



时空聚类分析的普适性方法

邓敏*, 刘启亮, 王佳璆, 石岩

中南大学测绘与国土信息工程系, 长沙 410083

* 通信作者. E-mail: dengmin028@yahoo.com

收稿日期: 2010-10-10; 接受日期: 2011-03-08

国家高技术研究发展计划 (批准号: 2009AA12Z206)、国家自然科学基金 (批准号: 40871180) 和中南大学前沿研究计划 (批准号: 2010QYZD002) 资助项目

摘要 时空聚类分析是时空数据挖掘领域近年来研究的热点问题, 对于揭示时空要素的发展变化趋势、规律以及本质特征具有重要意义. 目前, 时空聚类分析的研究仍还初步, 缺乏具有普适性的时空聚类分析方法. 为此, 本文首先建立了一套时空聚类分析的普适性理论方法框架. 进而, 借助时空统计学、智能计算等工具, 提出了一种时空一体化的时空聚类方法. 该方法很好地顾及了时空数据的时空耦合、时空相关与时空异质特征, 避免了过多人为主观因素的干扰, 时空聚类结果具有较好的可靠性. 通过采用中国大陆地区 42 年 (1951~1992) 年平均气温时空数据进行分析, 验证了本文提出的理论与方法的可行性与有效性.

关键词 时空聚类 时空统计学 时空相关性 时空异质性 时空数据挖掘

1 引言

时空聚类分析作为时空数据挖掘的一个主要研究内容, 是计算机科学与地球信息科学领域交叉研究中的一个最前沿、最具挑战的研究课题^[1,2]. 时空聚类分析旨在从时空数据库中发现具有相似特征的时空实体集合 (即时空簇), 亦是传统的聚类分析从空间域到时空域的进一步扩展. 时空聚类在全球气候变化^[3,4]、公共卫生安全^[5,6]、地震监测分析^[7,8] 以及犯罪热点分析^[9] 等领域具有重要应用价值, 有助于更好地发现和分析地理现象发展变化的趋势、规律与本质特征.

国内外学者虽然已经针对时空聚类分析开展了初步的研究, 但依然面临两方面难点: (1) 时空数据具有时空耦合的特点. 时空数据是空间维与时间维的有机组合, 空间数据与时序数据的一些性质并不能直接推广到时空域. 例如, 空间数据不存在过去、现在、将来的区别, 而时空数据却具有这种特性; (2) 时空数据具有时空相关性与时空异质性两个重要性质. 时空数据的分布在时间和空间上具有关联性, 也会随着时间和空间的变化而产生差异. 此外, 当前的时空聚类分析研究仍没有形成一套具有普适性的理论方法框架. 为此, 本文将从时空数据的特征出发, 建立一套时空聚类分析的理论框架, 同时发展一种顾及时空相关性与异质性的时空聚类新方法. 下面, 首先对现有的相关研究进行简要回顾, 并提出本文的研究策略.

2 相关研究回顾及本文的研究策略

2.1 相关研究回顾

现有的时空聚类方法主要包括时空扫描统计方法^[5,6]、基于密度的方法^[4,7,8]以及基于时空距离的方法^[10~12]。最近, Kisilevich 等^[2]对时空聚类分析方法进行了较为系统的阐述, 本文将对部分典型的时空聚类方法进行具体分析。时空扫描统计方法在流行病时空聚集分析研究中具有重要应用, 其主要思想是采用一个预设的时空扫描窗口, 即由空间距离定义半径、时间间隔定义高度的圆柱体, 以每个时空实体为中心进行扫描, 借助统计检验的方法确定疾病爆发的聚集区域。该方法一方面需要预先假定数据的概率分布模型, 且结果受扫描窗口的影响较大, 另一方面不能详细描述时空簇的位置和形状信息^[8]。基于密度的方法是将经典的 DBSCAN(density-based spatial clustering of applications with noise) 算法^[13]进一步扩展到时间维。Birant 等^[4]和 Wang 等^[7]分别提出了两种类似的时空聚类方法, 并且都简称为 ST-DBSCAN。它们的区别在于, Birant 等的方法是采用相邻时间点定义时间邻近, 并在 DBSCAN 定义空间邻域的基础上, 进一步定义了非空间属性邻域, 从而可以同时顾及时空实体的空间属性与非空间属性进行聚类。然而, 该方法定义时空邻近关系涉及过多的主观因素, 并且参数设置过多(4个), 增加了用户实际使用的难度。Wang 等的方法旨在发现地震数据的时空聚集性, 其仅通过引入时间半径来对 DBSCAN 方法进行扩展。最近, Pei 等^[8]发展了一种基于窗口邻近的时空聚类方法, 其主要思想是定义一个圆柱形的时空邻近域, 在每个时空邻近域内区分聚集部分和噪声部分, 并采用 DBSCAN 密度相连的策略将聚集部分连接成时空簇。该方法只能针对仅存在两种 Poisson 分布(簇与噪声)的时空数据进行时空聚类操作, 且无法顾及时空实体的非空间属性。基于时空距离的方法可以区分为两类, 一类是从时间、空间两个角度分别定义时空邻近实体^[10,11], 这类方法通常需要人为设置阈值, 其适于发现同种类型时空实体的在时间上的连续变化情况, 但难以用来探测时空簇; 另一类是综合定义时空耦合距离^[12], 这类方法主要用于地震序列的时空聚集性发现, 在实际中时空属性的融合比较困难。Wu 等^[14]将时间序列视为空间位置的非空间属性进行聚类分析, 实质上割裂了时空要素的时空耦合特性。

综上所述, 可以发现目前尚缺乏一种能够同时顾及时空实体的非空间属性和时空实体时空耦合特性的时空聚类方法。为此, 下面将发展一种时空聚类分析的新策略。

2.2 本文的研究策略

聚类分析旨在将最相似的实体归入同一簇, 而不同簇中实体的差异性尽可能大, 主要包括 3 方面研究内容^[15]: (1) 数据的聚集趋势估计, 即判断数据能否进行聚类分析; (2) 聚类方法设计; (3) 聚类结果有效性评价。在地理空间中, 时间和空间上的相关性是时空实体的基本特征, 也是进行时空聚类分析的前提。若实体间没有相关性, 则不会产生明显的聚集现象^[16]。时空聚类旨在将时空相关性较强的时空实体聚在同一簇, 时空聚类过程中必须充分考虑实体间的相关性^[1]。因此, 时空聚类分析可以归纳为 3 个步骤。

首先, 需要对时空数据进行探索性分析, 掌握时空数据的特性。其主要包括: (1) 时空相关性分析, 判断时空数据是否可以进行时空聚类分析; (2) 时空平稳性分析, 分析时空数据的时空异质特征。其次, 根据时空数据的具体特点发展专门性的时空聚类方法。最后, 需要对时空聚类分析的结果进行分析和评价。时空聚类方法研究中的核心问题是如何确定时空邻近域, 本文借助 Delaunay 三角网与时空自回归移动平均模型^[17](STARIMA, space-time autoregressive integrated moving average) 中时空延迟

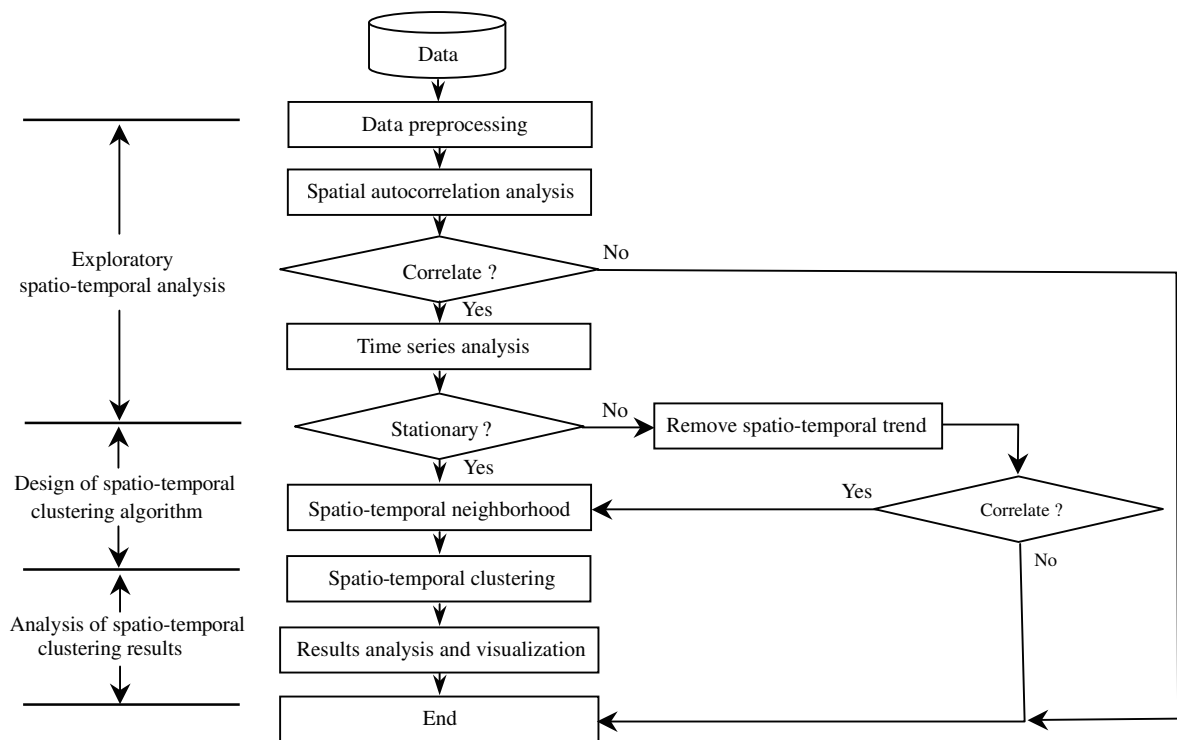


图 1 时空聚类分析的理论方法框架

Figure 1 The unified framework for spatio-temporal clustering

算子来构建一体化的时空邻近域. 由于 STARIMA 模型仅是针对时空平稳数据, 而对时空非平稳的时空数据, 需要剔除非线性的时空趋势获得平稳的时空序列后, 再构建时空邻近域. 进而, 借助基于密度聚类的思想完成时空聚类, 并对结果进行可视化分析. 整个时空聚类分析的理论方法框架如图 1 所示, 下面将分别阐述之.

3 时空聚类分析

3.1 时空数据探索性分析

时空数据探索性分析主要目的在于对时空数据的时空相关性与时空异质性 (非平稳性) 进行分析, 为时空聚类方法设计提供依据. 在进行时空数据探索性分析时, 为了保证分析结果的可靠性, 首先需要对数据进行必要的变换, 并对缺失数据进行相应处理. 进而, 可以借助 Moran'I 指数、Geary'C 指数、变差函数等方法^[18]对数据的空间相关性进行判断. 若时空数据在空间上不相关, 则表明数据表现为随机分布, 不存在聚集趋势, 故无法进行聚类分析. 最后, 对时空数据的平稳性进行检验, 空间平稳性可以借助空间趋势分析等方法进行判断, 而时间平稳性则可以通过时间序列分析的方法 (如自相关分析) 进行确定.

3.2 时空邻近域构建

一个时空实体的时空邻近域包括相邻时间点上所有空间邻近的实体. 如图 2 所示, 表示了 $A, B, \dots$,

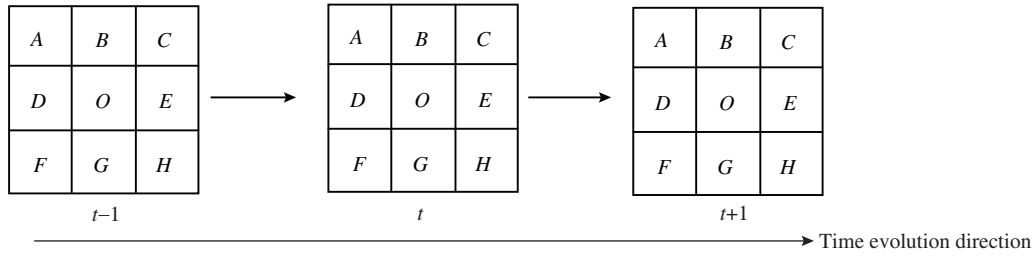


图 2 时空邻近域示意图

Figure 2 Sketch map of the spatio-temporal neighborhood

O 9 个空间实体在连续时间点 $t-1, t, t+1$ 上的演变情况, 若以八邻近方式定义空间邻近, 以直接时间相邻定义时间邻近, 一个时空实体定义为 $\langle \text{实体名}, \text{时间点} \rangle$ 的二元组, 如 t 时刻 A 实体表示为 $\langle A, t \rangle$. 则时间点 t 上 O 实体 $\langle O, t \rangle$ 的时空邻近实体集可以定义为: $\text{STN} = \{ \langle A, t-1 \rangle, \langle B, t-1 \rangle, \langle C, t-1 \rangle, \langle D, t-1 \rangle, \langle O, t-1 \rangle, \langle E, t-1 \rangle, \langle F, t-1 \rangle, \langle G, t-1 \rangle, \langle H, t-1 \rangle, \langle A, t \rangle, \langle B, t \rangle, \langle C, t \rangle, \langle D, t \rangle, \langle E, t \rangle, \langle F, t \rangle, \langle G, t \rangle, \langle H, t \rangle, \langle A, t+1 \rangle, \langle B, t+1 \rangle, \langle C, t+1 \rangle, \langle D, t+1 \rangle, \langle O, t+1 \rangle, \langle E, t+1 \rangle, \langle F, t+1 \rangle, \langle G, t+1 \rangle, \langle H, t+1 \rangle \}$. 可见, 时空邻近域构建需要获取两个重要参数: (1) 时间邻近范围, 即时间窗口; (2) 空间范围, 即空间邻近域. 根据时空数据探索性分析的结果, 本文借助时空统计学与 Delaunay 三角网, 分别针对时空平稳与时空非平稳的数据, 提出一种一体化时空邻近域构建方法.

首先考虑时空平稳的情况, 通过引入 STARIMA 模型中的时间延迟算子来进行时间窗口的获取. STARIMA 模型是一种线性时空动态预测模型, 其具体表达形式如下:

$$Z(t) = \sum_{k=1}^p \sum_{h=0}^{m_k} \phi_{kh} \mathbf{W}^{(h)} Z(t-k) - \sum_{l=1}^q \sum_{h=0}^{n_l} \theta_{lh} \mathbf{W}^{(h)} \varepsilon(t-l) + \varepsilon(t), \quad (1)$$

式中: p 为时间自回归阶数; q 为时间移动平均阶数; m_k 为第 k 个时间自回归项的空间阶数; n_l 为第 l 个时间移动平均项的空间阶数; k, h 分别为时间延迟和空间延迟; ϕ_{kh} 为在时间延迟 k 和空间延迟 h 处的自回归参数; θ_{lh} 为在时间延迟 l 和空间延迟 h 处的移动平均参数; $\varepsilon(t)$ 为随机误差.

从公式 (1) 可以发现, STARIMA 模型中采用时空延迟算子来反映时空相关性, 某个时间点 t 上, 特定空间位置上的非空间属性 $Z(t)$ 受到以前时间点上相同空间位置和空间邻近位置上非空间属性的共同影响, 这一性质充分顾及了时空数据的时空耦合特性. 时间延迟 k 表示某个时空实体受过去时间影响的最大范围, 空间延迟 h 表示时空实体在空间上受到其邻近实体影响的最大范围. 时间延迟 k 可以通过计算时空自相关函数与时空偏相关函数获得. 其中, 时空自相关函数可以表达为^[15]

$$\begin{aligned} \rho_{h0}(k) &= \frac{\gamma_{h0}(k)}{\sqrt{\sigma_h(0)\sigma_0(0)}} = \frac{\text{Cov}([\mathbf{W}^{(h)} Z(t)], [\mathbf{W}^{(0)} Z(t+k)])}{\sqrt{\text{var}(\mathbf{W}^{(h)} Z(t)) \cdot \text{var}(\mathbf{W}^{(0)} Z(t))}} \\ &= \frac{T}{T-k} \cdot \frac{\sum_{i=1}^N \sum_{t=1}^{T-k} [\mathbf{W}^{(h)} Z_i(t)][\mathbf{W}^{(0)} Z_i(t+k)]}{\sqrt{\sum_{i=1}^N \sum_{t=1}^{T-k} [\mathbf{W}^{(h)} Z_i(t)]^2 \cdot \sum_{i=1}^N \sum_{t=1}^{T-k} [\mathbf{W}^{(0)} Z_i(t)]^2}}, \end{aligned} \quad (2)$$

式中: $\rho_{h0}(k)$ 为时空自相关系数; $\gamma_{h0}(k)$ 为时空协方差; k 为时间延迟期数; h 为空间延迟期数; $\mathbf{W}^{(h)}$ 为空间延迟期数为 h 的空间权重矩阵; $\mathbf{W}^{(0)}$ 为空间延迟期数为 0 的空间权重矩阵, 即一个单位矩阵; T 为时间序列长度; N 为空间实体数量.

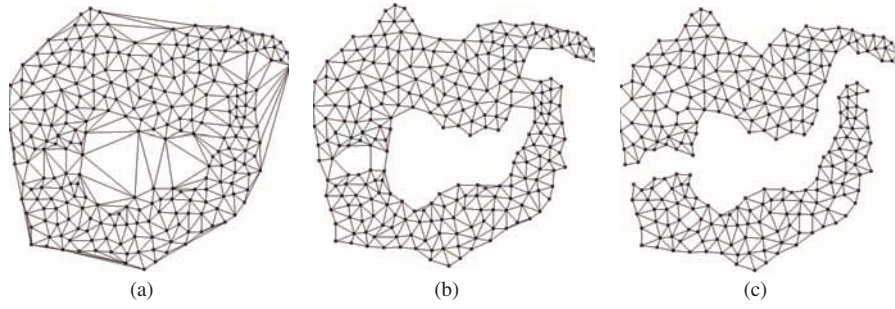


图 3 空间邻近域构建

Figure 3 The construction of a spatial neighborhood. (a) Delaunay triangulation; (b) delaunay triangulation after removing global long edges; (c) delaunay triangulation after removing local long edges

时空偏相关系数通过求解时空 Yule-Walker 方程组获得, 表达为^[19]

$$\rho_h(k) = \sum_{k=1}^{\rho} \sum_{h=1}^{m_k} \varphi_{kh} \rho_{h-k}(k), \quad (3)$$

式中: φ_{kh} 为时空偏相关系数; m_k 为第 k 个时间自回归项的空间阶数; k, h 分别为时间、空间延迟期数; ρ 为方程数.

为了能够顾及连续空间序列中的空间变异特征, 本文采用变差函数来分析连续空间序列中的变异结构, 构建空间权重随距离变化的函数. 因此, 对于一个时空实体, 只需考虑该时空实体与其空间变程区域内和时间延迟 k 范围内所有实体的相关程度. 根据公式 (2) 和 (3) 计算获得时空自相关函数和偏相关函数表, 通过观察时空自相关函数表与时空偏相关函数表的截尾情况, 可以获得时间延迟 k . 然而, STARIMA 模型是一种时空预测模型, 是根据过去时间的观察值来预测未来的非空间属性值, 故时间延迟与时间邻近具有一定的区别. 以图 2 为例, 时间延迟只包括前一个时间点, 而不考虑下一个时间点. 但是, 时间延迟算子实际上可以用于确定时空邻近域的时间窗口半径, 即一个时空实体受前 k 个时间点 (或时间段) 上时空实体的影响, 同时该实体也对后 k 个时间点 (或时间段) 上的时空实体有影响, 而这些具有较明显时空自相关特性的时空实体才真正构成了一个时空邻近域.

时间窗口确定后, 进一步需要获取实体间的空间邻近关系. Delaunay 三角网是定义实体间空间邻近关系的有效工具, 但是直接利用 Delaunay 三角网来构建空间邻近关系在边界处往往存在较大的误差, 如图 3(a) 所示. 为此, 下面通过提取 Delaunay 三角网中边长的统计特征, 施加整体和局部两个层次的距离约束, 来自适应的构建实体间的空间邻近关系.

定义 1 对于 Delaunay 三角网中的任一顶点 P_i , 整体边长约束条件记为 $\text{Global_Constraint}(P_i)$, 表达为

$$\text{Global_Constraint}(P_i) = \text{Global_Mean} + \frac{\text{Global_Mean}}{\text{Mean}(P_i)} \times \text{Global_Variation}, \quad (4)$$

式中: Global_Mean 表示 Delaunay 三角网中所有边长的均值; $\text{Mean}(P_i)$ 表示 Delaunay 三角网中与 P_i 直接相连的所有边的边长均值; Global_Variation 表示 Delaunay 三角网中所有边长的方差.

定义 2 给定一个图 G , P_i 为 G 的一个顶点, 则所有到 P_i 的路径小于或等于 k 的边, 构成 P_i 的 k 阶邻域, 记为 $NN_G^k(P_i)$.

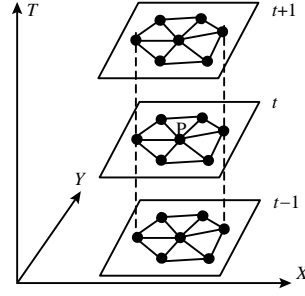


图 4 时空邻近域

Figure 4 A spatio-temporal neighborhood

定义 3 对于图 G 中任一顶点 P_i , 局部边长约束条件记为 $\text{Local_Constraint}(P_i)$, 表达为

$$\text{Local_Constraint}(P_i) = \text{Local_Mean}(P_i) + 2 \times \frac{\sum_{i=1}^N \text{Variation}(P_i)}{N}, \quad (5)$$

式中: $\text{Local_Mean}(P_i)$ 表示图 G 中 P_i 的二阶邻域边长均值; $\text{Variation}(P_i)$ 表示图 G 中 P_i 的一阶邻域内边长的方差; N 表示 G 中顶点数.

在原始 Delaunay 三角网中依次将长度大于 $\text{Global_Constraint}(P_i)$ 和 $\text{Local_Constraint}(P_i)$ 的边长删除, 即可获得实体间的空间邻近关系. 如图 3 所示, 图 3(b) 为删除长度大于 $\text{Global_Constraint}(P_i)$ 的所有边后的结果, 图 3(c) 为进一步删除长度大于 $\text{Local_Constraint}(P_i)$ 的所有边后的结果, 也是空间实体间邻近关系的最终表达结果. 因此, 一个时空实体的时空邻近域可以定义如下.

定义 4 对于时空实体 $\langle P, t \rangle$, 其时空邻近实体为以时间延迟 k 为时间窗口半径, 并且在空间上有公共边的所有时空实体, 记为 $\text{STN}_k(P)$. 图 4 表示了 $k = 1$ 时, $\langle P, t \rangle$ 的时空邻近域.

以上是时空数据表现为时空平稳时构建时空邻近域的方法, 对于时空非平稳的数据则无法直接按照上述步骤构建时空邻近域, 即时空自相关函数与时空偏相关函数没有明显的截尾. 针对非平稳的时空序列而言, 其一般可以分解为两个部分: 大尺度上确定的趋势和小尺度上随机的变异^[18]. 大尺度的趋势可以表达为空间位置与时间的非线性函数, 小尺度的时空变异表示了时空相关的误差项, 对于非平稳的时空序列而言, 移除大尺度的趋势后, 剩余部分一般是平稳的, 可以对其时空相关性进行分析. 神经网络非线性拟合的优势可以用于非平稳时空序列的时空趋势提取^[20], 因此, 针对非平稳的时空序列, 本文借助 BP(back propagation) 神经网络^[21] 提取并剔除非平稳的时空趋势, 获得平稳的时空序列, 再采用上述方法构建时空邻近域.

3.3 时空聚类方法

在时空邻近域构建的基础上, 需要进一步顾及非空间属性进行时空聚类. 任一时空实体 $\langle P, t_i \rangle$ 的非空间属性表示为 $\text{Att}(\langle P, t_i \rangle)$, 两个时空实体 $\langle P, t_i \rangle, \langle Q, t_j \rangle$ 之间的非空间属性距离表达为 $\text{Dist}_{\text{Att}}(\langle P, t_i \rangle, \langle Q, t_j \rangle)$, 采用 Euclid 距离进行度量 (若包含多个非空间属性则预先进行归一化处理). 进一步, 引入时空直接可达、时空相连、时空核、扩展时空核等概念.

定义 5 (时空直接可达) 如果 $\text{Dist}_{\text{Att}}(\text{Att}(\langle P, t_i \rangle), \text{Att}(\langle Q, t_j \rangle)) \leq \varepsilon_{\text{direct}}$ 且 $\langle P, t_i \rangle, \langle Q, t_j \rangle$ 互为时空邻居, 则称 $\langle P, t_i \rangle, \langle Q, t_j \rangle$ 时空直接可达, 记为 $\langle P, t_i \rangle \leftrightarrow \langle Q, t_j \rangle$.

定义 6 (时空相连) 给定时空实体集合 STSET , $\text{Avg}(\text{STSET})$ 表示 STSET 中所有时空实体非空间属性均值. 对于一个时空实体 $\langle Q, t_j \rangle$, 若 $\text{Dist}_{\text{Att}}(\text{Avg}(\text{STSET}), \text{Att}(\langle Q, t_j \rangle)) \leq \varepsilon_{\text{connect}}$, 且

$\exists \langle P, t_i \rangle \in \mathbf{STSET}$, $\langle P, t_i \rangle$ 与 $\langle Q, t_j \rangle$ 互为时空邻居, 则称 $\langle Q, t_j \rangle$ 与 $\mathbf{STSET}$ 时空相连, 记为 $\langle Q, t_j \rangle \oplus \mathbf{STSET}$.

定义 7 (时空核) 对于未归入任何簇的时空实体 $\langle P, t_i \rangle$, 若在其时空邻近域内与其满足时空直接可达的时空实体数量大于时空邻居总数的一半, 则称 $\langle P, t_i \rangle$ 为一个初始时空核.

定义 8 (扩展时空核) 给定当前聚类集合 $\mathbf{CLUSET}$, 对于其中一个时空核 $\langle Q, t_j \rangle$, 其时间序列的非空间属性平均值记为 Avg_T . 若令 Avg_T 代替 $\text{Att}(\langle Q, t_j \rangle)$, 仍满足 $\langle Q, t_j \rangle \oplus \mathbf{CLUSET}$, 则称 $\langle Q, t_j \rangle$ 为一个扩展时空核.

基于上述定义, 本文的时空聚类步骤可以总结如下:

1) 选取一个时空核, 若其与时空邻近域内邻近时间点上、空间位置相同的时空实体均满足时空直接相连则将其视为一个初始时空核.

2) 针对一个初始时空核, 在空间上, 根据 Delaunay 三角网构建的邻近关系, 由近到远; 同时依据时间的由近及远, 加入一个满足时空直接可达的时空实体, 初始时空簇开始生成.

3) 按照步骤 2) 中原则进行时空簇扩展, 针对每个扩展时空核, 加入所有满足时空直接可达与时空相连的时空实体, 直到一个时空簇生成完成.

4) 重新选取一个未归入任何时空簇且未被标记为孤立点的初始时空核, 重复步骤 1)~3), 直到所有时空实体被归入某一时空簇, 或被标记为孤立点时, 时空聚类结束.

对于时空非平稳的数据, 邻近时空实体间的非空间属性差异较小且具有渐变性, 导致传统基于密度方法难以保证时空簇中实体非空间属性的相似性, 且易受时间序列上的个别局部异常值干扰. 时空相连与扩展时空核的概念综合考虑时空实体非空间属性整体与局部的差异性, 以及实体在时间序列上的偏离程度, 有利于获得更合理的时空聚类结果. 此外, 本文提出的时空聚类方法的时间复杂度主要来自两个部分: 时空邻近域搜索与时空簇生成. 对于包含 n 个时空实体的数据库, 借助 Delaunay 三角网构建时空邻近域的策略复杂度约为 $O(n \log_2 n)^{[16]}$, 优于 ST-DBSCAN 的复杂度 $O(n^2)$; 时空簇生成的复杂度与 ST-DBSCAN 类似, 也是近似线性, 故本文时空聚类方法的复杂度约为 $O(n \log_2 n)$. 需要指出的是, 本文提出的整个时空聚类分析框架的复杂度要高于 $O(n \log_2 n)$, 因为还包括其他预处理及时空分析步骤, 但是这些步骤已经被证明可以应用于大型时空数据库^[18,20]. 因此, 本文所提出的时空聚类分析框架亦可以应用于大型时空数据库.

4 中国陆地区域气温数据时空聚类分析

4.1 数据描述与预处理

本文选取中国陆地区域 42 年 (1951~1992) 的年平均温度时空数据库进行实例分析. 气象监测数据是一类典型的时空数据, 在全球气候变化研究中具有核心作用, 实际应用十分广泛. 气象监测数据一般结构较为复杂, 采用这种类型数据进行时空聚类分析研究可以有效检验时空聚类方法的性能. 同时, 由于目前尚缺乏有效的时空聚类结果定量评价方法, 采用我国气温监测时空数据库进行实例分析, 可以借助一定的先验知识对聚类结果进行可靠性分析. 原始数据源于 1951 年以来中央气象局长期预报整理的我国国际气象交换站年平均温度数据库, 共包含 194 个站点. 时空数据的预处理主要是剔除和修补部分缺失数据. 为了避免大范围插值带来的误差和干扰, 本文对于缺失较为严重的站点资料不予采用 (如数据缺失 10 年以上或连续缺失 5 年以上); 对于个别缺失数据采用时间序列移动平均的策略进行修补. 在这种情况下, 选取 137 个记录较为完整的气象站数据进行时空聚类分析研究. 气象站



图 5 我国国际气象交换站空间分布

Figure 5 Spatial distribution of weather stations in China

表 1 数据库属性表

Table 1 The attribute table of the meteorological database

Variables	Type	Description of the character
Name	Float	The name of the meteorological station
Time	Date time	Established date of station
Longitude	Integer	WGS_1984 coordinate system
Latitude	Integer	WGS_1984 coordinate system
Elevation	Integer	Height above sea level
Temperature	Integer	Annual average temperature

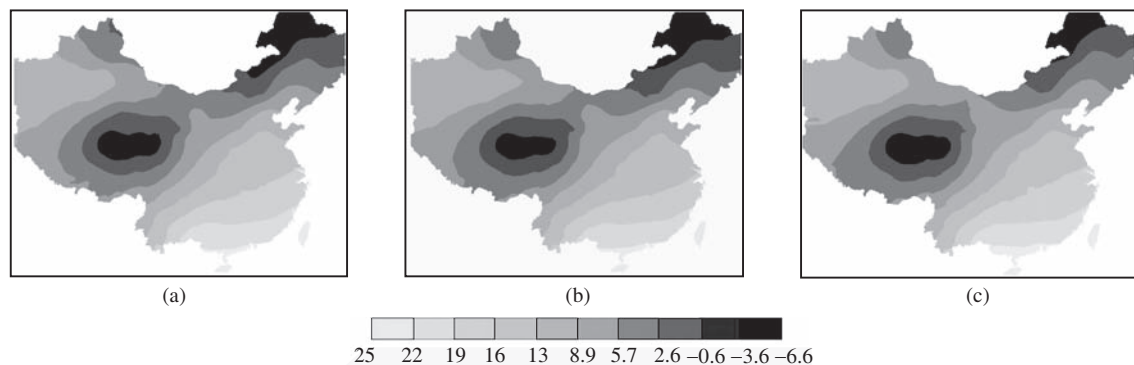


图 6 1970, 1980 和 1990 年气温空间分布

Figure 6 Spatial distributions of temperatures in the year 1970, 1980 and 1990. (a) 1970; (b) 1980; (c) 1990

点空间分布如图 5 所示, 数据库主要属性列于表 1.

4.2 时空数据探索性分析

首先对数据的空间相关性进行检验, 这里借助变差函数来分析空间相关性. 进而, 采用普通 Kriging 方法进行插值. 如图 6 所示, 给出了部分年份平均气温插值的结果. 针对每一年分别计算其块金值 C_0 ,

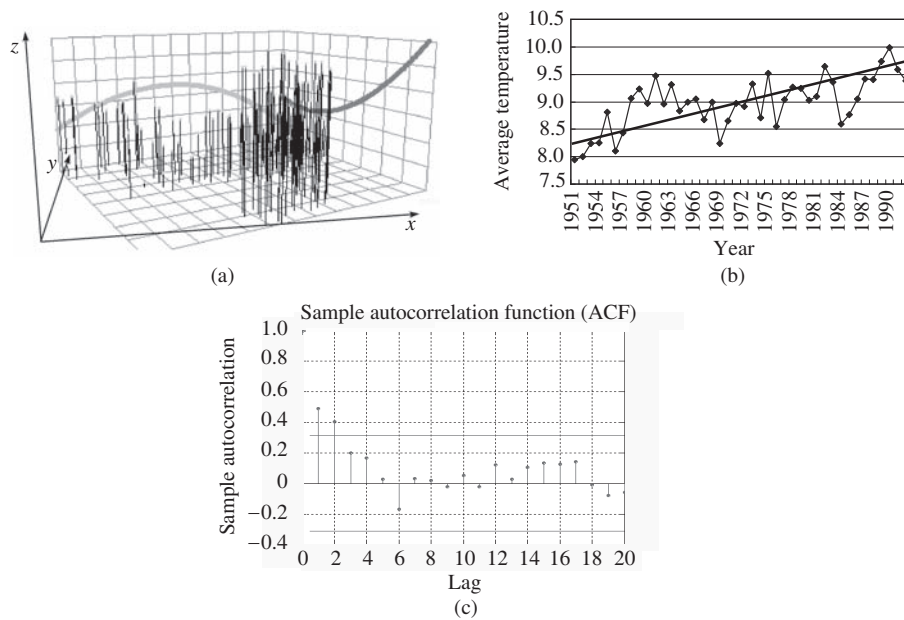


图 7 时空平稳性分析

Figure 7 Space-time stationary analysis. (a) Spatial trend analysis; (b) temporal trend analysis; (c) time autocorrelation analysis

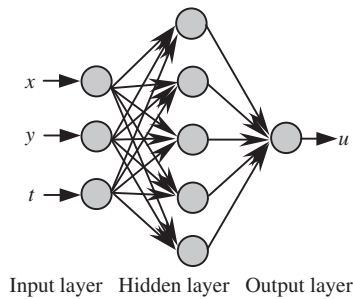


图 8 BP 神经网络结构

Figure 8 The structure of BPANN

拱高 C , 取平均值 $C_0 = 1.58$, $C = 8.56$, 基台值 $C_0 + C = 10.14$. $C_0/(C_0 + C) \approx 15.6\%$. 基台值与块金值的比例可以用来表明变量的空间相关的程度, 若比例小于 75%, 则证明变量间具有较明显的空间相关性^[22]. 上述分析表明, 气温变量具有较强的相关性, 可以进行聚类分析.

进而, 分析数据的时空平稳性, 从图 6 中亦可观察气温由北到南逐渐升高具有趋势性. 为了更直观地反映气温变化的总体规律, 采用趋势分析的方法进行分析. 图 7(a) 显示了所有站点 1951~1992 年平均气温的空间趋势图, X 轴表示东西方向, Y 轴表示南北方向, Z 轴表示气温, 可见气温在东西方向上呈倒 U 型趋势, 而从北到南方向气温上升趋势明显, 故可以判断空间域上的数据是非平稳的. 针对时间域, 研究区域 1951~1992 年平均气温的演变趋势如图 7(b) 所示, 具有一定的上升趋势, 图 7(c) 为时间序列的自相关函数图, 从第 5 期时间延迟后, 自相关函数值趋近 0, 表明时间序列具有较弱的时间平稳性. 综上分析可以判定, 研究时空数据是非平稳的, 需要先剔除时空趋势.

下面采用一个 3 层的 BP 神经网络, 网络结构如图 8 所示. 一个输入层, 包括 3 个输入: 经度 x , 纬度 y 和时间 t (这里采用相对时间 1~42 年), 输入数据进行标准化处理; 一个隐层, 包括 5 个 sigmoid

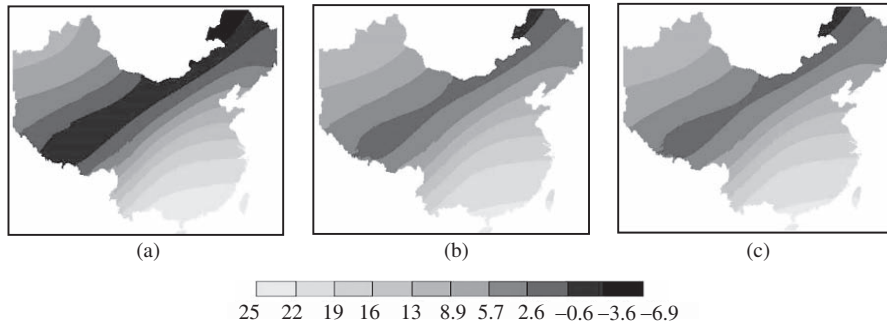


图 9 非平稳时空趋势提取结果

Figure 9 Large scale spatio-temporal trends. (a) 1970; (b) 1980; (c) 1990

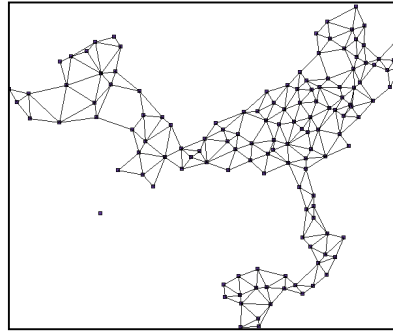


图 10 空间邻近关系

Figure 10 The construction of spatial neighborhood

函数; 一个输出层, 表示空间位置 (x, y) 在时间 t 上的年平均气温. 网络进行训练收敛后, 得到各年的时空趋势 (图 9 给出了部分年份的结果), 并从原始数据中移除, 进一步对残差进行空间相关分析. 变程的平均值 $r = 1364$ km, 块金值、拱高的平均值为 $C_0 = 3.2$, $C = 8.4$, $C_0/(C_0 + C) \approx 27.5\%$, 具有较强的空间相关性.

4.3 时空聚类

首先构建时空邻近域, 采用 STARIMA 建模获取时间窗口时, 借助 Gauss 变差函数定权^[20], 权重可以定义为空间延迟为 h 的函数:

$$\begin{cases} W(h) = [(C_0 + C) - \gamma(h)] / (C_0 + C_1) = [C \cdot e^{-3(h/r)^2}] / (C_0 + C), & h \leq r; \\ W(h) = 0, & h = 0 \text{ or } h > r, \end{cases} \quad (6)$$

式中: 参数 r , C , $C_0 + C$ 取各年结果的平均值, 分别为 $r = 1364$ km, $C = 8.4$, $C_0 + C = 11.6$.

根据公式 (2) 和 (3) 计算时空自相关函数与偏相关函数, 观察截尾获得时间延迟为 2, 即时间窗口半径为 2 年. 各站点空间邻近关系构建结果如图 10 所示.

本文的时空聚类方法需要输入两个参数: $\varepsilon_{\text{direct}}$ 和 $\varepsilon_{\text{connect}}$, 分别用于判断时空直接可达与时空相连. 时空聚类时一方面要求同一时空簇中实体的非空间属性差异尽可能少; 另一方面时空簇的数目不能过多. 当聚类参数设置过低时, 虽然时空簇内部的差异较小, 但是过多较小的时空簇缺乏实际的意

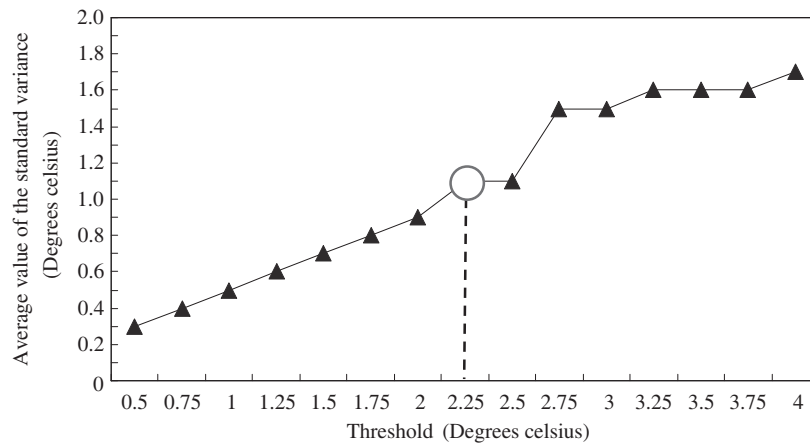


图 11 参数分析

Figure 11 The determination of parameters

义; 参数设置过高时, 虽然时空簇的数量减少, 但是簇内部的差异过大. 因此, 需要在时空簇的数量与时空簇内部非空间属性的差异性之间寻找平衡, 本文参数设置时采取了反馈调节的策略, 即进行多次聚类操作并记录每个聚类结果时所有空间簇中非空间属性标准差的平均值, 绘制变化曲线. 开始时随着参数逐渐增大, 非空间属性标准差缓慢增大, 即簇内实体非空间属性差异逐渐增大; 当参数增大到一定程度时, 非空间属性标准差将达到一个临界值, 并出现一定的跃迁, 随后再增大参数时非空间属性的平均标准差将保持稳定, 接近整体的标准差. 当参数使非空间属性标准差达到临界值时, 可以视为时空簇的数量与时空簇内部差异性之间的平衡点, 故可以通过寻找非空间属性标准差平均值出现拐点时的参数设置作为最佳的聚类参数. 如图 11 所示, 当 $\varepsilon_{\text{direct}} = 2.25^{\circ}\text{C}$, $\varepsilon_{\text{connect}} = 2.25^{\circ}\text{C}$ 时标准差曲线出现了较明显的拐点, 故采用该参数进行时空聚类分析, 本文方法时空聚类结果如图 12(a)~(c) 所示.

4.4 时空聚类结果分析与比较

图 12(a)~(c) 中显示了不同视角下的时空聚类结果, 共发现 12 个较大的时空簇 $C_1 \sim C_{12}$. 进而, 需要分析时空聚类结果的有效性, 目前定量评价时空聚类结果有效性的方法还缺乏深入研究. 然而, 中国陆地区域气温的总体分布具有一定的先验知识, 可以用于对时空聚类结果进行评价. 从整体上分析时空聚类结果可以发现, 由南到北主要分布 $C_1, C_2, \dots, C_{10}$ 等 10 个时空簇, C_1 (平均气温 24.1°C) 主要包括海南岛区域, 是我国热带区域; C_2 (平均气温 21.1°C) 主要包括北回归线到长江以南区域, 属于亚热带; C_3 (平均气温 16.5°C) 主要包括长江以北到秦岭淮河一线, 是亚热带与暖温带的缓冲地带; C_4 (平均气温 12.8°C) 主要包括淮河以北到阴山一线, 属于典型的暖温带; C_5 (平均气温 8.8°C) 与 C_{10} (平均气温 8.2°C) 主要包括辽东半岛北部与内蒙古高原中部, 属于典型中温带; C_6 (平均气温 5.5°C) 与 C_9 (平均气温 1.6°C) 主要包括内东北地区东部与蒙古高原东部, 是中温带与寒温带的过渡地带; C_8 (平均气温 -1.6°C) 位于我国最北部, 属于寒温带. C_{11} (平均气温 2.6°C) 位于河西走廊地区, 由于海拔较高与一般中温带地区相比气温较低; C_{12} (平均气温 11.4°C) 位于新疆南部. 可见, 上述时空簇间的气温具有较为明显的差异. 针对局部区域进行分析可以发现, 相邻时空簇之间一般没有截然的界限, 而是存在一定的重叠现象, 这也充分体现了气温变化的渐变性. 新疆北部与甘肃省附近存在较多小的时空簇和孤立站点, 表明这一地区的温度变化较为复杂, 与西北地区地形变化较大的实际情况比较吻合.

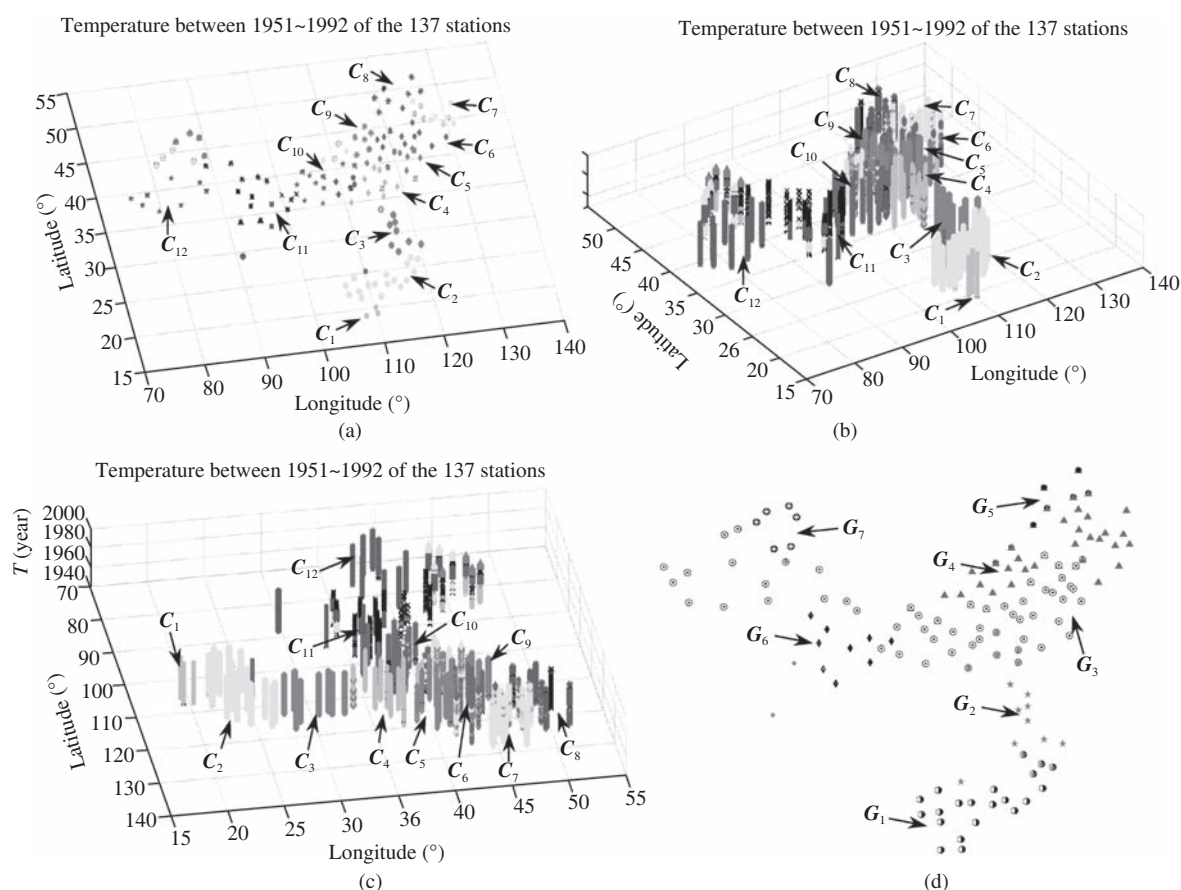


图 12 时空聚类结果 (×— 表示孤立点)

Figure 12 Spatio-temporal clustering results (×— outliers). (a) Clustering results viewed from vertical perspective; (b) clustering results viewed from 45° perspective; (c) clustering results viewed from horizontal perspective; (d) clustering result by ST-DBSCAN

为了进行比较, 本文亦给出了 Birant 等^[4]提出的 ST-DBSCAN 算法的聚类结果, 如图 12(d) 所示. ST-DBSCAN 算法需要 4 个输入参数: 空间距离参数 Eps_1 , 非空间属性距离参数 Eps_2 , 密度阈值 $MinPts$ 以及非空间属性差异阈值 $\Delta\epsilon$. 根据文献^[4]给出的启发式参数确定方法, 上述 4 个参数分别确定为 $Eps_1 = 468$ km, $Eps_2 = 3.4^\circ\text{C}$, $MinPts = 8$, $\Delta\epsilon = 3.4^\circ\text{C}$. 分析 ST-DBSCAN 的聚类结果可以发现, 其主要发现 7 个较大的时空簇 $G_1 \sim G_7$. 其中 G_3 包含了从新疆到东部沿海的广大区域, 认为内陆与沿海具有相同的气温分布模式, 这显然是不合理的; G_1 主要包含了本文方法所发现的时空簇 C_1 和 C_2 , G_2 主要包含了本文方法所发现的时空簇 C_3 和 C_4 , G_4 主要包含了本文方法所发现的时空簇 C_6 , C_7 和 C_9 , 从上文时空簇的统计分析结果容易发现 ST-DBSCAN 算法发现的时空簇中实体气温差异较大, 难以详细反映我国气温的时空分布情况. ST-DBSCAN 算法聚类结果不够理想的主要原因可以归结为两点: (1) 由于气象站空间分布不均匀, 确定聚类参数时存在误差, 进而导致时空邻近域的确定不够准确; (2) 可能受到时间序列上的个别局部异常值干扰, 如 G_4 的东西两部分由一个“颈”连接, 但这两部分的气温整体差异比较明显, 通过实际分析发现该处“颈”位置站点在时间序列上有几个局部的异常值, 该异常值恰好与周围其他簇的气温值近似, 故导致了错误的聚类结果.

从以上分析与比较可见, 本文提出方法的气温时空聚类结果更加符合我国气温的整体分布规律, 一方面证明了时空聚类结果的可靠性, 可以推广到其他时空数据的时空聚类操作; 另一方面对我国气候演变与气温分布模式研究具有一定的参考意义. 进一步分析一些较小的时空簇与孤立点, 对于研究我国上世纪后半叶气候变化的突变性将具有一定的指导意义.

5 结论

本文从时空数据的基本特征出发, 建立了一套时空聚类分析的理论方法框架, 并提出了一种时空一体化的时空聚类方法, 成功应用于中国陆地区域 42 年 (1951~1992) 年平均气温数据库, 证明了其有效性. 与已有方法相比, 本文方法主要具有以下两个优点: 1) 提出了时空聚类分析的理论方法框架, 该框架易于推广用于其他类型的时空数据, 具有很强的普适性; 2) 很好地顾及了时空数据的时空耦合、时空相关与时空异质特征, 避免了过多人为主观因素的干扰, 时空聚类结果具有较好的可靠性.

进一步的工作将主要集中在以下 3 个方面: 1) 采用本文提出的理论方法研究全球大尺度下的气象资料时空聚类分析, 如全球海表气温、海表气压, 探索全球气候变化的产生机理; 2) 发展多尺度的时空聚类方法, 本文研究很大程度上是基于较大的时空尺度, 需要研究不同尺度下的时空聚集现象, 这对于更加全面地反映地理现象发展变化规律具有重要意义; 3) 时空聚类结果的定量评价.

致谢 感谢中国气象局以及气象数据共享网为本文提供的气象时空数据; 感谢英国伦敦大学程涛教授对本文提供的指导意见; 感谢许兵同学为本文时空聚类结果可视化所作的编程工作.

参考文献

- 1 Shekhar S, Vatsavai R R, Celik M. Spatial and Spatio-temporal Data Mining: Recent Advances, Next Generation of Data Mining. New York: CRC Press, 2009. 549–578
- 2 Kisilevich S, Mansmann F, Nanni M, et al. Spatio-Temporal Clustering, Data Mining and Knowledge Discovery Handbook. 2nd ed. New York: Springer Press, 2010. 855–874
- 3 Steinbach M, Tan P N, Kumar V, et al. Clustering earth science data: goals, issues and results. In: Kamath C, ed. Proceedings of the 4th KDD Workshop on Mining Scientific Datasets in conjunction with 7th ACM SIGKDD International Conference on Knowledge and Data Mining. San Francisco: ACM Press, 2001. 1–8
- 4 Birant D, Kut A. ST-DBSCAN: an algorithm for clustering spatial-temporal data. Data Knowl Disc, 2007, 60: 208–221
- 5 Gaudart J, Poudiougou B, Dicko A, et al. Space-time clustering of childhood malaria at the household level: a dynamic cohort in a mali village. BMC Public Health, 2006, 6: 1–13
- 6 Kulldorff M, Heffernan R, Hartman J, et al. A space-time permutation scan statistics for disease outbreak detection. PLoS Med, 2005, 2: 216–224
- 7 Wang M, Wang A P, Li A B. Mining spatial-temporal clusters from geo-database. Lect Notes Artif Intell, 2006, 4093: 263–270
- 8 Pei T, Zhou C H, Zhu A X, et al. Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise. Int J Geogr Inf Sci, 2010, 24: 925–948
- 9 Grubestic T H, Mack E A. Spatio-temporal interaction of urban crime. J Quant Criminol, 2008, 24: 285–306
- 10 Jacquez G M. A k nearest neighbour test for space-time interaction. Stat Med, 1996, 15: 1935–1949
- 11 Kulldorff M, Hjalmarsson U. The Knox method and other tests for space-time interaction. Biometrics, 1999, 55: 544–552
- 12 Zaliapin I, Gabrielov A, Keilis-borok V, et al. Clustering analysis of seismicity and aftershock identification. Phys Rev Lett, 2008, 018501: 1–4

- 13 Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J W, Fayyad U M, eds. Proceedings of the 2nd the International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press, 1996. 226–231
- 14 Wu T S, Song G J, Ma X J, et al. Mining geographic episode association patterns of abnormal events in global earth science data. *Sci China Ser E-Tech Sci*, 2008, 51: 155–164
- 15 Bezdek J C, Pal N R. Some new indexes of cluster validity. *IEEE Trans Syst Man Cy*, 1998, 28: 301–315
- 16 Estivill-castro V, Lee I. Multi-level clustering and its visualization for exploratory spatial analysis. *Geoinformatica*, 2002, 6: 123–152
- 17 Martin R L, Oeppen J E. The identification of regional forecasting models using space-time correlation functions. *Trans Inst Brit Geogr*, 1975, 66: 95–118
- 18 Haining R P. *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press, 2003. 183–201
- 19 Kamarianakis Y, Prastacos P. Space-time modeling of traffic flow. *Comput Geosci-UK*, 2005, 31: 119–133
- 20 Cheng T, Wang J Q, Li X, et al. A hybrid approach to model nonstationary space-time series. *Int Arch Photogramm Remote Sens Spatial Inf Sci*, 2008, XXXVII: 195–202
- 21 Cheng T, Wang J Q, Li X. Accommodating spatial associations in DRNN for space-time analysis. *Comput Environ Urban Syst*, 2009, 33: 409–418
- 22 Wang S Q, Zhu S L, Zhou C H. Characteristics of spatial variability of soil thickness in China(in Chinese). *Geogr Res*, 2001, 20: 161–169 [王绍强, 朱松丽, 周成虎. 中国土壤土层厚度的空间变异性特征. *地理研究*, 2001, 20: 161–169]

A general method of spatio-temporal clustering analysis

DENG Min*, LIU QiLiang, WANG JiaQiu & SHI Yan

Department of Surveying and Geo-informatics, Central South University, Changsha 410083, China

*E-mail: dengmin208@tom.com

Abstract Spatio-temporal clustering has been a hot topic in the field of spatio-temporal data mining and knowledge discovery. It can be employed to uncover and interpret developmental trends of geographic phenomenon in the real world. However, existing spatio-temporal clustering methods seldom consider both spatio-temporal autocorrelations and heterogeneities among spatio-temporal entities, and the coupling in space and time has not been well highlighted. In this paper, a unified framework for the clustering analysis of spatio-temporal data is proposed, and a novel spatio-temporal clustering algorithm is developed by means of a spatio-temporal statistics methodology and intelligence computation technology. Our method is applied successfully to finding spatio-temporal cluster in China's annual temperature database for the period 1951~1992.

Keywords spatio-temporal clustering, spatio-temporal statistics, spatio-temporal autocorrelation, spatio-temporal heterogeneity, spatio-temporal data mining



Deng Min, Deng Min was born in 1974. He received his Ph. D. degree from Asian Institute of Technology (AIT), Bangkok, Thailand in 2004. Currently, he is a full professor at the Department of Surveying and Geo-informatics, Central South University. Prof. Deng has published over 100 journal papers and two authored research monographs. His research interests include spatio-temporal data mining, reasoning and analysis.