基于密度的轨迹时空聚类分析

吴 笛[1,2],杜云艳[1*],易嘉伟[1,2],魏海涛[1,3],莫 洋[1,2]

(1. 中国科学院地理科学与资源研究所 资源与环境信息系统国家重点实验室,北京 100101;
2. 中国科学院大学,北京 100049; 3. 山东科技大学测绘科学与工程学院,青岛 266510)

摘要:通过轨迹聚类分析挖掘物体移动模式的空间分布和时间特征,对于认识运动的形成机制,预测运动的未来发展具有重要的意义。目前,轨迹聚类研究主要关注物体的空间位置变化,时空聚类中时间约束一般只是作为辅助信息,并不真正参与聚类。本文提出基于密度的轨迹时空聚类方法,在聚类过程中同时考虑轨迹包含的时空信息,在空间聚类的基础上提出了轨迹线段时间距离的度量方法和阈值确定原则,对时空邻域密度进行聚类分析,挖掘物体的时空移动模式。实验对南海涡旋轨迹进行时空聚类分析,得到了涡旋典型移动模式的空间分布和时间特征,验证了基于密度的轨迹时空聚类方法的有效性。加入时间约束后,移动通道主要发生缩短、分裂和消失的变化。和空间聚类相比,轨迹时空聚类可有效地划分发生在同一位置不同时间的轨迹,得到的聚类结果更加细化,移动模式更加准确,有利于物体的移动模式做更深入的分析。

关键词:轨迹聚类;时空数据挖掘;涡旋;南海

DOI:10.3724/SP.J.1047.2015.01162

1 引言

聚类分析是按照相似程度划分数据,并保持类间距离最大,类内距离最小。轨迹聚类则是聚类分析在时空轨迹上的扩展,其目的是基于空间或时间相似性,把具有相似行为的时空对象划分为一类,通过聚类可以发现物体的移动模式,分析移动规律,甚至预测未来的运动行为[1-2]。

现有的轨迹聚类方法大多从空间位置出发,以不同的数据组织方式挖掘物体的移动规律。Dykes和Mountain[3]用场表达轨迹,通过格网内的轨迹密度划分出活跃区域。Camargo等[4]用矢量线表达完整轨迹,通过历史数据建立回归模型,计算轨迹和模型间的相似性进而聚类得到移动模式。Lee等[5]也是将轨迹表达为线型矢量,但按照特征点打断轨迹,把子轨迹作为研究对象,提出了“划分-聚合”的轨迹聚类方法。物体活跃的空间位置和时间息息相关,上述聚类方法仅从空间位置出发,分析台风、涡旋等变化迅速且受季节性因素影响的时空现象时,无法有效地揭示其时空模式。

这类现象中,时间是移动模式的一个重要特征,分析轨迹的移动时间特点,对于认识其运动机理具有重要的意义。然而,空间和时间属于不同的维度,具有不同的度量方式,简单地用相同的方法进行处理并不合理有效。现有的移动物体时空变化研究主要按照2种思路:(1)先从空间位置入手,挖掘物体在各个时段的聚类模式,再按照时间顺序进行连接,得到完整的移动规律[6-8]。时间地理学领域对于人类活动规律的挖掘按照这种方式展开;Shaw等[7]对时空路径进行聚类分析,在每个采样时间内对物体的空间位置进行聚类,得到聚集区域,再按照时间顺序连接这些区域,得到概括的时空路径。(2)先从时间入手,发现可能呈现移动规律的时间区间,再对相应的部分进行聚类,得到发生在该时间的移动模式[9-11]。Nanni和Pedreschi[10]提出的TF-OPTICS算法,是先探测轨迹上模式挖掘的有效时间区间,鉴此,对轨迹进行OPTICS聚类。这2种思路中时间都只是作为一个辅助信息,对数据进行约束或选择,并不真正参与聚类,存在很大的局限性:逐个时间段聚类,只有物体按照相同的采样频

收稿日期:2015-04-29;**修回日期**:2015-05-27.
基金项目:国家自然科学基金项目“基于海洋要素场的涡旋过程数据建模与可视化”(41371378)。
作者简介:吴笛(1990-),女,硕士生,研究方向为时空数据挖掘。E-mail: wudi@lreis.ac.cn
***通讯作者**:杜云艳(1973-),女,博士,研究员,研究方向为时空建模与推理。E-mail: duyy@lreis.ac.cn

率活动时，才能准确得到时空规律；探测时间区间，则要求物体规律性移动的意义较为明显，这样才有利于判断出重要的时间区间。目前，从轨迹角度将时间和空间同时参与聚类分析的研究较少，但是，在时空数据挖掘领域已有相关研究。Pei等[12]提出的时间窗口k阶邻近距离同时考虑时空属性，考察两点在时间窗口内的空间邻近性。点与点之间的空间和时间距离度量较为简单，一些轨迹模式挖掘方法是将轨迹表达为一系列的点，用各点间的时间间隔度量时间维的变化[13]。因此，仍然从线型轨迹角度进行时空分析，关注物体移动过程的变化，如果有合适的方法度量线状轨迹间的空间和时间距离，就可用类似的思路解决轨迹的时空聚类问题。

本文采用基于密度的聚类方法对轨迹进行分析，综合考虑物体移动的时空信息，在现有密度的部分轨迹线段聚类方法的基础上进行改进，定义线段时间距离的度量方法，并确定其阈值和空间邻域一同构建线段的时空邻域。基于时空邻域内的轨迹密度进行聚类，得到物体移动模式的空间分布和时间特征，并利用南海涡旋轨迹数据进行多组实验。

2　基于密度的轨迹时空聚类方法

为了发现物体的时空移动规律，本文提出了基于密度的轨迹时空聚类方法，在Lee等[5]轨迹线段空间聚类方法的基础上进行改进，综合考虑轨迹的时空信息，挖掘物体移动模式的空间分布和时间特征。

Lee等的方法从空间位置出发对轨迹进行聚类，主要包括轨迹划分或简化，线段空间聚类和表征轨迹提取3个步骤，流程如图1所示。该方法参与聚类的对象为线段，因此，首先需将轨迹按照一定的原则处理为线段。研究移动物体完整生命过程的位置变化时，将轨迹按照起止点简化为线段（OD方式）；研究移动物体部分生命过程的移动模式时，将轨迹按照特征点划分为多条线段（MDL方

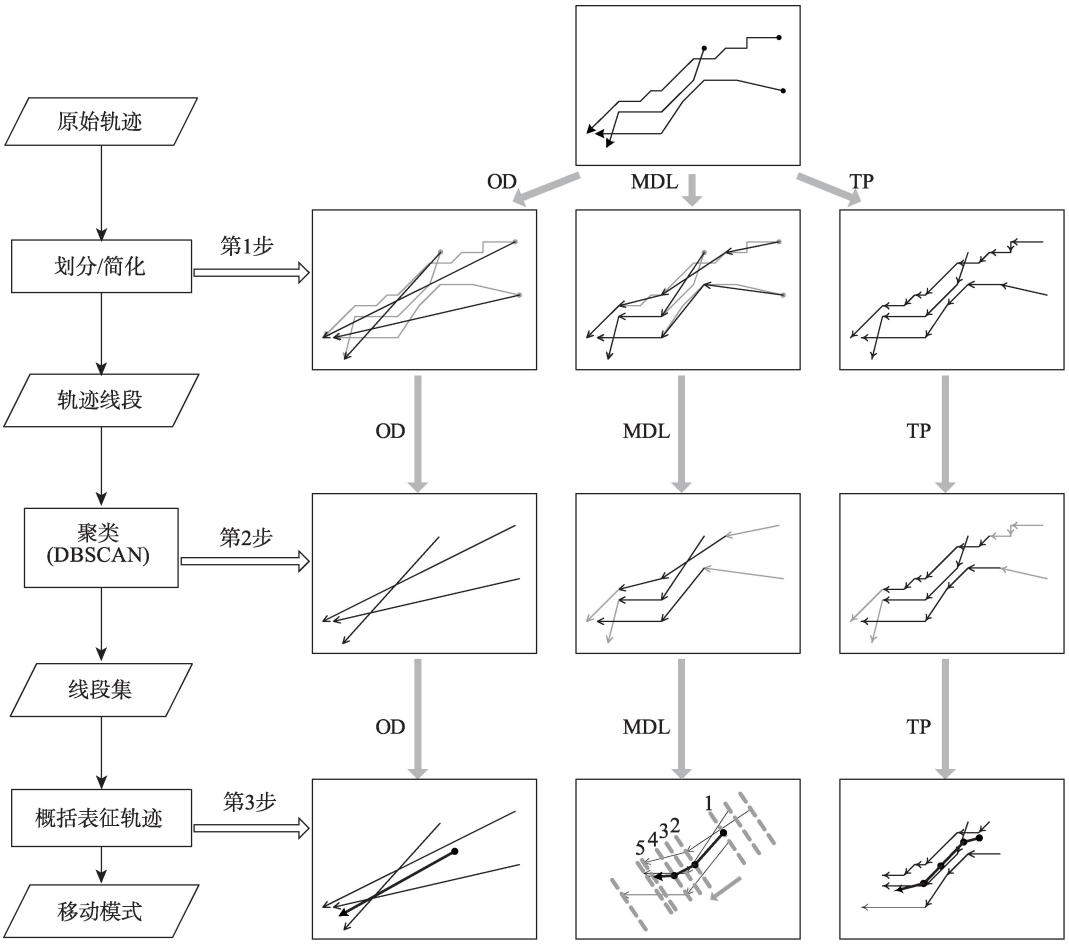


图1　轨迹空间聚类流程图

Fig. 1　A flowchart showing the process of trajectory spatial clustering

式);研究移动物体瞬时移动时,将轨迹按照采样节点划分为多条线段(TP方式)。聚类采用基于密度的DBSCAN算法,定义了线段的空间距离和空间邻域,实现了算法在线数据上的扩展。从聚类中概括表征轨迹采用线扫描的方式(图1 MDL方法第3步),确定各扫描线上的平均位置,再依次连接,得到聚类中心轨迹。

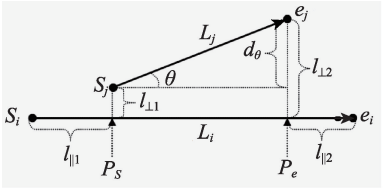
基于密度的轨迹时空聚类主要是在聚类(第2步)时从空间扩展到时空,定义轨迹线段间的时间距离和时间邻域,与空间邻域一起构建时空邻域,考察线段在时空邻域内的密度,进而实现聚类。本文详细介绍轨迹时空聚类中的几个重要概念。

2.1 空间距离

线段间的相对位置关系通过二者间的相对距离、相对长度,以及相对角度唯一确定,因此,Lee等[5]分别定义垂直距离、平行距离和角度距离,并将三者之和作为线段间的空间距离。其具体定义如图2所示,指定较长的线段为L_i,较短的为L_j,P_s、P_e分别是L_j的起止点(s_j, e_j)在L_i上的投影点,垂直距离($d_\perp$)、平行距离($d_{/\!/}$)和角度距离(d_θ)的计算如图2公式所示。其中,对于无方向的线段,角度距离计算公式中θ的取值范围仅为[0, π/2]。

2.2 时间距离

线段的时间区间由线段2个端点对应的时间确定,度量时空线段的时间距离,即度量对应时间区间之间的时间距离。将时间在一维的轴上展开(图3),存在时间相交和相离的2种情况。样本间距离的度量有多种方式,常见的欧氏距离、夹角余弦等主要用来度量样本点或矢量间的差异性,汉明距离用来度量字符串间的差异性,杰卡德距离用来度量2个集合间的差异性[14]。其中,杰卡德距离是由集合A和集合B相同部分和相异部分的大小确定,其距离$J(A, B)$的计算公式如式(1)所示,取值范围为[0, 1]。



$$d_\perp=\frac{l_{\perp 1}^2+l_{\perp 2}^2}{l_{\perp 1}+l_{\perp 2}}$$

$$d_{/\!/}=\min(l_{/\!/1}, l_{/\!/2})$$

$$d_\theta=\begin{cases}\|L_j\|\times\sin\theta, & 0\leqslant\theta<\pi/2\\ \|L_j\|, & \pi/2\leqslant\theta\leqslant\pi\end{cases}$$

$$d_{ij}=d_\perp+d_{/\!/}+d_\theta$$

图2 线段空间距离度量

Fig. 2 Spatial distance between line segments

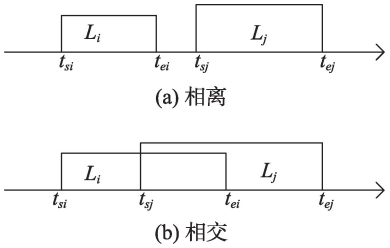


图3 时间距离度量

Fig. 3 Temporal distance between line segments

$$J(A,B)=1-\frac{|A\cap B|}{|A\cup B|} \tag{1}$$

时间区间距离的度量主要从二者是相离还是相交,相离、相交部分的大小等方面进行考察,与杰卡德距离度量2个集合差异性的思路类似。与一般集合差异性度量不同,时间距离需考察相离和相交2种情况,对于相离的时间区间也要考察相离部分的长短,因此更为复杂。

2个时空线段的时间跨度分别为T_i、T_j。首先定义时间区间的时间差(ΔT_{ij}),如式(2)所示:没有重叠区域时(图3(a)),二者相离的时间;存在重叠区域时(图3(b)),二者相交的时间;规定相离为正,相交为负。

$$\Delta T_{ij}=\max(t_{si},t_{sj})-\min(t_{ei},t_{ej}) \tag{2}$$

时间差仅能在一定程度上反映出2个时间区间之间的差异性;时间跨度也是时间区间差异性的重要体现,例如,不同长度的2组线段时间区间相离的时间差相同时,长时间跨度的一组线段应具有更短的时间距离。因此,综合考虑时间差和时间跨度,将时间距离定义为式(3)。

$$\begin{cases}T_i=t_{ei}-t_{si}\\ T_j=t_{ej}-t_{sj}\\ disT_{ij}=\Delta T_{ij}/(T_i+T_j)\end{cases} \tag{3}$$

时间距离是由时空线段的时间跨度和时间差共同决定,2个时空线段相离时,相距的时间越短,其自身的时间跨度越大,二者的时间距离越小;2个时空线段相交时,重叠的时间越长,其自身的时间跨度越小,二者的时间距离也越小;从相离到相交,时间距离在整个取值区间内的变化是一致的。理论上时间距离的取值区间为整个实数域。

研究周期性活动的物体时,例如,多年的涡旋轨迹,通常忽略发生周期间的差异,划归为同一周期中进行分析,此时时间区间不再是一维展开的直线,而是首尾相接的环线。这种情况下,时间差的

计算需相应发生变化（式（4）），式中，T_p为移动物体的运动周期。时间距离的计算公式保持不变，此时时间距离的取值受到周期长度的限制，不再是整个实数域，而是缩小到某个实数区间。

$$\Delta T_{ij}' = \min(\Delta T_{ij}, T_P - T_i - T_j - \Delta T_{ij}) \quad (4)$$

2.3 时空邻域

时空邻域是将空间邻域和时间邻域相结合确定的时空范围，先从空间和时间角度分别确定邻域范围，再将二者有机结合。空间邻域阈值ε_s的确定沿用空间聚类中的最小熵方法。在最小熵理论[15]下，空间阈值对应的划分结果已是最优，时间阈值如果仍然按照这种方式将无法得到有效的阈值结果，因此需从数据本身出发考虑。这里通过时间距离的累积概率分布确定阈值：计算全部时空线段两两之间的时间距离，从小到大划分区间统计并绘制时间距离的概率分布图和累积概率分布图，将累积概率达到45%时对应的时间距离作为阈值ε_t，由此确定时间邻域范围。

时间距离阈值的大小反映同一类线段在时间上连续变化的程度：阈值越大，轨迹线段间同时发生的比例越小，移动模式存在的时间越松散；阈值越小，轨迹线段间同时发生的比例越大，移动模式存在的时间越紧凑。时间距离阈值确定指标，对应累积概率达到45%，是由涡旋轨迹多组时空聚类实验比较得到。一般认为累积概率分布达到50%对应时间距离的平均水平，聚类中阈值高于该标准，可得到更为聚集的结果。

空间邻域和时间邻域确定后，二者共同构建时空邻域实现聚类。基于密度的DBSCAN聚类，涉及邻域大小（ε）和最小线个数（*MinLns*）2个参数，邻域大小系由时间和空间邻域阈值共同决定，最小线个数由邻域大小估计得到。

基于时空邻域密度的DBSCAN聚类时，从任一线段出发，计算与其他所有线段间的空间距离与时间距离；统计同时满足空间阈值（ε_s）、时间阈值（ε_t）范围的线段个数，并与最小线个数（*MinLns*）进行比较；当时空邻域范围内的线段数目大于给定的最小线个数（*MinLns*）时，该线段即为核心线段，形成一个聚类，其邻域内的直接密度可达线段也将聚到该类中，再对这些线段按照同样的方式依次进行聚类扩展，得到最终的聚类结果。

3　南海涡旋轨迹时空聚类应用分析

3.1 南海涡旋轨迹数据

南海涡旋轨迹数据是从近20 a（1992年10月至2012年3月）海表面异常（SLA）的高度计融合产品（融合了TOPEX/Poseidon、Jason-1、ERS-1/2和Envisat）中识别、跟踪得到[16]，产品的时间分辨率为7 d，空间分辨率为1/3°×1/3°。涡旋轨迹数据库记录了每个涡旋的空间位置、开始时间、结束时间，以及每7 d的状态，包括涡旋半径、强度、动能等信息。为了检验轨迹时空聚类方法的有效性，本文选择了南海区域816条（暖涡398条，冷涡418条）长距离移动的强过程涡旋轨迹展开实验。

按照SLA数据的空间分辨率绘制格网，计算每个格网的轨迹密度（经过该格网的轨迹个数与轨迹总数之比），可发现南海涡旋高频率的活动大体分布在3个区域：南海北部、南海中部和南海西南部（图4）。

涡旋轨迹高密度区域的分布一定程度上，反映出涡旋存在的移动模式，但是，从独立的单元出发，不能体现移动的连续性、方向性。因此，需在此基础上对涡旋轨迹进行聚类分析，找到轨迹高密度的连通区域，并提取表征轨迹，得到涡旋的移动模式。

3.2 南海涡旋轨迹空间聚类结果

实验分别采用MDL、OD和TP 3种方式，对暖涡（AE）和冷涡（CE）轨迹进行空间聚类，各组实验的轨迹数目和聚类参数如表1所示。3组实验得到的聚类结果和移动模式的空间分布如图5所示。

图5中各个模式的支持率（参与构成模式的轨迹数与轨迹总数之比）用百分数标记在模式旁，MDL和TP结果图中灰色的背景为涡旋轨迹密度，红线和蓝线分别表示发生在夏季风期和冬季风期的构成移动模式的原始部分轨迹；OD结果图中彩色背景为涡旋产生和消亡位置的核密度分布，红色和蓝色越深的区域表示暖涡和冷涡产生得越多，棕色越深的区域表示涡旋消亡得越多，灰线表示参与构成移动模式的轨迹线段。

3组结果整体而言较为相似，通过这3组实验结果可以看出，南海涡旋的活动存在区域性的特征。从涡旋移动模式角度，可以将南海划分为北部（16.5°N以北）、中部（11°~16.5°N）和南部（11°N以南）3个部分，涡旋在这3个区域内存在不同的移动模式，暖涡和冷涡的移动通道也不尽相同。

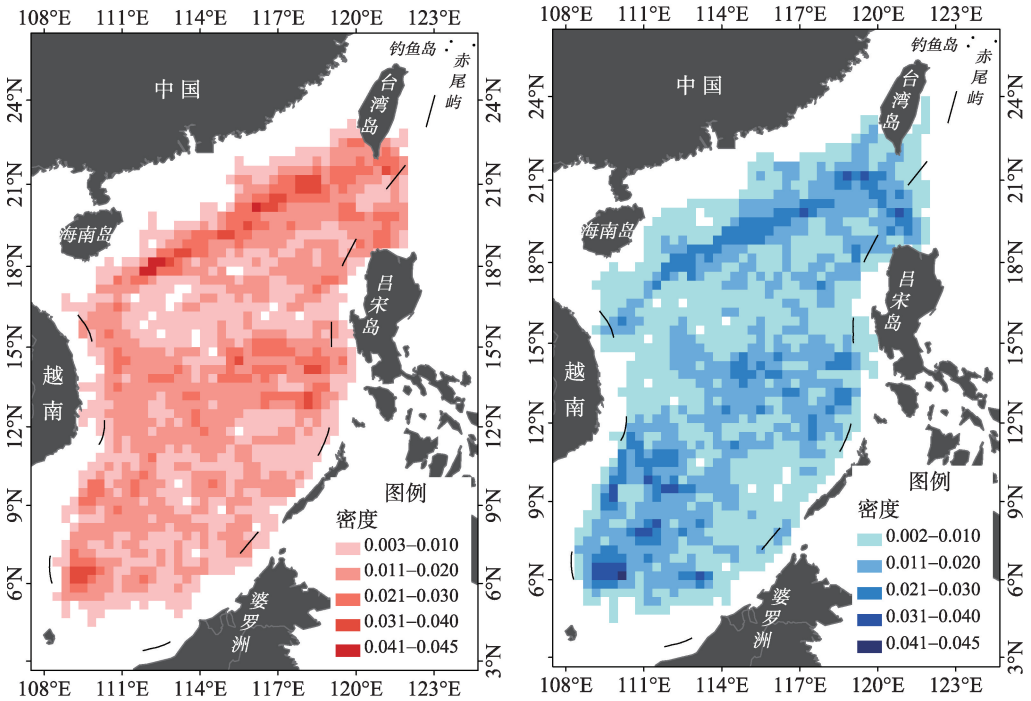


图4 南海涡旋轨迹密度分布

Fig. 4 Trajectory density of eddies in the South China Sea

表1 轨迹聚类数目和聚类参数

Tab. 1 Number of trajectories and clustering parameters

	MDL-AE	MDL-CE	OD-AE	OD-CE	TP-AE	TP-CE
轨迹数目	398	418	398	418	1077	1185
ε_s	74	79	124	124	79	77
MinLns	6	6	5	5	7	7

暖涡在南海北部的移动模式主要沿北部陆架，从吕宋海峡开始，沿陆架向西南方向延伸，直到海南岛东南部，南海西边缘结束，这也是暖涡在南海最显著的移动通道。除此之外，MDL方式在南海北部118°E以东还得到了2个较弱，移动方向相反的模式，从吕宋岛西北角开始，分别向西南和西北方向展开。南海中部，暖涡的移动通道从吕宋岛西南侧开始，穿越中央海盆，一直向西延伸，最终到达越南以外海域。南海南部，暖涡的长距离移动模式并不显著，主要集中在南海的西南边缘，3种方式在该区域的结果有明显的差别，OD方式的通道开始于海洋深处，而TP方式则形成了2个较弱的模式。

冷涡在南海北部的典型移动模式也是沿北部陆架展开，但相比于暖涡整体更靠近海洋深处。在北部118°E以东，冷涡还存在一个从吕宋岛北侧开始，跨越吕宋海峡，向西北方向延伸的通道。南海中部，冷涡的移动移动模式开始的位置靠近吕宋岛西南角，同样经过中央海盆向西延伸，并明显呈现西北向的偏转，没有到达南海西边界就已经结束。南海南部，冷涡形成了西南方向的移动通道，从南海中部偏南开始，向西延伸到靠近西边界，再转向南，直到南海西南边缘结束。

按照MDL、OD和TP 3种方式划分或简化轨迹，得到的聚类结果虽然整体而言较为相似，但还是存在一些差别。OD方式通过轨迹起止点进行简化，忽略了活动的细节，仅考虑涡旋产生和消亡的位置，得到的移动模式体现涡旋大体的移动方向，从产生的高密度区域移动到消亡的高密度区域。TP方式通过轨迹采样节点进行划分，充分考虑了涡旋的瞬时活动，能反映轨迹简化时忽略的信息，但得到的移动模式包含一定的随机性。MDL方式通过特征点进行划分，既对轨迹进行了简化，又保留了一些重要的变化细节。相较于前2种方式，其反映的涡旋移动模式更具有代表性。

3.3 南海涡旋轨迹时空聚类结果

空间聚类结果仅反映出了涡旋移动模式的空间分布，然而受到季节性因素的影响，涡旋会在不同的时间存在不同的移动规律。因此，对涡旋轨迹进行时空聚类分析，挖掘涡旋的时空移动模式。实

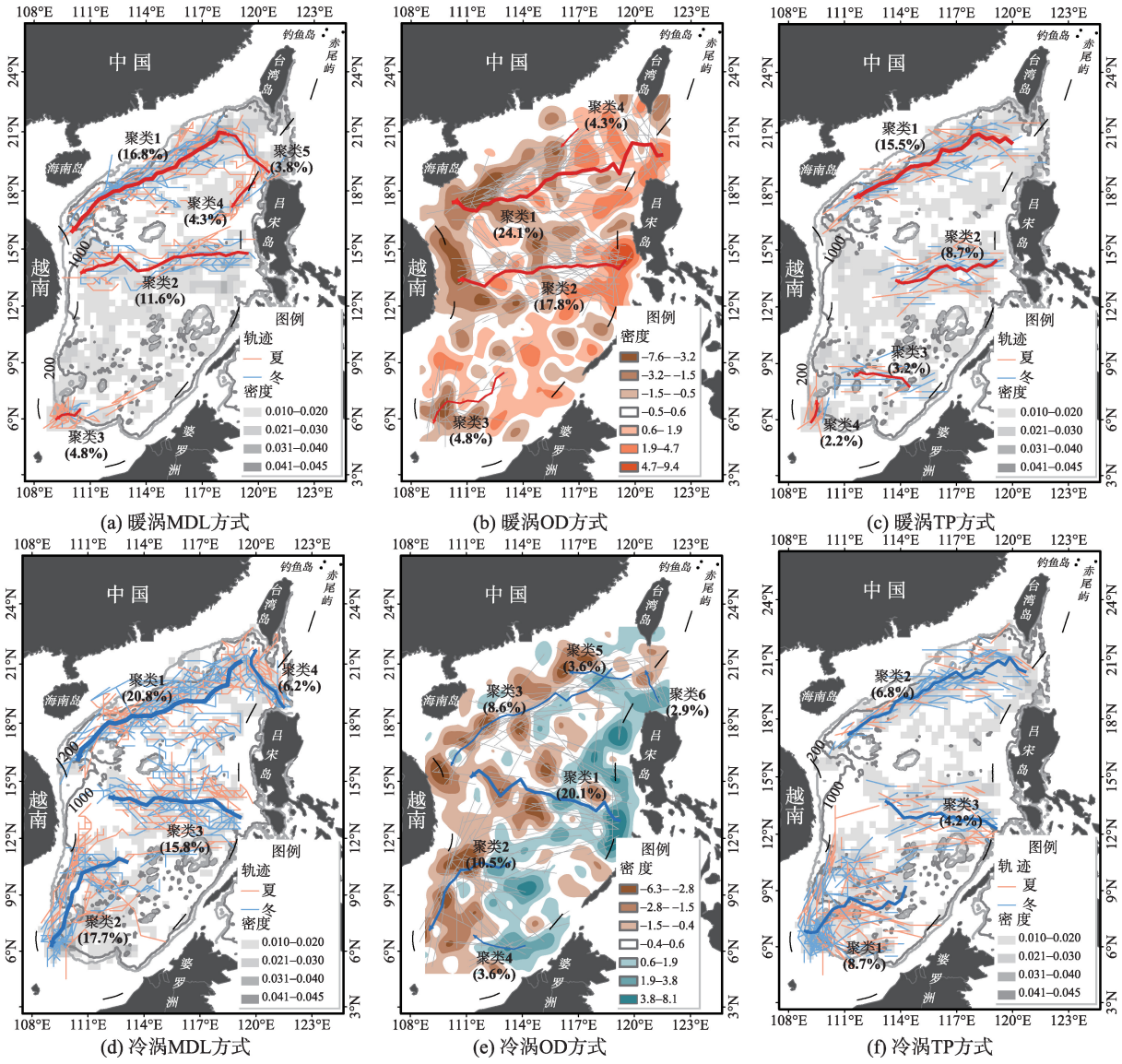


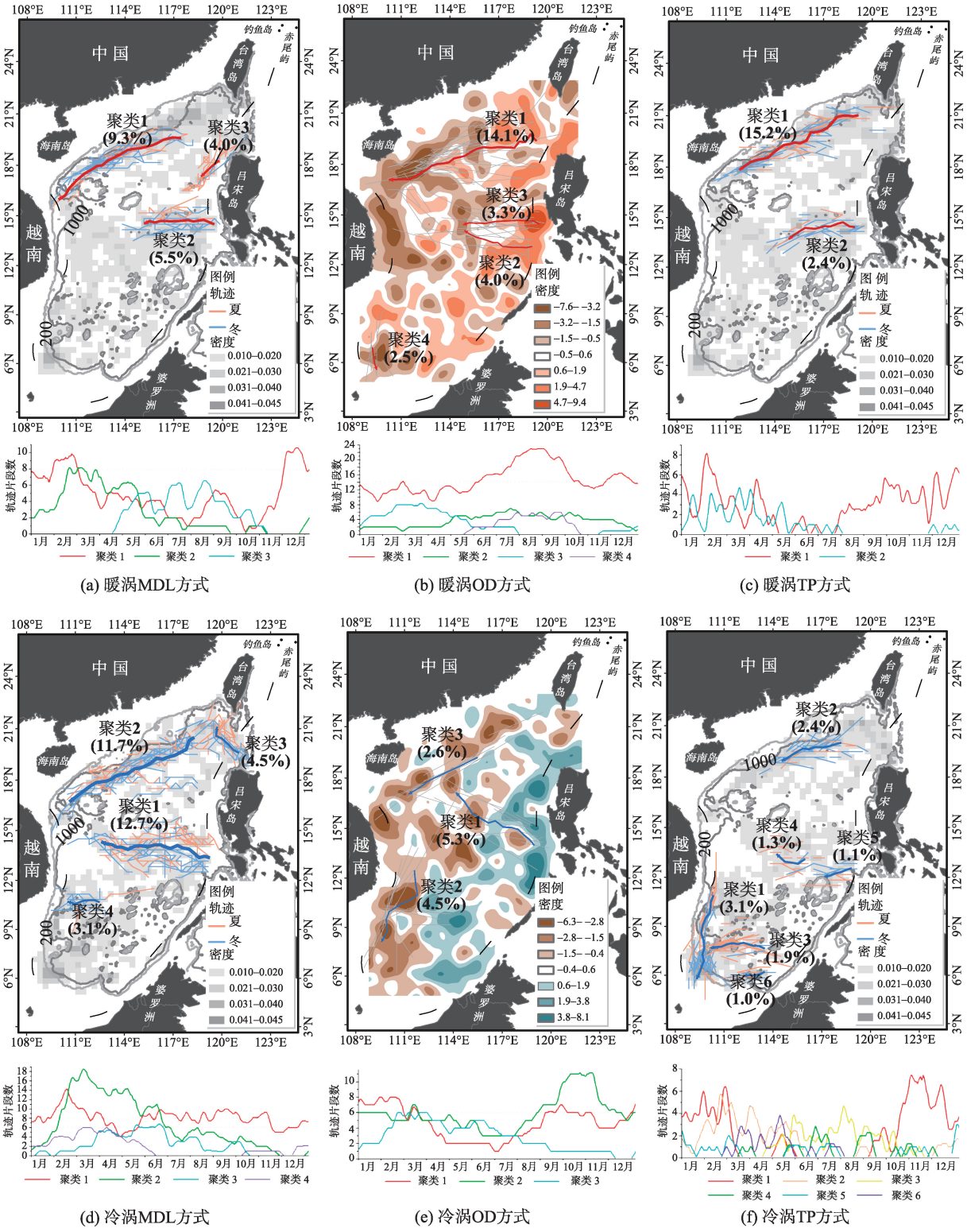
图5　南海涡旋轨迹空间聚类结果

Fig. 5　Spatial clustering results of eddies' trajectories in the South China Sea

验数据与空间聚类一致，涡旋轨迹以天为单位（d）记录发生的时间，忽略涡旋发生年份的差异，因此，轨迹的时间区间是一年内的循环时间，轨迹持续时间最短为7 d，这样时间距离的范围为[−0.5,12.6)。仍然采用MDL、OD和TP 3种方式对暖涡（AE）和冷涡（CE）轨迹进行时空聚类，时间距离阈值依次为0.9、0.1和4。其中，涡旋相邻2节点构成的轨迹线段其时间跨度基本均为7 d，和其他2组数据相比，时间跨度较小，也相对固定。将构成各个移动模式的轨迹线段对应的时间区间在时间轴上累计，得到各个模式的时间分布图，峰值位置即可反映模式呈现的时间。3组时空聚类结果如图6所示。

按照MDL、OD和TP 3种方式划分或简化轨迹，得到的3组结果从不同角度反映涡旋的时空移动规律，结果也并不完全相同，特别表现在冷涡的移动模式。由于涡旋的移动过程受到多种因素复杂的作用，涡旋整个移动过程体现出较为统一的时空规律比较困难，时空移动规律多集中在部分位置或部分时间，因此，以MDL方式得到的移动模式为主，综合3组聚类结果，得到南海涡旋移动模式的空间分布和时间特征。

暖涡在南海北部沿陆架的移动模式仍然是暖涡最显著的时空移动特征，从吕宋海峡外开始向西南延伸到海南岛东南部。冬夏季风期均有涡旋沿该通道移动，主要出现在冬季风期的12–次年2月，以及夏季风期的8–9月。南海中部，暖涡从吕宋岛



(a) 暖涡MDL方式 (b) 暖涡OD方式 (c) 暖涡TP方式

(d) 冷涡MDL方式 (e) 冷涡OD方式 (f) 冷涡TP方式

图6 南海涡旋轨迹时空聚类结果

Fig. 6 Spatiotemporal clustering results of eddies' trajectories in the South China Sea

西南侧开始,经过中央海盆西向传播,结束在海盆中央,该通道主要发生在2-5月。南海南部,暖涡在MDL和TP方式中都没有表现出典型的时空移动模式,仅OD方式中得到了一条位于西南边缘出现在夏季的较弱的通道。

冷涡在南海北部也存在沿北部陆架的典型移

动模式，但相对集中在114°~117° E之间，主要发生在2−5月，冬季风期的后期并进入冬夏之交。南海中部，3种方式得到的冷涡的移动通道存在差异，整体而言相比于暖涡更靠南，呈现向西北方向偏转的趋势，涡旋沿该通道的移动在全年均有发生。南海南部，3组聚类结果存在明显的不同，MDL方式在越南以外海域11° N附近，得到了一个较短的向西延伸的通道，主要存在于2−4月；OD和TP方式的结果有一些相似之处，特别是TP方式的结果得到了一些复杂的局部模式：一个南向的移动通道发生在12−次年2月间，从越南外海11° N开始一直到南海西南部边缘，西向8° N附近一个与之垂直的主要移动通道，主要发生在夏季开始的5−8月。

3.4 2种聚类结果的比较

轨迹时空聚类方法针对空间聚类忽略了时间特征的缺点进行改进，基于密度的聚类过程中，将空间信息和时间信息相结合，构造时空邻域，考察轨迹线段在时空邻域内的密度大小，进而实现聚类。其聚类结果和空间聚类结果相比，可把同一区域发生在不同时间的轨迹划分开，不仅能反映出移动通道存在的时间规律，相应的空间分布规律也更加准确。

轨迹时空聚类方法得到的涡旋移动模式合理有效。不同时间南海涡旋会呈现不同的移动模式，主要是受到季风气候和黑潮等大尺度环流的影响[17−19]。Yuan等[20]对夏秋季南海北部暖涡的活动进行分析，发现暖涡在夏季从吕宋岛西北部开始逐渐移动到北部陆架坡，进入秋冬季风时期沿着陆架坡向西南移动，这与时空聚类提取的南海北部暖涡的移动模式基本相同。Chow等[21]对东沙冷涡进行了深入的研究，发现东沙冷涡的活动开始于东沙附近并沿陆架向西南移动，主要发生在冬季或春季。Du等[22]在提取的涡旋中对东沙冷涡的移动轨迹进行分析，验证了Chow等[21]对东沙冷涡移动特点的分析，南海北部冷涡的这一特点也与时空聚类的结果一致。Chen等[23]通过对多年涡旋传播速度进行矢量合成，得到平均涡旋传播速度场，发现南海北部涡旋主要沿陆架西南向传播，南海中部涡旋呈准西向传播，整体与时空聚类的模式较为相似。

比较南海涡旋的空间聚类结果和时空聚类结果可发现，移动通道主要发生了3种变化：缩短、分裂和消失。时空聚类结果中，移动模式缩短的变化较为普遍，例如，MDL方式提取的南海北部和中部的移动模式都明显缩短；由一个移动模式分裂为2个或多个移动模式，主要表现在TP方式提取的冷涡南海中部和南部的移动模式，以及OD方式提取的暖涡南海中部的移动模式；空间聚类移动模式在时空聚类结果中，不再呈现主要出现在MDL方式提取结果在南海南部的变化。发生这些变化的原因主要与轨迹在空间和时空的不同密度分布相关：移动模式缩短表明，空间移动模式中存在一个主要的连续时空高密度区域，时空邻域的限制将该区域凸显出来；移动模式分裂表明，空间高密度区域是由多个时空高密度区域组合而成，时空邻域限制将其划分到不同的类别；移动模式消失表明，空间移动模式中没有明显连通的时空高密度区域，轨迹线段虽然分布在邻近的位置，但其发生时间没有明显的规律。本文以TP方式提取的冷涡南海南部的移动模式为例，重点分析移动模式分裂的变化，从参与构成模式的轨迹线段入手展开进一步的分析，揭示了发生该变化的原因。

从数量上看，移动模式由1条（图5(f) 聚类1）分裂为3条（图6(f) 聚类1、3、6），参与的轨迹线段也相应地划分到了不同的类中。从空间分布来看，空间聚类结果中参与的轨迹线段较分散，综合得到了一个西南向的移动通道；而时空聚类结果中分裂成的3个移动通道有不同的位置和走向，其中，最典型的1个移动模式位于南海西边界并向南延伸，另2个位于该模式的东侧垂直于该模式的走向并向西延伸。从时间分布来看（图7），空间聚类结果存在多个峰值，分裂得到的3个时空移动模式将时间规律分别显现出来，南向的模式（红线所示）主要发生在冬季的12−次年2月，西向靠北的模式（绿线所示）发生在夏季的5−8月，西向靠南的模式（紫线所示）主要发生在季风期转换的3−5月。这说明空间移动模式实际上是多个连续时空高密度区域的综合，通过时空邻域的约束，将这几个区域划分开来，对应形成了多个时空移动模式，得到的移动规律更加的具体细化。

4　结论

本文采用基于密度的轨迹聚类方法，挖掘移动物体的时空移动模式。在现有基于密度的部分轨迹聚类方法的基础上进行改进，综合考虑轨迹的时间和空间信息，提出了轨迹线段的时间距离度量方式和阈值确定原则，和空间邻域一同构建时空邻

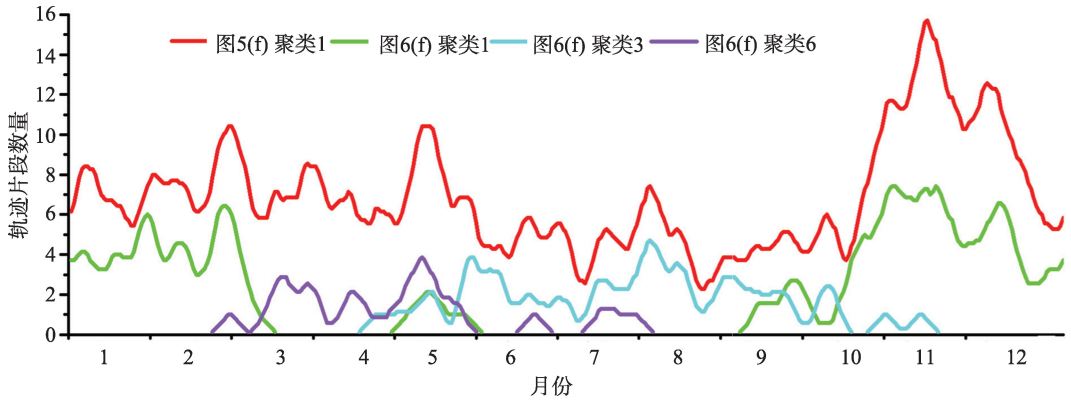


图7 南海南部冷涡TP方式的空间移动模式与时间移动模式时间分布对比

Fig. 7 Time distribution of movement patterns from clustering cold eddies trajectory partitions in the southern South China Sea comparing between the spatial and spatiotemporal results

域,基于时空邻域内的轨迹密度实现聚类。实验利用南海涡旋轨迹数据,验证基于密度的轨迹时空聚类算法的有效性,得到了南海涡旋在北部、中部和南部3个区域不同典型移动模式的空间分布和时间特征。相比于空间聚类,时空聚类可有效地将发生在同一区域不同时间的涡旋轨迹区分开来,得到的聚类更细化,移动通道更准确,有利于对涡旋的移动规律做更进一步的分析。比较空间聚类和时空聚类得到的移动模式,发现加入时间约束后,移动通道主要发生缩短、分裂和消失的变化,而构成移动模式的轨迹线段的时空密度分布差异是产生这些变化的主要原因。

如何有效地将时空信息相结合,是时空数据挖掘中不断探索的问题。本文基于密度的轨迹时空聚类方法提供了一种研究思路,其中,提出的时间距离计算方法,主要从时间的连贯性角度出发,并不强调2个线段在时间维度的先后关系,这样的定义对于空间和时间相互约束的模式挖掘可能并不合适,针对时空信息结合的问题,今后还应做更广泛深入的探讨。

参考文献:

[1] Han J, Lee J-G, Kamber M. An overview of clustering methods in geographic data analysis[A]. In Miller H J, Han J. Geographic data mining and knowledge discovery [M]. London: CRC Press, 2009:149-187.

[2] Jeung H, Yiu M, Jensen C. Trajectory pattern mining[A]. In Zheng Y, Zhou X. Computing with spatial trajectories [M]. New York: Springer, 2011:143-177.

[3] Dykes J A, Mountain D M. Seeking structure in records of spatio-temporal behaviour: visualization issues, efforts and applications[J]. Computational Statistics & Data Analysis, 2003,43(4):581-603.

[4] Camargo S J, Robertson A, Gaffney S, *et al*. Cluster analysis of western North Pacific tropical cyclone tracks[C]. The 26th conference on hurricanes and tropical meteorology, 2004:250-251.

[5] Lee J-G, Han J, Whang K-Y. Trajectory clustering: a partition- and- group framework[C]. Proceedings of the the 2007 ACM SIGMOD international conference on Management of data, 2007:593-604.

[6] Benkert M, Djordjevic B, Gudmundsson J, *et al*. Finding Popular Places[M]. In: Tokuyama T. Algorithms and Computation. Berlin, Heidelberg: Springer, 2007:776-787.

[7] Shaw S-L, Yu H, Bombom L S. A space-time GIS approach to exploring large individual-based spatiotemporal datasets[J]. Transactions in GIS, 2008,12(4):425-441.

[8] Shoshany M, Even-Paz A, Bekhor S. Evolution of clusters in dynamic point patterns: with a case study of Ants' simulation[J]. International Journal of Geographical Information Science, 2007,21(7):777-797.

[9] D'auria M, Nanni M, Pedreschi D. Time-focused density-based clustering of trajectories of moving objects[C]. Proceedings of the Workshop on Mining Spatio- Temporal Data (MSTD-2005), 2005:14.

[10] Nanni M, Pedreschi D. Time-focused clustering of trajectories of moving objects[J]. Journal of Intelligent Information Systems, 2006,27(3):267-289.

[11] Spaccapietra S, Parent C, Damiani M L, *et al*. A conceptual view on trajectories[J]. Data & Knowledge Engineering, 2008,65(1):126-146.

[12] Pei T, Zhou C, Zhu A X, *et al*. Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise[J]. International Journal of Geographi-

cal Information Science, 2010,24(6):925-948.

[13] Giannotti F, Nanni M, Pinelli F, *et al*. Trajectory pattern mining[C]. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007:330-339.

[14] Tan P-N, Steinbach M, Kumar V. Data[A]. In: Introduction to data mining[M]. Boston: Pearson Addison Wesley, 2006:5-18.

[15] Shannon C E. A mathematical theory of communication [J]. Bell System Technical Journal, 1948,27(3):379-423.

[16] Yi J, Du Y, He Z, *et al*. Enhancing the accuracy of automatic eddy detection and the capability of recognizing the multi-core structures from maps of sea level anomaly [J]. Ocean Science, 2014,10(1):39-48.

[17] Cai S, Su J, Gan Z, *et al*. The numerical study of the South China Sea upper circulation characteristics and its dynamic mechanism, in winter[J]. Continental Shelf Research, 2002,22(15):2247-2264.

[18] Metzger E J, Hurlburt H E. The nondeterministic nature of kuroshio penetration and eddy shedding in the South China Sea[J]. Journal of Physical Oceanography, 2001,31 (7):1712-1732.

[19] Su J. Overview of the South China Sea circulation and its influence on the coastal physical oceanography outside the Pearl River Estuary[J]. Continental Shelf Research, 2004,24(16):1745-1760.

[20] Yuan D, Han W, Hu D. Anti-cyclonic eddies northwest of Luzon in summer-fall observed by satellite altimeters[J]. Geophysical Research Letters, 2007,34(L13610).doi:10.1029/2007GL029401).

[21] Chow C-H, Hu J-H, Centurioni L R, *et al*. Mesoscale Dongsha cyclonic eddy in the northern South China Sea by drifter and satellite observations [J]. Journal of Geophysical Research, 2008,113(C4).

[22] Du Y, Yi J, Wu D, *et al*. Mesoscale ocean eddies in the South China Sea from 1992 to 2012: evolution processes and statistical analysis[J]. Acta Oceanologica Sinica, 2014, 33(11):36-47.

[23] Chen G, Hou Y, Chu X. Mesoscale eddies in the South China Sea: mean properties, spatiotemporal variability, and impact on thermohaline structure[J]. Journal of Geophysical Research, 2011,116(C06018).

Density-Based Spatiotemporal Clustering Analysis of Trajectories

WU Di[1,2], DU Yunyan[1*], YI Jiawei[1,2], WEI Haitao[1,3]and MO Yang[1,2]

(1. *State Key Lab of Resources and Environmental Information System*, *IGSNRR*, *CAS*, *Beijing* 100101, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*;
3. *Shandong University of Science and Technology*, *Qingdao* 266510, *China*)

Abstract: Trajectory clustering, which aims to uncover the meaningful spatial distributions and temporal variations of moving objects, is of much importance in understanding potential dynamic mechanisms and predicting future development. However, placing many focuses on locational changes, many studies have made limited use of the time dimension in trajectories. This paper presents a density-based clustering method, which integrates time and space information in identifying significant migrating paths from trajectory datasets. Definition of temporal distances between any line segments decomposed from trajectories as well as the criterion of distance threshold selection is provided in detail. The experiments conducted on ocean eddies in the South China Sea demonstrate the effectiveness of this method in obtaining spatiotemporal migrating patterns. The migrating paths in the results are shortened, or separated into parts, or they turn insignificant as the effect of including time component in density clustering, which reveal more specific movement characteristics in the temporal domain covered by spatial clustering. This advantage facilitates the analysis of objects moving along the same path while displaying distinct time patterns.

Key words: trajectory clustering; spatiotemporal data mining; ocean eddies; the South China Sea

***Corresponding author:** DU Yunyan, E-mail: duyy@lreis.ac.cn