文章编号:1671-8860(2011)07-0847-06　　**文献标志码**:A

一种基于场论的层次空间聚类算法

邓　敏[1,2]　彭东亮[1]　刘启亮[1,2]　石　岩[1]

(1　中南大学测绘与国土信息工程系,长沙市麓山南路 932 号,410083)
(2　湖南省地理空间信息工程技术研究中心,长沙市麓山南路 932 号,410083)

摘　要:从空间数据场的角度出发,提出了一种基于场论的层次空间聚类算法(简称 HSCBFT)。该算法是通过模拟空间实体间的凝聚力来描述空间实体间的相互作用,进而采取层次凝聚的策略进行聚类。通过实验分析可以发现,层次空间聚类算法具有如下优势:① 空间聚类簇中各空间实体很好地满足了空间邻近且专题属性相似的要求;② 能发现任意形状的空间簇,且具有良好的抗噪性;③ 输入参数较少。

关键词:空间聚类;场论;凝聚力;空间数据挖掘
中图法分类号:P208

空间聚类是当前地球信息科学与计算机科学领域共同关注的热点问题之一[1-4],旨在将空间数据库中的空间实体划分成具有一定意义的若干簇,使得同一簇中的实体尽可能相似,而不同簇中实体间的差异尽可能大。目前,空间聚类技术已广泛应用于遥感图像分类、热点分析、制图综合及地震空间分布模式挖掘等众多应用领域,主要用于揭示空间数据的分布规律,以及探测空间异常点。现有的空间聚类算法大致可以分为:① 基于划分的聚类方法,如 k-Means[5]、k-Mediods[6]、等;② 基于层次的聚类方法,主要有 BIRCH[7]、CURE[8]、CHAMELEON[9]、AMOEBA[10] 等;③ 基于密度的聚类方法,例如 DBSCAN[11]、OPTICS[12]、DENCLUE[13]、ADBSC[14]、LDBSC[15]、FTSC[16]等;④ 基于网格的聚类方法,代表算法有 STING[17]、WaveCluster[18]等;⑤ 基于图论的聚类方法,如 ZEMST[19]、AUTOCLUST[20] 等;⑥ 混合聚类方法,代表算法有 CLIQUE[21]、NN-Density[22]等。其中,层次空间聚类方法采用递归策略,依据一定的度量准则对空间数据进行合并或分裂,直到获得指定的聚类结果,可以有效地反映空间数据分布的层次结构,对于认识和解释复杂的地学现象具有重要意义。因此,层次空间聚类方法一直是空间聚类研究中的主要内容之一。传统的 Single-Link、Complete-Link、Average-Link 等凝聚式层次聚类方法分别采用两簇之间的最小距离、最大距离、均值距离作为合并依据,聚类结果容易出现"球型偏见"问题,无法获得任意形状的空间簇,且聚类结果易受噪声影响。改进的层次聚类算法有 BIRCH、CURE、ROCK、CHAMELEON、AMOEBA 等。虽然这些方法的聚类结果质量有所提高,但依然存在输入参数过多、难以适应空间数据的空间分异特性等缺陷。更为重要的是,上述层次聚类方法是针对传统事务性数据库提出的,难以同时满足空间聚类中空间邻近且专题属性相似的要求。

综上所述,层次空间聚类方法一方面需要综合顾及空间邻近与专题属性的相似;另一方面需要适应空间数据分布的复杂性(如任意形状、密度不均匀、噪声点等)。因此,本文在空间数据场的基础上,通过模拟空间实体间的凝聚力作用,提出了一种基于场论的层次空间聚类算法(简称 HSCBFT)。

收稿日期:2011-04-28。
项目来源:国家 863 计划资助项目(2009AA12Z206);地理空间信息工程国家测绘局重点实验室开放研究基金资助项目(201015);江苏省资源环境信息工程重点实验室(中国矿业大学)开放研究基金资助项目(JS200901);江西省数字国土重点实验室开放研究基金资助项目(DLLJ201005);中南大学前沿研究计划资助项目(2010QYZD002)。

1 基于场论的层次空间聚类算法

在地理空间中,各空间实体间存在一定的依赖与联系。空间数据场理论[23]认为,空间实体在其周围一定范围内产生一个虚拟的物理场,实体间通过这种物理场产生的凝聚力作用互相联系。因此,采用空间数据场来描述空间实体间的相互作用比起传统的距离度量方式具有更明显的物理意义。空间数据场的核心在于场强函数的定义,目前的场强函数定义方法主要包括高斯函数法[13,24]、分段梯形函数法[25]以及外部约束法[16]等。其中外部约束法定义的凝聚场在适应空间数据分异特性、顾及空间实体间邻近关系以及减少人为参数设置等方面具有一定的优势,为此,本文引入了凝聚场[16]来描述空间实体间的相互作用。

1.1 凝聚场

基于外部约束的凝聚场,其核心思想在于采用Delaunay三角网与Voronoi图来约束数据场的衰减,即在一个实体的直接Voronoi邻近区域内,场强函数与距离的平方满足反比关系;而在直接Voronoi邻近区域之外,场强函数迅速衰减到可以忽略的程度。因此,根据凝聚场的定义,一个实体只与其直接Delaunay邻近实体间有较为明显的作用,这种作用表现为凝聚力。如图1所示,p_1的直接Delaunay邻近实体为p_2、p_3、p_4、p_5、p_6,故p_1只与p_2、p_3、p_4、p_5、p_6具有明显的凝聚力作用。

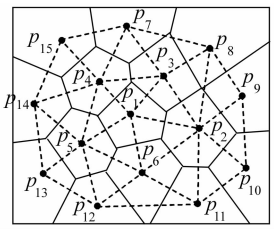


图1 凝聚场

Fig. 1 Aggregation Field

然而,文献[16]定义的凝聚场仅考虑了空间属性的差异,而没有顾及专题属性,因此,本文进一步将专题属性纳入凝聚场的定义。本文借鉴文献[26-27]的策略,将空间属性与专题属性归一化后分别计算空间属性距离与专题属性距离,再进行加权融合,表达为:

$$D(p,q)=w_1\sqrt{(x_p-x_q)^2+(y_p-y_q)^2}+w_2\sqrt{\sum_{k=1}^{n}(A_{pk}-A_{qk})} \tag{1}$$

式中,A_{pk}表示空间实体p的第k维专题属性;w_1、w_2分别表示空间属性与专题属性权值,可以根据实际情况设置,默认情况为$w_1=w_2=0.5$。进一步地,对于任意空间实体p,其直接Voronoi邻近区域定义为DNV(p),其直接Delaunay邻近实体定义为ND(p),则凝聚场的场强函数可以表达为:

$$E_p=k\frac{1}{D(p,x_i)2\sigma},k=\begin{cases}1,x_i\in \mathrm{DNV}(p)\\0,x_i\notin \mathrm{DNV}(p)\end{cases} \tag{2}$$

式中,E_p为场源p(亦是一个实体)在空间上产生的凝聚场的场强;k为凝聚场辐射因子;x_i为任意一个空间位置;$D(p,x_i)$为实体p与x_i的复合距离;σ为衰减因子。根据场强函数的定义,空间实体p、q间的凝聚力可以表达为:

$$F_p(p,q)=E_p m_q=k\frac{1}{D(p,q)2\sigma}$$

$$m_q=\frac{m_q}{D(p,q)2\sigma},k=\begin{cases}1,q\in ND(p)\\0,q\notin ND(p)\end{cases} \tag{3}$$

式中,m_q为实体q的质量,考虑到可以将空间点实体均视为单位质点,故令m_q为1;$D(p,q)$为实体p与q的加权距离,定义如式(1);σ表示衰减因子。

1.2 算法描述

空间簇间的凝聚操作是凝聚法层次空间聚类的核心内容,必须顾及空间簇的整体特征,具体体现在以下两个方面:① 从整体上顾及空间簇的形状,而不是仅通过部分代表点;② 顾及空间簇专题属性的整体差异,因为专题属性的空间分布可能是非均匀的,经常具有渐变趋势。为了满足上述两个要求,本文在空间实体间凝聚力的基础上定义了不同簇间的凝聚力,并通过空间簇专题属性平均水平来约束专题属性在空间上的非均匀性,最后综合定义了基于场论层次空间聚类算法的凝聚系数。

定义1 簇间凝聚力。对于两个空间簇C_i、C_j,簇间凝聚力定义为这两个空间簇实体间凝聚力的平均值,记为$F_C(C_i,C_j)$,表达为:

$$F_C(C_i,C_j)=\frac{\sum F_p(p,q)}{n},\quad p\in C_i,q\in C_j\ \text{且}\ q\in ND(p) \tag{4}$$

式中,$F_p(p,q)$为实体p和实体q之间的凝聚力;n为空间簇C_i、C_j实体间凝聚力作用数目。

定义2 专题属性平均水平。对于任意簇C,其专题属性平均水平定义为簇中所有实体专题属性的平均值,记为$\overline{A(C)}$,表达为:

$$\overline{A(C)}=\frac{\sum_{i=1}^{n}A(p_i)}{n},p_i\in C \tag{5}$$

式中,$A(p_i)$为空间实体 p_i的专题属性值;n 为空间簇 C 中的实体数。

定义 3　凝聚系数。对于两个空间簇 C_i、C_j,其凝聚系数定义为簇间凝聚力与专题属性差异的比值,记为 $AR(C_i,C_j)$,表达为:

$$AR(C_i,C_j)=\frac{F_C(C_i,C_j)}{|\overline{A(C_i)}-\overline{A(C_j)}|} \tag{6}$$

根据以上定义,进行层次聚类时每次合并凝聚系数最大的簇,即每次对簇间凝聚力较大且专题属性平均水平差异较小的两个簇进行合并。进而,基于场论的层次空间聚类算法如下。

1) 剔除空间异常点。为了避免空间异常点的影响,本文采取与文献[7]类似的思想,即每次合并两个凝聚系数最大的簇,空间异常点所在的簇生长较为缓慢或者难以与其他实体合并,将形成若干孤立的小簇,实际中可以采取删除这些小簇的策略剔除空间异常点。文献[7]建议在空间簇数为实体总数的三分之一时,将较小的簇删除(例如仅含一个或两个实体的簇)。在实际中,当空间数据分布不均匀时这种处理方法极有可能产生误判。对于一个包含 n 个实体的空间数据库,其空间簇数[28]将不超过$\sqrt{n}$,本文通过实验发现当空间簇数达到 1.5～2.5 倍$\sqrt{n}$时,且删除实体数目小于等于 3 的空间簇,效果较好。所有异常点被剔除后,将不再参与凝聚聚类,且重新建立空间实体间的 Delaunay 邻近关系。

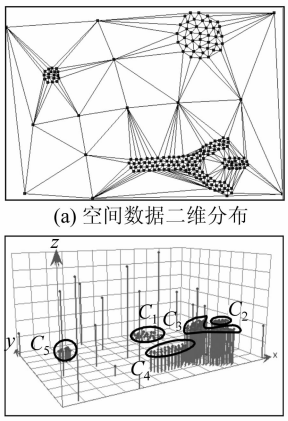
2) 凝聚聚类。每次合并两个凝聚系数最大的两个簇,直到满足预设聚类空间簇数目或所有实体聚为一个空间簇,聚类数目一般由用户设定或采用有效性评价指标辅助选取。聚类的过程将同时生成一棵树状图,体现了聚类的层次结构。

2　实验分析

本文分别采用一组模拟实验与一个实际应用实例来说明本文提出的层次空间聚类方法的可行性与有效性。

2.1　模拟算例分析与比较

模拟数据是在文献[11]的二维模拟数据库的基础上进一步添加了一维专题属性,其中图 2(a)表达了模拟数据的二维几何分布;图 2(b)在二维几何分布的基础上添加了一维专题属性(Z 轴高度表示了专题属性的数值大小),共设置了 5 个空间簇 C_1～C_5以及 20 个异常点。



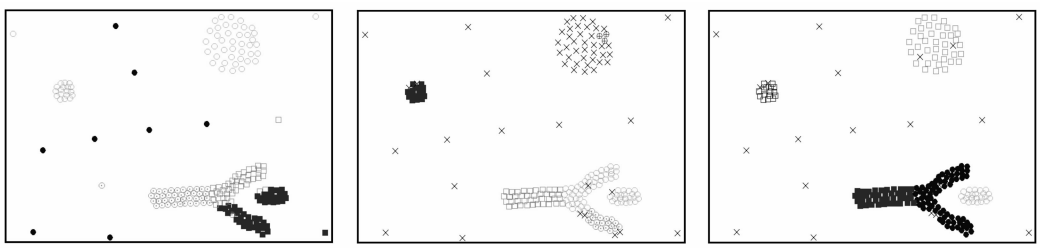
(a) 空间数据二维分布

(b) 专题属性可视化

图 2　模拟数据库(n=207)

Fig. 2　Simulation Database (n=207)

如图 3(a)和 3(b)分别为 k-Means、CURE 算法的聚类结果;图 3(c)为本文提出的基于场论的层次空间聚类算法(HSCBFT 算法)的聚类结果。其中 k-Means 与 CURE 算法中距离函数的定义如式(1),CURE 算法中参数设置选用文献[7]中的建议值,即代表点 Rep = 10,收缩因子α=0.5。比较3个算法的聚类结果可以发现:



(a) k-Means 算法聚类结果　(b) CURE 算法聚类结果　(c) 本文 HSCBFT 算法聚类结果

图 3　模拟数据聚类结果比较分析(×-异常点)

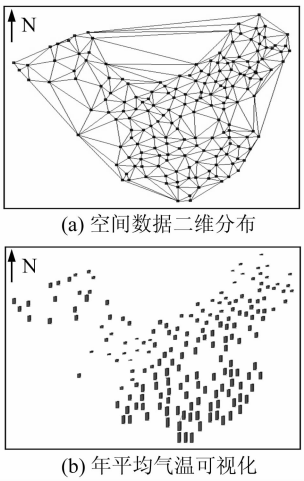
Fig. 3　Different Clustering Results from Different Algorithms on Simulation Data(×-outlier)

① 本文提出的 HSCBFT 算法可以正确识别数据集中的空间簇,能够将空间邻近且专题属性相似的空间实体聚为一类;② k-Means 与 CURE 算法的聚类结果将空间上不邻近的实体错误地聚在一

起,同时还可以发现 CURE 算法在空间数据分布不均匀的情况下,在密度较稀疏的区域可能误判较多的异常点;③ HSCBFT 算法能够识别任意形状的空间簇,并且具有较好的抗噪能力。

2.2 实际算例分析

进一步采用 2002 年全国 183 个气象站点的年平均气温数据,利用 HSCBFT 算法挖掘我国气温的空间分布模式。图 4(a)表达了 2002 年各气象站点的几何分布和邻近关系;图 4(b)可视化表达了各气象站点 2002 年平均气温值。限于篇幅,这里仅给出簇数 $k=5,8,11,13$ 时的四个实验结果,如图 5 所示。将聚类结果与气温数据的实际空间分布情况(如图 3(b))进行比较可以发现:① 整体上较好地反映了我国气温由南到北的梯度变化,如图 5(b)中簇 A_1、A_2、A_3、A_4 的平均气温依次为 21.94 ℃、17.21 ℃、10.23 ℃、3.78 ℃,由南到北逐渐降低;② 能够比较准确地反映气温分布的局部特征,如图 5(d)中簇 B_1(内蒙古高原北部)、B_2(大兴安岭南部)、C_1(大兴安岭北部)、C_2(长白山区)的平均温度分别为5.90 ℃、3.42 ℃、0.88 ℃、4.36 ℃,这些区域虽然空间上邻近但海拔变化较大,导致气温差异明显,体现了局部的差异性;③ 可以反映气温分布的层次性。例如,随着聚类簇数的增加,图 5(b)中的 A_4 簇逐渐分解为图 5(c)中的 B_1、B_2、B_3 簇,最后得到图 5(d)中的 B_1、B_2、C_1、C_2 几个空间簇,较好反映了气温分布由粗到精的逐步细化过程,在很大程度上体现了空间数据分布的层次性。

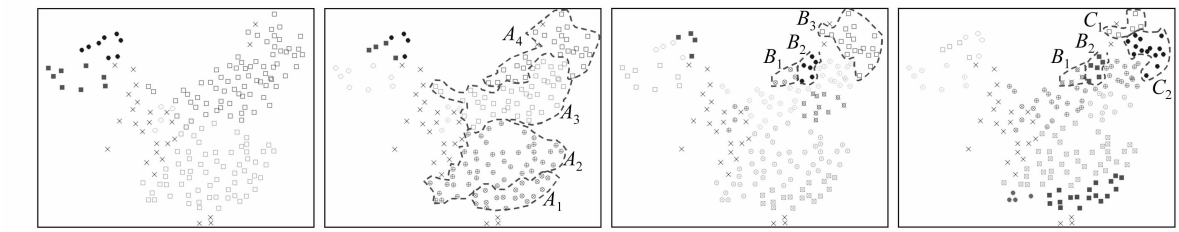


(a) 空间数据二维分布

(b) 年平均气温可视化

图 4 实际数据库($n=183$)

Fig. 4 Real-world Database($n=183$)



(a) k=5 (b) k=8 (c) k=11 (d) k=13

图 5 基于场论的层次空间聚类算法结果(×-异常点)

Fig. 5 Clustering Results of HSCBFT (×-outlier)

3 结 语

空间聚类作为空间数据挖掘的一种重要手段,在地理信息科学、遥感图像处理、制图学、气象学等领域具有重要的应用。层次空间聚类是反映空间数据的多尺度特性的有效手段,可以用来进行深层次的空间数据分析[29]。针对当前层次空间聚类算法中存在的局限,本文从空间数据场的角度出发,提出了一种基于场论的层次空间聚类算法。通过实验分析与比较发现,本文提出的层次聚类算法具有如下优点:① 聚类结果能很好地满足空间邻近且专题属性相似的要求;② 能发现任意形状的空间簇,且具有良好的抗噪性;③ 能够较好反映空间数据分布的层次性。

进一步的研究工作将主要集中在以下两个方面:① 提高凝聚场定义的准确性。采用 Delaunay 三角网约束定义的凝聚场,在边界处可能存在一定的误差,出现一些无意义的长边,采用一定的统计方法对其进行修剪将进一步提高凝聚场的应用效果。② 进一步应用本文方法进行空间数据多尺度挖掘与分析的研究。

参 考 文 献

[1] 李德仁,王树良,李德毅,等. 论空间数据挖掘和知识发现的理论和方法[J]. 武汉大学学报·信息科学版, 2002, 27(3):221-233

[2] 裴韬,周成虎,骆剑承,等. 空间数据知识发现研究进展评述[J]. 中国图像图形学报, 2001, 6(9):854-860

[3] 李斐，王殊伟，柯宝贵.应用聚类分析方法进行实测重力数据的选点优化[J]. 武汉大学学报·信息科学版,2009，34(3)：257-260

[4] 王海军，张德礼. 基于空间聚类的城镇土地定级方法研究[J]. 武汉大学学报·信息科学版,2006,31(7):628-631

[5] Macqueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]. The 5th Berkeley Symposium on Mathematical Statistics and Probability，Berkeley，1967

[6] Ng R，Han J. Efficient and Effective Clustering Method for Spatial Data Mining[C]. The 1994 International Conference on very Large Databases (VLDB'94)，Santiago，1994

[7] Zhang T，Ramakrishnan R，Livny M. BIRCH：An Efficient Data Clustering Method for very Large Databases[C]. The International Conference Management of Data，Montreal，Canada，1996

[8] Guha S，Rastogi R，Shim K. CURE：An Efficient Clustering Algorithm for Large Databases[C]. The ACM-SIGMOD International Conference on Management of Data (SIGMOD'98)，Seattle，Washington，1998

[9] Karypis G，Han E H，Kumar V. Chameleon：Hierarchical Clustering Using Dynamic Modeling[J]. IEEE Computer，1999，32(8)：68-75

[10] Estivill-Castro V，Lee I. Multi-level Clustering and Its Visualization for Exploratory Spatial Analysis [J]. GeoInformatica,2002，6(2)：123-152

[11] Ester M，Kriegel H P，Sander J，et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]. The 2nd the International Conference on Knowledge Discovery and Data Mining，Portland，OR，1996

[12] Ankerst M，Breunig M ，Kriegel H P，et al. OPTICS：Ordering Points to Identify the Clustering Structure[C]. The 1999 ACM-SIGMOD International Conference on Management of Data (DIGMOD'99)，Philadelphia，PA，1999

[13] Hinneburg A，Keim D. An Efficient Approach to Clustering Large Multimedia Database with Noise [C]. The International Conference on Knowledge Discovery and Data Mining (KDD' 98)，New York，1998

[14] 李光强，邓敏，刘启亮，等. 一种适应局部密度变化的空间聚类方法[J]. 测绘学报，2009,38(3)：255-263

[15] 刘启亮，李光强，邓敏. 一种基于局部分布的空间聚类算法[J]. 武汉大学学报·信息科学版，2010，35(3)：373-377

[16] 邓敏，刘启亮，李光强，等. 一种基于场论的空间聚类算法[J]. 遥感学报，2010，14(4):702-717

[17] Wang W，Yang J，Muntz R. STING：A Statistical Information Grid Approach to Spatial Data Mining [C]. The 1997 International Conference on very Large Databases (VLDB'97)，Athens，Greece，1997

[18] Sheikholeslami G，Chatterjee S，Zhang A. Wave Cluster：A Multi-resolution Clustering Approach for Very Large Spatial Databases[C]. The 24th International Conference on very Large Databases，New York，1998

[19] Zahn C T. Graph-theoretical Methods for Detecting and Describing Gestalt Clusters[J]. IEEE Transaction on Computers，1971，C20：68-86

[20] Estivill-Castro V，Lee I. AUTOCLUST：Automatic Clustering via Boundary Extraction for Mining Massive Point-data Sets[C]. The Fifth International Conference on Geo-computation，Chicago，2000

[21] Agrawal R，Gehrke J，Gunopulos D，et al. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications[C]. The ACM-SIGMOD International Conference on Management of Data (SIGMOD'98)，Settle WA，1998

[22] Pei T，Zhu A X，Zhou C H，et al. A New Approach to the Nearest-neighbor Method to Discover Cluster Features in Overlaid Spatial Point Processes[J]. International Journal of Geographical Information Science，2006，20(2)：153-168

[23] 王树良. 基于数据场与云模型的空间数据挖掘和知识发现[D]. 武汉:武汉大学，2002

[24] 淦文燕，李德毅，王建民. 一种基于数据场的层次聚类算法[J]. 电子学报，2006，34(2)：258-262

[25] Nosovskiy G I，Liu Dongquan，Sourina Olga. Automatic Clustering and Boundary Detection Algorithm Based on Adaptive Influence Function[J]. Pattern Recognition，2008，41(9)：2 757-2 776

[26] 李新运，郑新奇，闫弘文. 坐标与属性一体化的空间聚类方法研究[J]. 地理与地理信息科学，2004，20(2)：38-40

[27] 焦利民，刘耀林，刘艳芳. 区域城镇基准地价水平的空间自相关格局分析[J]. 武汉大学学报·信息科学版,2009,34(7)：873- 877

[28] 于剑，程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学(E辑)，2002,32(2)：274-280

[29] 汪闽. 空间聚类挖掘方法研究[D]. 北京:中国科学院地理科学与资源研究所，2003

第一作者简介:邓敏,教授,博士生导师,现从事时空数据挖掘、推理与分析方面的研究,发表论文 100 余篇。

E-mail:dengmin208@tom.com

A Hierarchical Spatial Clustering Algorithm Based on Field Theory

DENG Min[1,2] *PENG Dongliang*[1] *LIU Qiliang*[1,2] *SHI Yan*[1]

(1 Department of Surveying and Geo-informatics, Central South University, 932 South Lushan, Changsha 410083, China)

(2 Research Center of Foundation Geographic Information Engineering, Hunan Province, 932 South Lushan Road, Changsha 410083, China)

Abstract: In this paper, a hierarchical spatial clustering algorithm based on field theory (HSCBFT in abbreviation) is proposed. The field theory of spatial data is firstly employed to describe the interaction among spatial entities. Then, the agglomerative strategy is utilized to find clusters at different levels. Two experiments are preformed to illustrate three advantages of our algorithm. i) It can commendably meet the requirement that clustered entities are close to each other and similar in thematic attribute; ii) It can also discovery clusters with arbitrary shape and is robust to outliers; iii) It needs to input fewer parameters.

Key words: spatial clustering; field theory; aggregation force; spatial data mining

About the first author: DENG Min, professor, Ph. D supervisor. His major research areas include spatio-temporal data mining, reasoning and analysis. He has published over 100 journal papers.

E-mail: dengmin208@tom.com

(上接第842页)

A New Method of Data in Local Area Network Brought into ITRF

GAO Le[1,2] *CHENG Yingyan*[1] *ZHENG Zuoya*[1,2] *ZHAO Chunmei*[1]

(1 Chinese Academy of Surveying and Mapping, 16 Beitaiping Road, Beijing 100830, China)

(2 Shandong University of Science and Technology, 579 Qianwangang Road, Qingdao 266510, China)

Abstract: The variance-covariance matrices of station coordinates can be rigorously transformed between two reference frames at the same epoch, ignoring the influence of velocities and 7 transform parameters rates. By the method, local area data can be brought into ITRF and avoid data processing repeat simultaneity. Moreover it's convenient to use directly international data products existed in the Internet. Analyzing the results of the example, discover that the transformed coordinates deviation can excel 0.1 mm level, which can meet the transformation request between two reference frames. This paper can be used for reference to unite data in different reference frames.

Key words: ITRF; local reference frame; local area data; variance-covariance matrices; rigorous transformation method

About the first author: GAO Le, postgraduate, he is mainly engaged in the research on GNSS data processing of geodesy.

E-mail: sdgaole@163.com