地理时空三向聚类分析方法的构建与实践

程昌秀[1,2,3],宋长青[1,2],吴晓静[1,2],沈 石[1,2],高培超[1,2],叶思菁[1,2]
(1. 北京师范大学地表过程与资源生态国家重点实验室,北京 100875; 2. 北京师范大学地理科学学部,北京 100875; 3. 国家青藏高原科学数据中心,北京 100101)

摘要: 随着地理数据获取能力的不断提升,地理数据体量呈指数增长,数据种类、数据性质更加多元化。对数据的有效甄别和归类成为理解地理现象时空特征、演化过程和行为机制的关键。传统聚类方法面临数据体量大、维数高、质量差的挑战,加之对地理空间与时间关联分析的需求,对聚类方法改进和提升研究的要求越来越迫切。本文介绍了从单向到三向聚类构建思路的变革。单向聚类是仅在样本或属性方向上进行聚类,易忽视非常相似的局部特征、易犯"横看成岭侧成峰"的错误。双向聚类是基于数据矩阵内元素值的相似性,形成一个子矩阵分割方案,使子矩阵内元素相似度尽可能高,子矩阵间元素相似度尽可能低,从而实现行列两方向的同时聚类,避免了单向聚类的不足。鉴于双向聚类难以满足地理研究超出双向的解译需求,本文提出并研发了一个全新的三向聚类方法,给出了运用该方法开展地理时空格局过程探测的流程,总结了如何根据研究涉及的"空间—时间—尺度—属性"构建三维数据体;最后,展示了三向聚类的地理实践案例。结果表明:① 三向聚类是一种大数据时代探测地理数据时空分异规律的有效方法,可以解决数据维度高、质量低等问题;② 面对不同的地理问题,三向聚类在算法层面上是通用的,不同之处仅在于:根据不同问题涉及的空间、时间、尺度、属性的不同,构建不同的数据体;不同数据体聚类得到的不同结果回答不同的地理问题;③ 三向聚类可以实现地理数据的时空分异规律多方向、多尺度、多层次的联合解译,揭示地理特征时空尺度叠加效应。最后,论文强调根据地理问题组织数据的重要性,期待未来能够提升三向聚类在多空间尺度、多属性方面的地理研究实践。

关键词: 三向聚类;空间—时间—尺度—属性;联合解译;时空局部相似性;时空分异
DOI: 10.11821/dlxb202005002

1 引言

地理学是研究地理要素或地理综合体空间格局、时间演变过程和驱动机制的一门学科[1-3]。地理数据是测度地理特征基本的量化指标,由于地理数据具有重要的空间与时间属性,因而它是理解地理特征的核心基础。随着技术的不断进步,获取地理数据的能力不断增强,以大数据为代表的新型数据为地理学提供了难得的机遇[4]。地理研究的核心内容之一是对地理数据的分析,即通过数据集群的时空特征,认识地理现象的格局特征、演化过程和行为机制。聚类分析是识别数据集合内部结构与数值分异规律的重要工具,

收稿日期: 2020-02-06; **修订日期:** 2020-04-22
基金项目: 国家重点研发计划(2019YFA0606901); 中国科学院战略性先导科技专项(XDA23100303) **[Foundation:** National Key R&D Program of China, No.2019YFA0606901; Strategic Priority Research Program of the Chinese Academy of Sciences, No.XDA23100303]
作者简介: 程昌秀(1973-), 女, 新疆人, 教授, 主要从事地理时空数据分析等研究。E-mail: chengcx@bnu.edu.cn
通讯作者: 宋长青(1961-), 男, 黑龙江人, 教授, 主要从事地理学研究范式、地理区域综合研究方法等研究。E-mail: songcq@bnu.edu.cn

因此地理数据的聚类分析是识别地理特征的关键。例如，解焱等[5]用聚类结果开展了中国生物地理区划的实证研究；结果表明：利用聚类结果辅助区划界线的确定可以减少区划对研究者科学知识及经验的依赖，其研究结果更具有客观性。王秀红[6]、郑度等[7]、宋辞等[8]也分别肯定了聚类在地理时空分异规律研究中的作用。

随着大数据时代的到来，传统聚类方法面临挑战。近年基因、文本分析等领域开始逐步采用双向聚类分析数据，并取得重要进展；但三向聚类的思想及算法国内尚未报道。过去几十年，传统聚类在科学数据的分类研究中起了重要作用。随着大数据时代的到来，数据的收集越来越容易、数据种类越来越多、体量越来越大。参与分析的数据不仅存在数据量大，而且存在数据维度高（属性多）、数据质量低（例如，测量不够精确、数据缺失或稀疏）等特点，向传统聚类方法提出挑战。最早推进聚类方法进行时代变革的是基因领域。Hartigan等最早提出“块聚类”的双向聚类思想[9]。自Cheng等提出了面向高维基因数据的双向聚类方法[10]以来，双向聚类算法不断发展完善，并广泛应用于基因分析[11-12]、医学[13-15]、文本分析与自然语言理解[16-17]等领域；近期双向聚类在生态群落领域的研究也初见端倪[18]。

随着地理数据获取能力的不断提升，地理数据也存在量大、维度高、质量参差不齐等特点。同时，从浩如烟海的地理数据中解读时空分异规律时，易犯“横看成岭侧成峰”的错误，且存在“远近高低各不同”的跨尺度解译需求。地理现象或事件是在空间、时间、尺度、属性上的综合体现，“如何从上述4个角度联合解译地理特征”成为地理数据分析领域面临的时代挑战。鉴于双向聚类在地理领域鲜有应用，作者团队针对荷兰物候时空分异[19]、全球自然灾害事件时空分异[20]、中国春季物候时空分异模式[21]等问题，引入双向聚类方法，探索了其对“多空间—多属性”“多空间—多时间—单属性”等地理数据的解译能力。为进一步拓展地理研究的联合解译能力，作者团队提出并研发了一种全新的三向聚类算法[22-23]，并在北京城区$PM_{2.5}$时空分异模式研究中验证了三向聚类在不同时间尺度下对“多空间—多时间—单属性”的联合解释能力[23]。

2 聚类方法从单向到三向的变革

简单地说，聚类是对大量未知标注的数据集，按数据内在的相似性将其划分为多个类别（成为“簇”或“类”），并使类内数据尽可能相似、而类间数据尽可能不相似。

2.1 单向聚类

以层次聚类为代表的传统聚类方法是以样本为研究对象，将样本视为由属性构造的n维空间中的点，通过测量两点（样本）间的距离或矢量余弦相似度，将接近的样本聚为一类。传统的聚类方法可分为两种：一种是基于所有属性对样本进行聚类，例如根据一些经济指标对中国各城市进行分级；另一种是基于所有样本对属性进行聚类。这类仅对样本或属性进行聚类的方法称“单向聚类”（图1a）。

单向聚类在过去几十年科学分类的研究中起到了重要的作用。但是，来自基因、文本分析领域的大量实践发现了单向聚类的不足。① 单向聚类仅关注样本在“大部分”的属性上的相似程度，易造成部分相似极强的特征信息被忽略。特别是当属性数目（维数）过高时，单向聚类通常用主成分分析（Principal Components Analysis, PCA）进行降维并选择主要成分参与聚类，从而进一步证实了单向聚类重视主要信息而忽略次要信息的特点。在物种基因表达的研究中，单向聚类的这一缺点被充分暴露出来。例如，在成千上万的基因中，某些生物功能往往仅在少量基因片段上表现出极强的相似性，而关注

图1 单向聚类的意图及案例

Fig. 1 The demo and example of one-way clustering

大部分基因的单向聚类则无法分辨这些局部相似的信号。② 单向聚类仅从某个（行或列）角度观察数据，易犯“横看成岭侧成峰”的错误。以图1b所示数据片段为例，若对样本进行聚类，样本1和样本2会被聚成一类；若对指标进行聚类，指标1和指标2会被聚成一类；但联合观察样本和指标数值分布，不难发现：样本1~3在指标1~3上具有更强的相似度，然而单向聚类不擅长此类数据子集相似性的探测。③ 面对稀疏高维矩阵（数据缺失值多、维数高），距离函数或夹角余弦的计算及其有效性也面临挑战。首先，对于缺失的数值、名义量以及类型量，如何将其转化为数值参与距离和夹角的运算？其次，有研究表明在2~10维的低维空间中，“用欧氏距离来度量数据之间的相似性”是有意义的，但在更高维空间中欧氏距离就逐渐失去了其度量数据相似度的作用。

2.2 双向聚类

双向聚类是基于数据矩阵内元素值的相似性，形成一个子矩阵分割方案，使子矩阵内的元素尽可能相似，子矩阵间的元素相似度尽可能低，从而实现行列两方向的同时聚类。稀疏双向聚类、谱双向聚类和信息双向聚类是目前常见的双向聚类方法。下面以信息双向聚类为例，介绍双向聚类的特点。

信息双向聚类实质是将行（x）和列（y）视为两随机变量，通过不断移动调整样本在x向的位置，或属性在y向的位置，找到一种分割（聚类）方案，使该分割方案下概化后的数据分布与概化前的分布尽可能接近，以保证子矩阵内元素值尽可能相似、而子矩阵间元素差异尽可能大，如图2a所示。算法流程如下：① 根据用户给定样本向分类数（k）、属性向分类数（l），随机生成一个$k\times l$的分割（聚类）方案，将原始数据矩阵划分成$k\times l$子矩阵。② 将数据矩阵的行和列可视为两个随机变量，在不考虑$k\times l$分割的情况下，原始矩阵元素反映了一个非常精细的二元联合分布；若考虑$k\times l$分割方案，则将各子矩阵内所有元素都赋为该子矩阵所有元素样本数据的均值，形成了一个该聚类结果概化后的数据二元联合分布。③ 计算这两个分布的信息散度，用于衡量两分布的接近程度，即信息散度越小表示两分布越接近，分类后的信息量丢失也越少，聚类效果越好。④ 在当前的分割方案下，尝试着逐步把不同的行或列调换到其它分区中，形成新的$k\times l$分割方案；重复第②~③步；如果调换后的信息散度小于调换前的信息散度，则调换后的方案更优，保存为当前聚类方案。⑤ 重复执行第④步，直到找不到信息散度更小的方案，则得到了一个较优的数据分割方案。上面仅为信息双向聚类的基本框架，为了尽快收敛、跳出局部最优，还有很多算法值得加入、功能值得完善。

可见，双向聚类则是对数据块（数据子集）的聚类。双向聚类可以检测出图1b所示的样本、指标的局部相似性；也可以检测出图2b中亮绿色区域对应的基因在一组亲缘关系密切的亚种（例如细菌）中具有很强的活性。

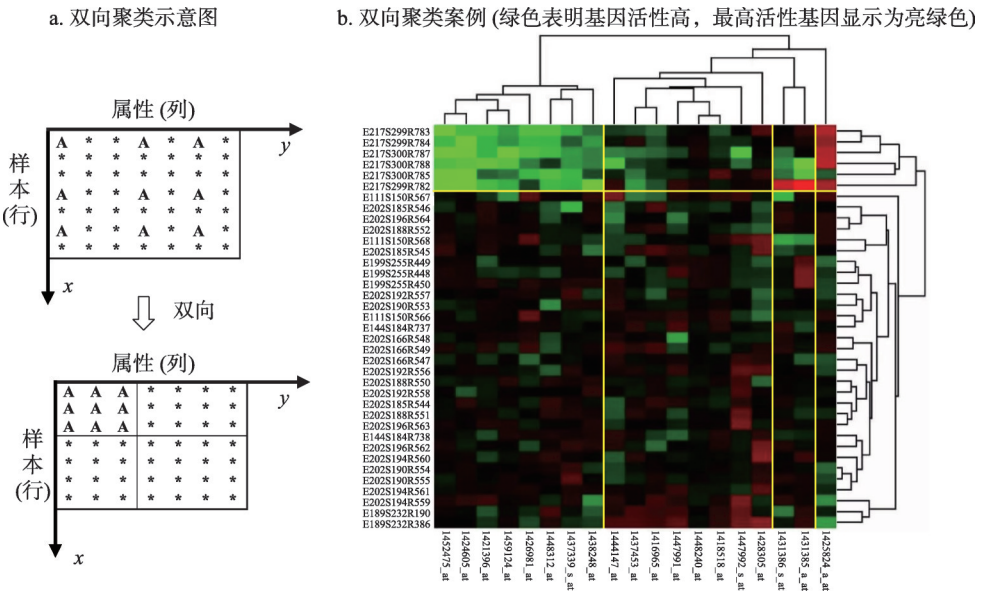


图2　双向聚类示意及案例

Fig. 2　The demo and example of co-clustering

双向聚类优势如下：① 联合解译：双向聚类通过对数据子集的聚类，可以实现样本和属性的联合解译，在地理研究中可以实现时间、空间、尺度、属性中任意两方向的联合解译[18, 20]；② 局部相似性的探测：双向聚类是基于数据块（子集）的聚类，只要在“部分”属性上存在相似性即可，故双向聚类可以解决一部分单向聚类无法分割的问题，对于发现地理现象的局部异质特征有意义[18]。

2.3 三向聚类

双向聚类可以从两个不同方向联合解译数据的分异性规律[20-21]。但是地理现象或事件通常是在时间、空间、尺度、属性（简称：空间—时间—尺度—属性）上的综合体现；因此地理研究往往需要对超出双向的数据进行联合解译。为此作者团队基于双向聚类思想，提出了三向聚类算法，研发了相应的程序包[22-23]。

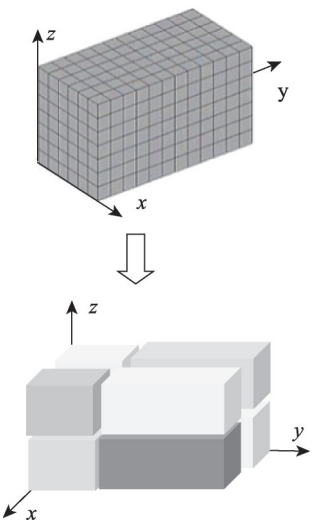


图3　三向聚类示图

Fig. 3　Demo of tri-clustering

三向聚类是将数据的行（x）、列（y）、高（z）视为随机变量；通过不断移动或调整研究对象在x、y、z方向上的位置，找到一种数据在三维空间的分割方案；聚类后使三维子数据体内元素尽可能相似、而子数据体间尽可能保持较大差异，如图3所示。其核心算法的伪代码如表1所示，算法思路与机器学习中逐步寻优的过程类似。为了避免陷入局部最优，可以选多个随机种子，重复执行表1的算法，最后选择信息损失最小的分割方案，作为聚类方案。

构成三向聚类数据体x、y、z维度可以分别从空间、时间、尺度、属性空间中选择3个不同方向进行组合。由于三向聚类可以探测数据在$\{x, y, z\}$联合空间中数据的局部相似和全局分异规律，因此可以实现地理现象多维度的解译[22-23]。此外，三向聚类也可以处理高维、低质量的数据，具有时代优

表1 三向聚类核心算法的伪代码

Tab. 1 Pseudo-code of tri-clustering

算法：基于信息散度(I-divergence)的立方体平均三向类算法

输入：O_0(数据立方体)，k(方向1的聚簇数量)，l(方向2的聚簇数量)，m(方向3的聚簇数量)，

输出：优化后的$k\times l\times m$的三向聚类结果

开始：

1. 初始化：基于原始数据O_0，方向1~3上分别被随机分为k、l、m个区域(聚簇)，该数据体和数据分割方案，记为O_i；

2. 对O_i各区域内数据求均值，并用均值代替区内各元素，形成该分割方案下聚类结果的概化数据体$\hat{O}_i$；

3. 计算信息散度(目标函数)：$f_i=D(O_i\|\hat{O}_i)$ /*f_i表征在该分割方案下的概化后的数据体($\hat{O}_i$)与概化前的数据体(O_i)的接近程度，值越小越接近；即数据子集内元素越相似、而数据子集间元素差异越大*/

4. 开始迭代：

4.1 以O_i数据体及其分割方案为基础，在行或列或高的方向上，按一定规则，逐步尝试将O_i中的数据向量在所属方向的不同区间移动或交换，形成新的数据体和分割方案，记为O_j；

4.2 对O_j各区域内数据求均值，并用均值代替区内各元素，形成该聚类结果的概化数据体$\hat{O}_j$；

4.3 计算信息散度：$f_j=D(O_j\|\hat{O}_j)$

4.4 若$f_j<f_i$，则$O_i=O_j$，$f_i=f_j$，并跳转到4、开始下一次迭代；否则，直接跳转到4、开始下一次迭代

5. 结束迭代(直到目标函数收敛)

结束

势。主要表现在以下3个方面：① 大数据时代，可能存在数据的维数高且仅关注少数维数据相似性的应用需求。② 大数据时代，数据可能存在较高比例的缺失值、甚至可能是稀疏矩阵。例如，2013年2月—2014年1月期间北京城区18个环境监测站$PM_{2.5}$观测值的缺失率达到22.54%。在传统基于距离函数的聚类中，如此多的缺失值难以参与运算。而三向聚类是基于数据块的聚类，其信息度量函数是相对熵，因此支持缺失值的聚类，当然也支持名义变量和类型变量的聚类。③ 大数据时代，地理数据维数可能达到数十、数百级别。受“维度效应”影响，传统聚类方法的聚类效果和运算效率都面临挑战。三向聚类专为高维矩阵设计，不受“维度效应”影响。文献[21, 23]分别对74154行×40列数据矩阵、18行×299列×24层数据体进聚类，结果表明三向聚类在聚类效果和运算效率上有较好表现。

图4 三向聚类研究流程图

Fig. 4 Workflow of using tri-clustering

3 三向聚类的应用流程与地理时空数据组织

3.1 应用流程

运用三向聚类方法解译地理特征的流程（图4）。首先，根据研究涉及的地理问题，确定地理时空密度、界定时空尺度、选择多种地理属性，组

织参与聚类的时空数据体；然后，采用三向聚类算法得到聚类结果；之后，将聚类结果定位到相应的时间、或空间、或尺度上制图，根据绘制的图或谱解译地理现象的时空分异规律。

三向聚类在算法层面上是通用的，不同之处仅在于：根据不同问题涉及的空间、时间、尺度、属性的不同，构建出不同的数据体；不同数据体聚出的结果回答不同的地理问题。

3.2 地理时空数据的组织

地理研究通常涉及不同尺度的空间、时间以及时空叠加的问题，特征属性通常涉及单属性、多属性两类。图5a和图5b分别给出了在单属性、多属性情况下，面向不同空间、时间、尺度问题的数据组织方案。其中，横、纵方向分别给出了对应的空间、时间与尺度问题；图中区域被实线分为左上、左下、右上、右下4个区域，其中后3个区域分别对应空间问题、时间问题、时空叠加问题；不同区内，面向尺度又用虚线进行分割。因此，每个单元格对应横、纵表头所联合描述的地理问题，单元格内的矩阵或数据体则是解决对应问题的数据组织形式。图5中①~⑤是二维数据表，适用于双向聚类；⑥~⑫是三维数据体，适用于三向聚类；对于表中右斜下方空白单元格对应的更复杂的问题，需要更高向的聚类方法。

结合已有相关的研究实践[20-21, 23]，图6给出了图5中②、④、⑧和⑫的典型实例方案。中国春季物候时空分异特征研究[21]采用图5中②的“多空间—多时间—单属性”数据组织方案，形成了如图6a所示的数据矩阵；其中灰色区域给出了1979—2018年期间中国领土内4万余个格网上每年紫丁香开花始期的序日。全球自然灾害频发率空间格局的研究[20]采用图5中④的“多空间—多属性”数据组织方案，形成了如图6b所示的数据矩阵；其中灰色区域给出了全球200余个国家地区，对应的11种自然灾害每万平方公里的发生率。北京城区$PM_{2.5}$在不同尺度下时空格局与过程研究[23]则采用了图5中⑧的“双时间尺度下—多空间—单属性”的数据组织方案，形成了如图6c所示的数据体；给出了不同监测站不同天不同小时$PM_{2.5}$的监测值；此外，为了验证三向聚类在多属性方面的分析能力，后续拟采用图5中⑫给出的“多时间—多空间—多属性”的数据组织方案，形成如图6d所示的数据体，探讨不同监测站在不同时间各污染物指标的时空分异及相互作用规律。

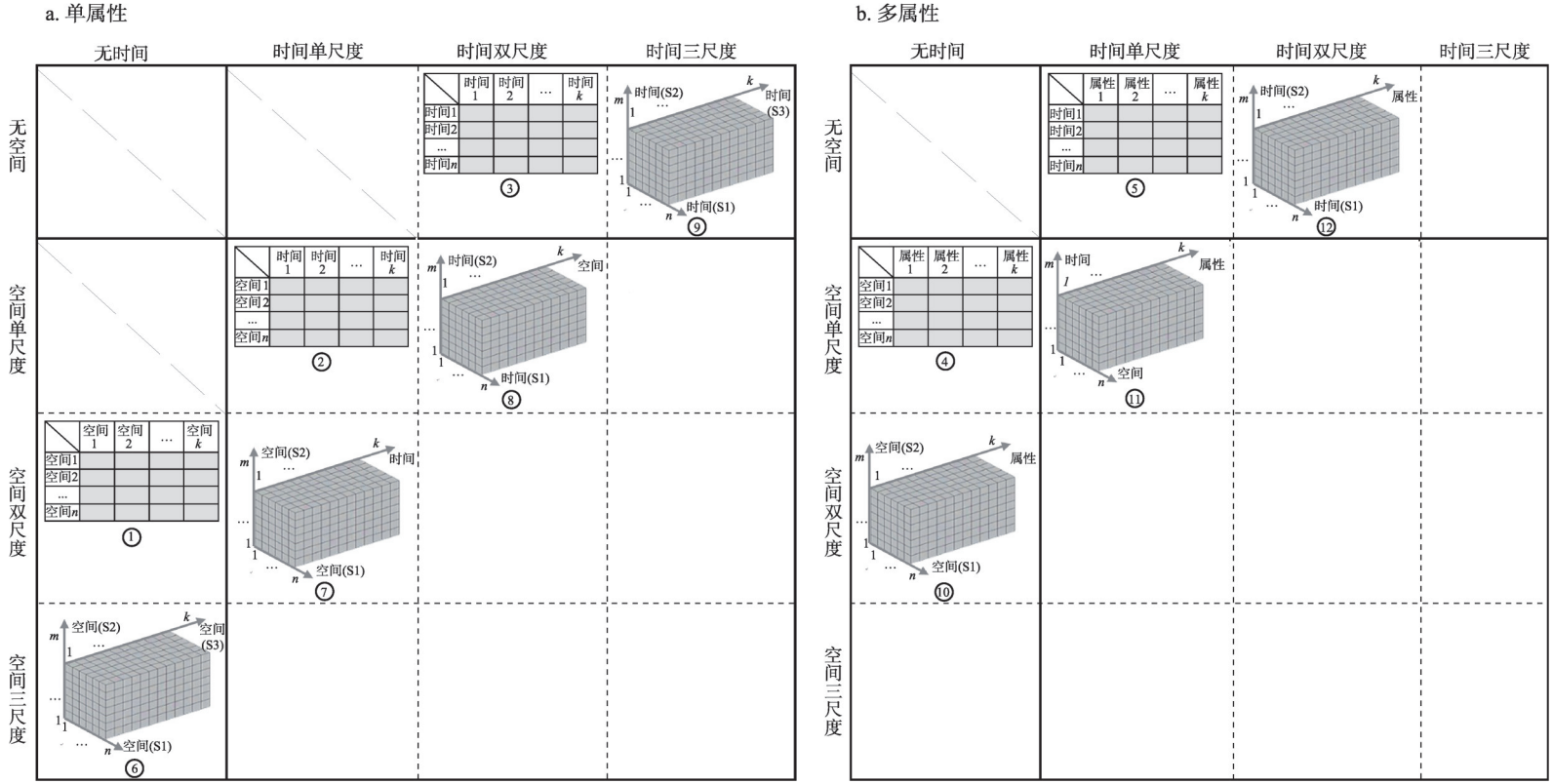
当然，目前三向聚类在空间上仅支持网格数据的多尺度研究，如何实现县、市、省、国家这类任意多边形数据的多尺度聚类，仍有待深入研究。

4 地理时空分异与叠加效应的解译实践

4.1 多方向、多尺度解译北京城区$PM_{2.5}$时空分异特征

以北京城区$PM_{2.5}$在不同尺度下的时空格局与过程研究[23]为例，对图6c的数据体进行三向聚类后，结果如图7a所示；其空间分层异质性统计量（SSH）q值[24]为0.87，表示聚类效果良好。图7a中不同的色块代表聚类结果中不同数据体子集，扁平化显示的各数据体子集的$PM_{2.5}$均值参见图7a中条形图例。该聚类结果在空间、天尺度、小时尺度上进行定位后，分别得到图7b~图7d。

4.1.1 多方向解译 所谓多方向解译是指既可以沿着聚类结果的各方向分别解译，也可以将多个方向联合在一起进行解译。特别是针对类内方差较小的数据子集，多方向联合解译能更体现时空一体化的分异解释。



注：a、b中部的区域被实线分为左上、左下、右上、右下4个区域，其中后3个区域分别对应空间问题、时间问题、时空叠加问题，不同区域内，面向尺度又用虚线进行分割；单元格内的二维表格适用于双向聚类，三维数据体则适用于三向聚类，白色区域对应更高维度数据需要更高向的聚类方法；图中灰色区域为对应的属性测量值。

图5　面向“空间—时间—尺度—属性”聚类数据组织方案

Fig. 5　Data organization schemes for clustering according to "space-time-scale-attribute"

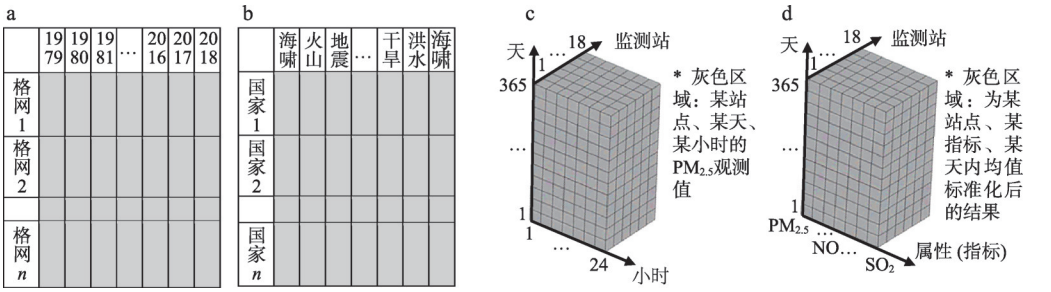


图6　根据不同地理问题组织的数据矩(体)的案例

Fig. 6　Examples of data matrices according to different geographic topics

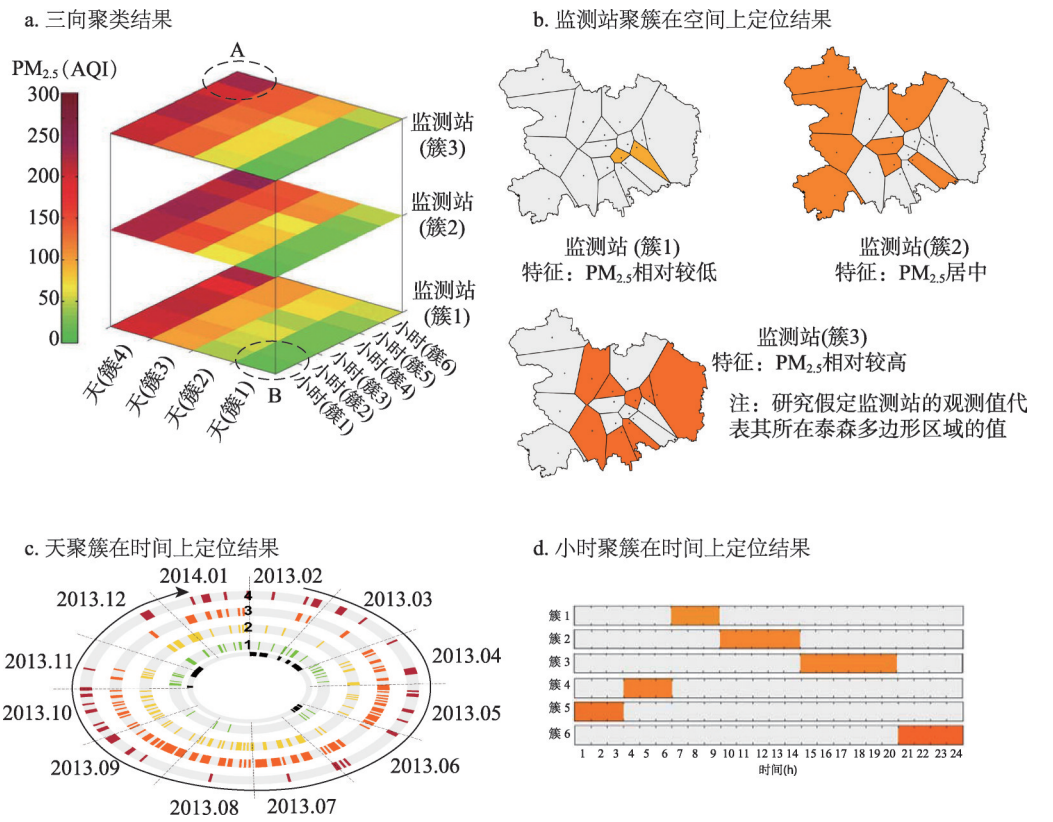


图7　$PM_{2.5}$三向聚类结果及其时空分异规律的解译(改自文献[23])

Fig. 7　Tri-clustering results of $PM_{2.5}$ and the interpretation of spatio-temporal differentiations (revised from [23])

① 各方向分别解译：仅从空间方向上看（图7b），前门东大街、美国大使馆区的$PM_{2.5}$值相对较低，北京西部的奥体、官园、万寿寺、天坛等区域的$PM_{2.5}$值居中，北四环北路、农展馆、东四、丰台花园、南三环西路、永定门大街等区域$PM_{2.5}$值较高[23]。当然，也可以分别从天尺度（图7c）或小时尺度上（图7d），解译聚类结果。

② 多方向联合解译：以图7a中$PM_{2.5}$值最高的数据块A为例，可以联合空间、天尺度、小时尺度3张图，联合解译数据块A；即这种污染最严重的情况通常发生在图7b监测站（簇3）显示地区，且时间上通常集中在图7c簇4所在的天上，且通常集中于图7d簇6所示的时间段内。当然，也可以用同样的逻辑解译图7a中$PM_{2.5}$值最低的数据块B。

4.1.2 多尺度解译 所谓多尺度解译，主要针对数据体中各向存在不同时间尺度的情况（图6c）；通过多尺度解译可以读出地理特征中“远近高低各不同”的分异规律。根据图7c和图7d，可知：① 北京城区$PM_{2.5}$在天尺度上的分异不显著：相对而言$PM_{2.5}$浓度较高的聚簇4主要分布在2013年10月，浓度较低的聚簇1主要分布在2013年4月—5月、2013年11月至2014年1月期间（图7c）[23]；② 北京城区$PM_{2.5}$在小时尺度上分异相对显著：$PM_{2.5}$值较高的簇5和簇6主要分布于晚21点到早4点之间；$PM_{2.5}$值较低的簇1和簇2主要位于早7点到午14点之间（图7d）[23]。

4.2 多层次解译1979年以来中国春季物候时空分异特征

所谓多层次嵌套解译是指先沿某方向解译聚类结果，然后针对某聚簇再深入分析它在其它方向上的分异规律。下面以中国1979年以来春季物候时空分异研究[21]为例，介绍多层次嵌套解译的实践。

4.2.1 “先空间再时间”的联合解译 对图6a的中国74154个格网上1979—2018年紫丁香开花始期序日进行双向聚类后，其SSH q值为0.91，表示聚类效果良好。对聚类结果可先进行空间定位，如图8a所示；再进行时间定位，如图8b所示。

图8a展示了中国1979年以来紫丁香开花始期序日呈现的3种格局（聚类结果）；根据开花序日特早与较早的分界线变化，可知格局1~3分别表示花始期序日不断提前的状态。将图8a空间解译的结果代入到图8b可以实现中国过去40年开花始期的时空格局与演化过程的联合解译，并将其时空演化可分为4个阶段：① 普遍较晚期（1978—1995年）：中国开花始期序日在格局1和2之间波动，开花始期普遍较晚，江西、新疆北部和内蒙古中部开花始期变化频繁；② 集聚提前期（1996—1998年）：中国开花始期序日集聚提前，呈现从格局1到格局2再到格局3的直线提前趋势，中国西南和东北区域开花始期都呈现大幅提前趋势；③ 波动提前期（1999—2012年）：中国开花始期序日在格局2

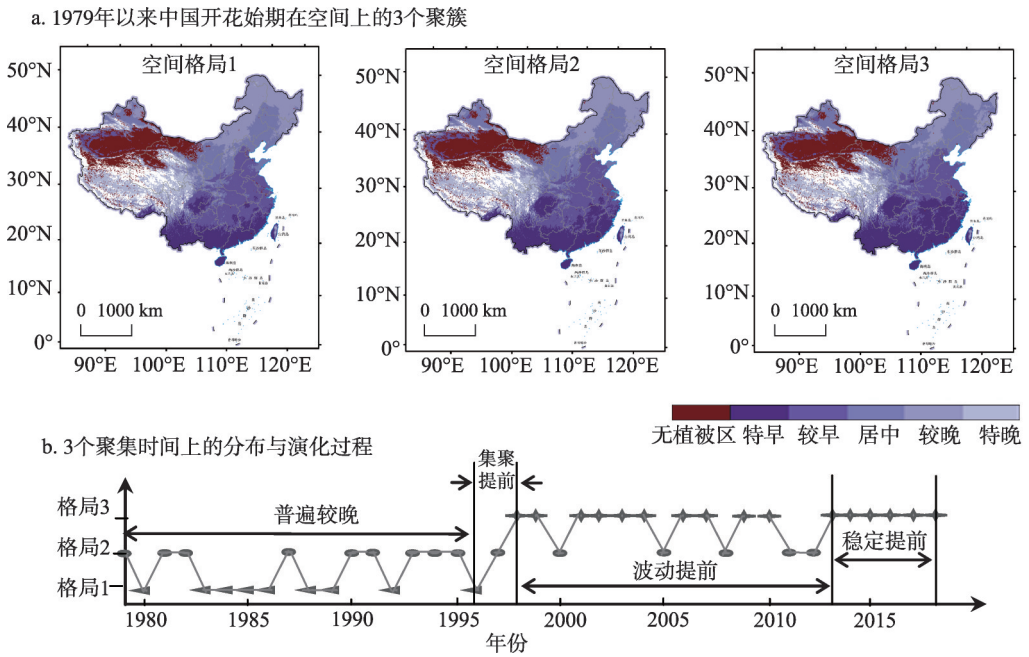


图8 多层次嵌套(先空间再时间)解译案例(改自文献[21])

Fig. 8 The example of multi-level nested (first space and then time) interpretation (revised from [21])

和格局3之间波动，开花始期呈现波动性提前，福建、湖南和黑龙江东部开花始期变化频繁；④ 稳定提前期（2013—2018年）：中国开花始期序日，呈现稳定提前状态[21]。

4.2.2 “先时间再空间”的联合解译　对于上述聚类结果，也可以采用先时间再空间的方式解译。从时间的聚类结果来看，开花始期的变动趋势被聚成15类；根据开花始期趋势线的波形，可将其概括为平稳、先平稳后波动、先波动后平稳、频繁波动和剧烈波动等5类，如图9a所示[22]。再根据如图9b示出的不同时间趋势对应的空间分布，可知：绿色区域（中国大部分地区）的开花始期基本处于平稳态，蓝色区域（贵州北部、湖南和湖北南部等区域）的开花始期呈现先平稳后波动上升的趋势，而橘黄区域（四川东部、湖南东南部和江西北部等区域）呈现先波动后平稳上升的趋势[21]。

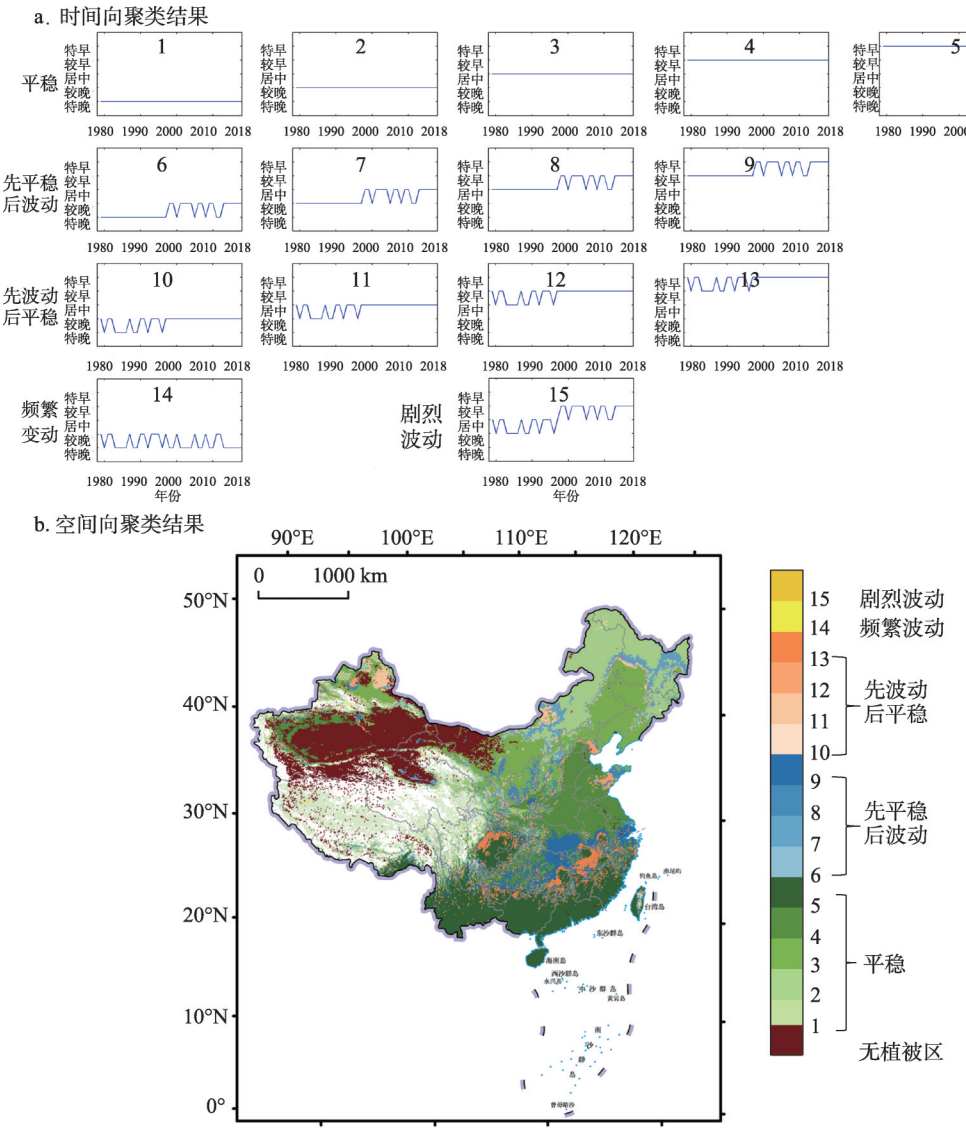


图9　多层次嵌套(先时间再空间)解译案例(改自文献[21])

Fig. 9　The example of multi-level nested (first time and then space) interpretation (revised from [21])

5 结论与讨论

本文将基因和文本分析领域双向聚类方法引入到地理学领域，分析了从单向到三向聚类构建思路的变革，系统辨析了三向聚类的优势，给出了利用三向聚类开展地理时空格局与过程研究的流程，并结合实践案例，展示了面向“空间—时间—尺度—属性”的数据三维矩阵组织思路，展示了如何多方向、多尺度、多层次嵌套地联合解译聚类结果、揭示地理特征时空分异的叠加效应。结论如下：① 三向聚类是一种大数据时代探测地理数据时空分异规律的有效方法，可以解决数据维度高、质量低等问题；② 面对不同的地理问题，三向聚类在算法层面上是通用的，不同之处仅在于：根据不同问题涉及的空间、时间、尺度、属性的不同，构建不同的数据体；不同数据体聚出的不同结果回答不同的地理问题；③ 三向聚类可以实现地理数据的时空分异规律多方向、多尺度、多层次的联合解译，揭示地理特征时空尺度叠加效应。

除理解三向聚类的核心思想和理论外，对地理问题的深入理解以及数据三维矩阵的组织范式是运用三向聚类开展地理问题研究的基础和关键。研究者可以参照图5给出的地理问题数据组织方案开展研究。如何探索和发展出更多适用不同地理问题的数据组织方案是未来研究的重要方向和基础之一，期待未来能够提升三向聚类方法在多空间尺度、多属性方面的地理研究实践。

参考文献(References)

[1] Fu Bojie. Geography: From knowledge, science to decision making support. Acta Geographica Sinica, 2017, 72(11): 1923-1932. [傅伯杰. 地理学: 从知识、科学到决策. 地理学报, 2017, 72(11): 1923-1932.]

[2] Song Changqing, Cheng Changxiu, Shi Peijun. Geography complexity: New connotations of geography in the new era. Acta Geographica Sinica, 2018, 73(7): 1189-1198. [宋长青, 程昌秀, 史培军. 新时代地理复杂性的内涵. 地理学报, 2018, 73(7): 1189-1198.]

[3] Wang Jinfeng, Ge Yong, Li Lianfa, et al. Spatiotemporal data analysis in geography. Acta Geographica Sinica, 2014, 69(9): 1326-1345. [王劲峰, 葛咏, 李连发, 等. 地理学时空数据分析方法. 地理学报, 2014, 69(9): 1326-1345.]

[4] Cheng Changxiu, Shi Peijun, Song Changqing, et al. Geographic big data: A new opportunity for geography complexity study. Acta Geographica Sinica, 2018, 73(8): 1397-1406. [程昌秀, 史培军, 宋长青, 等. 地理大数据为地理复杂性研究提供新机遇. 地理学报, 2018, 73(8): 1397-1406.]

[5] Xie Yan, Li Dianmo, John MacKinnon. Preliminary researches on bio-geographic divisions of China. Acta Ecologica Sinica, 2002, 22(10): 1599-1615. [解焱, 李典谟, John MacKinnon. 中国生物地理区划研究. 生态学报, 2002, 22(10): 1599-1615.]

[6] Wang Xiuhong. Application of multivariate statistical analysis in regionalization study. Scientia Geographica Sinica, 2003, 23(1): 66-71. [王秀红. 多元统计分析在分区研究中的应用. 地理科学, 2003, 23(1): 66-71.]

[7] Zheng Du, Ou Yang, Zhou Chenghu. Understanding of and thinking over geographic regionalization methodology. Acta Geographica Sinica, 2008, 63(6): 563-573. [郑度, 欧阳, 周成虎. 对自然地理区划方法的认识与思考. 地理学报, 2008, 63(6): 563-573.]

[8] Song Ci, Pei Tao. Research progress in time series clustering methods based on characteristics. Progress in Geography, 2012, 31(10): 1307-1317. [宋辞, 裴韬. 基于特征的时间序列聚类方法研究进展. 地理科学进展, 2012, 31(10): 1307-1317.]

[9] Hartigan J A. Direct clustering of a data matrix. Journal of the American Statistical Association, 1972, 67(337): 123-129.

[10] Cheng Y, Church G M. Biclustering of expression data. Eighth International Conference on Intelligent Systems for Molecular Biology. Menlo Park: AAAI Press, 2000: 93-103.

[11] Xiong Yun, Qiu Boren, Zhang Kun, et al. Gen-Cluster: An efficient gene expression data high dimensional clustering algorithm. Journal of Fudan University (Natural Science), 2008, 47(2): 135-146. [熊赟, 邱伯仁, 张坤, 等. Gen-Cluster:

一个基因表达数据的高维聚类算法, 复旦大学学报(自然科学版), 2008, 47(2): 135-146.]

[12] Liu Wei, Chen Ling. Parallel biclustering algorithm for gene expressing data. Journal of Chinese Computer Systems, 2009, 30(4): 683-689. [刘维, 陈崚. 基因表达数据的并行双向聚类算法. 小型微型计算机系统, 2009, 30(4): 683-689.]

[13] Wu Lei, Li Shu. Knowledge discovery on compatibility laws of TCM prescription for stroke disease based on biclustering method. Chinese Journal of Information on TCM, 2013, 20(11): 16-19. [吴磊, 李舒. 基于双向聚类方法的中医治疗中风病方剂配伍规律知识发现. 中国中医药信息杂志, 2013, 20(11): 16-19.]

[14] Xu Su, Li Wei. Analysis on the precision medicine research hotspots by biclustering. Medicine and Philosophy, 2015, 36(6B): 1-34. [徐速, 李维. 精准医学研究热点的双向聚类计量分析. 医学与哲学, 2015, 36(6B): 1-34.]

[15] Niu Yujin, Hu Yaping, Li Li. Biclustering econometric analysis of research hotspots of general medicine. Chinese General Practice, 2016, 19(36): 4428-4433. [牛玉敬, 胡亚平, 黎莉. 全科医学研究热点双向聚类计量分析. 中国全科医学, 2016, 19(36): 4428-4433.]

[16] Yao Qiang, Zhang Yan, Zhang Shijing. The application of biclustering in bibliometrics: A case study of performance evaluation of hospital. Journal of Intelligence, 2012, 31(3): 54-59. [姚强, 张研, 张士靖. 双向聚类在文献计量学中的应用初探: 以医院绩效评价为例. 情报杂志, 2012, 31(3): 54-59.]

[17] Su Pan, Wang Anni, Zhang Jie. Status quo and hot spots in studies on family caregivers: A bibliometric analysis. China Journal Medicine Library Information Science, 2017, 25(9): 34-42. [苏盼, 王安妮, 张杰. 基于文献计量学的家庭照顾者相关研究现状及热点分析. 中华医学图书情报杂志, 2017, 25(9): 34-42.]

[18] Fang Quan, Liu Yizhen, Lin Zhaohui, et al. Research on quercus variabilis community characteristics and diversity of Yunjun Mountain. Plant Science Journal, 2015, 33(3): 311-319. [方全, 刘以珍, 林朝晖, 等. 云居山栓皮栎群落特征及多样性研究. 植物科学学报, 2015, 33(3): 311-319.]

[19] Wu Xiaojing, Zurita-Milla R, Kraak M J. Co-clustering geo-referenced time series: Exploring spatio-temporal patterns in Dutch temperature data. International Journal of Geographic Information Science, 2015, 29(4): 624-642.

[20] Shen Shi, Cheng Changxiu, Song Changqing, et al. Spatial distribution patterns of global natural disasters based on biclustering. Natural Hazards, 2018, 92(3): 1809-1828.

[21] Wu Xiaojing, Cheng Changxiu, Qiao Cancan, et al. Spatio-temporal differentiation of spring phenology in China driven by temperatures and photoperiod from 1979 to 2018. Science China Earth Sciences, 2020, 63. Doi: https://doi.org/10.1007/s11430-019-9577-5. [吴晓静, 程昌秀, 乔灿灿, 等. 光温驱动下中国1979—2018年春季物候时空分异规律. 中国科学: 地球科学, 2020, 50. Doi: 10.1360/SSTe-2019-0212.]

[22] Wu X J, Zurita-Milla R, Izquierdo-Verdiguier E, et al. Triclustering georeferenced time series for analyzing patterns of intra-annual variability in temperature. Annals of the American Association of Geographers, 2018, 108(1): 71-87.

[23] Wu X J, Cheng C X, Zurita-Milla R, et al. An overview of clustering methods for geo-referenced time series: From one-way clustering to co- and tri-clustering. International Journal of Geographic Information Science, 2020. Doi: 10.1080/13658816.2020.1726922.

[24] Wang Jingfeng, Xu Chengdong. Geodetector: Principle and prospective. Acta Geographica Sinica, 2017, 72(1): 116-134. [王劲峰, 徐成东. 地理探测器: 原理与展望. 地理学报, 2017, 72(1): 116-134.]

Tri-clustering: Construction and practice of space-time integrated analysis tool

CHENG Changxiu[1,2,3], SONG Changqing[1,2], WU Xiaojing[1,2], SHEN Shi[1,2], GAO Peichao[1,2], YE Sijing[1,2]

(1. State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; 2. Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; 3. National Tibetan Plateau Data Center, Beijing 100101, China)

Abstract: With the improvement of geographic data acquisition capabilities, the volume of geographic data has been growing exponentially, and the data types as well as characteristics have become more diverse. The effective identification and classification of data has become the key to understand spatio-temporal patterns, evolutionary processes, and driving mechanisms of geographic phenomena. However, traditional clustering methods are facing some challenges, such as large amount, high-dimensionality and poor-quality of the data to be dealt with. Therefore, it is necessary to improve clustering methods. This paper first describes the transformation from one-way clustering to tri-clustering. One-way clustering methods perform the clustering analysis along with the samples or the attributes. They played an important role in previous studies, but ignored local features that are very similar. Co-clustering methods perform the submatrix partitioning scheme based on location similarity of elements within the data matrix. They avoid shortages of one-way clustering by realizing the clustering from both rows and columns, making similar elements into the same submatrix and dissimilar ones into different ones. However, they cannot satisfy multiple directions interpretations of geographical research since they do not support 3D panel data body. Then, we develop a new tri-clustering method, presents the workflow of using tri-clustering to spatio-temporal patterns' studies, and summarizes how to construct the 3D data matrix for clustering according to different aspects of 'space-time-scale-attribute' involved in the analysis. Finally, we show some practices of tri-cluster. The results show that: (1) Tri-clustering is an effective method to identify the spatio-temporal differentiation of geographic data in the era of big data by solving problems, i.e. data of high dimensionality and low quality. (2) Tri-clustering is universal in the algorithmic level when facing different geographic topics, but the differences rely on the 3D data matrices constructed according to different aspects of "space-time-scale-attribute" involved in the analysis. And, different data matrices are clustered to different results, which answer different topics. (3) Tri-clustering is able to interpret the spatio-temporal differentiation of geographic data in multiple directions, multiple scales, and multiple hierarchies, and thereby reveal the superposition effects of spatio-temporal scales of geographic features. Finally, we emphasize the significance of constructing 3D data matrices based on different geographic topics and expect that tri-clustering methods can enhance the ability to analyze geographic data with multiple spatial scales and attributes in the future.

Keywords: tri-clustering; space-time-scale-attribute; integrated interpretation; spatio-temporal local similarity; spatio-temporal differentiation