引文格式：孙铭炜,何虹儒,龚晓颖. 川滇地区速度场层次聚类最优模型构建与分析[J]. 地理空间信息, 2026, 24(2):128-132. doi:10. 3969/j.issn. 1672-4623.2026.02.028

川滇地区速度场层次聚类最优模型构建与分析

孙铭炜，何虹儒，龚晓颖*
（西南交通大学 地球科学与工程学院，四川 成都 611756）

摘　要：随着大地测量技术的快速发展，基于GNSS速度场聚类分析的方法已被用于识别构造块体。为了进一步分析不同层次聚类算法在具有旋转特征速度场中的聚类性能，对川滇区域的GPS速度场采用10种不同的层次聚类算法，并利用已有的地质资料和轮廓系数分析不同距离度量和连接方式组合下的聚类效果，获取每种连接方式的最优距离度量以及最优的算法参数组合。结果表明：曼哈顿距离、欧氏距离、马氏距离分别是单连接、全连接、均值连接最优的距离度量；欧氏距离+ward连接的组合方式的效果最优，对噪声和离群点有较好的鲁棒性，能够较好地分离具有旋转特性的速度场数据。

关键词：层次聚类；川滇地区；GNSS速度场
中图分类号：P228　　**文献标志码：**B　　**文章编号：**1672-4623（2026）02-0128-05

Construction and Analysis of the Optimal Model for Hierarchical Clustering of Velocity Field in Sichuan-Yunnan Region

SUN Mingwei, HE Hongru, GONG Xiaoying
(Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: With rapid advancements in geodetic technology, methods based on GNSS velocity field clustering analysis have been employed to identify tectonic blocks. In order to analyze the clustering performance of different hierarchical clustering algorithms in velocity fields with rotational characteristics, we applied ten clustering algorithms in GPS velocity fields of the Sichuan-Yunnan Region, utilized existing geological data and silhouette coefficients to analyze the effectiveness of various combinations of distance metrics and linkage methods, and obtained the optimal distance metric for each linkage method and the optimal combination of algorithm parameters. The results indicate that Manhattan, Euclidean, and Mahalanobis distances are the optimal metrics for single, complete, and average linkages, respectively. The combination of Euclidean distance with Ward linkage yielded the best performance, demonstrating robustness against noise and outliers and effectively separating velocity fields with rotational characteristics.

Key words: hierarchical clustering, Sichuan-Yunnan Region, GNSS velocity field

川滇地区位于青藏高原的东南边缘，是地壳形变和地震活动性最强烈的地域之一[1-2]。川滇地区的速度场整体上表现出顺时针旋转的特点，一直是地学研究人员关注的重点区域。虽然已有基于川滇地区速度场的聚类分析研究，但考虑到不同层次聚类算法的结果存在差异，为通过层次聚类分析科学有效的研究具有旋转型速度场的区域地壳运动特征，有必要构建针对川滇地区GNSS速度场的层次聚类最优模型。本文以川滇地区的GPS水平速度场为研究对象，采用10种不同的层次聚类算法，测试和分析不同距离度量和类间连接方式组合的性能，并构建川滇地区GPS速度场层次聚类最优模型。

1　层次聚类

1.1　层次聚类算法

层次聚类（hierarchical clustering）是一种结构性的聚类算法，它根据数据间的相似度大小获得层次架构，反复对数据进行分裂或者合并操作，进而将数据归类[3]。层次聚类算法能够全局且直观地展示簇间关系，目前被广泛应用于模式识别、数据挖掘、生物信息学等领域。

层次聚类中，数据集的划分可采用2种策略。“自下而上”的凝聚式（agglomerative）策略首先把每一个数据对象都视作一个独立的聚类簇，不断地将相似度最高的2个簇进行合并，当只剩最后一个簇或者达

项目来源：国家自然科学基金资助项目（42374036）；四川省自然科学基金资助项目（2023NSFSC0773）。
第一作者简介：孙铭炜（1999—），硕士研究生，研究方向为GNSS数据处理理论与方法，E-mail：smw15332015023@163.com。
通信作者：龚晓颖（1986—），博士，副教授，研究方向为GNSS精密定位与定轨，E-mail：xygong@swjtu.edu.cn。

到其他终止条件时输出聚类结果。“自上而下”的分裂式（Divisive）策略首先把所有数据对象视作一个簇，之后的每一步则是将最不相似的对象从同一个簇中分离，直到不能再进行分裂，即所有对象都被划分出来为止。考虑到凝聚方法的合并过程遵循“从小范围相似性到大范围一致性”的原则，更适合探究速度场数据的内在层次结构，且算法的实现更加简单和高效，因此本文采用凝聚式策略。

1.2　距离度量

聚类的依据一般是数据特征之间的相似性，聚类对象间的相似程度往往通过“距离”的远近来衡量，即数据特征之间的差距，比如密度、形态、状态、时间等。相似性的度量标准被称为距离度量（Metric），在本文中为水平速度矢量之间的距离。在层次聚类中，较常用的距离度量方式有曼哈顿距离、欧氏距离和马氏距离等。设 D 表示样本之间的距离；$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 表示2个向量，$\boldsymbol{n}$ 是向量维数，各距离度量方法的详细定义如下。

1.2.1　曼哈顿距离

曼哈顿距离是对两组多维数据对应变量之差的绝对值求和，即2点在多维空间中沿着坐标轴方向累加的距离。其公式如下：

$$D(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{n}\left|\boldsymbol{X}_i-\boldsymbol{Y}_i\right| \tag{1}$$

1.2.2　欧氏距离

欧氏距离是一种在多维空间中衡量2点间直线距离的标准方法，在处理二维或三维空间中的点对点距离计算时展现出极高的便捷性和直观性，已被广泛应用于各个领域。一般定义为：

$$D(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{Y}_i)^2} \tag{2}$$

1.2.3　马氏距离

马氏距离顾及数据变量的相关性，通过协方差矩阵来表达各种特性之间的联系，适用于多变量数据且变量之间存在相关性的情况。设 S 为样本的协方差矩阵，则马氏距离的公式如下：

$$D(\boldsymbol{X},\boldsymbol{Y})=\sqrt{(\boldsymbol{X}-\boldsymbol{Y})^T S^{-1}(\boldsymbol{X}-\boldsymbol{Y})} \tag{3}$$

1.3　类间连接方式

类间连接（Linkage）指在2个集合之间，根据数据点间的距离特点确定所要连接合并的集合的方法。层次凝聚聚类通常采用以下4种连接方式[4]。

1.3.1　单连接

单连接将2个集合间的距离定义为二者当中距离最近的数据点间的距离：

$$d=\min\{d(a,b):a\in A,b\in B\} \tag{4}$$

由于单连接仅仅考虑了距离最近的一组成员，即使2个集合的主体在特征空间上相距较远，在距离相近数据点对的作用下，2个集合仍然会被过早地合并。这会使得异常值对聚类结构的影响被放大，进而导致不规则的聚类形态，降低聚类准确性。

1.3.2　全连接

全连接使用两个集合间最远数据点的距离作为集合间的距离：

$$d=\max\{d(a,b):a\in A,b\in B\} \tag{5}$$

其合并决策参考集合之间最远的一组成员，合并后的聚类在内部具有较小的距离变化，因此，全连接可以在一定程度上避免个别离群影响。

1.3.3　均值连接

在均值连接中，2个集合间的距离定义为所有样本点对的距离平均值：

$$d=\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B}d(a,b) \tag{6}$$

该方法平衡了单连接和全连接的效果，既考虑整体样本间的平均相似性，又不像单连接那样容易受离群点影响，生成的聚类形态通常更为均衡。

1.3.4　ward连接

ward连接[5]基于平方误差准则，每次决策都选择使方差增量最小的2个聚类进行合并。设 $S(\boldsymbol{X})$ 代表集合 $\boldsymbol{X}$ 的方差，对于集合 A 与集合 B 而言，2个集合合并后的方差增量为：

$$\Delta S(A,B)=S(A\cup B)-[S(A)+S(B)] \tag{7}$$

其特点是不直接依赖于对象间距离，而是侧重于通过最小化群间方差增量来决定聚类合并的顺序，展现出对聚类结构形态的高度灵活性。

1.4　轮廓系数

为了定量评价聚类质量，本文引入轮廓系数[6]作为内部有效性指标。轮廓系数通过比较簇内的紧凑程度和簇间的分离程度来评估聚类质量。聚类结果中对象 i 的轮廓系数定义为：

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{8}$$

式中，$a(i)$ 为样本到同簇其他点的平均距离；$b(i)$ 为样本到其他簇的所有点的平均距离。为了评估整个聚类的质量，将所有样本的轮廓系数取平均值来代表整体的轮廓系数。轮廓系数的取值范围为[−1, 1]，数值越高聚类效果越佳。

2 数 据

本文使用的GPS速度场数据来自文献[7]，按照川滇地区的经纬度范围（97°~108°E，22°~35°N）进行筛选，删除单方向速度不确定度大于2 mm·a^{-1}的站点，最后选出位于研究区域内的557个站点。图1给出了川滇地区GPS水平速度场图和主要块体分布。

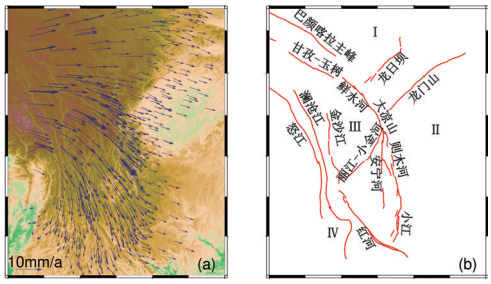


图1 川滇地区GPS水平运动速度场和主要块体分布

在针对川滇地区的地质构造活动研究中，多位学者依据各自的数据与分析方法，提出了不尽相同的活动块体划分方案[8]。

3 川滇地区速度场层次聚类最优模型

为了获取性能最优的层次聚类算法，本章首先研究在速度场聚类分析中适合不同连接度量的距离度量。然后，对采用最优距离度量的4种连接方式的层次聚类结果进行比较分析。为了更加准确地判断层次聚类结果的正确性，本文引入K-means算法的聚类结果作为参考，K-means算法基于欧氏距离定义不同样本点之间的相似性，通过最小化聚类中样本点到聚类质心的距离平方和，确保每个样本点均隶属于唯一聚类，与块体划分原则相符[9]。

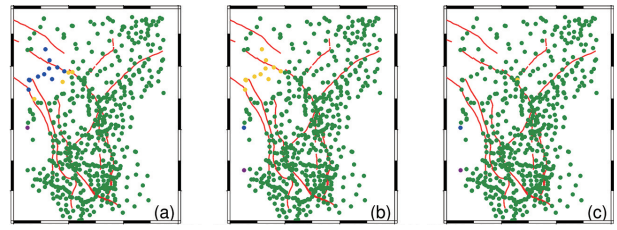
3.1 距离度量方法性能分析

本小节首先对采用不同距离度量的单连接、全连接和均值连接的聚类结果进行分析，研究在速度场聚类分析中最适合这3种连接方式的距离度量。由于ward连接度量需要使用欧氏距离计算聚类质心的几何精度，因此，在这一部分没有对ward连接度量进行讨论。各聚类结果的轮廓系数值由表1所示。

表1 各聚类结果的轮廓系数

	单连接	全连接	均值连接	ward连接	K-means
曼哈顿距离	0.259	0.335	0.406		
欧氏距离	0.147	0.380	0.394	0.472	0.501
马氏距离	-0.167	0.418	0.443		

图2展示了基于不同距离度量的单连接层次聚类结果。在采用单连接后，欧氏距离、曼哈顿距离、马氏距离在聚类结果中均倾向于将单个孤立点或少数离群点划分为一簇，不符合活动块体的定义。产生这种聚类结构的原因可能是单连接的制约，在聚类合并过程中，单连接算法将属于不同簇的对象间的距离最小值作为每对聚类间的距离，当个别对象之间的距离较近，单连接算法会优先合并这2个簇[4]。由于异常值通常距离整个数据集合较远，并且层次聚类在每一步中只合并2个簇，所以，在单连接算法的聚类结果中存在孤立点作为簇的情况，聚类精度下降。此外，3种单连接聚类结果的轮廓系数值分别为0.259、0.147和-0.167，明显偏低，说明聚类结构不合理。单连接对噪声和异常值过于敏感，不适合用于川滇地区速度场的聚类分析。



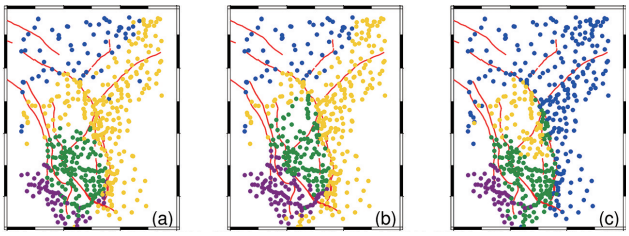
a 单连接+曼哈顿距离 b 单连接+欧氏距离 c 单连接+马氏距离

图2 单连接与不同距离度量组合的层次聚类结果在地理空间中的分布

图3给出了全连接结合不同距离度量的结果。在地理空间中，曼哈顿距离+全连接的层次聚类算法（图3a）识别出龙日坝断裂作为巴颜喀拉块体（蓝色）与黄色聚类的分界线，但黄色聚类涵盖了华南块体与川滇菱形块体中偏北的区域。紫色聚类为密支那-西盟块体，绿色聚类代表川滇菱形块体，二者以腾冲-澜沧断裂为界。图3a中川滇菱形块体的东西边界得到有效识别，南部则以腾冲-澜沧江断裂带为边界，这与李铁明[10]的观点一致；欧氏距离+全连接的结果（图3b）中，金沙江-红河断裂作为滇西南块体和川滇菱形块体的边界被识别出，巴颜喀拉块体与华南块体依然以龙日坝断裂为界，绿色聚类仍覆盖华南块体与川滇菱形块体以北的区域；马氏距离+全连接的结果（图3c）中，巴彦喀拉与华南块体没有被分离，合并为蓝色聚类，紫色聚类为密支那-西盟块体，而黄色聚类覆盖川滇菱形块体的一部分，绿色聚类分布在红河断裂带附近，其与紫色和黄色聚类之间的边界非常模糊，不符合块体边界的定义。

从距离度量的特点分析，曼哈顿距离是对二维速度向量间差值的绝对值求和，未顾及变量的相关性，可能无法捕捉某些结构，因此，聚类结果中存在明显的错分点。马氏距离考虑了数据的分布和相关性，其聚类结果在地理空间中较连贯，轮廓系数值达到

0.418，较其他2种距离更高，但依然存在部分错分点。欧氏距离为速度空间中的直线距离，在本文所用的GPS速度变化范围不大的情况下，能够很好地表达对象间的相似度，使用该距离度量的全连接聚类结果在地理空间中表现出较好的连贯性。



a 全连接+曼哈顿距离　b 全连接+欧氏距离　c 全连接+马氏距离

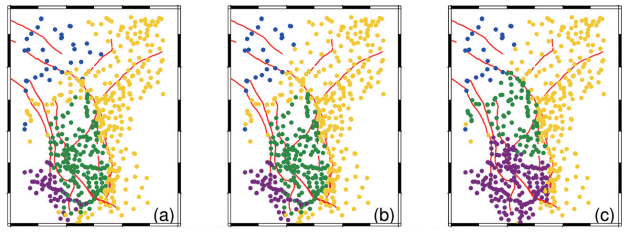
图3　全连接与不同距离度量组合的层次聚类结果在地理空间中的分布

从全连接的定义来看，它基于2个聚类中最远点之间的距离进行合并决策，允许聚类合并后存在微小的距离变化，能够抵抗部分噪声或异常值的影响。即使存在个别远离聚类中心的点，全连接方法也能基于大多数点对间的距离做出合并决策，保持聚类结果的稳定性。图4为基于不同距离度量的均值连接层次聚类结果。在地理空间中，曼哈顿距离与欧氏距离的结果（图4a、b）比较相似，紫色聚类覆盖密支那–西盟块体，绿色聚类形成川滇菱形块体，腾冲–澜沧断裂作为其南边界被识别出来，二者区别在于使用曼哈顿距离（图4a）识别出了龙日坝断裂作为巴颜喀拉与华南块体的分界线，表明该断裂两侧的速度存在差异[11]；马氏距离的结果（图4c）中，巴颜喀拉和华南块体没有得到有效识别，滇西南块体越过红河断裂，占据了川滇菱形块体的南部区域，这可能与红河断裂带的活动性减弱相关[12]。

相比前面的结果，均值连接的结果中远离聚类整体的孤立站点较少，轮廓系数均有所提升，分别为0.406、0.394和0.443，以马氏距离最优。与单连接和全连接相比，均值连接考虑了2个簇中所有点的平均距离，从而能够更全面地反映聚类间的相似度，在一定程度上平衡不同聚类的大小和形状，避免形成过大或过小的聚类。综上所述，均值连接方法能够在一定程度上适应数据的复杂性和多样性，提供较为紧凑合理的聚类结果。

3.2　类间连接方法性能分析

图5给出了K–means算法的聚类结果。由图可知，黄色聚类的边界与华南块体非常吻合，龙日坝断裂作为巴颜喀拉块体（蓝色）和华南块体（黄色）的边界被识别出来。此外，紫色和绿色聚类在东西方向



a 均值连接+曼哈顿距离　b 均值连接+欧氏距离　c 均值连接+马氏距离

图4　均值连接与不同距离度量组合的层次聚类结果在地理空间中的分布

的断裂带基本被有效识别，但两者间的边界并不明确，即紫色聚类不仅覆盖了滇西南块体，还占据了川滇菱形块体的偏南部区域，绿色聚类则覆盖川滇菱形，二者并未以红河断裂为界，可能是由于红河断裂活动性减弱[12]所导致。

曼哈顿距离+单连接组合的结果（图2a）受速度离群值影响较大，出现孤立点或者极少数点成簇的现象，不能作为块体划分的参考依据。

欧氏距离+全连接的结果（图3b）中，滇西南块体（紫色）与川滇菱形块体（绿色）的边界均被识别出，二者以澜沧江–红河断裂为界，蓝色点为巴颜喀拉块体，华南块体的边界——龙日坝、大凉山、安宁河、则木河、小江等断裂带均被识别。由于聚类数量的限制，位于川滇菱形块体上方、鲜水河断裂左侧的小部分站点被划分给华南块体，温扬茂[13]的层次聚类结果也表明此处存在速度阶跃，但还没有相关的块体划分案例。

马氏距离+均值连接的结果（图4c）中，龙日坝断裂未被识别出，巴颜喀拉块体（蓝色）与华南块体（黄色）没有得到有效区分。川滇菱形块体一般以丽江–小金河断裂为界细分为川西北和滇中块体[14]，但绿色聚类与紫色聚类的边界位于滇中块体的中部，川滇块体的南边界–红河断裂带和丽江–小金河断裂带没有被识别。

图6给出了ward连接的聚类结果在地理空间中的分布。欧氏距离+ward连接的结果与K–means算法的结果类似，其轮廓系数为0.472，在所有层次聚类方案中是最优的。此结果识别出了龙日坝断裂，温扬茂[13]的层次聚类结果也将其作为巴颜喀拉和华南块体的边界。此外，传统的块体边界如大凉山、安宁河、则木河、小江等断裂带均被清晰识别，但是红河断裂和丽江–小金河断裂仍被紫色和绿色聚类所覆盖，未被算法确定为块体边界。

根据K–means算法、欧氏距离+全连接、马氏距离+均值连接和欧氏距离+ward连接的聚类结果，发现

绿色和紫色聚类的边界位于红河断裂带和丽江–小金河断裂带之间，川滇菱形块体的南边界和内边界没有被识别。其原因可能是现今红河断裂的滑移速率偏低，过去3~5万年的红河断裂滑移率为1.1±0.4 mm·a^{-1}，且红河断裂的地震活动性不强[15-16]。丽江–小金河断裂的走滑速率为0.2±0.7 mm·a^{-1}，挤压速率为1.8±0.7 mm·a^{-1}[17]，相较其他断裂带而言，该断裂带的运动速率均偏低。较低的滑移率和较弱的地震活动表明断裂带周围的构造运动较为平缓，断裂带两侧GPS站点的水平速度差异不足以被聚类分析识别。此外，丽江–小金河断裂带是次级块体川西北和滇中块体的分界线，聚类的数量也可能是其未被识别的原因。

通过对4种连接度量的层次聚类算法性能进行比较分析，发现欧氏距离+ward连接得到的聚类结果在地理空间中呈现出清晰聚类边界，且与构造边界吻合较好，能够有效分离具有旋转特征的速度场数据，结合轮廓系数分析，认为欧氏距离+ward连接的层次聚类结果最优。ward连接遵循聚类合并后方差增量最小原则，在合并聚类时，选择能够使得合并后的聚类内部方差增量最小的2个聚类进行合并，保证了每一步形成的聚类尽可能地紧密。所以基于方差的连接度量对异常值具有更好的鲁棒性，比基于距离的方法更能抵抗噪声和离群点的影响。

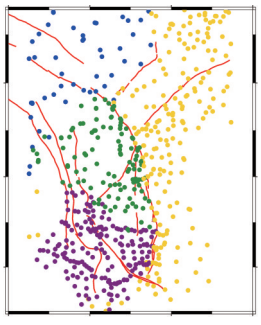


图5 K–means算法的聚类结果在地理空间中的分布

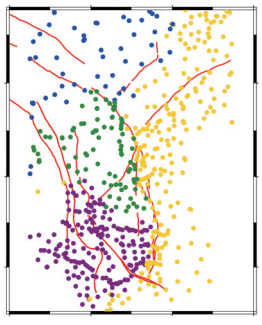


图6 ward连接+欧氏距离度量组合的聚类结果在地理空间中的分布

4 结 语

本文评估了多种层次聚类算法在川滇地区GPS速度场分析中的性能。结果表明，采用欧氏距离与ward连接法的组合效果最优，其聚类结果边界清晰、错分点少，与地质块体划分吻合度高，显著提升了聚类结构的稳定性和质量。其他方法如单连接和全连接受离群值或数据分布影响较大，表现不佳。

参考文献

[1] 张培震,邓起东,张国民,等. 中国大陆的强震活动与活动地块[J]. 中国科学:D辑,2003,33(B04):12-20

[2] 徐锡伟,闻学泽,郑荣章,等. 川滇地区活动块体最新构造变动样式及其动力来源[J]. 中国科学:D辑,2003,33(B04):151-162

[3] 章永来,周耀鉴. 聚类算法综述[J]. 计算机应用,2019,39(7):1869-1882

[4] 余恒,侯晓岚. 层次聚类算法的天文学应用[J]. 中国科学:物理学 力学 天文学,2022,52(8):118-131

[5] 何玉花,张东水,李燕婷,等. 基于聚类的黄河上游干流地区滑坡特征研究[J]. 地理空间信息,2017,15(11):60-64

[6] 孙林,刘梦含,徐久成. 基于优化初始聚类中心和轮廓系数的K-means聚类算法[J]. 模糊系统与数学,2022,36(1):47-65

[7] Wang M, Shen Z K. Presen-day Rustal Deformation of Continental China Derived from GPS and Its Tectonic Implications[J]. Journal of Geophysical Research: Solid Earth,2020,125(2):e2019JB018774

[8] 骆佳骥,崔效锋,胡幸平,等. 川滇地区活动块体划分与现代构造应力场分区研究综述[J]. 地震研究,2012,35(3):309-317

[9] Zhang K, Wei D. Implications from the Kinematic Pattern of the Longmenshan Region[J]. Tectonophysics, 2011, 504(1-4):57-64

[10] 李铁明,邓志辉,吕弋培. 川滇地区现今地壳形变及其与强震时空分布的相关性研究[J]. 中国地震,2003,18(2):30-45

[11] 徐锡伟,闻学泽,陈桂华,等. 巴颜喀拉地块东部龙日坝断裂带的发现及其大地构造意义[J]. 中国科学(D辑:地球科学),2008,58(5):529-542

[12] 吕江宁,沈正康,王敏. 川滇地区现代地壳运动速度场和活动块体模型研究[J]. 地震地质,2003,25(4):543-554

[13] 温扬茂,张国波,许才军. 川滇地区GPS速度场聚类分析[J]. 大地测量与地球动力学,2018,38(5):447-453

[14] 乔学军,王琪,杜瑞林. 川滇地区活动地块现今地壳形变特征[J]. 地球物理学报,2004,47(5):806-812

[15] Shi X, Sieh K, Weldon R, et al. Slip Rate and Rare Large Prehistoric Earthquakes of the Red River Fault, Southwestern China[J]. Geochemistry, Geophysics, Geosystems, 2018,19(7):2014-2031

[16] 王阎昭,王恩宁,沈正康,等. 基于GPS资料约束反演川滇地区主要断裂现今活动速率[J]. 中国科学(D辑:地球科学),2008,58(5):582-597

[17] 丁开华,许才军,邹蓉,等. 利用GPS分析川滇地区活动地块运动与应变模型[J]. 武汉大学学报(信息科学版),2013,38(7):822-827