毛宁，韩钟锐，陈石，李永波，2026. 基于无监督聚类的近震波形特征向量研究. 地震学报，**48**(6)：1−21. doi: 10.11939/jass.20260114.
Mao N, Han Z R, Chen S, Li Y B, 2026. Feature vector analysis of local earthquake waveforms based on unsupervised clustering. *Acta Seismologica Sinica*, **48**(6): 1−21. doi: 10.11939/jass.20260114.

基于无监督聚类的近震波形特征向量研究*

毛 宁[1,2)] 韩钟锐[1,2)] 陈 石[1,2),†] 李永波[1,2)]

1) 中国北京 100081 中国地震局地球物理研究所
2) 中国北京 100095 北京白家疃地球科学国家野外科学观测研究站

摘要 人工智能地震学通过对地震波形特征向量的提取，可高效完成多种数据处理任务，但由于 AI 模型的黑盒特点，模型的可解释性问题一直是研究热点. 本文使用地震波大模型对谛听地震学数据集中的 3 473 条震中距小于 200 km 的近震波形样本进行 512 维深度特征向量提取. 采用无监督聚类、特征空间最近邻一致性检验、波形包络与频谱分析以及空间分布对比方法，分析了地震波大模型处理近震波形的特征向量构建规律. 结果表明：同时排除同一事件和同一台站后，近震波形特征向量空间最近邻地理距离中位数为 881.7 km，台站分块置换检验未显示显著差异，134 个多台站事件的跨台站记录簇标签一致率显著高于随机基线，表明特征向量中仍保留一定的事件相关信息. 本文给出的 4 个聚类在震级、震中距、峰值归一化波形的相对 RMS 和主频上存在显著差异，而震源深度差异不显著；聚类空间分布与地形起伏和布格重力梯度背景呈现一定空间对应关系. 这说明地震波大模型的特征向量数据可以采用无监督聚类方法开展可解释性研究，未来可从近震波形中识别具有波形差异和空间相似性的事件群，为近震波形空间分区和区域构造背景解释提供数据驱动参考.

关键词 近震波形 地震波大模型 特征向量 谛听地震学数据集 无监督聚类
doi：10.11939/jass.20260114 **中图分类号**：P315.3+1 **文献标识码**：A

Feature vector analysis of local earthquake waveforms based on unsupervised clustering

Mao Ning[1,2)] Han Zhongrui[1,2)] Chen Shi[1,2),†] Li Yongbo[1,2)]

1) *Institute of Geophysics, China Earthquake Administration, Beijing* 100081, *China*
2) *Beijing Baijiatuan Earth Science National Observation and Research Station, Beijing* 100095, *China*

Abstract: Large seismic waveform models can efficiently support earthquake detection, phase picking, and event classification by transforming waveforms into high-dimensional feature vectors. However, the seismological interpretation of these intermediate representations remains difficult to establish owing to the black-box nature of deep neural networks. This study investigates whether feature vectors extracted from the pre-trained DiTingSFM-0.1B model retain

* **基金项目** 应急管理部重点科技计划项目(2025EMST040301)、国家自然科学基金原创项目(42450233)和中央级公益性科研院所基本科研业务费专项(DQJB26B38)共同资助.
收稿日期 2026-07-09 收到初稿，2026-09-23 决定采用修改稿.
作者简介 毛宁，在读博士研究生，主要从事地球物理场数据分析与人工智能方法应用研究，e-mail：maoning@cea-igp.ac.cn；†通信作者：陈石，博士，研究员，主要从事时变地球物理场理论方法研究与解释，e-mail：chenshi@cea-igp.ac.cn

measurable waveform characteristics, catalog parameters, event-related information, and spatial sampling patterns. Unsupervised clustering is adopted as a probe for representations rather than a tool to define intrinsic earthquake categories.

The dataset comprises 3 473 three-component local-earthquake waveform samples selected from DiTing Dataset 2.0, corresponding to 3 333 independent natural earthquakes recorded at 714 stations between February 2022 and February 2023. All samples have epicentral distances less than 200 km. Epicenters lie between approximately (18°N−52°N and 73°E−132°E), with local magnitudes ranging from M_L0.0−5.7. Each sample includes vertical, north–south, and east–west components sampled at 100 Hz. A 60 s waveform window is aligned relative to the P-wave arrival time. Prior to model inference, each trace is demeaned and normalized by its standard deviation computed over the full window. A 512-dimensional feature vector is extracted after global average pooling and prior to the task-specific classification head of the pre-trained DiTingSFM-0.1B model.

Four feature-input schemes were compared: raw model outputs, *Z*-score-standardized features, L_2-normalized features, and *Z*-score standardization followed by L_2 normalization. K-means++ clustering was implemented for candidate numbers of clusters ranging from 2 to 10. The silhouette coefficient, Calinski–Harabasz index, and Davies–Bouldin index were adopted to assess clustering structure. The median Hopkins statistic was 0.757 (95% interval: 0.754−0.761), indicating non-random aggregation within the feature space. Nevertheless, the gap statistic failed to identify a unique optimal number of clusters. The silhouette coefficient and Calinski–Harabasz index favoured a coarse two-cluster structure, while the Davies–Bouldin index supported further subdivision at $k=4$. Therefore, $k=4$ was adopted as an operational partition to compare waveform, spectral, catalog, and spatial characteristics, rather than interpreted as the unique count of intrinsic seismological classes.

The reproducibility of the $k=4$ partition was assessed via repeated initialization, independent-event resampling, and alternative clustering algorithms. Across 100 independent initializations, the median adjusted Rand index and normalized mutual information were 0.992 and 0.985, respectively. Event-level resampling yielded corresponding median values of 0.979 and 0.963, with cluster-specific Jaccard retention rates ranging from 0.981 to 0.987. Principal-component K-means and spherical K-means yielded results broadly consistent with the original partition, whereas the Gaussian mixture model and HDBSCAN failed to recover the same structure. Accordingly, the $k=4$ partition is reproducible within K-means and related distance-partitioning frameworks yet partially dependent on the choice of clustering algorithm. The four clusters contained 433, 1 604, 748, and 688 waveform samples, respectively.

A station-aware nearest-neighbour analysis was performed to evaluate geographic consistency within the feature space. After removing additional records belonging to the same earthquake, 38.1% of all feature-space nearest neighbours originated from the same station, substantially higher than the median value of 0.37% for random pairs. This finding demonstrates pronounced station dependence within the feature space. The analysis was repeated after excluding records from both the same earthquake and the same station. The observed median epicentral distance between nearest neighbours was 881.7 km. A station-blocked permutation test, which preserved each station's sample size and geographic sampling extent, yielded a median of the null distribution equal to 889.1 km (95% interval: 873.7−902.1 km). The observed distance was not significantly smaller than the null distribution ($p=0.203$).

A covariate-residualization analysis was further carried out to control for station identity, magnitude, epicentral distance, back azimuth, and recording time. The nearest-neighbour distance computed on residualized features was 1 197.3 km, compared with a station-blocked null median of 1 191.7 km; the difference was not statistically significant ($p=0.868$). Between station identity and cluster labels, the bias-corrected Cramér's V was 0.540 and normalized mutual information was 0.182, indicating a non-negligible yet partial association. For 134 earthquakes recorded at multiple stations, 62.3% of the 146 cross-station record pairs were assigned to the same cluster, significantly exceeding the random baseline median of 28.8%. Station-grouped holdout validation also yielded an adjusted Rand index of 0.971 ± 0.013 and normalized mutual information of 0.954 ± 0.016. These results show that feature-space structure is strongly modulated by station and sampling conditions yet not solely determined by station identity. The representations retain measurable event-related waveform information, though nearest-neighbour analysis alone cannot independently prove that they encode propagation paths or regional medium structure.

Waveform envelopes, power spectral densities, catalog parameters, and waveform statistics were compared across the four clusters. This analysis adopted a processing workflow independent of the standardization applied to model inputs. Each trace of the waveform in digital counts was centered by subtracting its median and normalized by its maximum absolute value. Relative root-mean-square (RMS) values were computed for the full 60 s window, the P-S window, and the 10 s post-S-wave window. These dimensionless quantities describe waveform fluctuations relative to the trace peak and do not represent absolute ground-motion amplitude or seismic wave energy. Power spectral density was estimated via Welch's method, and dominant frequency was defined as the frequency corresponding to the maximum average three-component power spectral density.

For statistical comparisons, a single record was retained per independent earthquake, and station-blocked permutations were used to construct null distributions. Back azimuth was parameterized using its sine and cosine components, and the Holm correction was applied across all statistical tests. Differences in back azimuth were minor, whereas magnitude and epicentral distance exhibited significant inter-cluster differences, with effect sizes of 0.336 and 0.069, respectively. Focal depth showed no significant inter-cluster difference. The full-window, P-S-window, and post-S-wave-window relative RMS yielded effect sizes of 0.361, 0.300, and 0.176, while dominant frequency had an effect size of 0.435.

Clusters C_0 and C_1 generally featured lower magnitudes, intermediate epicentral distances, relatively low dominant frequencies, and larger relative RMS values. Cluster C_2 was characterized by shorter epicentral distances, moderately larger magnitudes, smaller relative RMS values, and substantially higher dominant frequencies. Cluster C_3 typically exhibited larger epicentral distances and magnitudes, intermediate relative RMS values, and stronger spectral decay. These results demonstrate that this operational feature-space partition corresponds to measurable differences in peak-normalized waveform envelopes, relative temporal fluctuations, spectral structure, magnitude, and epicentral distance. Since instrument responses and station gains were not uniformly corrected, these statistics are interpreted only as relative waveform differences under the present processing conditions.

Six low-dimensional baseline models were constructed to examine whether the $k=4$ partition could be reproduced using a small set of variables. These baselines comprised three single-

variable cases (feature-vector norm, full-window relative RMS, dominant frequency), alongside simple catalog-parameter, waveform-statistic, and combined baselines. Agreement with the 512-dimensional clustering remained limited: the adjusted Rand index did not exceed 0.226, and normalized mutual information did not exceed 0.292. Hence, none of the tested individual variables or low-dimensional combinations directly reproduced the four-cluster partition. Nonetheless, this low agreement does not prove that high-dimensional representations contain richer or more physically meaningful information; discrepancies may alternatively arise from unmeasured waveform attributes, station and instrument characteristics, high-dimensional noise, or feature-space geometry.

Cluster epicenters were further compared with topography and horizontal gradients derived from WGM2012 Bouguer gravity anomalies. The transition between Cluster C_0 and the remaining clusters lies broadly near the north-northeast-trending Greater Khingan–Taihang–Wuling gravity-gradient belt. This pattern suggests a plausible spatial correspondence between cluster distribution and regional geophysical background. However, station-blocked permutation and covariate-residualization analyses revealed no significant epicentral proximity after accounting for station-specific sampling biases. The association between clustering results, topography, and Bouguer gravity gradients should thus be treated as an exploratory spatial correspondence rather than direct evidence that the model identifies tectonic boundaries.

Overall, unsupervised clustering offers a powerful approach for probing feature vectors generated by large seismic waveform models. Representations from the pre-trained DiTingS-FM-0.1B model reflect observable differences in magnitude, epicentral distance, peak-normalized waveform features, relative RMS, and dominant frequency. Their spatial organisation is also strongly linked to station composition and geographic sampling, and cross-station consistency indicates partial retention of event-related information. Since the records used in this study overlap in source properties and time window with a subset of the model's pre-training corpus, these results characterise representations formed under the present data distribution and should not be interpreted as proof of generalisation to fully unseen data. Further validation using independent time periods, regions, and stations, combined with uniformly corrected instrument responses and explicit event–station propagation paths, is required to disentangle the respective contributions of earthquake source, propagation path, station response, instrumental characteristics, and sampling structure.

Key words: local earthquake waveform; large seismic waveform model; feature vector; DiTing seismological dataset; unsupervised clustering

引言

近震波形是区域地震监测中最常见、信息量也最丰富的数据类型. 与仅使用震相到时、震级或震源位置的传统目录分析相比, 完整三分量波形包含了震源破裂过程、传播路径介质、近台场地条件以及仪器响应共同作用后的综合结果(Aki, Richards, 2002). 因此, 若能够在大量近震波形之间建立稳定的相似性度量, 就有可能从波形本身识别具有共同传播路径、构造背景或场地效应的事件群. 我国大陆构造单元复杂, 活动地块、活动断裂和构造应力场空间变化显著(许忠淮等, 1989; 邓起东等, 2002; 张培震等, 2003), 为利用波形特征研究近震空间分布提供了良好的区域背景.

对于近震事件而言，地震波形可看作震源时间函数、传播路径格林函数和场地响应共同作用的结果，因此波形中包含的信息远不止震相到时或震级等少数参数．传统波形分析通常需要依赖人工选取的时域、频域或震相特征，近年来，深度学习方法为直接从连续地震波形中自动提取有效信息提供了新的技术途径．卷积神经网络、残差网络和注意力模型能够在端到端框架下学习波形的局部形态、时频变化和长程依赖关系，已广泛应用于地震检测、震相拾取和事件识别等任务，并在大规模数据集上显著提高了自动化处理性能（Perol *et al*，2018；Ross *et al*，2018a，b；Zhu，Beroza，2019；Mousavi *et al*，2020）．随着模型规模、训练数据和基准工具的发展，机器学习的应用也逐渐从单一检测或拾取任务，扩展到地震活动识别、波形表征学习以及固体地球科学的数据驱动发现（Bergen *et al*，2019；Beroza *et al*，2021；Mousavi，Beroza，2022）．

已有研究已将深度特征与无监督聚类用于地震信号分析．Mousavi 等（2019）利用深度卷积自编码器学习地震信号特征，并在特征空间中开展无监督聚类；Seydoux 等（2020）将深度散射网络与高斯混合模型结合，用于连续地震记录中地震信号和背景噪声的检测与聚类．近年来，SeisCLIP 等地震基础模型进一步表明，经过大规模预训练的模型能够为事件分类、定位和震源机制分析等下游任务提供通用特征表示（Si *et al*，2024）．这些研究说明，深度特征聚类和地震基础模型已具有较好的研究基础．与上述工作相比，本文并不旨在提出新的深度聚类网络，而是基于谛听地震基础模型（DiTing seismic foundation model，缩写为 DiTingSFM）0.1B 预训练版本输出的 512 维中间特征向量，分析该特征与近震波形、目录参数和全国尺度空间分布之间的关系，从模型表征分析角度探讨特征向量中可能保留的地震学信息．

在此基础上，模型从波形中学习到的特征向量可能同时受到 P 波和 S 波到时、频谱衰减、震中距、震级、场地响应和传播路径效应等因素影响，是多种波形属性共同作用后的综合编码．因此，利用特征向量开展地震学分析的关键，不仅在于获得高维特征向量本身，还在于判断这些特征向量中是否存在稳定的事件群结构，以及这些结构是否能够在地理空间分布、原始波形和频谱特征上得到解释．

无监督聚类为识别高维特征中的事件群结构提供了直接方法．K-means 是最常用的划分式聚类方法之一，其基本思想是在特征空间中最小化簇内平方误差，从而将样本划分为若干相对紧凑的簇（MacQueen，1967；Lloyd，1982）．为减小初始聚类中心选择对结果的影响，K-means＋＋采用基于距离平方加权的概率抽样方法选择初始中心，使初始中心在特征空间中尽可能分散，从而提高聚类稳定性和收敛质量（Arthur，Vassilvitskii，2007）．DeepCluster 将深度神经网络特征与 K-means 聚类相结合，通过交替进行特征学习和聚类分配，实现了无人工标签条件下的视觉特征学习（Caron *et al*，2018），说明无监督聚类可作为分析深度特征空间结构的重要工具．本文不进行模型再训练或伪标签迭代优化，而是将 K-means＋＋作为表征探针，用于分析 DiTingSFM-0.1B 预训练模型输出特征向量中的潜在分组结构．但聚类指标并不必然对应地质解释的最优结果，仍需结合地理空间分布、原始波形、频谱特征和区域构造背景进行综合判断．

综上所述，地震波大模型能够从近震波形中提取高维特征向量，但这些特征向量是否保留可检验的波形、目录参数和空间分布信息，仍是地震波大模型表征分析中的关键问题．本文以谛听数据集 2.0 中震中距小于 200 km 的 3 473 条近震波形样本作为研究对象，基于 DiTingSFM-0.1B 预训练模型输出的 512 维特征向量，采用 K-means＋＋无监督聚类、特征空

间最近邻一致性检验、簇内波形包络与频谱分析以及空间分布对比，分析近震波形特征向量的表征特征及其地震学含义，以期为近震波形特征向量的表征分析及其区域空间分布研究提供方法参考.

1 数据

本研究使用谛听数据集2.0中的天然地震波形记录. 谛听数据集是面向人工智能地震学研究构建的大规模地震波形数据资源，为地震波形智能处理提供了重要数据基础(Zhao *et al*, 2023). 谛听数据集2.0主要来源于中国地震台网中心的震相观测报告和国家测震台网数据备份中心的事件波形资料，包含中国大陆及邻区天然地震事件的三分量波形记录，并提供P, S震相到时、初动极性、震级、震中距、方位角等配套标注信息. 数据集中的事件标签由省级地震局编目专家人工编目和复核，波形记录完整，信噪比计算中不存在NaN或Inf值，绝大部分人工标注的P, S震相到时与理论计算到时具有较好的一致性；同时，数据集中还对典型干扰信号进行了噪声标注，为数据质量控制和智能模型训练提供了可靠基础(赵明等，2023).

为突出近台波形中震源、传播路径和场地响应对波形形态的影响，同时减少远距离传播、多路径效应和复杂区域震相对聚类结果的干扰，本文筛选震中距小于200 km的事件-台站三分量近震波形记录作为研究对象. 最终选取了2022年2月至2023年2月期间研究区714个地震台站记录的3 473条近震波形样本，对应3 333个天然地震事件，其中，3 199个事件对应1条台站记录，128个事件对应2条台站记录，6个事件对应3条台站记录. 地震震中分布于(18°N—52°N, 73°E—132°E)范围内，震级范围为M_L0.0—5.7. 事件空间分布见图1.

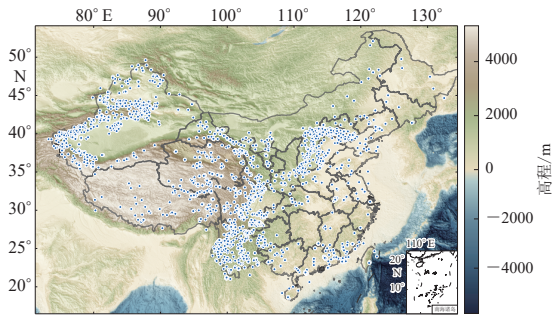

图1 研究所用近震事件震中分布及研究区构造背景

Fig. 1 Epicentral distribution of the local earthquakes and topographic setting of the studied area

每条地震波形样本包含*Z*, N和E三个分量，分别对应垂直向、南北向和东西向，采样率为100 Hz. 以P波到时为基准截取60 s波形窗口，共6 000个采样点，P波到时统一位于第200个采样点，即波形起始后2 s. 对近震事件的分析表明，对于一定距离范围内的中小地震事件，60 s波形片段通常能够覆盖事件的主要波形过程(Jia *et al*, 2022). 因此，本文所选波形窗口能够较完整地保留P波、S波及其后续主要振动信息，为后续特征向量提取和无监督聚类分析提供数据基础.

为满足DiTingSFM-0.1B模型推理的输入要求，分别减去每条样本的*Z*, N和E三个分量在60 s窗口内的时间均值，再除以相应的时间标准差，实现逐记录、逐分量标准差归一化；当标准差为0时，将除数置为1. 本文推理前未额外实施线性去趋势、带通滤波、重采样或仪器响应校正，也未使用数据增强. 原始波形为数字计数值，经过标准差归一化后，模型输入为无量纲波形. 特征提取前对输入数据和输出特征矩阵进行有效性检查，未发现NaN

或 Inf 等非有限值.

2 方法

本文方法围绕地震波大模型特征向量的提取与无监督聚类展开. 首先，利用 DiTingSFM-0.1B 预训练模型将三分量近震波形样本转换为 512 维特征向量；随后对特征矩阵进行有效性检查，并构建原始特征、*Z*-Score 标准化特征、L_2 归一化特征和 *Z*-Score＋L_2 归一化特征四种输入方案；最后，对每种输入方案分别在不同簇数下开展 K-means＋＋无监督聚类，采用轮廓系数、卡林斯基-哈拉巴斯(Calinski-Harabasz，缩写为 CH)指数和戴维斯-博尔丁(Davies-Bouldin，缩写为 DB)指数评价聚类结构，为后续特征向量表征分析提供基础.

2.1 特征提取

本文采用 DiTingSFM-0.1B 预训练模型提取三分量近震波形的 512 维特征向量. 如图 2 所示，模型整体可概化为波形片段嵌入、时序表征编码、事件级特征会聚和预训练分类输出四个模块. 输入数据为经过逐记录、逐分量标准差归一化后的三分量近震波形，波形片段嵌入模块采用一维卷积层 Conv1d(3，512)，卷积核大小和步长均为 50，将 3×6 000 的时间序列压缩为约 120 个 512 维片段 token. 该过程相当于对连续波形进行固定长度分片，并将各片段映射到统一特征空间，从而在降低序列长度的同时保留局部波形形态信息.

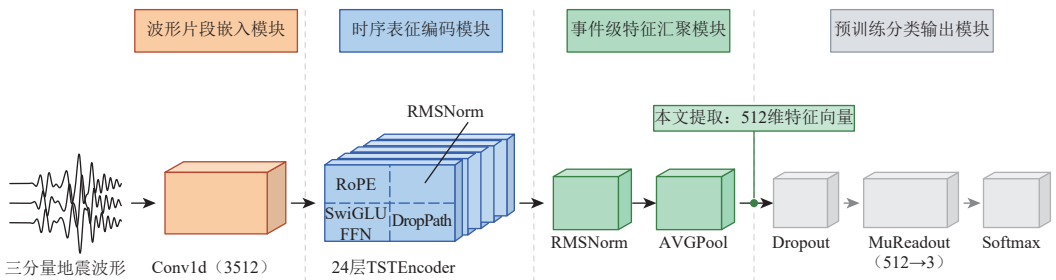

图 2　DiTingSFM-0.1B 预训练模型结构

Fig. 2　Architecture of the DiTingSFM-0.1B pre-trained model

时序表征编码模块由 24 层 TSTEncoder 堆叠而成，每层主要包含多头自注意力、旋转位置编码、均方根(root mean square，缩写为 RMS)归一化(RMSNorm，均方根层归一化)、SwiGLU 前馈网络和 DropPath 正则化等结构. 自注意力机制能够对序列内部的长程依赖关系进行建模，是 Transformer 及其时间序列变体的核心组成部分(Vaswani *et al*，2017；Nie *et al*，2023). 在此基础上，旋转位置编码、RMS 归一化和随机深度等结构分别用于增强序列位置信息表达、提高训练稳定性和改善模型泛化能力(Huang *et al*，2016；Zhang，Sennrich，2019；Su *et al*，2024). 编码器输出经 RMS 归一化后，通过全局平均池化会聚为事件级特征向量.

推理在评估模式和无梯度环境下完成，Dropout 在推理过程中关闭. 模型编码器之后依次为全局平均池化、Dropout 和三分类投影头，其中三分类投影头用于波形类型识别相关任务. 本文提取全局平均池化之后、三分类投影层之前的 512 维特征向量，不使用最终

Softmax 分类概率. 实现过程中, 特征钩子设置在池化层后的 Dropout 输出端. 由于评估模式下 Dropout 为恒等映射, 保存特征与全局平均池化输出一致. 最终, 3 473 条近震波形样本被表示为一个 3 473×512 的特征矩阵, 并作为后续聚类分析的输入.

2.2 聚类算法

在获得每个样本对应的 512 维特征向量后, 本文采用无监督聚类方法分析近震波形在特征空间中的潜在分组结构. 该思路与 DeepCluster 中深度模型特征向量结合无监督聚类的基本思想具有相似性, 即利用聚类结果作为探针, 分析预训练模型输出特征向量中的潜在结构.

聚类前首先对 512 维特征进行有效性检查, 确认特征矩阵中不存在 NaN 或 Inf 等异常值. 为检验特征数值尺度和向量模长对聚类结果的影响, 本文构建了四种输入特征方案: 原始特征、*Z*-Score 标准化特征、L_2 归一化特征以及 *Z*-Score+L_2 归一化特征(Jain, 2010). 其中: *Z*-Score 标准化用于消除不同特征维度均值和尺度差异, 使各维特征具有零均值和单位标准差; L_2 归一化用于统一每个样本特征向量的长度, 使聚类过程更关注特征方向差异. 本文同时保留未经 *Z*-Score 标准化和 L_2 归一化处理的原始特征方案, 用于考察模型原始输出的数值尺度和向量模长对聚类结构的影响. 由于模型输入已经过逐记录、逐分量标准差归一化, 故本文不将特征向量的数值大小或模长直接解释为原始波形的绝对幅值或能量.

本文采用 K-means++算法对 512 维特征向量进行聚类. K-means 是最常用的划分式聚类方法之一, 其基本思想是在给定簇数 k 的条件下, 将样本划分为 k 个互不重叠的簇, 并通过迭代更新簇中心与样本归属, 使簇内平方误差和最小(MacQueen, 1967). 设第 i 个波形样本的特征向量为 z_i, 第 c 个簇为 C_c, 其簇中心为 μ_c, 则目标函数为:

$$J=\sum_{c=1}^{k}\sum_{z_i\in C_c}\|z_i-\mu_c\|_2^2, \tag{1}$$

式中 $\|\cdot\|_2^2$ 表示欧氏距离的平方. K-means++通过概率化初始化策略优先选择彼此距离较远的样本点作为初始簇中心, 从而提高初始中心的分散性, 改善算法收敛稳定性, 并降低陷入较差局部最优的概率(Arthur, Vassilvitskii, 2007).

为减轻初始簇中心随机性对聚类结果的影响, 本文对四种输入特征方案分别在簇数 $k\in[2, 10]$ 时进行 K-means++聚类. 对于每一种特征方案和每一个候选 k, 均独立重复运行 20 次, 每次采用不同随机种子, 最大迭代次数为 300, 收敛阈值为 10^{-4}, 并保留簇内平方误差和最小的一次结果作为该方案和该簇数下的代表性聚类结果.

随后分别计算轮廓系数、CH 指数和 DB 指数, 以评价不同特征处理方案和不同簇数下的聚类结构(Caliński, Harabasz, 1974; Davies, Bouldin, 1979; Rousseeuw, 1987). 其中: 轮廓系数综合衡量样本与本簇内其它样本的相似性及其与最近邻簇之间的分离程度, 取值范围为 [−1, 1], 轮廓系数越接近 1, 表示聚类结构越清晰; CH 指数通过比较簇间离散度和簇内离散度来评价聚类效果, 数值越大通常表示簇间分离越明显、簇内越紧凑; DB 指数刻画簇内平均散布程度与簇中心间距离的相对关系, 数值越小表示聚类结果在簇内紧凑性与簇间分离度上的综合表现越好.

为进一步评价聚类结构, 本文开展聚类倾向、Gap 统计量、重复初始化和事件重抽样稳定性检验. 首先, 将原始特征向量经主成分分析(principal component analysis, 缩写为 PCA)降至保留 95% 累计方差的 122 维空间, 重复 50 次计算 Hopkins 统计量, 每次随机抽取 10% 的样

本；同时在簇数 $k\in[1, 10]$ 时计算 Gap 统计量，每个候选 k 采用 20 组参考数据，并根据一个标准误差准则评价簇数（Hopkins，Skellam，1954；Tibshirani *et al*，2001）.

对 $k=2$ 和 $k=4$ 分别采用不同随机种子独立运行 K-means＋＋100 次，通过调整兰德指数（adjusted rand index，缩写为 ARI）和归一化互信息（normalized mutual information，缩写为 NMI）评价给定簇数下的初始化稳定性（Hubert，Arabie，1985；Vinh *et al*，2010）. 随后以独立事件为抽样单位，每次随机抽取 80% 的独立事件重新聚类，共重复 100 次，并计算重抽样结果相对于全样本结果的 ARI，NMI 和各簇 Jaccard 保持率，以评价聚类结构对样本组成的敏感性.

此外，采用 PCA 降维后的 K-means＋＋、基于余弦距离的球面 K-means、对角协方差高斯混合模型和层次化基于密度的含噪声空间聚类（Hierarchical Density-Based Spatial Clustering of Applications with Noise，缩写为 HDBSCAN）进行替代方法比较，并通过 ARI 和 NMI 评价其与原始特征 $k=4$ 结果的一致程度. 上述分析仅用于评价聚类结构对抽样、距离度量和聚类模型的敏感性，不使用地形、重力或其它区域背景资料确定簇数.

3 结果

本文将 DiTingSFM-0.1B 预训练模型输出的 512 维特征向量作为分析对象，围绕“特征向量是否保留可检验的波形、目录参数和空间分布信息”展开结果分析. 首先，比较不同特征预处理方案下的聚类指标和稳定性变化，评价特征空间中的潜在分组结构，并确定最终聚类方案，采用 t 分布随机近邻嵌入（t-distributed stochastic neighbor embedding，缩写为 t-SNE）展示聚类结果在特征向量空间中的分布；随后通过特征空间最近邻一致性检验、簇内波形包络与频谱分析以及空间分布对比，分析特征向量聚类结果在特征空间、波形参数和区域背景中的表现.

3.1 簇数选择及 t-SNE 可视化

图 3 给出了原始特征、*Z*-Score 标准化特征、L_2 归一化特征以及 *Z*-Score＋L_2 归一化特征四种输入方案下，轮廓系数、CH 指数和 DB 指数随簇数 k 的变化. 由于不同特征预处理方式会改变特征空间中的距离关系，四种方案下的指标值并不能简单作为完全等价的绝对优劣比较，而主要用于分析聚类结构对特征数值尺度和向量模长的敏感性.

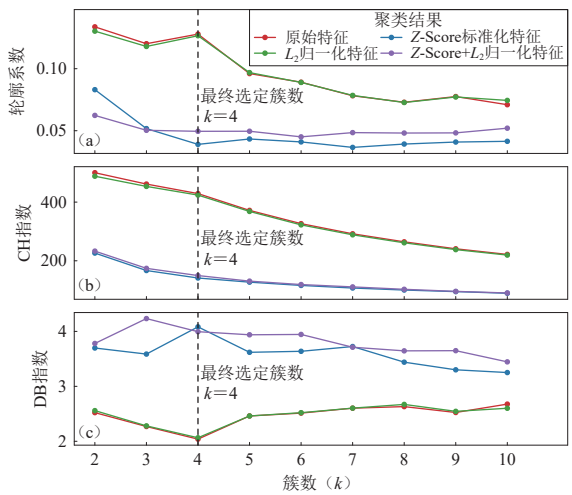

图 3　轮廓系数(a)、CH 指数(b)和 DB 指数(c)随簇数 k 的变化

Fig. 3　Variations in the silhouette coefficient (a), CH index (b), and DB index (c) with the cluster number k

从图 3 可以看出，在原始特征向量空间中，轮廓系数和 CH 指数均在 $k=2$ 时达到最大，提示特征向量中存在较明显的主导二分结构；同时，DB 指数在 $k=4$ 时达到最小，表明在候选簇数范围内，$k=4$ 对应的聚类结果在簇内紧凑性与簇间分离度之间具有较好的平衡. 类似地，在仅进行 L_2 归一化的特征空间中，$k=2$ 仍表现出最高的轮廓系数和 CH 指数，而 $k=4$ 对

应最低的 DB 指数. 相比之下, 在 *Z*-Score 标准化及 *Z*-Score+L_2 归一化后, $k=2$ 的优势更加突出, 说明当原始特征中的数值尺度差异被削弱后, 特征空间主要表现为更粗粒度的二分结构.

上述结果表明, 不同预处理方式会改变特征空间的距离结构, 模型原始输出的数值尺度和向量模长也是影响聚类划分的因素. 由于模型输入已经过逐记录、逐分量标准差归一化, 这些信息不能直接解释为原始波形的绝对幅值或能量, 但可能综合反映归一化后波形的形态、频谱结构、持续时间和信噪比等差异. 为保留模型原始输出的距离结构, 并与标准化和归一化方案进行敏感性对照, 本文采用原始 512 维特征向量作为后续聚类输入.

经 PCA 降至保留 95% 累计方差的 122 维空间后, Hopkins 统计量中位数为 0.757, 95% 区间为 0.754—0.761, 表明降维后的特征空间具有一定的非随机聚集结构. 不同簇数评价指标给出的结果并不完全一致: 轮廓系数和 CH 指数在 $k=2$ 时较高, 反映特征空间中较明显的粗粒度分异; DB 指数则在 $k=4$ 时达到较低值. Gap 统计量在候选范围内未给出明确且唯一的最优簇数. 由此可见, 现有指标能够支持特征空间具有分组结构, 但不足以确定唯一的自然类别数.

考虑到 $k=2$ 主要描述特征空间的整体分异, 难以进一步比较其内部不同波形表征的差别, 本文在保留这一粗粒度认识的基础上, 采用 $k=4$ 作为后续分析的操作性细分方案. 该选择主要用于考察不同特征向量子集在波形包络、频谱结构和目录参数上的差异, 不将四个簇视为边界明确、彼此独立的客观地震学类别, 也不以其与地形或重力背景的对应关系作为簇数选择依据.

稳定性分析表明, $k=4$ 的聚类划分在不同初始化和独立事件重抽样条件下具有较高的一致性, 独立事件重抽样结果的 ARI 和 NMI 中位数分别为 0.979 和 0.963, 各簇 Jaccard 保持率中位数为 0.981—0.987. PCA 降维后的 K-means++和球面 K-means 的结果也与原始划分较为一致, 而高斯混合模型和 HDBSCAN 未恢复出相同的四簇结构. 这说明 $k=4$ 的聚类结果在 K-means 及相近距离聚类框架下具有较好的可重复性, 但仍存在一定的聚类模型依赖性. 后续分析均以这一操作性细分方案为条件进行解释.

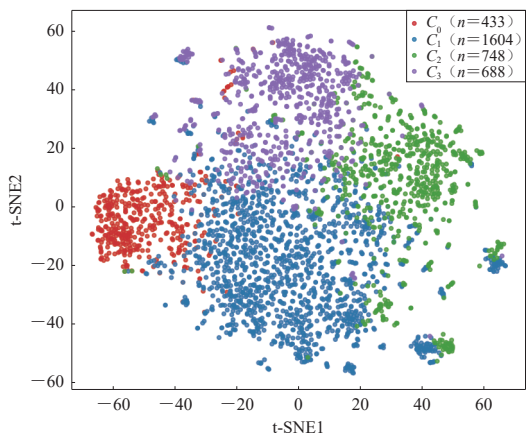

图 4 聚类结果的 t-SNE 可视化

横、纵坐标分别为 t-SNE 降维得到的第一和第二嵌入坐标, 均为无量纲相对坐标; 每个散点表示一条近震波形样本, 不同颜色表示 K-means++在原始 512 维特征向量空间中得到的簇标签. 坐标值仅表示样本在二维嵌入空间中的相对位置, 不对应具体物理量

Fig. 4 t-SNE visualization of the clustering results

The horizontal and vertical axes represent the first and second embedding coordinates obtained using t-SNE, respectively, and are dimensionless relative coordinates. Each point represents a local earthquake waveform sample, and the colors indicate the

最终 $k=4$ 聚类结果的样本数量及主要事件参数统计见表 1. C_1 样本数最多, 包含 1 604 个波形样本; C_0 样本数最少, 包含

表 1　各簇波形样本、独立事件和台站数量及主要参数

Table 1　Number of waveform samples, independent events, and stations in each cluster, together with the main parameters

簇	样本个数	独立事件个数	台站个数	震中经度范围/°E	震中纬度范围/°N	震中距中位数/km	震级中位数/M_L	深度中位数/km
C_0	433	417	207	76.94—127.65	18.47—44.83	65.00	1.3	10.0
C_1	1 604	1 564	413	74.04—130.36	21.11—52.11	84.95	1.4	10.0
C_2	748	735	292	74.54—130.87	18.64—51.56	57.20	2.0	10.0
C_3	688	668	289	73.25—132.49	19.78—49.62	106.00	2.4	10.0

表 2　不同簇数下 K-means＋＋聚类的初始化及事件重抽样稳定性

Table 2　Initialization and event resampling stability of K-means＋＋ clustering for varying cluster numbers

簇数k	输入特征	重复次数	两两比较次数	ARI均值±标准差	ARI中位数	NMI均值±标准差	NMI中位数	重抽样ARI中位数及95%区间	重抽样NMI中位数及95%区间	各簇Jaccard保持率中位数
2	原始512维特征向量	100	4 950	0.621±0.461	0.958	0.609±0.451	0.912	0.976 (0.805—0.998)	0.950 (0.708—0.995)	C_0: 0.990, C_1: 0.984
4	原始512维特征向量	100	4 950	0.985±0.061	0.992	0.981±0.055	0.985	0.979 (0.952—0.995)	0.963 (0.933—0.990)	C_0: 0.982, C_1: 0.986, C_2: 0.987, C_3:0.981

433 个波形样本. 各簇在样本数量、震中分布、震中距和震级等方面呈现不同特征, 而震源深度中位数均为 10.0 km, 为后续波形特征分析和空间分布解释提供了基础.

为直观展示最终聚类结果在特征空间中的分布, 本文进一步采用 t-SNE 方法将原始 512 维特征向量降至二维进行可视化(Van Der Maaten, Hinton, 2008). t-SNE 仅用于可视化展示, 不参与 K-means＋＋聚类计算、簇数选择和聚类稳定性评价. 本文设置困惑度为 30, 学习率为自动模式, 最大迭代次数为1 000, 初始嵌入采用 PCA 结果, 距离度量采用欧氏距离, 随机种子设为 2 026 万零 626. 图 4 显示, 四个簇在二维嵌入空间中呈现出一定的分离和过渡特征, 直观反映了 $k=4$ 聚类结果在特征向量空间中的分布形态; 聚类稳定性仍以表 2 中的重复初始化检验结果为依据.

3.2　特征空间一致性及台站敏感性检验

为检验特征向量空间中的邻近关系是否对应震中位置的地理邻近关系, 并评估台站因素对该关系的影响, 本文开展最近邻一致性及台站敏感性分析. 首先, 在原始 512 维特征向量空间中排除同一独立地震事件的其它记录后, 为每条波形样本寻找最近邻样本, 并统计最近邻来自同一台站的比例. 作为对照, 在相同分析范围内为每条样本随机抽取一个不同事件的记录, 重复 5 000 次, 构建同台站随机配对比例的零分布.

随后, 同时排除同一事件和同一台站的其它记录, 重新确定每条样本的特征空间最近邻, 并计算最近邻样本对应震中之间的球面大圆距离, 地球半径取 6 371 km. 为控制不同台站的区域采样范围、台网覆盖和样本空间密度差异, 采用台站分块置换构建零分布(Good, 2005). 每次置换均在同一台站内部随机重排震中位置标签, 从而保持各台站的样本数量、震中位置集合及地理采样范围不变, 同时固定特征向量、聚类标签和最近邻关系. 置换重复 5 000 次, 以最近邻地理距离中位数作为统计量, 并采用单侧置换检验判断实际距离是否小于零分布. 本文置换检验的 p 值均采用加 1 校正, 即

$$p_{perm}=\frac{b+1}{B+1}, \tag{2}$$

式中，B 为置换次数，b 为置换统计量至少与实际统计量同样极端的次数. 对于同台站最近邻比例和跨台站簇标签一致率，以置换统计量大于或等于实际统计量为极端方向；对于最近邻地理距离，以置换统计量小于或等于实际统计量为极端方向. 本文取 $B=5\,000$，因此当 5 000 次置换中无一次达到或超过实际统计量的极端程度时，最小可分辨 p 值为 1/5 001，约为 2.0×10^{-4}，不记为 0.

此外，采用偏差校正的克拉默 V 系数(以下简称 Cramér's V)和归一化互信息量化台站身份与簇标签之间的关联. 对于 134 个具有多台站记录的事件，进一步统计不同台站记录之间的簇标签一致率，并通过标签置换构建随机基线. 最后，以台站为分组单位进行 20 次训练-验证划分，每次随机留出 20% 的台站，在其余台站样本上重新进行 $k=4$ 聚类，并计算留出台站样本与全样本聚类结果之间的 ARI 和 NMI，用于评价聚类结构的跨台站稳定性.

为进一步检验震级、震中距、反方位角和记录时间等已知协变量能否解释特征空间的地理邻近关系，本文进行了协变量残差化敏感性分析. 以台站身份、震级及其二次项、震中距及其二次项、反方位角的正弦和余弦分量，以及记录时间的线性项和年周期分量作为解释变量，采用岭回归对 512 维特征向量逐维拟合. 拟合采用按独立事件分组的五折交叉方式，利用模型预测值与原始特征向量之差构建残差特征向量. 在残差特征空间中再次排除同一事件和同一台站的记录，并重复最近邻搜索和台站分块置换检验. 各簇检验结果采用 Holm 方法进行多重检验校正.

由图 5 可以看出，在排除同一独立地震事件后，全体样本中有 38.1% 的特征空间最近邻来自同一台站，明显高于随机配对零分布中位数 0.37% 及其 95% 区间 0.20%—0.60%. C_0，C_1，C_2 和 C_3 的同台站最近邻比例分别为 16.6%，35.6%，42.8% 和 33.4%，也均明显高于对应随机基线，各组实际比例均高于全部 5 000 次随机配对结果，即 $b=0$，对应 $p_{\rm perm}=1/5\,001\approx2.0\times10^{-4}$. 该结果表明，同一台站记录的波形在特征向量空间中具有明显聚集倾向，台站场地、仪器频响、噪声背景及区域采样范围可能对特征空间结构产生重要影响.

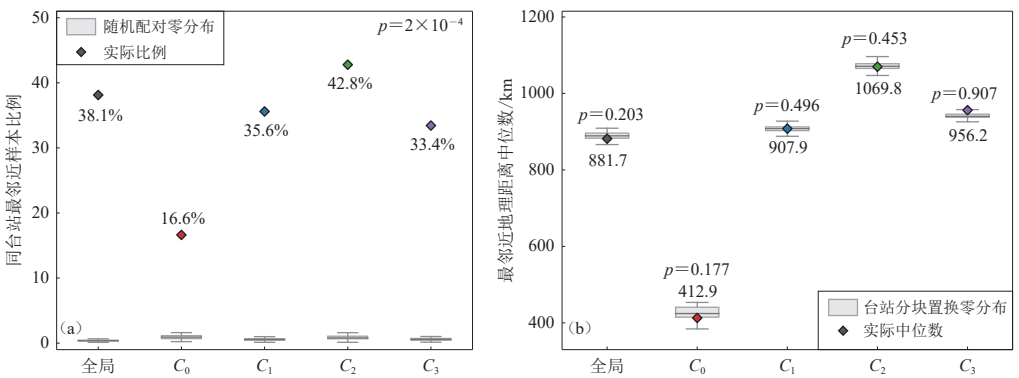

图 5 特征空间最邻近关系的台站敏感性检验

(a) 排除同一事件后，特征最邻近样本来自同一台站的比例；(b) 同时排除同一事件和同一台站后的最邻近地理距离. 单侧置换检验采用 $p=(b+1)/(B+1)$ 计算，其中 $B=5\,000$；图 a 以置换比例大于或等于实际比例为极端方向，图 b 以置换距离中位数小于或等于实际值为极端方向；箱线图的箱内横线、箱体和虚线分别表示零分布

的中位数、四分位距和 1.5 倍四分位距范围内的非异常值范围

Fig. 5　Station-sensitivity tests of nearest-neighbor relationships in the feature space

(a) Proportion of feature-space nearest neighbors recorded at the same station after excluding records from the same event; (b) Geographic distances between nearest neighbors after excluding records from both the same event and the same station. One-sided permutation p-values were calculated as $p=(b+1)/(B+1)$, with $B=5\,000$; the upper tail was used for the same-station proportion in Fig. a, whereas the lower tail was used for the median geographic distance in Fig. b; In the box plots, the center lines, boxes, and whiskers represent the medians, interquartile ranges, and ranges of non-outlier values within 1.5 times the interquartile range of the null distributions, respectively

同时排除同一事件和同一台站后，全体样本的最近邻地理距离中位数为 881.7 km. 在保持各台站震中位置集合和地理采样范围不变的台站分块置换中，零分布中位数为 889.1 km，95% 区间为 873.7—902.1 km，实际统计量位于置换零分布的 95% 区间内，单侧置换检验未显示显著差异（$p_{\mathrm{perm}}=0.203$）. 各簇内部也未显示显著差异，C_0—C_3 的 p_{perm} 分别为 0.177，0.496，0.453 和 0.907. 进一步对台站身份、震级、震中距、反方位角和记录时间进行残差化后，全体样本的最近邻地理距离中位数为 1 197.3 km，台站分块置换零分布中位数为 1 191.7 km，95% 区间为 1 179.0—1 203.1 km，差异仍不显著（$p_{\mathrm{perm}}=0.868$）. 各簇的检验结果经 Holm 校正后也均未达到显著水平，最小校正 p 值为 0.083.

事件的台站组成与簇标签的偏差校正 Cramér's V 为 0.540，归一化互信息为 0.182，说明簇标签与台站组成之间存在不可忽略但并非完全对应的关联. 另一方面，134 个多台站事件共形成 146 个跨台站记录对，其中 62.3% 的记录对被分入同一簇，明显高于置换零分布中位数 28.8%（95% 区间为 21.9%—36.3%），且高于全部 5 000 次置换结果，即 $b=0$，对应 $p_{\mathrm{perm}}=1/5\,001\approx 2.0\times10^{-4}$，表明特征向量中仍保留一定的事件相关信息. 经 20 次台站分组留出验证，得到 ARI 为 0.971±0.013，NMI 为 0.954±0.016，说明整体聚类结构在留出台站子集的检验条件下具有较高稳定性，但该结果并不等同于完全消除了台站响应的影响.

综上，特征空间最近邻关系表现出明显的台站依赖性. 在排除同一事件和同一台站，并进一步控制震级、震中距、反方位角和记录时间后，未检测到超出台站地理采样条件的显著震中邻近关系. 因此，现有最近邻结果主要反映特征向量与台站条件及观测采样结构之间的统计关联，不能单独作为模型编码传播路径或区域介质结构的证据. 另一方面，多台站事件的跨台站簇标签一致性显著高于随机基线，说明特征向量并非仅由台站因素决定，其中仍保留一定的事件相关波形信息.

3.3　簇内特征分析

特征向量是抽象的高维数值表示，不能直接等同于具体物理量. 为考察不同特征簇是否对应可观测的波形和目录参数差异，本文从波形包络、频谱结构、目录参数和波形统计指标等方面进行对比分析. 其中，波形包络、功率谱密度、主频和相对均方根采用独立于模型推理的数据处理流程计算.

波形统计从数字计数波形出发，对于每条记录，分别将 Z，N 和 E 三个分量减去各自在 60 s 窗口内的中位数，再除以该分量在同一窗口内的最大绝对值，得到逐记录、逐分量峰值归一化的无量纲波形：

$$\tilde{x}_c(n)=\frac{x_c(n)-\mathrm{median}[x_c]}{\max_n\left|x_c(n)-\mathrm{median}[x_c]\right|},\tag{3}$$

式中，c 为波形分量索引，$c\in\{\mathrm{Z},\ \mathrm{N},\ \mathrm{E}\}$，分别表示垂直、南北和东西分量；$n=0,\ \cdots,\ 5\ 999$；$x_c(n)$ 和 $x_c(n)$ 分别表示第 c 个分量在第 n 个采样点处的原始数字计数值和峰值归一化后的无量纲波形值，当分母为 0 时，将其置为 1. 需要说明的是，该处理仅用于波形统计分析，与模型推理过程中采用的逐分量标准差归一化不同.

对于任一时间窗 W，三分量相对均方根定义为：

$$\mathrm{RMS}_W=\sqrt{\frac{1}{3N_W}\sum_{c\in\{Z,\ N,\ E\}}\sum_{n\in W}\tilde{x}_c^2(n)},\tag{4}$$

式中 N_W 为时间窗内的采样点数. 60 s 全窗取 $W=[0,\ 6\ 000)$，P-S 窗取 $W=[n_\mathrm{P},\ n_\mathrm{S})$，S 后窗取 $W=[n_\mathrm{S},\ \min(n_\mathrm{S}+1\ 000,\ 6\ 000))$，即 S 波到时后 10 s，超出记录末端时截断. 上述 RMS 均由峰值归一化波形计算，为无量纲相对量，主要反映不同时间窗内波形相对于各分量峰值的持续起伏程度，不表示绝对地面运动强度或波形绝对能量.

功率谱密度采用 Welch 方法计算，采样率为 100 Hz，分段长度为 512 个采样点，相邻分段重叠 256 个采样点，并对三个分量的功率谱密度取平均；主频定义为平均功率谱密度最大

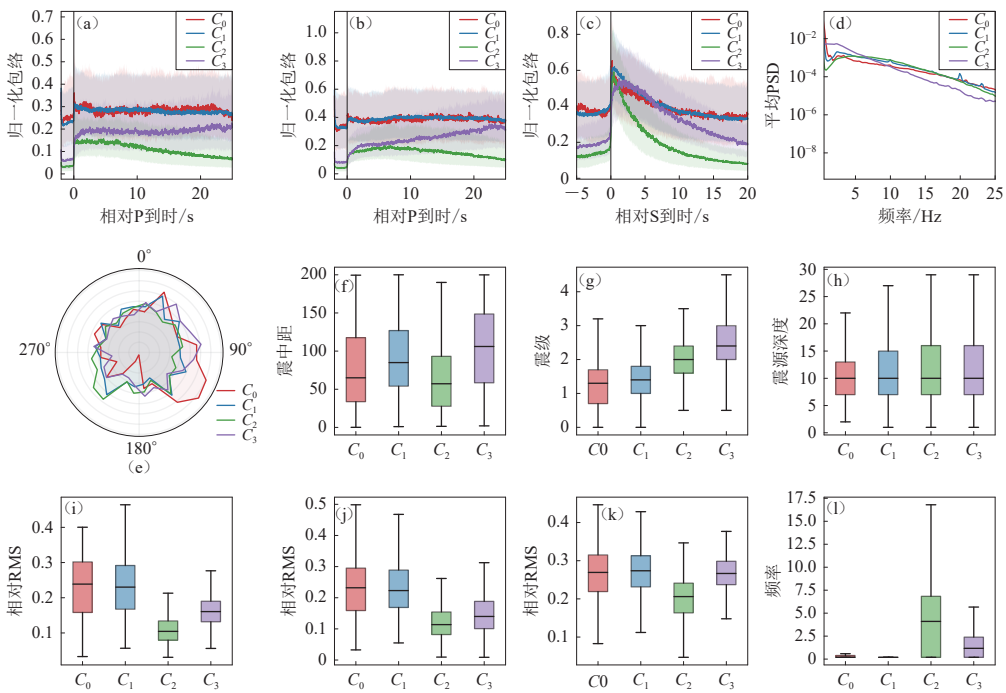

图 6 簇内波形包络、频谱及簇间差异分析

(a) Z 分量 P 对齐包络；(b) 水平分量 P 对齐包络；(c) 水平分量 S 对齐包络；(d) 平均功率谱密度 PSD；(e) 反方位角；(f) 震中距；(g) 震级；(h) 震源深度；(i) 60 s 全窗相对 RMS；(j) P-S 窗相对 RMS；(k) S 后窗相对 RMS；(l) 主频. 箱线图虚线表示 1.5 倍四分位距范围内的非异常值范围

Fig. 6 Intra-cluster waveform envelopes and spectra and inter-cluster difference analysis

(a) P-aligned Z-component envelope; (b) P-aligned horizontal-component envelope; (c) S-aligned horizontal-component envelope; (d) Mean power spectral density; (e) Back azimuth; (f) Epicentral distance; (g) Magnitude; (h) Focal depth; (i) Relative RMS over the full 60 s window; (j) Relative RMS in the P-S window; (k) Relative RMS in the post-S window; (l) Dominant frequency. In the box plots, the whiskers indicate non-outlier values within 1.5 times the interquartile range.

值对应的频率．Z 分量包络由 Z 分量 Hilbert 包络获得，水平分量包络由 N，E 两个水平分量 Hilbert 包络的平方和开方得到．对于目录参数和波形统计指标，为减小同一事件多台站记录及台站内样本相关性的影响，统计检验时首先对每个独立事件随机保留一条记录，随后采用 5 000 次台站分块置换，构建 Kruskal-Wallis 检验统计量（以下简称 H 统计量）的零分布，H 统计量用于衡量不同簇样本秩分布的总体差异，数值越大通常表示簇间差异越明显．并按第 3.2 节所述的加 1 校正方法计算右尾置换 p 值，将置换 H 值大于或等于实际 H 值定义为极端方向．反方位角属于圆周变量，因此本文将其转换为 sin（Baz）和 cos（Baz）两个分量后分别进行检验；其余连续变量采用 Kruskal-Wallis 检验比较四个簇之间的总体差异．对全部 9 项总体检验的 p 值采用 Holm 方法进行多重检验校正，并同时报告 H 统计量和 ε^2 效应量（Kruskal，Wallis，1952）．

表 3　不同聚类类别事件参数和波形统计指标的稳健总体检验结果

Table 3　Robust overall test results for event parameters and waveform statistical metrics for different cluster categories

指标	H值	自由度	Holm校正p值	ε^2
sin（Baz）	26.39	3	0.085	0.007
cos（Baz）	4.23	3	0.570	＜0.001
震中距	231.86	3	0.001 8	0.069
震级	1 120.93	3	0.001 8	0.336
震源深度	7.34	3	0.269	0.001
60 s全窗相对RMS	1 206.26	3	0.001 8	0.361
P-S窗相对RMS	1 000.89	3	0.001 8	0.300
S后窗相对RMS	588.51	3	0.001 8	0.176
主频	1 452.39	3	0.001 8	0.435

图 6a—c 给出了各簇波形包络统计结果．其中，图 6a 为 Z 分量 P 对齐包络，图 6b 为水平分量 P 对齐包络，图 6c 为水平分量 S 对齐包络；横轴分别为相对 P 波或 S 波到时，0 s 处虚线表示对齐到时，曲线表示各簇包络中位数，阴影表示 25%—75% 分位区间．可以看出，不同簇在 P 波前后及 S 波后续时窗的包络形态存在差异：C_0 和 C_1 整体包络幅值较高，C_2 包络幅值较低；在 S 对齐水平包络中，C_0，C_1 和 C_3 的 S 后包络衰减相对缓慢，C_2 衰减较快．图 6d 给出了各簇平均功率谱密度，采用 Welch 方法计算，采样率为 100 Hz，并对三分量功率谱密度取平均．结果显示 C_0 和 C_1 相对偏低频，C_2 在中高频部分更突出，C_3 的频谱整体衰减较明显．具体统计结果见表 3．

事件参数方面，图 6e—h 比较了反方位角、震中距、震级和震源深度在四个簇中的分布．反方位角以圆周分布图表示，并转换为 sin（Baz）和 cos（Baz）进行统计检验．经独立事件筛选、台站分块置换和 Holm 校正后，sin（Baz）和 cos（Baz）的校正 p 值分别为 0.085 和 0.570，效应量分别为 $\varepsilon^2=0.007$ 和 $\varepsilon^2<0.001$，表明反方位角在四个簇之间的总体差异较弱．震中距和震级的簇间差异仍达到显著水平，Holm 校正 p 值均为 0.001 8，效应量分别为 $\varepsilon^2=0.069$ 和 $\varepsilon^2=0.336$．其中，C_2 震中距整体较短，C_3 震中距较长；C_0 和 C_1 震级整体较低，C_2 中等偏高，C_3 最高．震源深度未显示显著簇间差异（$H=7.34$，校正 $p=0.269$，$\varepsilon^2=0.001$），因此仅认为现有目录深度未表现出明显的簇间分异．

波形统计结果见图 6i—l 和表 3．经独立事件筛选、台站分块置换和 Holm 校正后，60 s 全窗相对 RMS、P-S 窗相对 RMS、S 后窗相对 RMS 和主频的簇间差异仍达到显著水平，Holm 校正 p 值均为 0.001 8；其 H 统计量分别为 1 206.26，1 000.89，588.51 和 1 452.39，对应效应量 ε^2 分别为 0.361，0.300，0.176 和 0.435．其中，C_0 和 C_1 的 60 s 全窗相对 RMS 及 P-S 窗相对 RMS 整体较高，C_2 在三项相对 RMS 指标上均较低，C_3 的 S 后窗相对 RMS 相对较高．这些差异表示峰值归一化后波形在不同时间窗内的相对持续起伏程度不同，不能解释为绝对

振幅或绝对波形能量的差异. C_0 和 C_1 的主频中位数约为 0.20 Hz, C_2 约为 4.10 Hz, C_3 约为 1.17 Hz, 表明 C_2 具有更突出的高频特征. 结合波形包络和平均功率谱密度结果可见, 不同特征簇不仅在模型特征向量空间中具有差异, 也对应峰值归一化波形的包络形态、相对持续起伏和频谱结构差异.

综合来看, 四个聚类类别不仅在特征向量空间中可分, 也在波形包络、频谱结构和关键事件参数上表现出系统差异. C_0 可概括为低震级、中等震中距、低频且波形起伏较强的一类; C_1 为样本数最多的一类, 整体表现为低震级、中等震中距、低频和较高相对 RMS; C_2 具有较短震中距、中等偏高震级、较低相对 RMS 和显著高频特征; C_3 则表现为较大震中距、较高震级、低频衰减较明显和中等相对 RMS. 这些差异说明, 模型输出的特征向量并非只是在数学空间中形成分组, 而是与可观测的波形形态、频谱结构和事件参数差异相对应.

为检验 512 维特征向量聚类结果是否主要由少数变量驱动, 本文构建了六组低维基线: 特征向量模长、60 s 全窗相对 RMS 和主频三个单项基线; 由震级、震中距、60 s 全窗相对 RMS、主频和反方位角组成的简单参数基线; 由三项相对 RMS、主频、频谱质心、低频-中频-高频频带功率及其比值组成的波形统计特征基线; 以及由震级、震中距、震源深度、反方位角的正弦和余弦分量与上述波形统计特征组成的组合基线. 所有低维基线均采用中位数填补缺失值并进行 *Z*-Score 标准化, 随后使用一致的 K-means++流程进行 $k=4$ 聚类, 即采用欧氏距离、20 次不同随机初始化、最大迭代次数 300 和收敛阈值 10^{-4}, 并通过 ARI 和 NMI 衡量其与 512 维特征向量聚类结果的一致程度.

结果显示, 特征向量模长单项基线的 ARI 和 NMI 分别为 0.214 和 0.292, 60 s 全窗相对 RMS 单项基线分别为 0.116 和 0.169, 主频单项基线分别为 0.226 和 0.264; 波形统计特征组合基线分别为 0.170 和 0.276, 目录参数-波形统计特征组合基线分别为 0.195 和 0.274. 由震级、震中距、60 s 全窗相对 RMS、主频和反方位角组成的简单参数基线的 ARI 和 NMI 分别为 0.159 和 0.223. 上述结果表明, 所检验的单项变量和低维组合均不能直接复现 512 维特征向量的四簇划分, 但这种低一致性本身不能证明高维特征包含更丰富或更具物理意义的信息. 两类划分的差异还可能来自未纳入的波形属性、台站和仪器信息、高维噪声以及特征尺度和距离结构等因素, 仍需结合独立数据和进一步的表征探测加以区分.

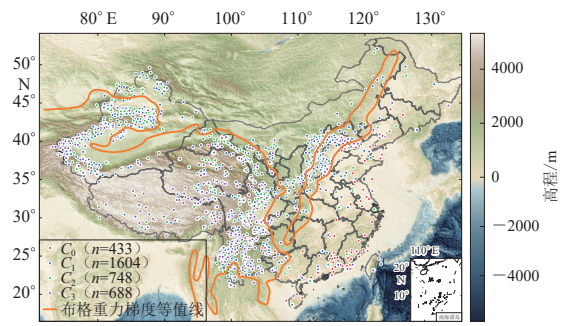

图 7 聚类结果空间分布

橙色实线为基于 WGM2012 布格重力异常数据计算得到的主要布格重力梯度等值线, 用于表示研究区主要重力场变化背景

Fig. 7 Spatial distribution of clustering results

Orange solid lines denote the major Bouguer gravity-gradient contours calculated from WGM2012 Bouguer gravity anomaly data, representing the main gravity-field variation background of the study area

3.4 空间分布及区域背景对比

为进一步分析特征向量聚类结果的空间分布特征, 本文将各簇样本震中位置绘制在地形图上, 并引入全球重力图 2012(World Gravity Map 2012, 缩写为 WGM2012)中的布格重力异常数据计算布格重力梯度等值线, 将其作为区域地球物理背景参考, 用于观察聚类空间分布与地形、重力场变化之间的空间对应现象.

首先对布格重力异常进行 75 km 尺度平滑，以削弱局部短波长异常对梯度计算的影响，突出区域尺度重力变化特征；随后将经纬度间距转换为水平距离，分别计算东西向和南北向重力异常变化率，并得到水平梯度幅值：

$$|\nabla g| = \sqrt{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial y}\right)^2}, \tag{5}$$

式中，x 和 y 数值单位均转换为 km，因此水平梯度单位为 $10^{-8}\ s^{-2}$. 本文选取研究区水平梯度幅值的全局 60% 分位数作为等值线阈值，该阈值约为 $0.265\times10^{-8}\ s^{-2}$. 为突出研究区尺度较大的重力场变化线索，并减少短小零散等值线对图面判读的干扰，仅保留沿线累计长度不小于 6 500 km 的连续等值线段，最终保留两条主要布格重力梯度等值线，以橙色实线绘制在图 7 中.

四个簇均涉及较多台站，C_0，C_1，C_2 和 C_3 分别包含 207，413，292 和 289 个台站，最大单台样本占比分别为 4.6%，2.9%，4.3% 和 3.2%. 第 3.2 节结果表明，各簇的台站组成仍存在一定差异，因此图 7 中的空间分布还需同时考虑台站组成和地理采样范围的影响.

从图 7 可以看出，四个聚类在空间上表现出一定的区域分异特征. C_0 主要分布于研究区东部和东南部部分区域，并与其它簇之间呈现相对清楚的空间过渡；C_1 分布范围较广，在华北、西北及部分中部地区均有出现；C_2 和 C_3 更多分布于西部、青藏高原及其邻区. 结合表 1 和图 6 可知，不同簇在震级、震中距、峰值归一化波形的相对 RMS 和主频等方面也存在差异，说明聚类空间分布并非孤立现象，而与波形频谱特征和事件参数差异共同出现.

同时，布格重力梯度等值线与部分聚类转换区域呈现一定空间对应的现象. 特别是 C_0 与其它簇之间的空间过渡带，大致位于北北东方向的大兴安岭—太行山—武陵山重力高梯度带附近(马宗晋等，2006). 该重力高梯度带通常被认为反映了中国大陆东部与中西部之间地壳密度结构和构造背景的显著变化. 结合图 7 可以认为，特征向量聚类结果的空间分异与区域重力场和地形背景之间存在一定对应现象，但这种对应关系目前仍属于定性空间对比.

近震波形同时受到震源过程、传播路径、台站场地和仪器响应等因素影响，不同区域的速度结构、衰减条件和场地响应可能改变波形包络、频谱组成和持续时间，并进一步反映在模型特征向量中. 然而，第 3.2 节的台站分块置换和协变量残差化检验均未显示独立而显著的震中邻近关系，说明图 7 中的空间分异还可能受到台站地理采样范围、台网覆盖和事件参数分布的影响. 因此，聚类结果与地形及布格重力梯度背景之间的关系仅作为探索性空间对应现象，其具体地球物理含义仍需结合独立区域数据、台站场地条件和事件-台站传播路径进一步验证.

4 讨论与结论

地震波大模型输出的特征向量是连接原始波形与下游智能处理任务的中间表示，分析其中包含的地震学信息有助于判断模型结果的物理含义和适用范围. 本文以谛听数据集 2.0 中震中距小于 200 km 的 3 473 条近震波形样本为数据基础，将 DiTingSFM-0.1B 预训练模型输出的 512 维特征向量作为分析对象，在特征向量空间中开展 K-means＋＋无监督聚类，并通过特征空间最近邻关系、波形包络与频谱差异以及空间分布对应关系，间接检验特征向量中是否保留与震源、传播路径和区域背景有关的信息. 主要结论如下：

1）特征空间最近邻关系表现出明显的台站依赖性. 排除同一事件后，全体样本中

38.1% 的最近邻样本来自同一台站，明显高于随机配对基线. 同时排除同一事件和同一台站后，最近邻地理距离中位数为 881.7 km，在台站分块置换中与零分布差异不显著；进一步控制台站身份、震级、震中距、反方位角和记录时间后，残差特征的最近邻地理距离仍未显示显著差异. 上述结果表明，原始特征空间中的地理邻近现象在较大程度上受到台站条件和观测采样结构影响，不能据此单独推断模型编码了传播路径或区域介质信息. 多台站事件的跨台站记录簇标签一致率显著高于随机基线，说明特征向量中仍保留一定的事件相关波形信息.

2）簇内特征分析表明，四个聚类在波形包络、频谱结构和关键事件参数上均表现出差异. 其中，震级、震中距、峰值归一化波形的相对 RMS 和主频是区分不同簇的主要指标，现有目录深度未显示显著簇间差异. C_0 可概括为低震级、中等震中距、低频且波形起伏较强的一类；C_1 为样本数最多的一类，整体表现为低震级、中等震中距、低频和较高相对 RMS；C_2 具有较短震中距、中等偏高震级、较低相对 RMS 和显著高频特征；C_3 则表现为较大震中距、较高震级、低频衰减较明显和中等相对 RMS. 多组单项和组合基线与 512 维特征向量聚类结果的一致性均较低，其中 ARI 不超过 0.226，NMI 不超过 0.292，说明当前四簇划分不能由所检验的单一变量或低维参数组合直接复现. 但该结果并不意味着 512 维特征向量必然包含更丰富或更具物理意义的信息，其差异来源仍可能涉及其它波形属性、台站和仪器条件、高维噪声及特征尺度等因素.

3）空间分布与地形及布格重力梯度等值线的对比表明，聚类结果与区域地球物理背景之间存在一定空间对应关系. 特别是 C_0 与其它簇之间的空间过渡带，大致位于北北东方向的大兴安岭—太行山—武陵山重力高梯度带附近. 研究区东部沉积层较发育，西部地形较高且基岩出露较多，这种区域差异可能通过传播路径格林函数、场地响应和频谱衰减等因素影响近震波形，并进一步反映在模型提取的特征向量中. 这一结果表明，特征向量聚类的空间分异与区域地形和重力场背景之间存在一定联系，其具体构造含义仍需结合更多地质与地球物理资料进一步验证.

综上，基于无监督聚类的特征向量分析可为地震波大模型表征分析提供一种可行途径. 本文将预训练模型输出的高维特征向量作为直接分析对象，通过无监督聚类识别其潜在结构，并结合特征空间邻近关系、波形与频谱差异以及空间分布背景，分析其中可能保留的地震学信息. 结果表明，DiTingSFM-0.1B 预训练模型输出的 512 维特征向量反映了震级、震中距、峰值归一化波形的相对 RMS 和主频等可观测差异，也呈现出与台站组成和震中空间分布相关的结构，其中空间邻近关系受到台站地理采样范围的明显影响.

需要指出的是，本文仍存在一定局限性. 首先，本文的基本分析单元为事件-台站三分量近震波形样本，特征向量同时受到震源、传播路径、台站场地条件和仪器响应等因素影响. 由于本文未对不同台站记录实施仪器响应校正和增益统一，相对 RMS、主频及频谱衰减可能同时受到仪器频响、台站场地、传播路径和震源过程的影响. 因此，这些指标仅用于描述现有数据处理条件下不同特征簇的相对波形差异，不应被直接解释为绝对地面运动强度或单一震源物理机制. 虽然本文开展了台站分组留出法验证与台站分块置换检验，但尚未采用完全独立区域或事件-台站路径分组进行外部验证，也未系统控制区域地震活动密度和目录完备度差异，因此仍难以完全区分震源、传播路径、台站场地和采样结构对特征向量的相对贡献.

其次，DiTingSFM 的预训练语料约包含 2 600 万条三分量波形，其中 DiTing2 数据覆盖时

段为 2020 年 3 月至 2023 年 2 月．本文所用记录来自 DiTing2 的 2022 年 2 月至 2023 年 2 月，与 DiTingSFM-0.1B 的预训练数据在数据来源和时间范围上存在重叠．当前分析旨在揭示模型在既有数据分布下形成的内部表征结构，而不是检验完全独立数据上的预测泛化能力．因此，聚类及特征近邻关系可能同时反映可迁移的波形、震源和传播路径信息，以及特定事件、台站、仪器或区域的记忆成分．未来需要在预训练未覆盖时段、独立区域及台站隔离条件下重复相同分析，以检验这些表征规律的稳定性和可迁移性．

再次，图 7 中布格重力梯度等值线与聚类空间分布之间的关系主要基于图面叠加和定性对比，尚未进一步采用样本到梯度带距离、空间重合率或置换统计量进行定量检验；同时，本文尚未引入精细活动断裂、沉积层厚度、速度结构或地质分区矢量数据，因此对聚类空间分异的构造解释仍具有一定概化性．后续研究可结合独立时段、独立区域或留一台站样本，以及统一仪器响应后的物理量、台站坐标、事件-台站射线路径、台站场地条件、区域速度结构、活动断裂和地质分区资料，对聚类空间分异与区域地球物理背景之间的关系进行定量验证，并进一步揭示特征向量中震源、路径和场地响应信息的来源及其适用范围．

谛听数据集 2.0（DOI: 10.12 080/nedc.11.ds.2023.0002；CSTR: 12166.11.ds.2023.0002）为本研究提供了地震波形数据，DiTingSFM-0.1B 预训练模型（https://diting4eqs.cn）为提取地震波形特征向量提供了技术支持，全球通用大洋水深图（General Bathymetric Chart of the Oceans，缩写为 GEBCO）（https://www.gebco.net/）提供了研究区地形网格数据，WGM2012（https://bgi.obs-mip.fr/grids-and-models-2）提供了研究区布格重力异常数据，审稿人为本文的撰写提出了建设性意见和建议，作者在此一并表示感谢．

参　考　文　献

邓起东，张培震，冉勇康，杨晓平，闵伟，楚全芝，2002．中国活动构造基本特征 [J]．中国科学（D 辑），**32**（12）：1020–1030．

Deng Q D，Zhang P Z，Ran Y K，Yang X P，Min W，Chu Q Z，2003．Basic characteristics of active tectonics of China[J]．*Science in China Series D：Earth Sciences*，**46**（4）：356–372（in Chinese）．

马宗晋，高祥林，宋正范，2006．中国布格重力异常水平梯度图的判读和构造解释 [J]．地球物理学报，**49**（1）：106–114．

Ma Z J，Gao X L，Song Z F，2006．Analysis and tectonic interpretation to the horizontal-gradient map calculated from Bouguer gravity data in the China mainland[J]．*Chinese Journal of Geophysics*，**49**（1）：106–114（in Chinese）．

许忠淮，汪素云，黄雨蕊，高阿甲，1989．由大量的地震资料推断的我国大陆构造应力场 [J]．地球物理学报，**32**（6）：636–647．

Xu Z H，Wang S Y，Huang Y R，Gao A J，1989．The tectonic stress field of Chinese continent deduced from a great number of earthquakes[J]．*Chinese Journal of Geophysics*，**32**（6）：636–647（in Chinese）．

张培震，邓起东，张国民，马瑾，甘卫军，闵伟，毛凤英，王琪，2003．中国大陆的强震活动与活动地块 [J]．中国科学（D 辑），**33**（S1）：12–20．

Zhang P Z，Deng Q D，Zhang G M，Ma J，Gan W J，Min W，Mao F Y，Wang Q，2003．Active tectonic blocks and strong earthquakes in the continent of China[J]．*Science in China Series D：Earth Sciences*，**46**（S2）：13–24（in Chinese）．

赵明，肖卓伟，陈石，张博，汪小厉，唐淋，郭凯，马莉，杨红艳，张怀，2023．谛听数据集 2.0：中国地震台网多功能大型人工智能训练数据集 [EB/OL]．https：//www.esdc.ac.cn/article/144．

Zhao M，Xiao Z W，Chen S，Zhang B，Wang X L，Tang L，Guo K，Ma L，Yang H Y，Zhang H，2023．Diting dataset 2.0：Multi-functional large-scale artificial intelligence training data set from Chinese seismic network[EB/OL]．https：//www.esdc.ac.cn/article/144（in Chinese）．**（查阅网上资料，未找到本条文献英文翻译信息，请确认）．**

Aki K, Richards P G, 2002. *Quantitative Seismology*[M]. 2nd ed. Sausalito: University Science Books: 3−6.

Arthur D, Vassilvitskii S, 2007. k-means++: The advantages of careful seeding[C]//*Proceedings of the* 18*th Annual ACM-SIAM Symposium on Discrete Algorithms*. New Orleans: Society for Industrial and Applied Mathematics: 1027−1035.

Bergen K J, Johnson P A, De Hoop M V, Beroza G C, 2019. Machine learning for data-driven discovery in solid Earth geoscience[J]. *Science*, **363**(6433): eaau0323.

Beroza G C, Segou M, Mousavi S M, 2021. Machine learning and earthquake forecasting-next steps[J]. *Nat Commun*, **12**(1): 4761.

Caliński T, Harabasz J, 1974. A dendrite method for cluster analysis[J]. *Commun Stat*, **3**(1): 1–27.

Caron M, Bojanowski P, Joulin A, Douze M, 2018. Deep clustering for unsupervised learning of visual features[C]//*Proceedings of the* 15*th European Conference on Computer Vision-ECCV* 2018. Munich: Springer: 139−156.

Davies D L, Bouldin D W, 1979. A cluster separation measure[J]. *IEEE Trans Pattern Anal Mach Intell*, **PAMI-1**(2): 224-227.

Good P, 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*[M]. 3rd ed. New York: Springer: 33−65.

Hopkins B, Skellam J G, 1954. A new method for determining the type of distribution of plant individuals[J]. *Ann Bot*, **18**(2): 213–227.

Huang G, Sun Y, Liu Z, Sedra D, Weinberger K Q, 2016. Deep networks with stochastic depth[C]//*Proceedings of the 14th European Conference on Computer Vision - ECCV 2016*. Amsterdam: Springer: 646−661.

Hubert L, Arabie P, 1985. Comparing partitions[J]. *J Classif*, **2**(1): 193–218.

Jain A K, 2010. Data clustering: 50 years beyond K-means[J]. *Pattern Recognit Lett*, **31**(8): 651–666.

Jia L Z, Chen H F, Xing K, 2022. Rapid classification of local seismic events using machine learning[J]. *J Seismol*, **26**(5): 897–912.

Kruskal W H, Wallis W A, 1952. Use of ranks in one-criterion variance analysis[J]. *J Am Stat Assoc*, **47**(260): 583–621.

Lloyd S, 1982. Least squares quantization in PCM[J]. *IEEE Trans Inf Theory*, **28**(2): 129–137.

MacQueen J, 1967. Some methods for classification and analysis of multivariate observations[C]//*Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume* 1: *Statistics*. Berkeley: University of California Press: 281−297.

Mousavi S M, Zhu W Q, Ellsworth W, Beroza G, 2019. Unsupervised clustering of seismic signals using deep convolutional autoencoders[J]. *IEEE Geosci Remote Sens Lett*, **16**(11): 1693–1697.

Mousavi S M, Ellsworth W L, Zhu W Q, Chuang L Y, Beroza G C, 2020. Earthquake transformer: An attentive deep-learning model for simultaneous earthquake detection and phase picking[J]. *Nat Commun*, **11**(1): 3952.

Mousavi S M, Beroza G C, 2022. Deep-learning seismology[J]. *Science*, **377**(6607): eabm4470.

Nie Y Q, Nguyen N H, Sinthong P, Kalagnanam J, 2023. A time series is worth 64 words: Long-term forecasting with transformers[C]//*Proceedings of the* 11*th International Conference on Learning Representations*. Kigali: OpenReview.

Perol T, Gharbi M, Denolle M, 2018. Convolutional neural network for earthquake detection and location[J]. *Sci Adv*, **4**(2): e1700578.

Ross Z E, Meier M A, Hauksson E, 2018a. P wave arrival picking and first-motion polarity determination with deep learning[J]. *J Geophys Res*: *Solid Earth*, **123**(6): 5120–5129.

Ross Z E, Meier M A, Hauksson E, Heaton T H, 2018b. Generalized seismic phase detection with deep learning[J]. *Bull Seismol Soc Am*, **108**(5A): 2894–2901.

Rousseeuw P J, 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. *J Comput Appl Math*, **20**: 53–65.

Seydoux L, Balestriero R, Poli P, De Hoop M, Campillo M, Baraniuk R, 2020. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning[J]. *Nat Commun*, **11**(1): 3972.

Si X, Wu X M, Sheng H L, Zhu J, Li Z F, 2024. SeisCLIP: A seismology foundation model pre-trained by multimodal data for multipurpose seismic feature extraction[J]. *IEEE Trans Geosci Remote Sens*, **62**: 5903713.

Su J L, Ahmed M, Lu Y, Pan S F, Wen B, Liu Y F, 2024. RoFormer: Enhanced transformer with rotary position embedding[J]. *Neurocomputing*, **568**: 127063.

Tibshirani R，Walther G，Hastie T，2001. Estimating the number of clusters in a data set *via* the gap statistic[J]. *J Roy Stat Soc Ser B*：*Stat Methodol*，**63**(2)：411–423.

Van Der Maaten L，Hinton G，2008. Visualizing data using t-SNE[J]. *J Mach Learn Res*，**9**(86)：2579–2605.

Vaswani A，Shazeer N，Parmar N，Uszkoreit J，Jones L，Gomez A N，Kaiser Ł，Polosukhin I，2017. Attention is all you need[C]//*Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach：Curran Associates Inc.：6000−6010.

Vinh N X，Epps J，Bailey J，2010. Information theoretic measures for clusterings comparison：Variants，properties，normalization and correction for chance[J]. *J Mach Learn Res*，**11**(95)：2837–2854.

Zhang B，Sennrich R，2019. Root mean square layer normalization[C]//*Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vancouver：Curran Associates Inc.：1110.

Zhao M，Xiao Z W，Chen S，Fang L H，2023. DiTing：A large-scale Chinese seismic benchmark dataset for artificial intelligence in seismology[J]. *Earthquake Sci*，**36**(2)：84-94.**（查阅网上资料，不确定 Earthquake Science 这个是否是国内期刊英文版，不确定作者是否要全拼，请确认）**.

Zhu W Q，Beroza G C，2019. PhaseNet：A deep-neural-network-based seismic arrival-time picking method[J]. *Geophys J Int*，**216**(1)：261–273.