixin Zhai offered helpful suggestions. Yihang Li and Shizhong Li participated in discussion. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA19090102, in part by the National Key Research and Development Programs of China under Grant 2018YFB0505300, and Grant 2017YFB0503703, in part by the Guangxi Science and Technology Major Project under Grant AA18118025, and in part by Wespace Academic Research Funding (2021).

Data Availability Statement: Restrictions apply to the availability of these data. The data are not publicly available due to privacy.

Acknowledgments: Weixin Zhai and Bo Chen make equal contribution to this manuscript. The authors would like to thank Xinyan Zhu, Professor at Wuhan University, and Lingjia Liu, Lecturer at Jiangxi Normal University for their suggestions and technical guidance. The authors are sincerely grateful for the comments of the anonymous reviewers and members of the editorial team.

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

References

- Li, L.; Goodchild, M.F. Automatically and accurately matching objects in geospatial datasets. *Adv. Geo Spat. Inf. Sci.* **2012**, *10*, 71–79.
- Saalfeld, A. *Automated Map Conflation*; University of Maryland: College Park, MD, USA, 1993.
- Xavier, E.M.; Ariza-López, F.J.; Ureña-Cámara, M.A. A survey of measures and methods for matching geospatial vector datasets. *ACM Comput. Surv.* **2016**, *49*. [[CrossRef](#)]
- Kim, J.O.; Yu, K.; Heo, J.; Lee, W.H. A new method for matching objects in two different geospatial datasets based on the geographic context. *Comput. Geosci.* **2010**, *36*, 1115–1122. [[CrossRef](#)]
- Ruiz, J.J.; Ariza, F.J.; Ureña, M.A.; Blázquez, E.B. Digital map conflation: A review of the process and a proposal for classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [[CrossRef](#)]
- Girres, J.; Touya, G. Quality assessment of the French OpenStreetMap dataset. *Trans. GIS* **2010**, *14*, 435–459. [[CrossRef](#)]
- Bisheng, Y.; Yunfei, Z.; Xuechen, L. A probabilistic relaxation approach for matching road networks. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 319–338.
- Devogele, T. On spatial database integration. *Int. J. Geogr. Inf. Syst.* **1998**, *12*, 335–352. [[CrossRef](#)]
- Walter, V.; Fritsch, D. Matching spatial data sets: A statistical approach. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 445–473. [[CrossRef](#)]
- Hampe, M.; Sester, M.; Harrie, L. Multiple representation databases to support visualisation on mobile devices. *ISPRS J. Photogramm. Remote Sens.* **2004**, *35*, 135–140.
- Rauschenbach, U.; Jeschke, S.; Schumann, H. General rectangular fisheye views for 2D graphics. *Comput. Graph.* **2001**, *25*, 609–617. [[CrossRef](#)]
- Ying, S.; Li, L.; Gao, Y.R.; Min, Y. Probabilistic matching of map objects in multi-scale space. In Proceedings of the 25th International Cartographic Conference, Paris, France, 3–8 July 2011.
- 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](#)]
- Taylor, E.G.; Hummel, J.E. Finding similarity in a model of relational reasoning. *Cog. Sys. Res.* **2009**, *10*, 229–239. [[CrossRef](#)]
- Xu, Y.; Xie, Z.; Chen, Z.; Wu, L. Shape similarity measurement model for holed polygons based on position graphs and Fourier descriptors. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 253–279. [[CrossRef](#)]

16. Tong, X.; Shi, W. A probability-based multi-measure feature matching method in map conflation. *Int. J. Remote Sens.* **2009**, *30*, 5453–5472. [[CrossRef](#)]
17. Fan, H.; Yang, B.; Zipf, A.; Rousell, A. A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. *Int. J. Geogr. Inf. Sci.* **2015**, *30*, 748–764. [[CrossRef](#)]
18. Zhang, X.; Zhao, X.; Molenaar, M.; Stoter, J.; Kraak, M.; Ai, T. Pattern classification approaches to matching building polygons at multiple scales. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2012**, 19–24. [[CrossRef](#)]
19. Ruiz-lendinez, J.J.; Ariza-lopez, F.J.; Ureña-cámara, M.A. Automatic positional accuracy assessment of geospatial databases using line-based methods. *Surv. Rev.* **2013**, *45*, 332–342. [[CrossRef](#)]
20. 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](#)]
21. Deng, M.; Li, Z.L.; Chen, X.Y. Extended Hausdorff distance for spatial objects in GIS. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 459–475.
22. Tong, X.; Liang, D.; Jin, Y. A linear road object matching method for conflation based on optimization and logistic regression. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 824–846. [[CrossRef](#)]
23. Revell, P.; Antoine, B. Automated matching of building feature of differing levels of detail: A case study. In Proceedings of the 24th International Cartographic Conference, Santiago, Chile, 15–21 November 2009.
24. Qi, H.B.; Li, Z.L.; Chen, J. Automated change detection for updating settlements at smaller-scale maps from updated larger-scale maps. *J. Spatial Sci.* **2010**, *55*, 133–146. [[CrossRef](#)]
25. Winter, S. Location similarity of regions. *ISPRS J. Photogramm. Remote Sens.* **2000**, *55*, 189–200. [[CrossRef](#)]
26. Gössehn, G.; Sester, M. Change detection and integration of topographic updates from ATKIS and geoscientific data sets. In Proceedings of the International Conference Next Generation Geospatial Infrastructure, Boston, MA, USA, 19–21 October 2003.
27. Wang, H.; Tang, X.; Qiu, B.; Wang, W. Geometric matching method of area feature based on multi-weighted operators. *Geomat. Inf. Sci. Wuhan Univ.* **2013**, *38*, 1671–8860.
28. Ai, T.; Cheng, X.; Liu, P.; Yang, M. A shape analysis and template matching of building features by the Fourier transform method. *Comput. Environ. Urban Syst.* **2013**, *41*, 219–233. [[CrossRef](#)]
29. Deng, M.; Li, Z. A statistical model for directional relations between spatial objects. *Geoinformatica* **2008**, *12*, 193–217. [[CrossRef](#)]
30. Luo, G.; Zhang, X.; Qi, L.; Guo, T. The fast positioning and optimal combination matching method of change vector object. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 1285–1292.
31. 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](#)]
32. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, USA, 1993.
33. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; Chapman and Hall: New York, NY, USA, 1984.
34. Rish, I. Empirical study of the naive Bayes classifier. In *The IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*; IBM: New York, NY, USA, 2001; pp. 41–46.
35. Hsu, C.-W.; Lin, C.-J. Comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* **2002**, *13*, 415–425. [[CrossRef](#)] [[PubMed](#)]
36. McKenzie, G.; Janowicz, K.; Adams, B. A weighted multi-attribute method for matching user-generated points of interest. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 125–137. [[CrossRef](#)]
37. Juan, R.-L.; Manuel, U.-C.; Francisco, A.-L. Polygon and point-based approach to matching geospatial features. *ISPRS Int. J. Geoinf.* **2017**, *6*, 399. [[CrossRef](#)]
38. Pradhan, B.; Lee, S.; Buchroithner, M.F. A GIS-based back-propagation neural network model and its cross-application and validation for landslide susceptibility analyses. *Comput. Environ. Urban Syst.* **2010**, *34*, 216–235. [[CrossRef](#)]
39. Cobb, M.A.; Petry, F.E.; Shaw, K.B. Fuzzy spatial relationship refinements based on minimum bounding rectangle variations. *Fuzzy Sets Syst.* **2000**, *113*, 111–120. [[CrossRef](#)]
40. Hobbs, J.R. Granularity. In *Readings in Qualitative Reasoning about Physical Systems*; Weld, D.S., Klerer, J.D., Eds.; Morgan Kaufmann: San Mateo, CA, USA, 1990; pp. 542–545.
41. Stell, J.; Worboys, M. Stratified map spaces: A formal basis for multi-resolution spatial databases. *Int. Symp. Spat. Data Handl.* **1998**, 180–189.
42. Deng, M.; Li, Z.; Cheng, T. Rough-set representation of GIS data uncertainties with multiple granularities. *Acta Geod. Cartogr. Sin.* **2006**, *35*, 64–70.
43. Bin, J. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *Prof. Geogr.* **2012**, *65*, 482–494.
44. Cracknell, M.J.; Reading, A.M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput. Geosci.* **2014**, *63*, 22–33. [[CrossRef](#)]
45. Liu, Y.; Zhao, P.; Qin, D.; Li, G.; Chen, Z.; Zhang, Y. Driving intention identification based on long short-term memory and a case study in shifting strategy optimization. *IEEE Access* **2019**, *7*, 128593–128605. [[CrossRef](#)]
46. Zhang, L.; Sun, Z.; Zhang, C.; Dong, F.; Wei, P. Numerical investigation of the dynamic responses of long-span bridges with consideration of the random traffic flow based on the intelligent ACO-BPNN model. *IEEE Access* **2018**, *6*, 28520–28529. [[CrossRef](#)]
47. Doytsher, Y. A rubber sheeting algorithm for non-rectangular maps. *Comput. Geosci.* **2000**, *26*, 1001–1010. [[CrossRef](#)]

-
48. 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. [[CrossRef](#)]
 49. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.