A PROCESS-ORIENTED SPATIOTEMPORAL CLUSTERING METHOD FOR COMPLEX TRAJECTORIES

Jingyi Liu [1], Cunjin Xue [1*], Chengbin Wu [1] , Qing Dong [1]

[1] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China – (liujy2017, xuecj, wucb, dongq) @radi.ac.cn

Commission III, WG III/6

KEY WORDS: Spatiotemporal data mining, Complex trajectories, Clustering, Process-based, Marine abnormal variations

ABSTRACT:

Considering the critical role of trajectory data in Big Data era for dynamic geographical processes, human behaviour analysis and meteorological prediction, trajectory clustering has attracted growing attention. Many literatures have discussed the spatiotemporal clustering method of simple trajectories (i.e., has no branches, e.g. vehicle trajectories), yet there are few researches for clustering complex trajectories (i.e., has at least one split and/or merger and/or split-merger branch, e.g. ocean eddy trajectories, rainstorm trajectories). For addressing this issue, we propose a Process-Oriented Spatiotemporal Clustering Method (POSCM) for clustering complex trajectory data. The POSCM includes three parts: the first uses the semantic of process-sequence-state to represent the complex trajectories; the second proposes a Hierarchical Similarity Measurement Method (HSMM) to get the similarity between any two complex trajectories; in the last step, the complex trajectories clustering pattern is extracted through density-based clustering algorithm. Experiments on simulated trajectories are used to evaluate the POSCM and demonstrate the advantage by comparing against that of the VF2 algorithm. The POSCM is applied to the sea surface temperature abnormal variations trajectories from January 1950 to December 2017 in the Pacific Ocean. As shown in this case study, some new mined spatiotemporal patterns can provide new references for understanding the behaviours of marine abnormal variations under the background of the global change.

1. INTRODUCTION

Trajectories are a sequence recording the location and time of a moving object. Advanced data acquisition technologies provide a foundation for obtaining trajectories of different moving objects. In order to discover hidden and unknown knowledge from massive trajectory data, data mining is needed as the analysis method. As an important component of data mining tools, clustering analysis aims to find clusters of the same properties (Jain and Dubes, 1988). Interesting associations between global distribution patterns and data attributes can be found from the understanding of trajectory clusters (Hong et.al., 2017; Zheng et.al., 2018).

Most clustering studies focus on the trajectories which the spatial position and thematic attributes of objects are constantly changing, but its structure does not change during its lifespan, e.g., vehicle trajectories, this study calls them simple trajectories. In contrast, there is another kind of trajectories in the real world, which are produced by objects or phenomena with splitting, merging or splitting-merging behaviours during its lifespan (as shown in Figure 1), e.g., ocean eddy trajectories (Nan et al., 2011), rainstorm trajectories (Liu et al., 2014), this study calls them complex trajectories. However, the research on clustering of complex trajectories is relatively few, the reasons are as follows: (1) the traditional representation model has difficulty to represent complex trajectories with split and/or merger and/or split-merger branches; (2) the similarity measurement approaches of simple trajectories no longer applies for measuring the structural characteristics of complex trajectories.

For this issue, this paper presents a Process-Oriented Spatiotemporal Clustering Method (POSCM), which consists of new representation method and a new similarity measuring approach for complex trajectories, for clustering complex trajectories. The simulated complex trajectories dataset and the sea surface temperature abnormal variations trajectories from January 1950 to December 2017 in the Pacific Ocean are used to evaluate POSCM.

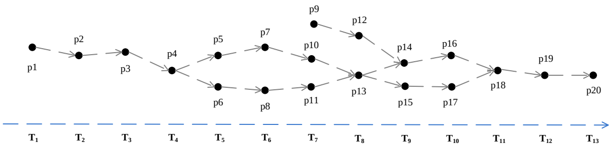


Figure 1. An example of complex trajectory.

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLII-2-W13-1237-2019

2. RELATED WORKS

2.1 Representation of trajectories

A compact and robust representation of the trajectories is important for mining the spatiotemporal movement patterns (Faisal et al., 2007). The conventional representation models for simple trajectories can be divided into three categories: vector, probabilistic distribution, and graph. The vector representation model is most commonly used, which is a series of points in a chronological order (Pakhira et al., 2004; Nanni and Pedreschi, 2007). The second representation model treats trajectories as samples of a probabilistic distribution, which can use distribution matching method to get the similarity (Li et al., 2006). The trajectories are represented as *G (N, E)* based on graph theory, where *N* represents a trajectory recording point, and *E* represents relationship edge between trajectory recording points (Mahrsi and Rossi, 2013; Wang et al., 2018). Obviously, the former two models are difficult to represent complex trajectories. The graph structure can well represent complex trajectories (Wang et al., 2018), but each trajectory point is contained in the graph, which is complicated and unnecessary for trajectory similarity measuring.

2.2 Similarity measurement for trajectories clustering

For measuring the similarity of the simple trajectories, there are many methods which consists of full-interval similarity measure, subinterval similarity measure, and single point similarity measure. The common measurements of the first kind methods are the Euclidean distance (Pakhira et al., 2004; Nanni and Pedreschi, 2007), the Hausdorff distance (Chen et al., 2011; Liu et al., 2014), Dynamic Time Warping (DTW) distance (Zhu and Zhu, 2017), or Minimum Boundary Rectangle (MBR) distance (Anagnostopoulos et al., 2006). The edit distance (Pelekis et al., 2012) and Longest Common Subsequence (LCSS) distance (Xie et al., 2017) are common measurements of the second methods. And the Fréchet distance (Brakatsoulas et al., 2005) is the single point similarity measurement.

It is difficult to apply these methods directly to the similarity measurement of complex trajectories. On the basis of graph matching theory, VF2 (Cordella et al. 2004) and graph edit distance (Riesen and Bunke, 2009) which measure the similarity between graphs, provided a new perspective to measure the similarity between complex trajectories. Wang et al (2018) have proposed a method for measuring topological structure similarity between complex trajectories inspired by VF2. However, there is still a lack of comprehensive measurement methods to measure the time, space, thematic attributes and structural features of complex trajectories.

3. METHODOLOGY

For mining cluster pattern of complex trajectories, we propose a Process-Oriented Spatiotemporal Clustering Method (POSCM), which includes the representation of complex trajectories, the construction of similarity measurement method, and clustering pattern extraction. Figure 2 shows the overall framework of POSCM. The following sections give a detailed description of key parts of the POSCM.

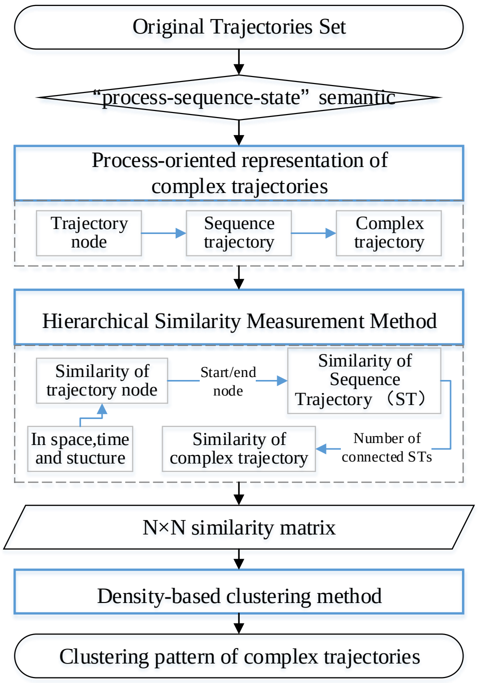

Original Trajectories Set
"process-sequence-state" semantic
Process-oriented representation of complex trajectories
Trajectory node
Sequence trajectory
Complex trajectory
Hierarchical Similarity Measurement Method
Similarity of trajectory node
Start/end node
Similarity of Sequence Trajectory (ST)
In space,time and stucture
Similarity of complex trajectory
Number of connected STs
N×N similarity matrix
Density-based clustering method
Clustering pattern of complex trajectories

Figure 2. The overall framework of the POSCM.

3.1 Process-oriented representation of complex trajectories

Inspired by a geographical process semantics (Yuan 2001; Xue et al., 2012; Yi et al., 2014), we proposed a new representation method for complex trajectories, i.e., the Process-oriented representation. According to the "process-sequence-state" semantic, the complex trajectories are abstracted into a process with several sequence trajectories i.e., *Tra (S1, S2, …Sn)*. And each sequence trajectory consists of a start node and an end node i.e., S_i *(SN, EN)*, where the start node is the start point, split point, merger point, or split-merger point of trajectory; the end node is the end point, split point, merger point, or split-merger point of trajectory. The start/end nodes (called trajectory node) are the basic element which depicts the spatial, temporal, and structure characteristics of the trajectory i.e., *SN/EN (x, y, T, nodetype, attribute)*, where the *nodetype* has five types: Production, Termination, Merging, Splitting, and Merging-splitting.

We use the complex trajectory shown in Figure 1 to illustrate the aforementioned representation method. Suppose the complex trajectory is *Tra*, where p_i is the trajectory point in T_i, and p1 and p9, p20 are the start points and end point, respectively. Based on the evolution characteristic of *Tra*, the sequence trajectories of *Tra* are as follows: $S_1(p_1, p_4)$, $S_2(p_4, p_{13})$, $S_3(p_{13}, p_{14})$, $S_4(p_{13}, p_{18})$, $S5(p_{14}, p_{18})$, $S6(p_{18}, p_{20})$. Taking p_4 as an example, the start/end nodes of each sequence trajectory can be represented as: $p_4(x_4, y_4, T_4,$ *Splitting)*. Therefore, the complex trajectory *Tra* can be represented as *Tra* $(S_1, S_2, S_3, S_4, S_5, S_6)$, and the spatiotemporal and structure features are implicit in the start/end nodes of sequential trajectories.

3.2 Hierarchical similarity measurement method

Based on the new representation of complex trajectories, a Hierarchical Similarity Measurement Method (HSMM) was developed for measuring the similarity of spatiotemporal structure and thematic characteristics between complex trajectories. The HSMM uses a hierarchical thought "similarity of trajectory node → similarity of sequence trajectories → similarity of complex trajectory" to get the similarity matrix between any two complex trajectories.

The first level is the similarity measurement for trajectory node (i.e., the start/end node as mentioned above). Euclidean distance is used to calculate spatial proximity based on spatial attribute (x, y). Considering the periodicity and temporal multiple scale of phenomena, the temporal similarity measurement must remove the influence of periodicity on the same time scale. The difference of thematic attributes is directly measured according to (*attributes*).

The second level of HSMM is for the sequence trajectories, which composes the basic spatial structure of complex trajectories. Each sequence trajectory consists of a start node and an end node, which is similar to Origin-Destination (OD) flow data (Zhu et al., 2018; Song et al., 2018). Based on the first level of HSMM, two sequence trajectories are more spatiotemporally similar if the spatial distance between their start nodes and their end nodes are smaller, and the periods during which they occur overlap more. If the *nodetype* of start nodes in two sequence trajectories is the same, and so as the end nodes, the spatial structure of these two sequence trajectories is similar.

The last level of HSMM measures the similarity between complex trajectories. Based on the similar pairs of sequence trajectories, we find the connected similarity structure, where each sequence trajectory can connect to the others. The maximum number of the sequence trajectories in the connected similarity structures is denoted by *MaxNum*, the similarity of two complex trajectories Tra_m (the number of sequence trajectories is *M*) and Tra_n (the number of sequence trajectories is *N*) can be defined as follows:

$$Similarity\ (Tra_m, Tra_n) = MaxNum\ /\ (M + N - MaxNum) \qquad (1)$$

3.3 Density-based clustering for pattern extraction

The density-based clustering algorithms, e.g., the DBSCAN, are widely used to extract trajectory clusters (Nanni and Pedreschi, 2006; Zhu and Guo 2014). The concepts of DBSCAN (i.e., core trajectory, density, reachability, etc) for complex trajectories are redefined and the two critical parameters: similarity threshold ε and density core threshold *Minpt*, are determined according to Nanni and Pedreschi (2006).

Based on the $n \times n$ similarity matrix from HSMM (n is the number of complex trajectories), ε and *Minpt*, the clustering cores and the density of each trajectory are obtained. We can extract clustering pattern of complex trajectories by connecting the clustering cores and their neighbours according to their density connectivity.

4. RESULTS AND DISCUSSION

4.1 Experiments on simulated datasets

For evaluating POSCM and demonstrating the advantage by comparing against that of the VF2 algorithm, we use simulated complex trajectories which contains noise (shown in Figure 3) for clustering experiment. VF2 is a graph-match algorithm by examining graph and subgraph isomorphism (Cordella et al. 2004). The clustering step of VF2 algorithm is implemented in the same steps as POSCM. The parameters of the two algorithms are the same.

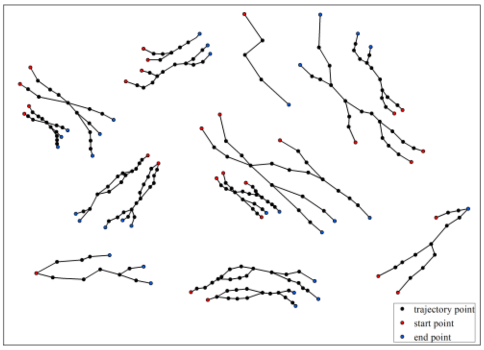


Figure 3. Simulated complex trajectories.

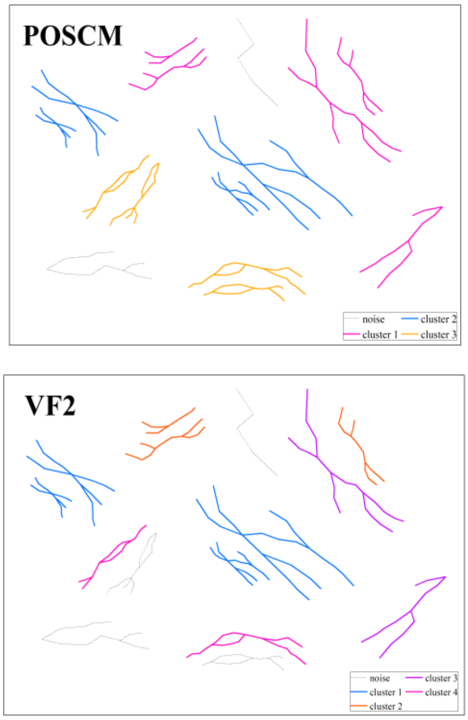


Figure 4. Clustering results based on structural similarity.

Figure 4 shows the clustering results of the two methods based on structural similarity. As shown in Figure 4, POSCM has better clustering performance in mining clustering patterns of complex trajectories with similar structures than VF2 algorithm, which is because that VF2 failed to detect the "partial isomorphism" (Wang et.al., 2018), and it requires that the in/out degree of two matching nodes must be consistent. Because of the comprehensive consideration of multi-attribute similarity in HSMM, we can also get clustering results of POSCM based on structural and spatial similarity (shown in the Figure 5).

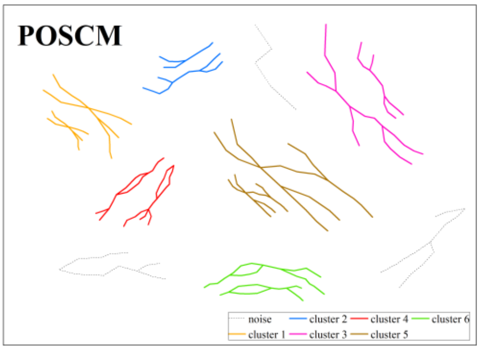
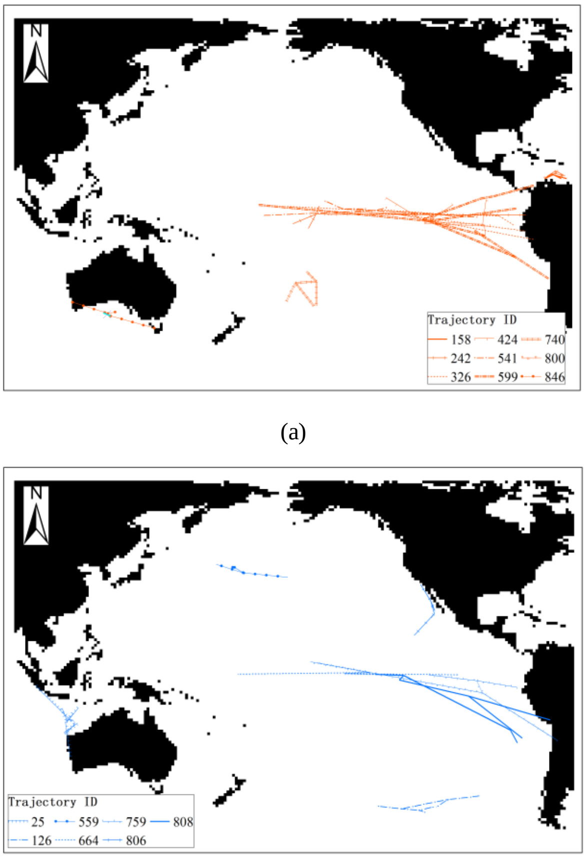


Figure 5. Clustering results of POSCM based on structural and spatial similarity.

4.2 Case study of SST abnormal variations trajectories

The Pacific Ocean from 100°E to 60°W and 50°S to 50°N is considered as a research area, and the Sea Surface Temperature (SST) Abnormal variations (SSTA) trajectories from January 1950 to December 2017 are obtained by the centroid movement of SST abnormal variation objects, which is explored using the dual-constraint spatiotemporal clustering approach (Liu et al., 2018). The purpose of the case study is conducted to illustrate applicability of POSCM to complex trajectories, therefore, only the SSTA complex trajectories are as the input data.

Figure 6 shows two clusters (named as Cluster 1 and Cluster 2) of SSTA complex trajectories based on structural similarity. Table 1 and Table 2 show the detailed information of the two clusters, respectively. The El Nino/ La Nina events are defined by the method in Mcphaden et.al (2006), and the main structure of Cluster 1/2 is the maximum similarity structure of all trajectories in it. Both clusters are mainly located in the eastern equatorial Pacific Ocean where is the sensitive region to El Nino and La Nina event. As shown in the Table 1/2, the lifespan of most trajectories is in the period of El Nino/ La Nina event, and the thematic attribute (above or below the mean of SST) of these trajectories in Cluster 1/2 are coincident. The main structure of Cluster 1/2 shows that the SSTAs have the behaviour of merging first and then splitting in this area during some El Nino events, and have two merging behaviours in this area during some La Nina events.



(a)

(b)

Figure 6. Clustering results of SSTA complex trajectories through POSCM.

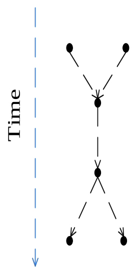
Trajectories in Cluster 1	Attribute of trajectories			Main structure of Cluster 1
	lifespan	*El Nino event*	*above or below the mean*	
158	1958.01-1958.08	√	above	
242	1965.02-1966.06	√	above	
326	1972.03-1973.02	√	above	
424	1982.04-1983.12	√	above	
541	1991.06-1992.07	√	above	
599	1997.02-1999.01	√	above	
740	2005.09-2006.05	×	above	
800	2009.12-2010.11	√	above	
846	2013.01-2013.05	×	above	

Table 1. Related information of Cluster 1.

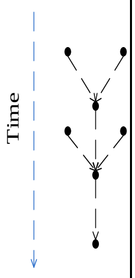
Trajectories in Cluster 2	Attribute of trajectories			Main structure of Cluster 2
	lifespan	*La Nina event*	*above or below the mean*	
25	1950.11-1952.03	√	below	
126	1956.06-1957.12	√	below	
559	1993.05-1993.11	×	below	
664	1998.07-2000.06	√	below	
759	2007.07-2008.06	√	below	
806	2010.06-2012.03	√	below	
808	2005.09-2006.05	√	below	

Table 2. Related information of Cluster 2.

5. CONCLUSION

In this paper, a Process-Oriented Spatiotemporal Clustering Method (POSCM) was proposed to extract clustering pattern of complex trajectories, which have at least one split and/or merger and/or split-merger branch during their lifetime. POSCM mainly focuses on solving the issues of representation and similarity measurement for complex trajectories clustering. For the former, POSCM presented a process-oriented representation method based on the “process-sequence-state” semantic, which is able to simplify the complex trajectories and ensure structure of the complex trajectories at the same time. Based on this representation of complex trajectories, POSCM developed a Hierarchical Similarity Measurement Method (HSMM) for measuring the similarity of any two complex trajectories in space, time, thematic attribute, and structure. And in the final step, POSCM clustering the complex trajectories through classical density-based clustering algorithm.

We used synthetic complex trajectories to evaluate POSCM and compare its performance against that of the VF2. From the results, POSCM is more suitable for measuring structural similarity of complex trajectories than VF2, furthermore, the HSMM in POSCM makes it possible for measuring comprehensive similarity of complex trajectories in space, time, thematic attribute, or structure.

Using SST abnormal variations trajectories data in the Pacific Ocean as the case study, the POSCM explored interesting patterns from the data. From the clustering results, we can find the similar behaviours of SST abnormal variations in the sensitive regions during some El Nino/ La Nina events, which may provide an important reference for research on the mutual response and driving mechanisms behind between global climate change and marine abnormal variations.

In conclusion, the POSCM can extract clustering patterns of complex trajectories based on multi-attribute similarity, which can contribute to a better understanding for the change behaviours of objects or phenomena. The future work will focus on further optimization of the POSCM, and the further analysis for the clustering pattern of SST abnormal variations trajectories.

ACKNOWLEDGEMENTS

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA19060103), by the National Natural Science Foundation of China (No. 41671401,41671445), and by the National key research and development program of China (No. 2016YFA0600304, 2017YFB0503605).

REFERENCES

Anagnostopoulos, A., Vlachos, M., Hadjieleftheriou, M., 2006. Global distance-based segmentation of trajectories. *in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM,* pp. 34–43.

Brakatsoulas, S., Pfoser, D., Salas, R., Carola, W., 2005. On map-matching vehicle tracking data. *Proceedings of the 31st International Conference on Very Large Data Bases*, pp: 853-864.

Chen, J., Wang, R., Liu, L., 2011. Clustering of trajectories based on Hausdorff distance. *Proc. IEEE International Conference on Electronics, Communications and Control*, pp. 1940-1944.

Cordella, L.P., Foggia, P., Sansone, C., 2004. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 26-10, pp:1367-1372.

Faisal, I.B., Ashfaq, A.K., Dan, S., 2007. Object Trajectory-Based Activity Classification and Recognition Using Hidden Markov Models. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, Vol. 16-7, pp. 1912-1919.

Hong, Z., Ying, C., Mahmassani, H.S., 2017. Recognizing Network Trip Patterns Using a Spatio-Temporal Vehicle Trajectory Clustering Algorithm. *IEEE Transactions on Intelligent Transportation Systems*, Vol 19-8, pp. 2548-2557.

Jain, A.K. and Dubes, R.C., 1988. Algorithms for clustering data, *Technometrics*, Vol. 32-2, pp. 227–229.

Li, X., Hu, W., Hu, W., 2006. A coarse-to-fine strategy for vehicle motion trajectory clustering, *in: 18th International Conference on Pattern Recognition (ICPR’06), IEEE,* Vol. 1, pp: 591–594.

Liu, J.Y., Xue, C.J., He, Y.W., Dong, Q., Kong, F.P., Hong, Y.L., 2018. Dual-constraint Spatiotemporal Clustering Approach for Exploring Marine Anomaly Patterns using Remote Sensing Products. *IEEE J-STARS*, Vol. 11-11, pp: 3963 - 3976.

Liu, M.Y., Tuzel, O., Ramalingam, S., Chellappa, R., 2014. Entropy-Rate Clustering: Cluster Analysis via Maximizing a Submodular Function Subject to a Matroid Constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36-1, pp:99-112.

Liu, W., Li, X., Rahn, D.A., 2016. Storm event representation and analysis based on a directed spatiotemporal graph model, *Taylor & Francis* pp. 1-14.

Mahrsi, M.K.E., Rossi, F., 2013. Graph-Based Approaches to Clustering Network-Constrained Trajectory Data. *New Frontiers in Mining Complex Patterns*, Vol. 7765, pp:124-137.

Mcphaden, M.J., Zebiak, S.E., Glantz, M.H., 2006. “ENSO as an integrating concept in earth science,” Science, vol. 314, no. 5806, pp. 1740–1745.

Nan, F., He, Z., Zhou, H., Wang, D., 2011. Three long-lived anticyclonic eddies in the northern South China Sea. *Journal of geophysical research-oceans*, Vol. 116-c5, pp. 879-889.

Nanni, M. and Pedreschi, D., 2006. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, Vol. 27 -3, pp:267–289. doi:10.1007/s10844-006-9953-7.

Pakhira, M.K., Bandyopadhyay, S., Maulik, U., 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, Vol. 37-3, pp:487-501.

Pelekis, N., Andrienko, G., Andrienko, N., Kopanakis, 2012. Visually Exploring Movement Data via Similarity-based Analysis. *Journal of Intelligent Information Systems*, Vol. 38- 2, pp: 343-391.

Riesen, K. and Bunke, H., 2009. Approximate graph edit distance computation by means of bipartite graph matching. *Image & Vision Computing*, Vol. 27-7, pp: 950-959.

Song, C., Pei, T., Ma, T., Du, Y., Shu, H., Guo, S., Fan, Z., 2018. Detecting arbitrarily shaped clusters in origin-destination flows using ant colony optimization. *International Journal of Geographical Information Science,* doi: 10.1080/13658816.2018.1516287.

Wang, H.M., Du, Y., Yi, J., Sun, Y., Liang, F., 2018. A new method for measuring topological structure similarity between complex trajectories. *IEEE Transactions on Knowledge and Data Engineering*, doi.org/10.1109/TKDE.2018.2872523.

Xie, X., Liao, W., Aghajan, H., Veelaert, P., Philips, W., 2017. Detecting Road Intersections from GPS Traces Using Longest Common Subsequence Algorithm. *International Journal of Geo-Information*, Vol. 6-1, pp:1-15.

Xue, C.J., Dong, Q., Xie, J., 2012. Marine spatio-temporal process semantics and its applications-taking the ENSO process and Chinese rainfall anomaly as an example, *Acta Oceanologica Sinica*, Vol. 33-2, pp.16-24.

Yi, J.W., Du, Y.Y., Liang, F.Y., Zhou, C. H., Wu, D., Mo, Y., 2014. A representation framework for studying spatiotemporal changes and interactions of dynamic geographic phenomena. *International Journal of Geographical Information Science*, Vol. 28-5, pp.1010-1027.

Yuan, M., 2001. Representing complex geographic phenomena in GIS. *Cartography and Geographic Information Science*, Vol. 28-2, pp: 83-96.

Zheng, L., Xia, D., Zhao, X., Tan, L., Li, H., Chen, L., Liu, W., 2018. Spatial–temporal travel pattern mining using massive taxi trajectory data. *Physica A: Statistical Mechanics and its Applications*, Vol. 501-1, pp. 24-41.

Zhu, D., Huang, Z., Shi, L., Wu, L., Liu, Y., 2018. Inferring spatial interaction patterns from sequential snapshots of spatial distributions. *Int. J. Geograph. Inf. Sci.*, vol. 32-4, pp. 783–805.

Zhu, H. and Zhu, L., 2017. Encrypted network behaviours identification based on dynamic time warping and k-nearest neighbour. *Cluster Computing*, Vol.1, pp:1-10.

Zhu, X. and Guo, D., 2014. Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS*, Vol. 18-3, pp. 421–435. doi: 10.1111/tgis.2014.18.issue-3.