A New Method for Measuring Topological Structure Similarity between Complex Trajectories

Huimeng Wang, Yunyan Du, Jiawei Yi, Yong Sun, and Fuyuan Liang

Abstract—We proposed a new framework to measure the similarity of topological structure between complex trajectories. There are three steps in the framework. A complex trajectory is first represented by a graph structure which consists of nodes and edges. Secondly, we developed a Comprehensive Structure Matching (CSM) algorithm to identify all common structures between the complex trajectories of interest. Thirdly, we used the Jaccard similarity coefficient to evaluate the similarity between two complex trajectories. We used synthetic graph data to evaluate the CSM method and examine its performance by comparing against that of the VF2 and the exact graph edit distance (EGED) algorithms. Results show that the CSM algorithm outperforms the EGED in terms of the computation efficiency. The CSM is more comprehensive than the VF2 algorithm as it further considers the partial isomorphism. We used the CSM algorithm to examine the 1993 to 2012 complex trajectories of anticyclonic eddies in the South China Sea (SCS). The CSM successfully found the complex trajectories that are similar to a thoroughly-studied ACE3 trajectory in the SCS. From the similar trajectories, we identified a dominant migrating path of the ocean eddies in the northern SCS. The CSM also successfully identified some new complex trajectories that propagated across the 18°N parallel in the SCS, which were not reported before. It also further identified multiple common structure models of the complex trajectories. These findings help us better understand the behaviors and the evolution of the mesoscale eddies in the SCS.

Index Terms—Complex trajectories, topological structure, graph isomorphism, similarity, ocean eddies

1 Introduction

Portal devices equipped with GPS-receivers and WiFi-sensors have produced a huge amount of spatial data representing the trajectories of moving objects. Trajectory analysis/pattern mining is one of the crucial research topics and significant progress has been made in recent years (e.g., [1], [2], [3], [4], [5],[6]). Similarity analysis is one of the most powerful trajectory data mining tools [2], [7], [8], [9] that can group and identify the movement patterns of moving objects (e.g., [5], [10], [11]).

A moving object could generate either a simple or a complex trajectory. A simple trajectory is usually produced by the object, such as an individual person, a vehicle, or an animal, that always moves as one piece in a space. It has a linear structure and without any branches. A complex trajectory is produced by the objects or phenomena that may change their structures while they are moving. For example, ocean eddies, oil spills, rain clouds, and air masses, may split into multiple components or merge with one or more counterparts nearby while they are moving (e.g., [12], [13]). A complex trajectory is nonlinear in terms of the structure and bears at least one split and/or merger branch.

The approaches that have been proposed to measure the similarity between trajectories could be generally categorized into three groups: space-based, space- and time-based, and attributes-based methods. The space-based methods evaluate the similarity between trajectories by measuring the Euclidean distance, the Hausdorff distance, or the Fréchet distance between the corresponding points along the trajectories [14], [15], [16]. The space- and time-based methods compare the Dynamic Time Warping [17], [18], [19] or the Longest Common Sub-Sequence [20] of the points to determine the similarity between trajectories. The attributes-based methods measure the similarity between trajectories by calculating the edit distance of one or more movement parameters such as the speed, acceleration, and direction [2], [20].

The afore-mentioned similarity-measuring approaches only work on the simple but not the complex trajectories. First, calculating the Euclidean distance requires exactly the same number of points along the two trajectories that are being compared, which is usually not true for most complex trajectories. Secondly, the Hausdorff distance does not consider the sequence of the points, which is particularly important in illustrating the structure of a complex trajectory. Both the Fréchet distance and the Dynamic Time Warping require a one-to-one relationship between every move of an object makes at a specific time and the points

- *H. Wang, Y. Du, and J. Yi are with the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China. E-mail: {wanghm, duyy, yijw}@lreis.ac.cn.*
- *Y. Sun is with the Shandong University of Science and Technology, Qingdao 266000, China. E-mail: ttsunyong@163.com.*
- *F. Liang is with the Department of Geography, Western Illinois University, Macomb, IL 61455. E-mail: F-Liang@wiu.edu.*

Manuscript received 9 Jan. 2018; revised 1 Aug. 2018; accepted 22 Sept. 2018. Date of publication 28 Sept. 2018; date of current version 10 Sept. 2019. (Corresponding author: Yunyan Du.)
Recommended for acceptance by F. Afrati.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2872523

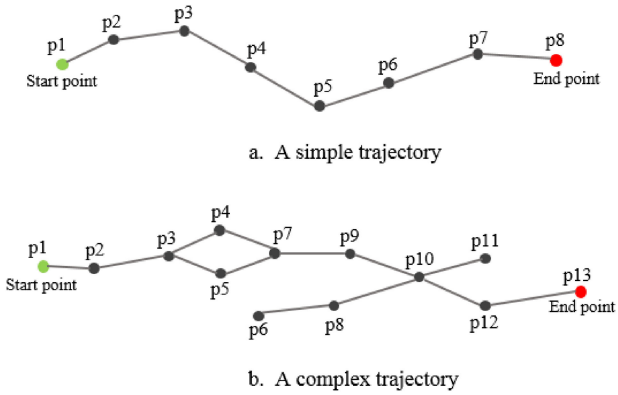


Fig. 1. The structural difference between (a) a simple trajectory and (b) a complex trajectory.

along the trajectory the object generated. Unfortunately, it is very common for a complex trajectory to have multiple points at a specific time when the object makes a move. Calculating the Longest Common Sub-Sequence and the edit distance requires a transformation of a trajectory to a regular string. However, it is really challenging, if not impossible, to convert a trajectory with branches into one string.

The topological structures of a trajectory may provide a new perspective to measure the similarity between complex trajectories. Similarities between complex trajectories could be inferred from measuring the similarity between graphs, which are widely used to describe the structures of objects or images [21], [22]. The similarity measurement of graphs has been widely used in many fields [23] such as matching chemical molecular structures and analyzing protein structures. Graph edit distance and isomorphism are the two classical approaches that have been used to match graphs to detect the common structures, as well as their corresponding attributes, of multiple graphs [24]. The exact graph edit distance (EGED) method measures the graph similarity by transforming one graph to another by adding, deleting, substituting nodes or edges. In practice, however, the EGED can only be used to study small graphs [23], [25], [26]. Another method, the VF2, was proposed to match large graphs by examining graph and subgraph isomorphism [22], [27]. However, the VF2 algorithm can only identify whether the structures of two graphs are the same (graph isomorphism) or whether one graph is a subgraph of another (subgraph isomorphism). Neither the EGED nor the VF2 methods can find the graph (subgraph) isomorphism and, at the same time, identify the common structures between graphs. Furthermore, the conventional methods usually decompose a complex trajectory into multiple simple trajectories before measuring the similarity. The decomposing process destroys the original structural features of a complex trajectory, and, as a result, the conventional similarity measurement methods actually fail to measure the global similarity of the entire trajectory.

We significantly expanded the VF2 algorithm and developed a Comprehensive Structure Matching (CSM) algorithm to measure the global similarity of the topological structures between complex trajectories. Unlike the convertional methods, the CSM algorithm measures the similarity between trajectories by comparing the graph structure as a whole. We evaluated the performance of the CSM algorithm by testing its performance on synthetically generated graphs. We then used the CSM algorithm to measure the similarity of complex trajectories that are generated by mesoscale ocean eddies in the South China Sea (SCS). As shown in this paper, the CSM algorithm is also able to identify the common structural patterns of ocean eddy trajectories. Such patterns have never been reported in previous researches and thus provide new insights of the ocean eddies in terms of their evolution processes.

The remainder of this paper is organized as follows: Section 2 defines the complex trajectories and presents the datasets that were used in this study. Section 3 shows how the CSM method measures the similarity of topological structures of complex trajectories. Section 4 is devoted to the algorithm simulation experiments, method comparison, and computational efficiency evaluation. Section 5 presents a real-world application and presents the new knowledge discovered in this study. Section 6 presents the major conclusions of this study and also briefly outlines our thoughts of the future research.

2 Definitions and Datasets

2.1 Definitions

In this study, we defined a trajectory as a series of points in a chronological order $(\mathrm{p_1},\ \mathrm{p_2},\ \mathrm{p_3}\ldots\ldots\ \mathrm{p_n})$. Each point along a trajectory is defined by its longitude, latitude, and a timestamp $(\mathrm{pi} = (\mathrm{x_i},\ \mathrm{y_i},\ \mathrm{t_i}))$.

A simple trajectory (Fig. 1a) has a linear structure and each point along it has a unique predecessor point (except for the starting point) and a unique successor point (except for the endpoint).

By contrast, a complex trajectory has a nonlinear structure and contains at least one branch resulting from either a split or a merger activity. The ratio between the number of branch points and the total number of points along that specific trajectory $(\eta,\ 0 < \eta <= 1)$ is used to measure the complexity of a complex trajectory. A point along a complex trajectory may have multiple predecessor and successor points. For example, a point on a major trajectory may split into two or even more points and such a split event would generate two or more branches along the major trajectory (Fig. 1b). Two or more points can also merge into one and the corresponding two or more branches would merge into one major trajectory. As illustrated in Fig. 1b, $\mathrm{p_3}$ splits into $\mathrm{p_4}$ and $\mathrm{p_5}$, which then merge into $\mathrm{p_7}$. Points $\mathrm{p_8}$ and $\mathrm{p_9}$ merge into $\mathrm{p_{10}}$, which then splits into $\mathrm{p_{11}}$ and $\mathrm{p_{12}}$. The total number of points along this complex trajectory is 13 and 9 out of them are on the branches. The trajectory complexity η therefore is 0.69.

2.2 The Datasets

In this study, we used both synthetic data and real-world complex trajectories of mesoscale eddy to test the CSM method. The synthetic dataset includes multiple synthetic graphs that have different numbers of nodes and different η. More information about the synthetic dataset is presented in Sections 4.2 and 4.3. The real-world complex trajectories are generated by mesoscale ocean eddies, which play an important role in marine biological environment and can even affect the migration of biological communities in the deep

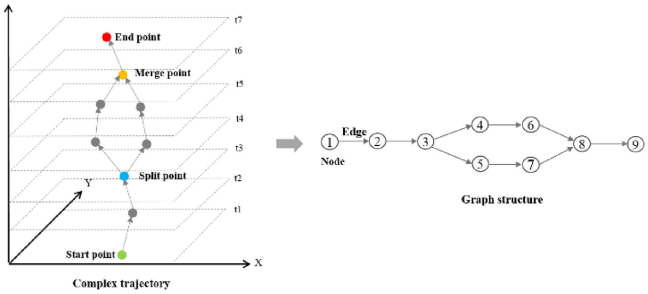


Fig. 2. A graph-structure representation of a complex trajectory.

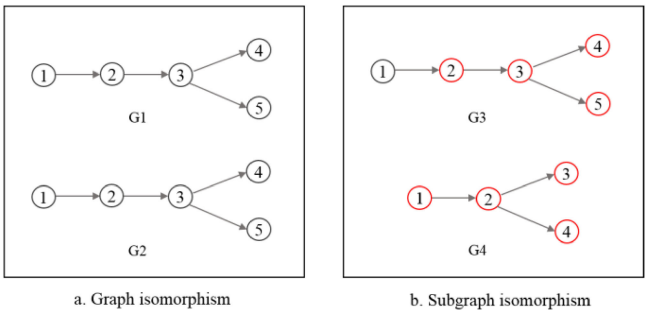


Fig. 3. An example of (a) graph isomorphism and (b) subgraph isomorphism.

sea along the vertical direction [29]. Previous studies have shown that eddies may split and merge during their lifespans (e.g., [12], [30], [31]) and generate complex trajectories. Mesoscale ocean eddies are very common in the SCS and have received much attention from the scientific community in the past decades (e.g., [32], [33], [34], [35], [36]). Other phenomena such as storms, soil spills, and rain clouds may generate very similar complex trajectory after similar splitting and merging activities when they are moving within a space (e.g., [12], [13], [28]). As a result, it is perfect to use the real-world eddies trajectories to test the CSM algorithm.

In this study, we examined the complex trajectories of the mesoscale ocean eddies that survive from January 1993 to December 2012 in the SCS. The trajectories were reconstructed from the Sea Level Anomaly (SLA) maps products [28], [35], which were produced from the spatial data acquired by the TOPEX/Poseidon, Jason 1, ERS-1, and ERS-2 satellites [37] and distributed by the AVISO (http://www.aviso.oceanobs.com). The SLA maps have a 1/4° spatial and a 1-day temporal resolution.

3 Methodology

Proposed by Cordella, the VF2 algorithm studies both graph and subgraph isomorphism to match large graphs [22]. The VF2 adopts a relatively simple order to match the graphs. It only selects the adjacent nodes of the matched nodes to execute the next match. In the choice of pruning strategy, the VF2 algorithm selects the nodes adjacent to the matched nodes as the alternative nodes. The VF2 greatly saves the time to search as it uses the relationship between the matching nodes on the waiting list and the matched nodes to curb the searching process. Studies have shown that the VF2 algorithm is more efficient to compare the graphs with a regular structure [22], [27]. However, the VF2 algorithm could only find the graph and the subgraph isomorphism but not the partial isomorphism between graphs.

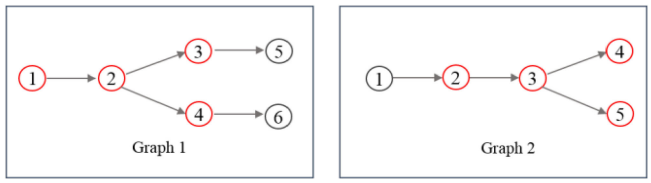


Fig. 4. An example of partial isomorphism.

Inspired by the VF2 algorithm, we developed a new framework to measure the similarity of topological structures between complex trajectories. Our framework includes three steps: representation of complex trajectories, comprehensive structure matching (CSM), and similarity measurement. A complex trajectory is first represented by a graph structure which consists of nodes and edges. We then used the CSM algorithm to identify all common structures (graph, subgraph, and partial isomorphism) between the trajectories and count the number of matching nodes. Lastly, we used the Jaccard similarity measure [38] to evaluate the similarity between the graphs and thus the trajectories.

3.1 Representation of Complex Trajectories

A graph usually contains multiple nodes and a set of edges that representing the relations between the nodes. In this study, we excluded the graphs with labeled nodes and/or edges and only considered the unlabeled graphs that indicate $\mathrm{Ln} = \mathrm{Le} = 1$ (Ln and Le represent a finite set of the nodes and edge labels, respectively). As a result, a graph is defined as a 2-tuple G (N, E) at time t, where N is a set of finite nodes and $\mathrm{E} \subseteq \mathrm{N} \times \mathrm{N}$ is a set of edges. As illustrates in Fig. 2, a node in the graph represents a point along a complex trajectory and an edge represents the relationship between the two points connected. The serial number of each point along the trajectory is used as the ID of the node. A complex trajectory thus is represented as a graph G(N, E). All nodes and edges of a trajectory are stored in N (G) and E (G), respectively.

3.2 Comprehensive Structure Matching

We then used the CSM algorithm to identify all matched nodes between the graphs, which represent different complex trajectories. The VF2 algorithm could identify the matched nodes by examining graph and subgraph isomorphism. For example, G1 and G2 in Fig. 3a are graph isomorphism as all nodes match. G4 in Fig. 3b is a subgraph of G3 as only four nodes match, i.e., nodes 1-4 in G3 match nodes 2-5 in G4, respectively.

However, the VF2 algorithm is not able to identify the partial isomorphism between complex trajectories. Fig. 4 shows an example of partial isomorphism between two graphs as nodes 1-4 in Graph 1 match nodes 2-5 in Graph 2, respectively. Obviously, it is important to examine the graph, subgraph, and partial isomorphism in order to measure the similarity between complex trajectories.

Inspired by the VF2 algorithm, we developed the CSM algorithm, which measure the similarity of topological structures of complex trajectories by identifying graph, subgraph, and partial isomorphism (Fig. 5). The algorithm first examines graph isomorphism and identifies the maximum

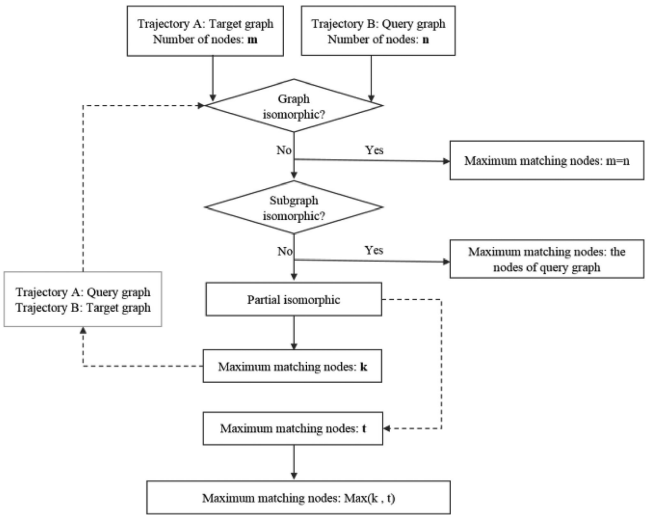


Fig. 5. The general workflow of the CSM algorithm.

number of the nodes that match between a target Graph A and a query Graph B. If no graph isomorphism is found, the algorithm would examine subgraph isomorphism and uses the number of the nodes in query Graph B as the maximum number of matched nodes. If no subgraph isomorphism is found, the CSM algorithm would examine the partial isomorphism and find the maximum number (*k*) of matched nodes. We then switch the target graph and the query graph, repeat the afore-mentioned procedure, and find the maximum number (*t*) of matched nodes. The CSM would then use the larger values of *k* and *t* as the maximum number of matched nodes in the case of partial isomorphism.

The pseudocodes in Table 1 illustrate how the CSM algorithm examines the partial isomorphism. It first computes the number of matched nodes (*Counts*) for each pair of a

TABLE 1
Pseudocodes of the CSM Algorithm

CSM Algorithm

Input: Target graph $\mathrm{G_t(N_n, E_n)}$, Query graph G_q $(\mathrm{N_m}, \mathrm{E_m})$,
Counts $= 0$, *Maxcounts* $= 0$, $\mathrm{M} = \emptyset$, $\mathrm{C} = \emptyset$, subsets $=\emptyset$
Output: *Maxcounts*
Algorithm Partial matching (G_t, G_q):
1: if GraphIsom (G_t, G_q) = true then
2: *Counts* = the number of nodes of G_t orG_q;
3: if SubgraphIsom (G_t, G_q) = true then
4: *Counts* = the number of nodes G_q;
5: else
6: for each node *j* in G_q do
7: if SubgraphIsom (G_{j_k}, G_t) = true then
8: Add the matched nodes into M;
9: *Counts*++;
10: else
11: if *Maxcounts*< *Counts* then
12: *Maxcounts* = *Counts*;
13: if Position (G_q, G_t, M) = true then
14: subsets = $\{[\mathrm{subG}_{q1}, \mathrm{subG}_{t1}] \ldots [\mathrm{subG}_{qi}, \mathrm{subG}_{ti}]\}$;
15: $\mathrm{M} = \emptyset$;
16: for each *i* in subsets do
17: *Maxcounts* + = Partial matching (subG_{qi}, subG_{ti});
18: return *Maxcounts*;

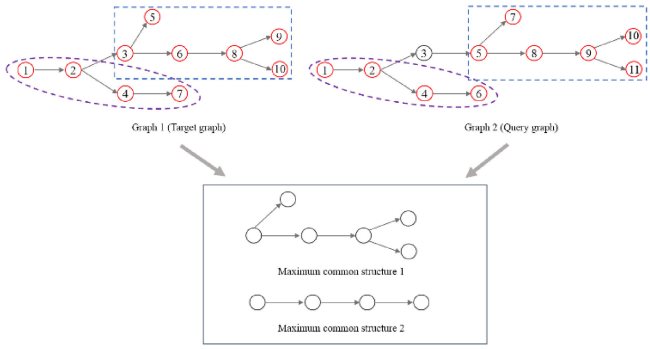


Fig. 6. An example illustrating how the partial matching algorithm works.

target graph (G_t) and a query graph (G_q). If G_t and G_q are isomorphic, *Counts* would be the number of the nodes in G_t or G_q. In fact, G_t and G_q would have the same number of nodes in this case. If G_t and G_q are not isomorphic, the algorithm would examine whether G_t and G_q are subgraph isomorphic or not. If G_q is a subgraph of G_t, *Counts* would be the number of the nodes in G_q. If G_t and G_q are not isomorphic, the CSM algorithm would start the partial matching process. The CSM algorithm would first find the maximal common subgraphs between G_t and G_q (Lines 6-19). Starting from the first node $(j = 1)$ in G_q (m), the CSM algorithm identifies a subgraph G_{j_k} $(1 <= k < m,\ 1 <= j < m$, m is the total number of nodes in G_q by adding k nodes and the associated edge in G_q starting from the jth node in G_q. The algorithm then compares G_{j_k} against the target graph G_t. If $\mathrm{G}_{j\_\mathrm{k}}$ and G_t are subgraph isomorphic, the algorithm would record the number of the matched nodes k and select the maximum k value as the number of matched nodes *Counts* between G_t and G_q. If the maximal common subgraphs are both in the front, middle, or rear of the G_t and G_q (Line 14), the CSM algorithm would identify the second maximum common subgraph of G_t and G_q except for the first one, until no maximum common subgraphs could be found between G_t and G_q.

The process of nodes matching, assumes that the degree of a node in the query graph should not be greater than that of the target graph. However, the assumption may not be always valid. As a result, we run one more pass of node mathching by swapping the target and query graphs. The CSM algorithm calculates the number of matched nodes between G_t and G_q before and after the switch, respectively (Fig. 6). The CSM uses the partial matching approach to find the maximum common structure (MCS) 1 between G_t and G_q. The MCS between the two graphs in Fig. 6 includes the 6 nodes bounded by the rectangle box. As the MCS 1 is both at the rear of the two graphs, the algorithm initiates a second matching process and finds the MCS 2, which contains the 4 nodes within the eclipse as shown in Fig. 6. All nodes in Graph 1 can find their matched counterparts in Graph 2 and thus the *Maxcounts* is 10. The CSM then switches the target and the query graphs, repeats the aforementioned procedure, and calculates the *Maxcounts*, which is also 10. In the end, the algorithm would accept the maximum value of the two *Maxcounts* values before and after the switch as the final comprehensive matching result. For the two graphs shown in Fig. 6, the final *Maxcounts* result would be 10.

TABLE 2
The Similarity Measurement Results of the CSM and VF2 Algorithms

Groups	Nodes	graph isomorphism (couples)		subgraph isomorphism (couples)		partial isomorphism (couples)	
		VF2	CSM	VF2	CSM	VF2	CSM
Group one	[10,100]	30	30	976	976	0	3944
Group two	(100,200]	19	19	691	691	0	4240
Group three	(200,300]	12	12	576	576	0	4362
Group four	(300,400]	10	10	243	243	0	4697
Group five	(400,500]	8	8	147	147	0	4795

3.3 Similarity Measurement

We then used the Jaccard similarity coefficient [38] to measure the similarity of two complex trajectories. The Jaccard similarity coefficient originally is used to measure the similarity and diversity of sample sets. It is defined as follows:

$$\mathrm{J(A,B)} = A \cap B / A \cup B = |A \cap B| / (|A| + |B| - |A \cap B|) \quad (0 \leq J(A,B) \leq 1). \tag{1}$$

In this study, $\mathrm{G_A}$ and $\mathrm{G_B}$ are the graphs representing the complex trajectories A and B, respectively. $|\mathrm{G_A}|$ and $|\mathrm{G_B}|$ are the total numbers of the nodes in $\mathrm{G_A}$ and $\mathrm{G_B}$, respectively. $\mathrm{G_A} \cap \mathrm{G_B}$ is the size of the intersection between $\mathrm{G_A}$ and $\mathrm{G_B}$, i.e., the maximum number (*Maxcounts*) of the matched nodes between the two complex trajectories. The Jaccard similarity coefficient can be rewritten as:

$$\mathrm{Similarity}(G_A, G_B) = Maxcounts / (|G_A| + |G_B| - Maxcounts) \tag{2}$$

If $\mathrm{G_A}$ and $\mathrm{G_B}$ are graph isomorphic, the Similarity $(\mathrm{G_A},\ \mathrm{G_B})$ would be 1. If $\mathrm{G_A}$ and $\mathrm{G_B}$ are subgraph isomorphic, i.e., $\mathrm{G_A}$ is a subgraph of $\mathrm{G_B}$, the Similarity $(\mathrm{G_A},\ \mathrm{G_B})$ would be $\mathrm{G_A}/\mathrm{G_B}$.

4 Method Comparison and Efficiency Analysis

We first used the synthetic graph data to evaluate and compare the performance of the CSM, the VF2, and the EGED algorithms. We then examined the computational efficiency of the CSM algorithm regards to the size and the complexity of the graphs.

4.1 Computation Complexity Evaluation

The computation time complexity of the CSM algorithm depends on two factors: the number of the nodes and the branches on the graphs. The CSM uses the same retrival method as the VF2 does. In the best-case scenario, the target and the query graphs both have N nodes and satisfy graph isomorphism and the computational complexity would be $O(N^2)$, which is the same as that of the VF2.

In the worst-case scenario, the CSM can only match up to two nodes in every round of search and it would need to explore all possible states. Assuming there are N nodes in the first graph, and after the first match, 2 nodes would be reduced from it. For the *i-th* match, the number of nodes would be $N - 2i$. According to [22], the complexity in the worst-case scenario of the VF2 is $O(N!N)$ and the total number of states would be:

$$N!N + (N-2)!(N-2) + \cdots + 2! \times 2. \tag{3}$$

It can be rewritten as:

$$\sum_{i=0}^{N} (N-2i)! \times (N-2i). \tag{4}$$

Therefore, the CSM would compare $N/2$ times to match all nodes and its computational complexity in the worst-case scenario would be $O(N!N^2)$. The worst-case complexity of VF2 is $O(N!N)$, which is better than the CSM.

4.2 Comparison of Algorithms

Experiments showed that the CSM algorithm runs very slow on the graphs with over 500 nodes (see Section 4.3), which are not very common in the real world. We generated five groups of synthetic graphs with different numbers of nodes (Table 2) and each group has roughly 100 graphs and each graph has less than 500 nodes. For each group, there are 4950 (C_{100}^2) couples that could be mutually calculated among the 100 graphs, among which graph/subgraph/partial isomorphism exists. Structurally, these synthetic graphs all have at least one split and/or merger branch, which is generated using three critical parameters, the number of nodes (N), the complexity η (see Section 2.1), and the number of graphs (z). The number of split and merger nodes, which equals to $N^*\eta$, were randomly generated. Different values of η $(0.2,\ 0.4,\ 0.6,\ 0.8)$ were used to generate graphs in different sizes (i.e., different number of nodes).

Table 2 shows the number of couples that the CSM and VF2 algorithms find isomorphic on the five groups of synthetic datasets. The CSM and VF2 identified exactly the same number of couples that are graph/subgraph isomorphic, showing that the CSM is as robust as the VF2. However, the CSM was able to find partial matching graphs whereas the VF2 can't.

We used the same synthetic graphs to compare the performance of the CSM against the EGED. The EGED measures the trajectory similarity by considering the entire structure of the graph. It computes the minimum edit operations which are needed to transform one graph into another by means of insertion, deletion and substitution of vertices and/or edges [24]. The EGED returns with and uses the edit distance to measure the similarity between two graphs.

TABLE 3
Paired Sample Correlation Analysis between the Numbers of Isomorphism that Were Identified by the CSM and EGED Algorithms

	N	Correlation	Sig.
CSM & EGED	1225	.801	.000

$$\text{Similarity}(G_A, G_B) = 1 - \textit{Edit distance}/\textit{Math.Max}(|G_A|, |G_B|) \quad (0 \le \text{Similarity}(G_A, G_B) \le 1), \tag{5}$$

where $|G_A|$ and $|G_B|$ are the total number of the nodes in graphs G_A and G_B, respectively.

We used 50 synthetic graphs with a small number of nodes to compare the performance of the CSM against the EGED as previous studies have shown that the EGED works only on rather small-size graphs [23], [24], [39] and it runs pretty slow. Totally 1225 couples of graphs were mutually calculated from the 50 synthetic graphs. The number of the nodes of these graphs ranges from 3 to 15 and the complexity η ranges from 0.2 to 1.

Tables 3 and 4 list the paired sample correlations and t-test results of the numbers of the isomorphism identified by the CSM and the EGED algorithms. The 0.801 correlation coefficient indicates that the two algorithms identified similar numbers of isomorphism from the same dataset. The t-test results show that the CSM method found a higher number of isomorphism than the EGED and the difference is statistically significant at the 0.05 confidence level.

The CSM algorithm can better reflect the similarity of the topological structure between complex trajectories than the EGED. As illustrated in Fig. 6, the CSM obtained a maximum number of matched nodes of 10 and a Jaccard similarity coefficient of 0.9. However, the edit distance of the EGED between the two graphs in Fig. 6 is three (by deleting one node and adding two nodes) and the similarity coefficient is 0.73. Such a difference is mainly due to the different ideas that are used in the CSM and the EDED to identify the similar graphs. The CSM mainly identifies all of the largest common topologies from the graphs. By contrast, the EGED determines the similarity between graphs by changing the structure of one graph into another and, as a result, the original structure of the graph might change.

The computation time of the CSM is also shorter than that of the EGED. Fig. 7 shows the computation time of the CSM and EGED when they were used to examine two synthetic graphs with a η of 0.6 and 0.8, respectively. The computation time of the EGED increases dramatically when the number of nodes in the graphs increases to more than 11. For the graphs with 11 or fewer nodes, the EGED algorithm

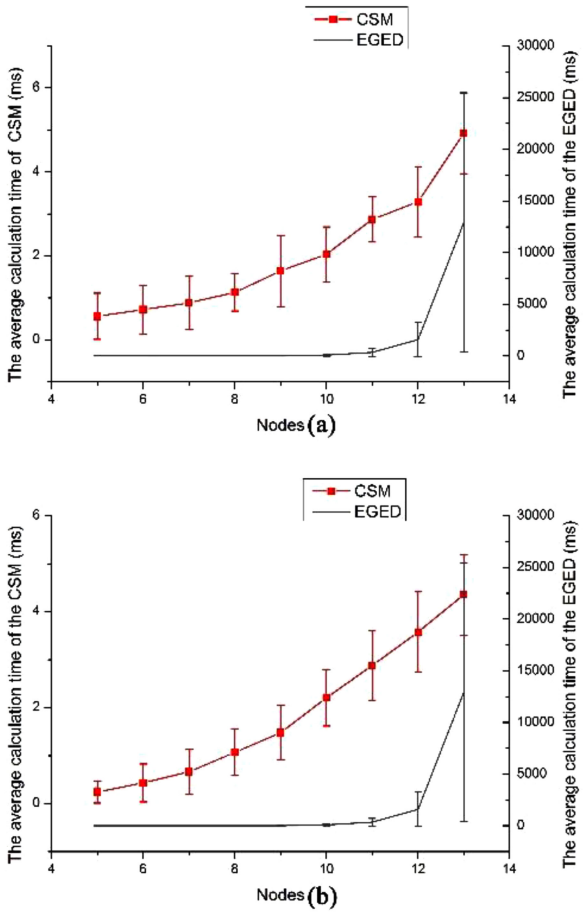


Fig. 7. The average calculation time of the EGED and CSM algorithms on graphs with (a) $\eta = 0.6$ and (b) $\eta = 0.8$.

is more than 100 or even 10,000 times slower than the CSM. In fact, our experiments also showed that the EGED algorithm can only handle the graphs with at most 15 nodes and no more than one branch $(\eta = 0.2)$.

4.3 Computation Efficiency Analysis

The performance of the CSM is also affected by the number of nodes and the complexity of the graphs. We generated nine groups of synthetic graphs with nodes from around 20 to 500 and a complexity η of 0.2, 0.4, 0.6, 0.8, respectively. For each complexity η, we used the CSM to measure the similarity of topological structures of 1,225 coupled graphs. As illustrated in Fig. 8, the computation time gradually increases as the number of nodes increases. However, the execution time decreases with increased complexity of the graphs that have the same number of nodes. In other words, in the case of the same number of nodes, the CSM finds similar trajectories

TABLE 4
Results of the Paired Sample T-Test on the Numbers of the Isomorphism that Were Identified by the CSM and EGED Algorithms

	Paired Differences					t	df	Sig. 2-tailed
	Mean	Std Deviation	Std Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
CSM -EGED	.0429	.09638	.00275	.03751	.04832	15.585	1224	.000

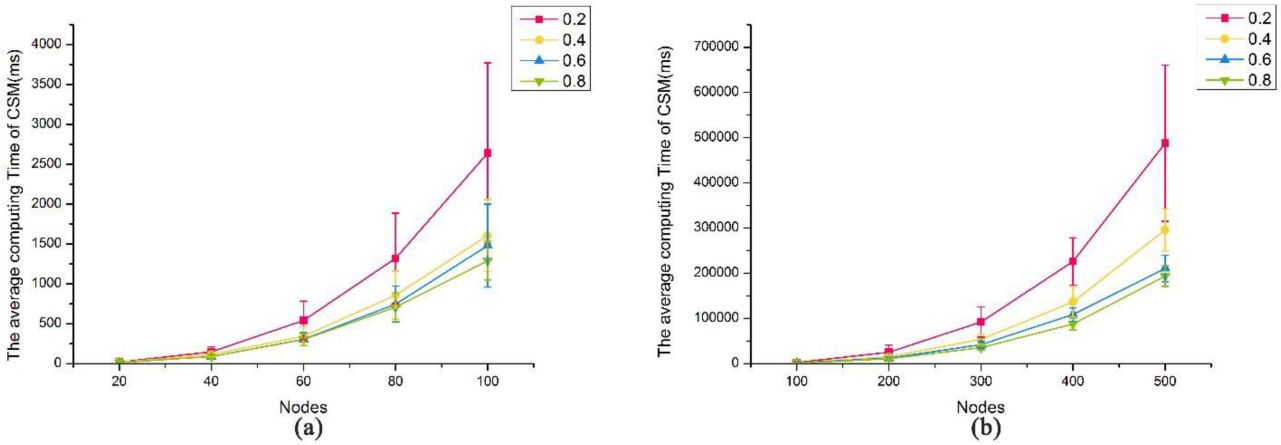


Fig. 8. The average computation time of the CSM on the graphs with the number of nodes from (a) 20 to 100 and from (b) 100 to 500.

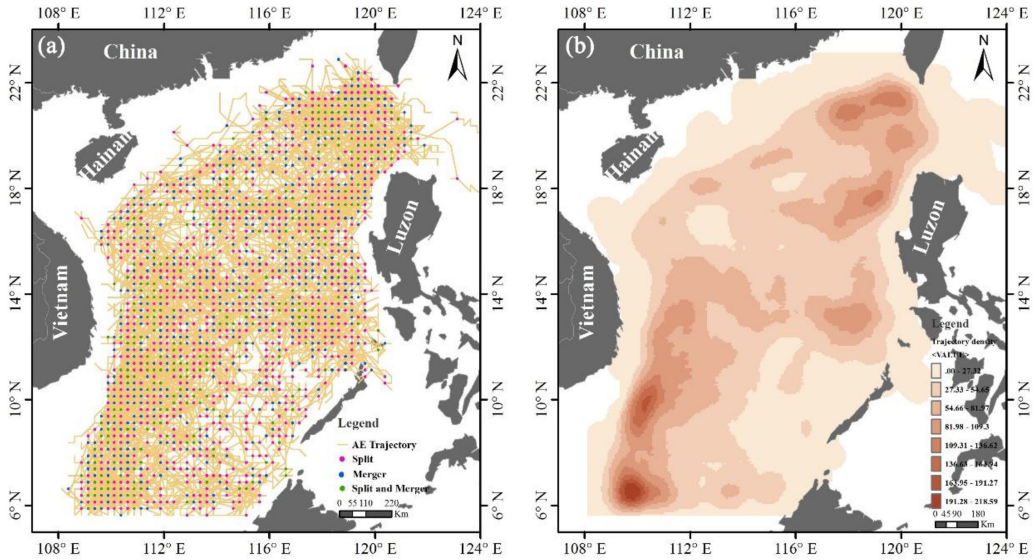


Fig. 9. (a) Distribution and (b) the line density map of the complex trajectories of the anticyclonic eddies (AE) in the SCS from 1993-2012.

faster when the trajectories have a high complexity and slower when the trajectories have a low complexity. This is because that the CSM would need to match fewer nodes each time when it is used to measure the similarity between the more complex trajectories, i.e., those with a large η. The matching process thus would end sooner. By contrast, when the CMS is used to measure the similarity between the less complex trajectories, i.e., those with a low η, it would need to match more nodes each time. Overall it would take more time to finish the matching process and find the maximum results between the less complex trajectories.

5 Applications and Discussion

Perfect isomorphism/exact match is not very common among real-world complex trajectories due to their consistent variations in time and spaces. We adopted the proposed CSM algorithm to examine the complex trajectories of mesoscale ocean eddies in the SCS, which, as illustrated in the following sections, are a perfect example to show the graph, subgraph, and partial isomorphism. The CSM also provides a new perspective to study the mesoscale ocean eddy trajectories and may reveal the common structural patterns of the trajectories.

5.1 Trajectories of Anticyclonic Eddies in the SCS

We used the CSM to examine 936 complex trajectories of the 1993 to 2012 anticyclonic eddies (AE), which at least split or merge one time during their lifespans. Fig. 9 shows the spatial distribution of the 936 AE complex trajectories (a) and the line density per square km across the SCS (b). The line density of the complex trajectories is higher around the Luzon Strait in the northern SCS, the west of the Luzon Island in the middle SCS, and the southwestern SCS.

5.2 Results and Discussion

We first used the CSM to study the complex trajectory of ACE3 in the northern SCS. The ACE3 is one of the three long-lived eddies that were reported by Nan et al. (2011). During its whole lifespan, the ACE3 splits and merges multiple times, which have been confirmed by in-situ observations [12].

The trajectory with an ID of 22670 in our database matches that of the ACE3. Both started near the northwest of the Luzon Island, then turned toward west, and finally decayed near the east of the Hainan Island. There is also difference between the two trajectories (Fig. 10a) mainly due to the different temporal scales of the SLA maps from which they were reconstructed. The trajectory 22670 in our

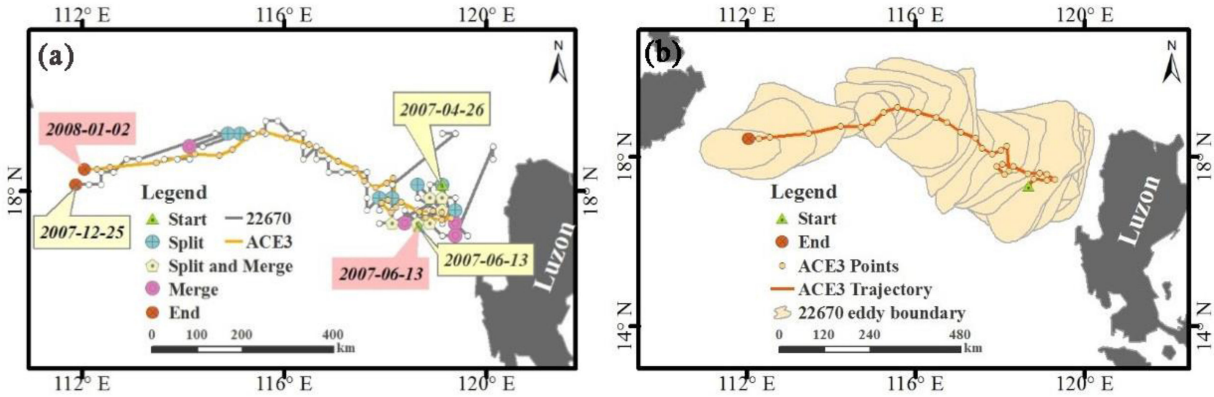


Fig. 10. The ACE3 trajectory and the trajectory 22670.

database and the ACE3 trajectory [12] were reconstructed from the daily and weekly SLA maps, respectively. Furthermore, the trajectory 22670 and the ACE3 trajectory are found at the same position on April 26, 2007 and on June 13, 2007 only. When the two trajectories don't perfectly line up, the points along the ACE3 trajectory are all within the outmost boundaries of the eddies that generated the trajectory 22670 at the corresponding time (Fig 10b). Meanwhile, the average Euclidean distance of the corresponding line segments between the two trajectories is 0.3, which is the least among the Euclidean distances that we can find between the ACE3 trajectory and any other trajectories in our database. As a result, we believe that the trajectory 22670 and the ACE3 trajectory are generated by the exactly same anticyclonic eddy in the SCS.

The CSM algorithm found 935 trajectories from our database that are similar to the trajectory 22670. Table 5 listed the top 20 most similar trajectories. All the top 20 trajectories were generated by long-lived anticyclonic eddies that last more than 110 days. During the period from 1992 to 2012, the CSM found at least one top 20 trajectory that is very similar to the trajectory 22670 every year except for 1999, 2002 and 2009 (Table 5), indicating that a ACE3-like eddy may develop in the northern SCS every year. The trajectories in 1992, 2002, and 2009 have 3 to 34 nodes whereas the trajectory 22670 has 64 nodes. The similarity coefficients between the complex trajectories in these three years and the trajectory 22670 are below 0.53 (the ratio between 34 and 64 in the case of subgraph isomorphism), which is less than the threshold value of 0.58 that we used to select the top-20 trajectories (Table 5).

TABLE 5
The Top 20 Most Similar Trajectories

Trajectory ID	Starting time-Ending time	Lifespan (days)	Similarity coefficient
30895	2012/05/16-2012/11/14	182	0.78
16858	2003/01/24-2004/04/13	414	0.71
1878	1994/04/04-1994/10/04	145	0.69
2989	1995/01/14-1995/06/10	185	0.68
11536	2000/08/28-2001/01/09	134	0.67
5500	1996/10/05-1997/05/11	218	0.67
18404	2004/09/26-2005/02/21	110	0.66
25308	2008/12/05-2009/04/06	122	0.65
7840	1998/04/25-1998/09/04	132	0.65
20454	2005/12/08-2006/04/15	128	0.64
6722	1997/07/26-1998/03/02	218	0.63
8449	1998/09/16-1999/03/03	168	0.62
12138	2001/01/05-2001/05/02	117	0.62
23529	2007/10/29- 2008/03/24	147	0.60
28268	2010/09/28-2011/01/21	115	0.60
12268	2001/02/02- 2001/06/07	125	0.60
17002	2003/11/27- 2004/05/09	164	0.59
29133	2011/03/31-2011/10/07	200	0.59
530	1993/04/16-1993/11/04	202	0.58
21023	2006/04/10-2006/08/28	140	0.58

Fig. 11 specifically illustrates each of the top 20 trajectories in Table 5 in comparison against the trajectory 22670. All the 20 trajectories are located in the northern SCS and shifted along the 18° N parallel. Along this parallel, 14 in situ observational stations have been deployed to explore the vertical structures of the eddies. Some argued that it is not easy to find an eddy that is cut through by the in situ observation section [12] [32]. The CSM has identified 5 anticyclonic eddy trajectories along the 18° N parallel in different years (Figs. 12-7, 12-8, 12-10, 12-11, and 12-18). One or more of them may be cut through by the in situ observation section. If that is the case, scientists would be able to study its vertical structures, which are important in unravelling the evolution of mesoscale ocean eddies in the northern SCS (e.g., [12], [34], [36]).

All the top 20 trajectories show a similar migration pattern as that of the trajectory 22670, starting near the west of the Luzon Island or the Luzon Strait, mainly propagating westward, then migrating along the SCS northern continental slope, and finally disappearing in the northwest SCS. Such a pattern is not different from what has been reported [40] - most mesoscale eddies in the SCS mainly propagate westward. However, one trajectory with an ID of 7840 in 1998 was found moving from north to south (Fig. 11-6). The trajectory 7840 started in spring and moved toward south mainly in summer. This might be caused by the remarkable El Nino in the summer of 1998, as a strong El Nino event could change the normal structure of the circulation field in the SCS and lead to the formation of multi-structure of the anticyclonic eddies in the west of the Luzon Island [36], [41].The trajectory density map (per 1/4 degrees) (Fig. 12a) shows a dominant migrating path along the northwest SCS continental shelf, where similar structures are commonly found. The path is consistent with that has been reported in previous studies (e.g., [35], [36], [42]). Fig 12b shows the

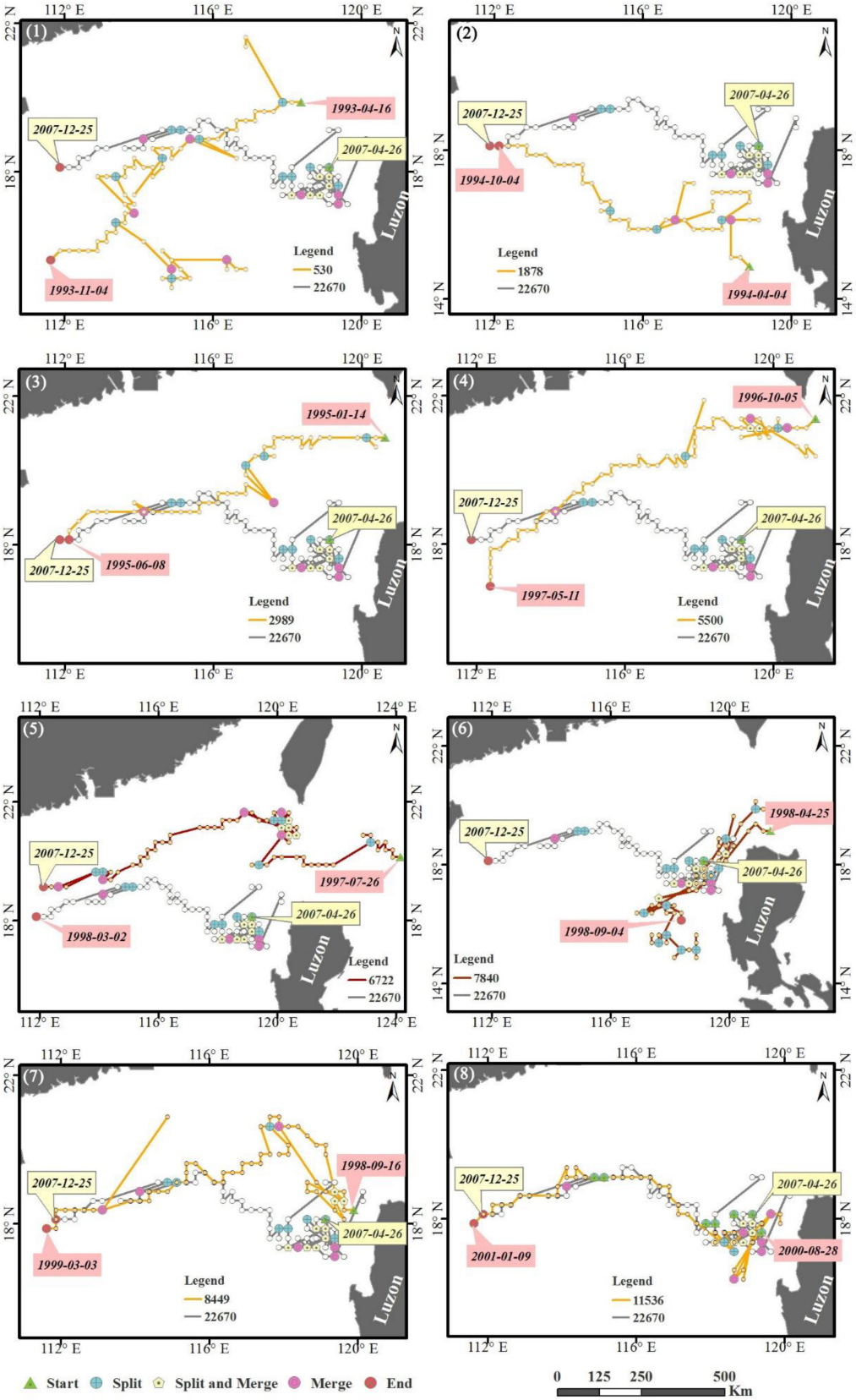


Fig. 11a. The top 20 trajectories that are most similar to the trajectory 22670.

density of the nodes, indicating where the top 20 complex trajectories split and merge. The split and merger activities mainly concentrated in the west of the Luzon Strait (R2) and the northwest of the Luzon Island (R3). About 12.21, 41.27, 37.79 percent of the nodes are related to the split and merger activities in the R1, R2, and R3 regions, respectively.The three regions (R1, R2, and R3) in Fig. 12b show dramatical difference in terms of trajectory topological structures.

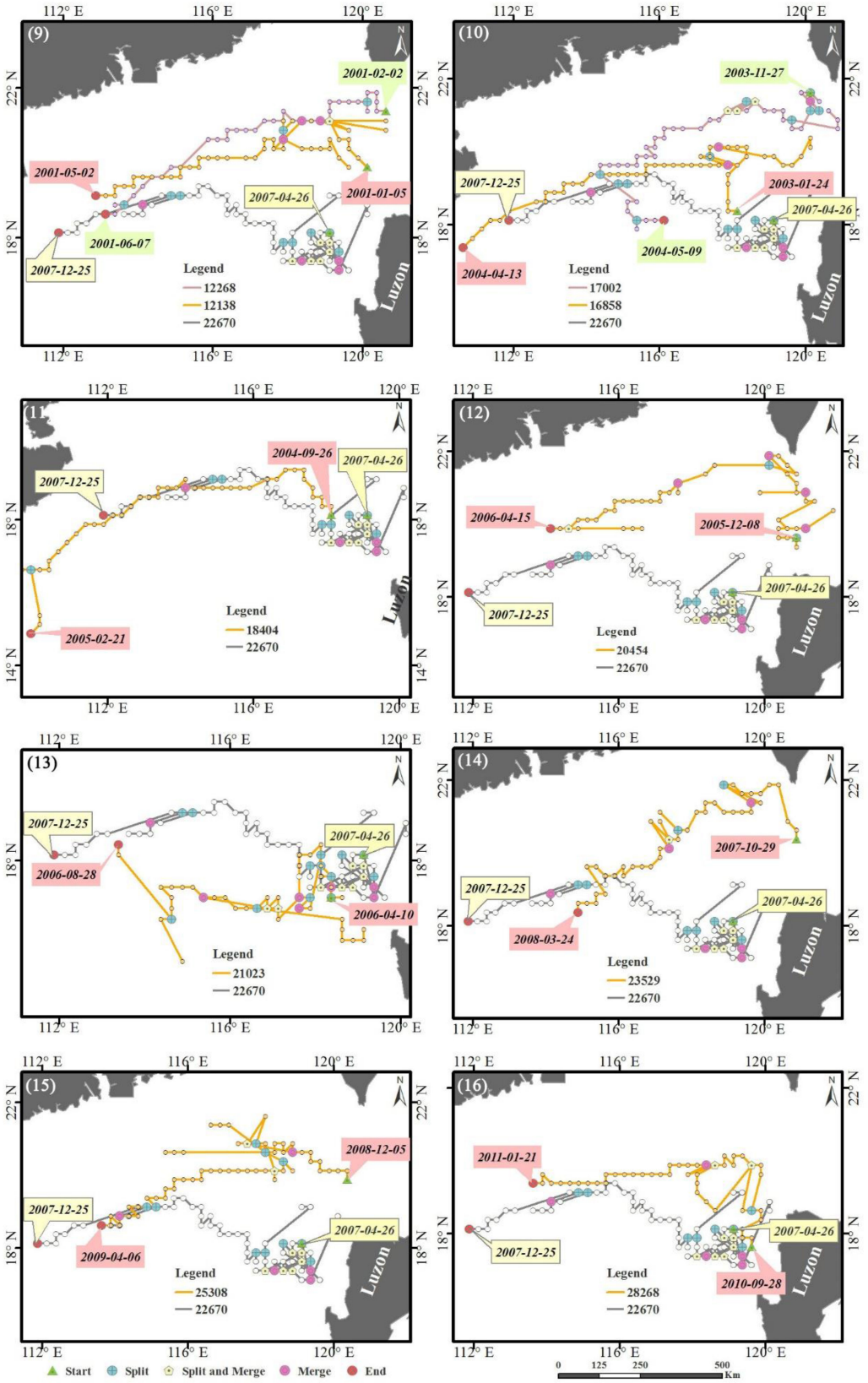


Fig. 11b. (continued).

Topologically linear trajectories are found in region R1, which is located in the east of the Hainan Island and along the SCS northern continental slope. In region R1, most split and merger events occurred in autumn and winter. The trajectories in region R3 show a split structure as there are two times of splits than mergers in this region. The trajectories in region R2 show very complex topological structures, usually starting with multiple linear structures then experiencing multiple merger and split processes. The ratio between the numbers of split and the merger events is 1.2 and most

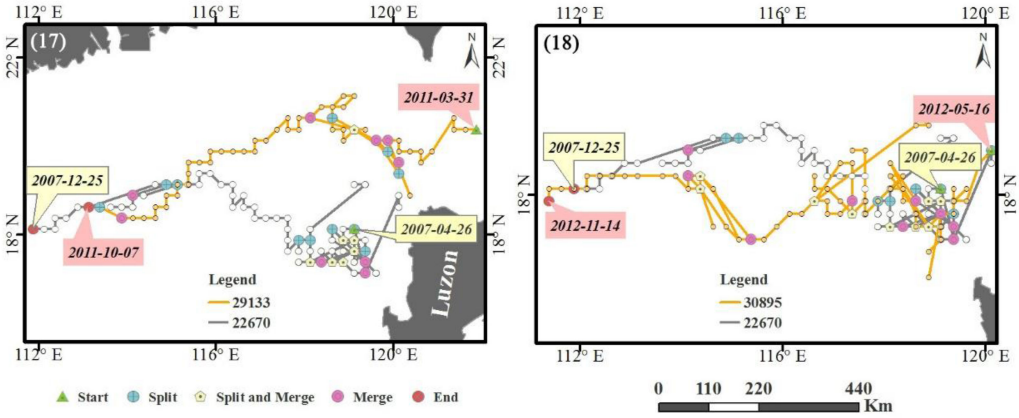


Fig. 11c. (continued).

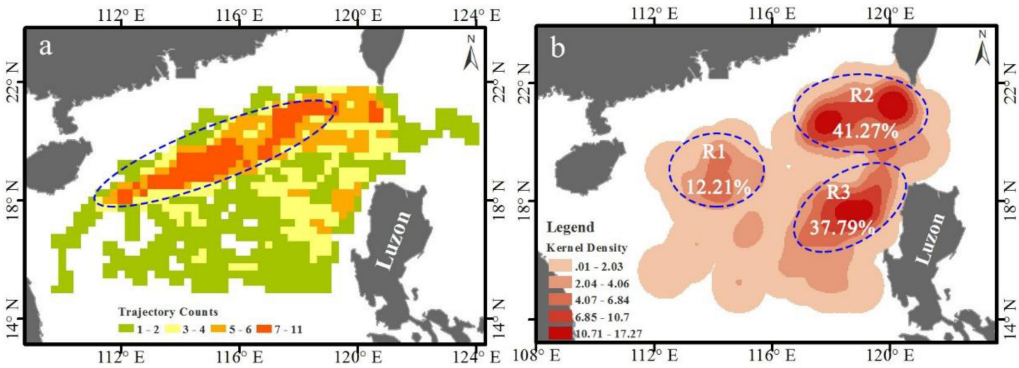


Fig. 12. The density map (a) of the top 20 trajectories and (b) of the nodes that suggest split and merge events along the top 20 trajectories.

of the splits and mergers in R2 occurred in autumn and winter.

In this study, we used the CSM framework to examine the large-scale complex trajectories of mesoscale ocean eddies and discovered new knowledge about the evolution of the trajectories. We believe that the method could also be applied to the motion at a small or even micro spatial scale. A single-cell trajectory may also have multiple branches [43] and the related studies of complex single-cell trajectories may reveal rare cell types and cryptic states [44]. Our method may be useful in identifying abnormal changes in gene expression and allelic imbalances by examining the similarity of the topological structures between complex single-cell trajectories. In the field of gesture modeling in multimedia, studies have shown that the hand motions could be represented by certain topological structures [45]. It is possible to use our method to measure the similarity of the topological structures between a hand gesture and the known gesture categories.

6 Conclusions

We proposed the CSM algorithm to measure the similarity of topological structures between complex trajectories, which represent the paths of the physical features that may split and merge while they are moving within a space. We used synthetic trajectory datasets to evaluate the CSM algorithm and compare its performance against that of the VF2 and the EGED algorithms. Results show that the CSM is successful in identifying the partial isomorphism of the topological structures between graphs. The CSM algorithm also outperforms the EGEG algorithms in terms of the computation efficiency in measuring the similarities between graphs.

We applied the CSM algorithm to study the complex trajectories of ocean eddies in the SCS. The algorithm successfully found complex trajectories that are similar to the trajectory of the well-studied ACE3 in the northern SCS. The similar complex trajectories identified by the CSM show a dominant westward migrating path of the ocean eddies in the northern SCS, which is consistent with previous research. The CSM also helps identify some common topological structures in the complex trajectories. These findings provide a new perspective to study the mesoscale eddies in the SCS.

It is also worth noting that the CSM algorithm currently only considers the topological structure when it is used to measure the similarity of trajectories. In the near future, we would incorporate other spatial and temporal factors such as the labels of the nodes and edges in the graph structure to measure the similarity between complex trajectories.

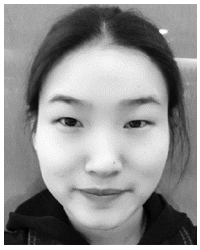
Acknowledgments

This work was supported in part by a grant from the National Science Foundation of China (41671445), National Key R&D Program of China (2017YFB0503605), and the National Science Foundation of China (41471330).

References

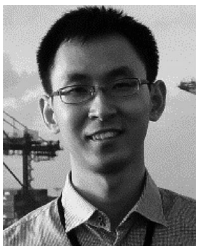
[1] M. C. González, C. A. Hidalgo, and A. L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–778, 2008.

[2] S. Dodge, P. Laube, and R. Weibel, "Movement similarity assessment using symbolic representation of trajectories," *Int. J. Geographical Inf. Sci.*, vol. 26, no. 9, pp. 1563–1588, 2012.
[3] T. Pei, S. Sobolevsky, C. Ratti, and S.-L. Shaw, "A new insight into land use classification based on aggregated mobile phone data," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 9, pp. 1988–2007, 2014.
[4] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–14, 2015.
[5] Y. Yuan and M. Raubal, "Measuring similarity of mobile phone user trajectories–a Spatio-temporal edit distance method," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 3, pp. 496–520, 2014.
[6] H. Wang, H. Wen, F. Yi, H. Zhu, and L. Sun, "Road Traffic anomaly detection via collaborative path inference from GPS Snippets," *Sens.*, vol. 17, no. 3, 2017, Art. no. 550.
[7] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*, New York, NY, USA: Taylor & Francis, pp. 352–366, 2001.
[8] H. P. Tsai, D. N. Yang, and M. S. Chen, "Mining group movement patterns for tracking moving objects efficiently," *IEEE Trans. Knowl. Data Eng.*, vol. 283, no. 2, pp. 266–281, Feb. 2010.
[9] A. C. Liefbroer and C. H. Elzinga, "Intergenerational transmission of behavioral patterns: how similar are parents' and children's demographic trajectories?" *Advances Life Course Res.*, vol. 17, no. 1, pp. 1–10, 2012.
[10] M. Buchin, S. Dodge, and B. Speckmann, "Context-aware similarity of trajectories," *Proc. Int. Conf. Geographic Inf. Sci.*, 2012, pp. 43–56.
[11] K. Buchin, M. Buchin, and M. V. Kreveld, "Finding long and similar parts of trajectories," *Comput. Geometry Theory Appl.*, vol. 44, no. 9, pp. 465–476, 2011.
[12] F. Nan, Z. He, H. Zhou, and D. Wang, "Three long-lived anticyclonic eddies in the northern South China Sea," *J. Geophysical Res.-Oceans*, vol. 116, no. c5, pp. 879–889, 2011.
[13] W. Liu, X. Li, and D. A. Rahn, *Storm Event Representation and Analysis Based on A Directed Spatiotemporal Graph Model*, New York, NY, USA: Taylor & Francis pp. 1–14, 2016.
[14] H. Alt, "The computational geometry of comparing shapes," *Efficient Algorithms*, vol. 5760, pp. 235–248, 2009.
[15] J. Chen, R. Wang, and L. Liu, "Clustering of trajectories based on Hausdorff distance," in *Proc. IEEE Int. Conf. Electron. Commun. Control*, pp. 1940–1944, 2011.
[16] D. UrKa, B. Kevin, and C. Francesca, "Analysis and visualisation of movement: An interdisciplinary review," *Movement Ecology*, vol. 3, no. 1, 2015, Art. no. 5.
[17] A. Kassidas, J. F. Macgregor, and P. A. Taylor, "Synchronization of batch trajectories using dynamic time warping," *Aiche J.*, vol. 44, no. 4, pp. 864–875, 1998.
[18] N. Vaughan and B. Gabrys, "Comparing and combining time series trajectories using dynamic time warping," *Procedia Comput. Sci.*, no. 96, pp. 465–474, 2016.
[19] M. Sharif and A. A. Alesheikh, "Context-awareness in similarity measures and pattern discoveries of trajectories: A context-based dynamic time warping method," *Mapping Sci. Remote Sens.*, vol. 54, pp. 426–452, 2017.
[20] N. Pelekis, Andrienko, G. Andrienko, and N. Kopanakis, "Visually exploring movement data via similarity-based analysis," *J. Intell. Inf. Syst.*, vol. 38, no. 2, pp. 343–391, 2012.
[21] H. Bunke and K. Shearer, "A graph distance metric based on the maximal common subgraph," *Pattern Recognit. Letters*, vol. 19, no. 3-4, pp. 255–259, 1998.
[22] L. P. Cordella, P. Foggia, and C. Sansone, "A (sub)graph isomorphism algorithm for matching large graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1367–1372, Oct. 2004.
[23] K. Riesen and H. Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," *Image Vis. Comput.*, vol. 27, no. 7, pp. 950–959, 2009.
[24] Z. Abu-Aisheh, R. Raveaux, and J. Y. Ramel, "An exact graph edit distance algorithm for solving pattern recognition problems," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 271–278.
[25] K. Riesen, *Structural Pattern Recognition with Graph Edit Distance*. New York, NY, USA: Springer, 2016.
[26] A. Fischer, C. Y. Suen, and V. Frinken, "Approximation of graph edit distance based on Hausdorff matching," *Pattern Recognit.*, vol. 48, no. 2, pp. 331–343, 2015.
[27] P. Foggia, C. Sansone, and M. Vento, "A performance comparison of five algorithms for graph isomorphism," in *Proc. 3rd IAPR TC-15 Workshop Graph-Based Representations Pattern Recognit.*, 2001, pp. 188–199.
[28] J. Yi, Y. Du, and F. Liang, "A representation framework for studying spatiotemporal changes and interactions of dynamic geographic phenomena," *Int. J. Geographical Inf. Sci.*, vol. 28, no. 5, pp. 1010–1027, 2014.
[29] D. K. Adams and L. S. Mullineaux, "Surface-generated mesoscale eddies transport deep-sea products from hydrothermal vents," *Sci.*, vol. 332, no. 6029, 2011, Art. no. 580.
[30] F. Fang and R. Morrow, "Evolution, movement and decay of warm-core Leeuwin Current eddies," *Deep-Sea Res. Part II*, vol. 50, no. 12, pp. 2245–2261, 2003.
[31] R. Rodríguez, A. Viudez, and S. Ruiz, "Vortex merger in oceanic tripoles," *J. Phys. Oceanography*, vol. 41, no. 6, pp. 1239–1251, 2011.
[32] D. Wang, H. Z. Xu, J. Lin, and J. Y. Hu, "Anticyclonic eddies in the northeastern South China Sea during winter 2003/2004," *J. Oceanography*, vol. 64, no. 6, pp. 925–935, 2008.
[33] G. X. Chen, Y. J. Hou, and X. Q. Chu, "Vertical structure and evolution of the Luzon Warm Eddy," *Oceanol. Limnol.*, vol. 28, no. 5, pp. 955–961, 2010.
[34] G. Chen, Y. Hou, and X. Chu, "Mesoscale eddies in the South China Sea: Mean properties, spatiotemporal variability, and impact on thermohaline structure," *J. Geophysical Res.*, vol. 116, no. c6, pp. 1–20, 2011.
[35] Y. Du, J. Yi, and W. Di, "Mesoscale oceanic eddies in the South China Sea from 1992 to 2012: evolution processes and statistical analysis," *Acta Oceanologica Sinica*, vol. 33, no. 11, pp. 36–47, 2014.
[36] X. Wang, W. Fang, and R. Chen, "The summer-fall anticyclonic eddy west of Luzon: Structure and evolution in 2012 and interannual variability," *J. Marine Syst.*, vol. 172, pp. 84–92, 2017.
[37] N. Ducet, P. Y. Le Traon, and G. Reverdin," Global high-resolution mapping of ocean circulation from TOPEX/Poseidon and ERS-1 and -2," *J. Geophysical Res. Oceans*, vol. 105, no. c8, pp. 19477–19498, 2000.
[38] P. Jaccard, "The distribution of the Flora in the Alpine Zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[39] X. Gao, B. Xiao, and D. Tao, "A survey of graph edit distance," *Pattern Anal. Appl.*, vol. 13, no. 1, pp. 113–129, 2010.
[40] P. Xiu, F. Chai, L. Shi, H. Xue, and Y. Chao, "A census of eddy activities in the South China Sea During 1993-2007," *J. Geophysical Res.-Oceans*, vol. 115, no. c3, pp. 1–15, 2010.
[41] W. D. Fang, J. J. Guo, and P. Shi, "Low frequency variability of South China Sea surface circulation from 11 years of satellite altimeter data," *Geophysical Res. Letters*, vol. 33, no. 22, pp. 2103–2112, 2006.
[42] D. Yuan, W. Han, and D. Hu, "Anti-cyclonic eddies northwest of Luzon in summer-fall observed by satellite altimeters," *Geophysical Res. Letters*, vol. 34, no. 13, pp. 256–260, 2007.
[43] X. Qiu, Q. Mao, and Y. Tang," Reversed graph embedding resolves complex single-cell trajectories," *Nature Methods*, vol. 14, no. 10, pp. 979–982, 2017.
[44] X. Qiu, A. Hill, and J. Packer," Single-cell mRNA quantification and differential analysis with Census," *Nature Methods*, vol. 14, no. 3, 2017, Art. no. 309.
[45] D. Tang, H. J. Chang, and A. Tejani, "Latent regression forest: Structured estimation of 3D hand poses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1374–1387, Jul. 2017.

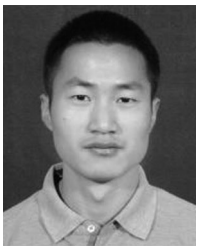


Huimeng Wang received the MS degree in geographic information science from the Shandong University of Science and Technology, in 2016. She is curremtly working toward the PhD degree in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. Her research interests include the development and application of geographic information science and spatiotem-poral data mining.

Yunyan Du is currently a professor with the State Key Laboratory of Resource and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS. Her main research activity is in the area of remote sensing and tempo–spatial modeling. She is a referee for *Deep Sea Research II*, *Environment and Urban Planning*, *PLUS One*, and *the Journal of Geographical Science*. She was the recipient of the Second National Advanced Prize of Science and Technology.



Jiawei Yi received the BS degree in geographic information science from Wuhan University, China, in 2010, and the MS degree in geographic information science from the University of Chinese Academy of Sciences, in 2013. He is currently working toward the PhD degree in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. His research interests include space-time representation and spatio-temporal data mining.



Yong Sun is currently working toward the PhD degree at the Shandong University of Science and Technology, Qingdao, China. His research interests include the development and application of geographic information science and spatio-tem-poral data mining.

Fuyuan Liang received the PhD degree in geography from the University of Georgia, Athens, GA, USA, in 2008. He is an associate professor of geography with Western Illinois University, Macomb, IL, teaching courses of geographic information science, remote sensing, and Landforms. His current research interests focused on geomorphometric and hydrological characteristics of karst landforms using GIS and remote sensing techniques.