文章编号：1000-0550（2023）00-0000-00　　DOI：10.14027/j.issn.1000-0550.2023.117

# 大数据背景下粒度分布沉积信息挖掘方法进展

袁瑞

长江大学地球物理与石油资源学院，武汉　430100

**摘　要**　【意义】沉积物颗粒的大小反映了颗粒的搬运方式、沉积过程和沉积环境等沉积因素，利用粒度分布数据揭示现代和古代沉积环境是沉积学研究的基础之一。经典的粒度分析方法一直存在定量化不足和多解性突出的缺陷。随着数学理论的完备和计算机的发展，非传统的粒度分布沉积学分析技术为定量表征沉积属性提供了新思路。【进展】系统梳理了沉积物粒级划分标准、粒度参数计算和传统沉积环境分析方法，重点介绍了粒度分布聚类和多重分形的基本原理和应用方法，对比论述了基于概率密度函数的单个粒度分布分解和基于端元模型的粒度分布数据集分解的次总体分离方法及工具。【结论与展望】最终归纳了粒度分布沉积学分析面临的问题及其大数据特点，展望了粒度分布沉积学研究的两个发展方向，包括粒度分布沉积信息的智能挖掘和大数据库的建设。在大数据背景下，粒度分布大数据技术将为深度挖掘沉积属性提供新引擎。

**关键词**　大数据；粒度分布；沉积信息；智能挖掘

**第一作者简介**　袁瑞，男，1987 年出生，博士，副教授，地球科学数据挖掘及测井地质，E-mail: yuanrui@yangtzeu.edu.cn

**中图分类号**　P512.2　**文献标志码**　A

## 0　引言

碎屑沉积物颗粒由复杂的沉积动力学机制驱动并受控于沉积水动力条件和沉积过程，颗粒大小和组合反映了颗粒的搬运方式和沉积环境等沉积因素，记录了原始沉积属性[1-4]。因此，由不同大小颗粒体积或重量百分比组成的粒度分布（grain-size distribution, GSD）频率和累积频率包含了丰富的沉积信息。从 1875 年开始，粒度分布就成为了河流、湖泊、海洋、沙漠和黄土等环境中研究现代沉积过程和沉积环境、推断古气候和古环境最广泛应用的基础数据之一[5-11]。Román-Sánchez *et al*.[12]甚至将粒度分布视为物源区母岩风化序列的“指纹”。

为了挖掘粒度分布中的沉积信息，早期形成了众多经典的粒度分布沉积学分析方法，诸如ϕ刻度、粒级划分标准、粒度参数计算方法和用于沉积环境分析的 Hjulstrom 图解[13]、Sahu 粒度判别函数[14]、CM 图版[8]、粒度指数[15]和概率累积曲线[16]等等。其中一些传统定性和半定量的分析方法目前仍然是沉积学相关研究的常用手段。后来，随着数学方法的不断完善，以粒度分布数据或参数作为变量的聚类（clustering）[17]、多重分形（multifractal）[18]、因子分析（factor analysis）和主成分分析（principal component analysis, PCA）[19]和趋势分析

**收稿日期**：2023-09-12；**收修改稿日期**：2023-11-14
**基金项目**：国家自然科学基金项目（42202113）[**Foundation:** National Natural Science Foundation of China, No. 42202113]

（sediment trend analysis, STA）[20]等方法有效解决了沉积学研究中遇到的困难。目前，沉积学家普遍认为，沉积物由来自不同沉积过程的次总体混合而成，粒度分布次总体具有更深层次沉积意义[21-24]。为了从粒度分布数据中分离、提取次总体，基于概率密度函数的单个粒度分布分解[23-25]及基于端元模型的粒度分布数据集分解获得了成功[26-27]。

筛析法、沉降法、场干扰法和图像法等碎屑沉积物粒度分析实验方法和设备日趋完善，粒度分布数据易获取、易整理、易存储，诸多地球科学研究单位和油田企业积累了海量的粒度分布数据。受研究方法及数据共享的限制，这些海量数据所包含的沉积属性并未得到充分挖掘。并且没有形成可开放获取的、包含粒度分布相关信息的数据库，难以与其他地球科学资料相互支撑使用。作为“第四科研范式”的大数据技术已经深入地球科学领域[28]。在大数据的背景下，机器学习（machine learning）、深度学习（deep learning）和人工智能（artificial intelligence）等方法将为粒度分布的沉积信息挖掘提供创新技术，激活粒度分布大数据的潜在科学价值[29-30]。尽管粒度分布的沉积学分析方法难以逐一论述、已发表文献无法全部统计，但是本文在介绍经典的粒度分布沉积学解释方法基础上，分析了聚类和多重分形技术在粒度分布研究中的基本原理和应用现状，阐述了基于正态（normal）[31]、偏正态（skew normal）[25]和威布尔（Weibull）[23]概率密度函数的单个粒度分布分解（single-sample unmixing, SSU）以及基于端元模型算法（end-member modelling algorithm, EMMA）的粒度分布数据集分解的基本方法及实例，最终展望了以粒度分布沉积信息智能挖掘及大数据库建设作为重点的发展方向和趋势。

## 1 传统沉积学解释方法

在已公开文献中，沉积学和地质学领域有关沉积物粒度分布解释的报道最早可以追溯到1875 年[5]。在随后的一个多世纪里，众多学者围绕粒度分布的沉积学分析进行了大量研究，包括沉积物颗粒大小划分标准、粒度分布参数定量化和沉积环境判别等等，一些经典的研究手段一直被沿用至今。

### 1.1 粒级划分标准

Krumbein[7]创造性地将粒度分布的自然刻度转换为以 2 为底的对数ϕ刻度（$\phi=-\log_2 d$，$d$ 为颗粒直径，单位 mm），成为粒度分布常用的另一种刻度形式。Udden[6]提出了以 $2^{1/2}$ 为间隔的粒级划分，并于 1914 年改进了划分标准和专业术语，把沉积物分为巨砾、砂砾、砂、粉砂和黏土 5 大类[32]。Wentworth[33]首次在砂岩中细分了极粗砂，用细砾、中砾和粗砾代替砂砾。参照 Udden 标准[6]，Lane[34]细分了 Wentworth 标准[33]中的砾石、粉砂和黏土，分别使用砂砾、中砾和巨砾划分 2 mm 以上粒级。Friedman *et al.* [35]将 Wentworth 标准[33]中的中砾

细分为 4 小类、粗砾细分为 2 小类、巨砾细分为 4 小类，使用极细中砾代替细砾。Blott *et al.*[36-37]在前人的基础上，将最小粒径下限调整至 122 nm，形成了粒径从-12 ϕ（巨碎屑）至 13 ϕ（极细黏土）每间隔 1 ϕ共 27 个粒级的完备划分标准。此外，石油地质和农业土壤等学科研究者根据行业特征，也形成了一些适用于各个领域或研究区的粒级划分标准[38-39]，整个地球科学碎屑物质的粒级划分标准难以完全统一（表 1）。同时，Folk[40]根据 Wentworth 标准[33]砂砾－砂－黏土或者砂－粉砂－黏土的占比，采用三角图可视化显示沉积物颗粒大小的组合特征，其他粒级划分标准也制定了相应的岩性三角图方案，用来命名主要和次要岩石名称。

**表 1　一些经典的粒级划分标准**

**Table 1　Some classical classification standards of grain-size scale**

| 粒级 | | | 岩石名称 | | | | |
|---|---|---|---|---|---|---|---|
| ϕ刻度 | 自然刻度 | Udden[6,32] | Wentworth[33] | Lane[34] | Friedman *et al.*[35] | Blott *et al.*[36-37] | 张昌民等[39] |
| <-12 | >4096 | | | | 极大巨砾 | 巨碎屑 | |
| -12~-11 | 2048~4096 | | | 极大巨砾 | | | |
| -11~-10 | 1024~2048 | | 巨砾 | 大巨砾 | 大巨砾 | 极大巨砾 | 巨砾 |
| -10~-9 | 512~1024 | | | 中巨砾 | 中巨砾 | 大巨砾 | |
| -9~-8 | 256~512 | | | 小巨砾 | 小巨砾 | 中巨砾 | |
| -8~-7 | 128~256 | 大粗砾 | 粗砾 | 大粗砾 | 大粗砾 | 小巨砾 | |
| -7~-6 | 64~128 | 中粗砾 | | 小粗砾 | 小粗砾 | 极小巨砾 | 粗砾 |
| -6~-5 | 32~64 | 小粗砾 | | 极粗砂砾 | 极粗中砾 | 极粗砂砾 | |
| -5~-4 | 16~32 | 极小粗砾 | 中砾 | 粗砂砾 | 粗中砾 | 粗砂砾 | 大中砾 |
| -4~-3 | 8~16 | 极粗砂砾 | | 中砂砾 | 中中砾 | 中砂砾 | 小中砾 |
| -3~-2 | 4~8 | 粗砂砾 | | 细砂砾 | 细中砾 | 细砂砾 | 细砾 |
| -2~-1 | 2~4 | 砂砾 | 细砾 | 极细砂砾 | 极细中砾 | 极细砂砾 | |
| -1~0 | 1~2 | 细砂砾 | 极粗砂 | 极粗砂 | 极粗砂 | 极粗砂 | 极粗砂 |
| 0~1 | 0.5~1 | 粗砂 | 粗砂 | 粗砂 | 粗砂 | 粗砂 | 粗砂 |
| 1~2 | 250~500 | 中砂 | 中砂 | 中砂 | 中砂 | 中砂 | 中砂 |
| 2~3 | 125~250 | 细砂 | 细砂 | 细砂 | 细砂 | 细砂 | 细砂 |
| 3~4 | 62.50~125 | 极细砂 | 极细砂 | 极细砂 | 极细砂 | 极细砂 | |
| 4~5 | 31.25~62.50 | 粗粉砂 | 粉砂 | 粗粉砂 | 极粗粉砂 | 极粗粉砂 | 粉砂 |
| 5~6 | 15.63~31.25 | 中粉砂 | | 中粉砂 | 粗粉砂 | 粗粉砂 | 泥质粉砂 |
| 6~7 | 7.81~15.63 | 细粉砂 | | 细粉砂 | 中粉砂 | 中粉砂 | 粉砂质泥 |
| 7~8 | 3.91~7.81 | 极细粉砂 | | 极细粉砂 | 细粉砂 | 细粉砂 | |
| 8~9 | 1.95~3.91 | 粗黏土 | 黏土 | 粗黏土 | 极细粉砂 | 极细粉砂 | 黏土 |
| 9~10 | 977~1953 | 中黏土 | | 中黏土 | 黏土 | 极粗黏土 | |
| 10~11 | 488~977 | 细黏土 | | 细黏土 | | 粗黏土 | |
| 11~12 | 244~488 | | | 极细黏土 | | 中黏土 | |
| 12~13 | 122~244 | | | | | 细黏土 | |
| >13 | <122 | | | | | 极细黏土 | |

注：表中岩石名称源自英文专业术语，名词类包括黏土－clay、粉砂－silt、砂－sand、砂砾－gravel、细砾－granule、中砾－pebble、粗砾－cobble、巨砾－boulder 或 bowlder、巨碎屑－megaclasts，形容词类包括极细－very fine、细－fine、中－medium、粗－coarse、极粗－very coarse、极小－very small、小－small、大－large、极大－very large。

## 1.2 粒度参数计算

为了定量化表达粒度分布，粒度均值（粒度整体大小）、分选性（粒度的分散程度）、偏度（频率曲线的不对称程度）和峰度（相对粒度均值的集中程度）是常用的 4 个参数。Krumbein *et al.*[41]提出了考虑整个颗粒组分的粒度参数矩法（moment method）计算公式，并产生了变形计算方法[1]。但是，矩法公式计算结果受粒度分布“尾部”影响较大，Inman[42]将统计学中常用的矩量近似图形类比法引入到粒度分布的累积频率曲线中，计算了粒度分布的 2 个偏度参数，其中第二个偏度对粒度分布“尾部”的偏斜性敏感。Folk *et al.* [43]对Inman[42]的方法进行了修改和完善，提出了著名的 Folk-Ward 图解法（graphical method），成为了另外一种广泛使用的粒度分布参数计算方法。矩法和图解法分别建立了粒度分布频率曲线形态描述标准（表 2）。两种方法计算的粒度均值和分选系数大小基本相同，但是偏度与峰度大小差异较大[44]。

**表 2　粒度分布矩法与图解法参数计算方法及其频率曲线形态描述标准**

**Table 2　Moment and graphical methods of GSD parameters and morphological description standards of frequency curve**

| 粒度参数及等级 | | Krumbein *et al.* [41] | | Folk *et al.* [43] | |
|---|---|---|---|---|---|
| 参数 | 等级 | 计算公式 | 划分标准 | 计算公式 | 划分标准 |
| 均值 | | $\bar{x}_\phi = \frac{\sum f m_\phi}{100}$ | | $M_z = \frac{\Phi_{16} + \Phi_{50} + \Phi_{84}}{3}$ | |
| 分选性 | 很好 | $\sigma_\phi = \sqrt{\frac{\sum f(m_\phi - \bar{x}_\phi)^2}{100}}$ | <0.35 | $\sigma_I = \frac{\Phi_{84} - \Phi_{16}}{4} + \frac{\Phi_{95} - \Phi_5}{6.6}$ | <0.35 |
| | 好 | | 0.35~0.50 | | 0.35~0.50 |
| | 中等—好 | | 0.50~0.70 | | 0.50~0.70 |
| | 中等 | | 0.70~1.00 | | 0.70~1.00 |
| | 差 | | 1.00~2.00 | | 1.00~2.00 |
| | 很差 | | 2.00~4.00 | | 2.00~4.00 |
| | 极差 | | >4.00 | | >4.00 |
| 偏度 | 极细偏 | $SK_\phi = \frac{\sum f(m_\phi - \bar{x}_\phi)^3}{100\sigma_\phi^3}$ | >1.30 | $Sk_I = \frac{\Phi_{16} + \Phi_{84} - 2\Phi_{50}}{2(\Phi_{84} - \Phi_{16})} + \frac{\Phi_5 + \Phi_{95} - 2\Phi_{50}}{2(\Phi_{95} - \Phi_5)}$ | 0.3~1.0 |
| | 细偏 | | 0.43~1.30 | | 0.1~0.3 |
| | 对称的 | | -0.43~0.43 | | -0.1~0.1 |
| | 粗偏 | | -1.30~-0.43 | | -0.3~-0.1 |
| | 极粗偏 | | <-1.30 | | -0.3~-1.0 |
| 峰度 | 极低峰 | $K_\phi = \frac{\sum f(m_\phi - \bar{x}_\phi)^4}{100\sigma_\phi^4}$ | <1.70 | $K_G = \frac{\Phi_{95} - \Phi_5}{2.44(\Phi_{75} - \Phi_{25})}$ | < 0.67 |
| | 低峰 | | 1.70~2.55 | | 0.67~0.90 |
| | 中峰 | | 2.55~3.70 | | 0.90~1.11 |
| | 尖峰 | | 3.70~7.40 | | 1.11~1.50 |
| | 极尖峰 | | >7.40 | | 1.50~3.00 |
| | 异常尖峰 | | | | > 3.00 |

注：$f$为粒度分布频率百分数；$m_\phi$为刻度中间值；$\Phi_5$, …, $\Phi_{95}$为累积频率曲线上取 5%, …, 95%时对应的粒径ϕ值。

## 1.3 沉积环境分析

为了利用粒度分布数据分析沉积意义，早期形成了众多经典定性、半定量方法。Hjulstrom[13]根据河流不同搬运方式下颗粒所需的流速，提出了碎屑颗粒的侵蚀、搬运、沉

积与水流速度关系的 Hjulstrom 图解。Friedman[45]利用粒度分布参数的散点交会图，区分了沙丘、海滩和河流沉积环境。Sahu[14]分析了世界各地大量沉积物粒度分布参数后，建立了 Sahu 粒度判别函数和成因图解，确定了风成沙丘、海滩、河流（三角洲）与浊流的定量判别标准。Passega[8]提出了著名的 CM 图版（C 为粒度分布累积频率为 1%对应的粒径，M 为粒度中值），依据粒度分布 C 和 M 值在图版中的集中分区性来判断沉积环境、搬运方式和水动力条件。Doeglas[15]使用粒度分布四分位数和尾部情况制定了粒度指数，用于区分沉积环境。Visher[16]将粒度分布累积频率数据在对数概率图纸上绘制概率累积曲线，依据曲线形态的线性分段，划分滚动、跳跃和悬浮次总体，并总结了河流、天然堤、浅滩、波浪改造和沙丘等沉积环境下牵引流、冲流和回流、波浪、潮汐水道、悬浮沉降以及浊流等沉积过程次总体的组合特征。诸多传统沉积学解释方法拥有各自的使用范围和不足（表 3），其中 CM 图版和概率累积曲线现今仍是常用的两种粒度分布沉积分析手段。

**表 3　粒度分布沉积环境分析传统方法适用性及注意事项**

**Table 3　Applicability of, and attention to sedimentary environment analysis traditional methods for GSD**

| 研究方法 | 适用性 | 注意事项 |
|---|---|---|
| Hjulstrom 图解[13] | 解释颗粒大小、水流速度与侵蚀、搬运和沉积的关系 | 仅适用于河流沉积环境 |
| 粒度参数散点图[45] | 根据粒度分布参数散点图区分不同沉积环境 | 需要大量沉积环境已知的样本人工建立分类线，分类线可能随着地区的不同而变化 |
| Sahu 判别函数[14] | 建立不同沉积环境粒度分布参数的线性判别函数 | 需要大量沉积环境已知的样本建立判别函数，判别函数系数和环境判别标准的适用性难以验证 |
| CM 图版[8] | 根据粒度分布 C 与 M 值的集中区域解释搬运方式、识别沉积环境 | 主要适用于牵引流和浊流环境中，定量性不足 |
| 粒度指数[15] | 依据粒度分布四分位数以及 1%和 99%分位数建立沉积环境分类表 | 分类命名较复杂，指数分类表适用性有待验证 |
| 概率累积曲线[16] | 基于单个粒度分布的概率累积曲线形态划分滚动、跳跃和悬浮次总体 | 无法完全分离滚动、跳跃和悬浮次总体对应的粒度组分 |

## 2　基于粒度分布整体的挖掘方法

粒度分布频率数据为沉积物中不同直径颗粒所占的体积或重量百分比。粒度分布数据结构为具有高度一致性的一维向量 $\boldsymbol{X}=(x_1, x_2, \ldots, x_n)$（$n$ 为粒度分布刻度个数）：（1）同一批次粒度分析实验获得的粒度分布刻度及其间隔一般相同；（2）若粒度分布为频率百分数，则 $\sum_{k=1}^{n} x_k = 100\%$；（3）若粒度分布为累积频率百分数，则 $x_n$=100%、$\sum_{k=1}^{n-1}(x_{k+1}-x_k) = 100\%$。数学和统计学新方法的完善为粒度分布的沉积学分析带来了新手段，包括基于粒度分布相似性的聚类、基于粒度分布频率数据结构的多重分形、基于降维思想的因子分析和主成分分析以及基于粒度分布空间参数变化的沉积趋势分析等等。其中因子和主成分分析从高维粒度分布数据中提取的第一主成分具有较好的沉积环境指示意义，但是无法建立其他主成分与沉积因素的确定性[19,46]；沉积趋势分析基于粒度分布参数，采用统计方法刻画不同时间或空间下粒度分布的差异性，一般需要的数据量较大，且没有固定的统计学方法[20,47-48]。以下主要介绍聚类和多重分形在粒度分布沉积学分析中的基本原理和应用。

### 2.1　聚类

聚类分析是常应用于模式识别（pattern recognition）、数据挖掘（data mining）和机器学习的一种非监督自动分类方法。基于样本之间的相似性，聚类分析在样本数据集内寻找若干个族类，同一族内样本之间相似、不同族类内样本之间相异。聚类分析的输入数据集可以是粒度分布频率或累积频率数据，也可以是从粒度分布中提取的参数，使用较灵活性。若在 $M$ 个粒度分布的聚类分析中，第 $i$ 个粒度分布的频率或者累积频率记为特征向量 $\boldsymbol{X}_i=(x_{i,1}, x_{i,2}, \ldots, x_{i,n})$，则两个粒度分布 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 之间的相似性 $d_{i,j}$ 常用闵可夫斯基距离度量：

$$d_{i,j}=\left[\sum_{k=1}^{n}\left(x_{i,k}-x_{j,k}\right)^{h}\right]^{1/h}$$

特别地，当 $h$=2 时，$d_{i,j}$ 为欧几里得距离；当 $h$=1 时，$d_{i,j}$ 为曼哈顿距离。$d_{i,j}$ 越小，粒度分布越相似，反之粒度分布越相异。

聚类的分类思想与粒度分布的沉积环境判别恰好吻合：相同沉积环境内的沉积物由相近的粒度组分组成，因此具有相似的粒度分布特征。随着聚类算法的不断丰富与改进，许多聚类分析方法已成功应用于粒度分布的沉积环境判别。李玉中等[49]根据崎岖列岛海区现代沉积物的分选系数、砂粒含量和黏粒含量 3 个参数，采用系统聚类（system clustering）分析方法划分现代沉积环境。Nelsonet *et al*.[17]从水槽模拟河床表面高清照片中提取沉积物粒度分布，对比分析了 *k*-均值聚类（*k*-means clustering）、凝聚聚类（agglomerative clustering）、谱聚类（spectral clustering）和模糊 *c*-均值聚类（fuzzy *c*-means clustering）的优缺点，认为粒度分布累积频率数据的聚类优于频率数据聚类。Ordóñez *et al*.[50]首次引入功能聚类（functional clustering）方法对西班牙阿维莱斯海湾内不同沉积环境的粒度分布进行聚类，与 *k*-均值聚类相比，功能聚类可在粒度分布频率数据内获得更好的聚类效果，且得到的族类分别与河流、海滩、风成沙丘和河湾沉积环境具有较强的内在关联。章婷曦等[51]对太湖西北部表层沉积物粒度 7 个参数进行 R 型聚类（R-type clustering），区分 4 种沉积环境。刘祥奇等[52]引用邻近传播聚类（affinity propagation clustering）分析了白洋淀地区露头剖面的粒度分布频率数据，从得到的 11 个簇类中总结了湖沼相、湖相、河流相以及洪积相的粒度分布特征。

在众多聚类方法中，系统聚类属于层次聚类（hierarchical clustering）的一种，其试图在不同层次对数据集进行划分，生成一系列嵌套的聚类树来完成聚类，单点聚类处在树的最底层，树顶层的根节点聚类覆盖了全部数据点，因此系统聚类在高维数据可视化聚类分析中应用较多。例如，针对来自鄱阳湖流域康山河江心洲 15 份河道和 12 份河漫滩沉积物粒度分布频率数据（图 1a），采用系统聚类方法，利用欧几里得距离计算相似性，可自动将数据集划分为两类，对应了两种不同的沉积环境（图 1b）。

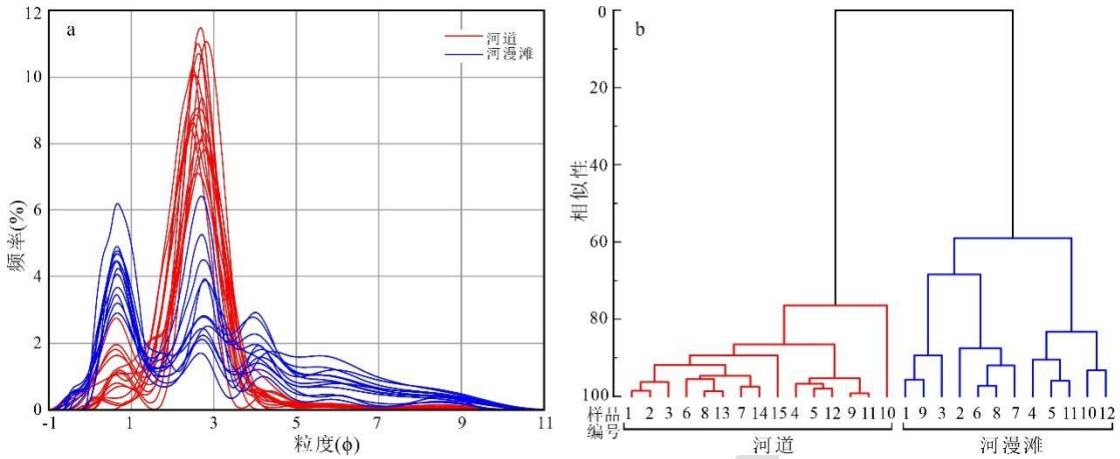


图 1　康山河江心洲 27 份沉积物ϕ刻度粒度分布频率数据及其系统聚类结果

（a）频率曲线；（b）系统聚类结果

Fig.1　Frequency data in ϕ-scale and system clustering for 27 GSDs for sediments from the central bar of the Kangshan River

(a) frequency curves; (b) system clustering result

## 2.2　多重分形

多重分形技术用于精细刻画一维和二维数据的复杂性和不均匀性，在信号处理、图像识别和化学等领域应用广泛。粒度分布的多重分形中，需要将频率数据的刻度规整到无量纲区间 $J$=[0, 5]上，运用二进制等分法将区间 $J$ 划分为 $N(\varepsilon)=2^k$ 个等距($\varepsilon=5\times2^{-k}$, $k$=1, 2, …, 6)子间距 $J_i$。子区间 $J_i$ 中粒径的概率密度用 $p_i(\varepsilon)$表示，利用 $p_i(\varepsilon)$构造配分函数族[18,53-56]：

$$u_i(q,\varepsilon)=\frac{p_i(\varepsilon)^q}{\sum_{i=1}^{N(\varepsilon)}p_i(\varepsilon)^q}$$

其中，$u_i(q, \varepsilon)$为第 $i$ 个子区间 $J_i$ 的 $q$ 阶概率，$q$ 为[-10, 10]区间内的整数。广义维数谱、奇异性指数和谱函数是多重分形理论中重要的 3 个参数。多重分形广义维数谱 $D(q)$从整体上体现多重分形的特征，其定义为[18,56]：

$$D(q)=\frac{1}{q-1}\lim_{\varepsilon\to0}\frac{\log\left(\sum_{i=1}^{N(\varepsilon)}p_i(\varepsilon)^q\right)}{\log\varepsilon},(q\neq1)$$

$$D_1=\lim_{\varepsilon\to0}\frac{\sum_{i=1}^{N(\varepsilon)}p_i(\varepsilon)^q\log p_i(\varepsilon)}{\log\varepsilon},(q=1)$$

其中，当 $q$<0 时，低密度或小概率分布信息被放大；当 $q$>0 时，高密度或大概率分布信息被放大。当 $q$=0 时，$D_0$ 为粒度分布的容量维数，与粒径的范围成正比；当 $q$=1 时，$D_1$ 为粒度分布的信息熵维数，与颗粒大小的密集度成反比；当 $q$=2 时，$D_2$ 为粒度分布的关联维数，与刻度间隔之间的频率均匀程度成正比[18,56]（图 2a）。

多重分形奇异性指数$\alpha(q)$与谱函数 $f(\alpha(q))$反映了多重分形的局部分维特征，其定义分别为[18,56]：

$$\alpha(q)=\lim_{\varepsilon\to 0}\frac{\sum_{i=1}^{N(\varepsilon)}u_i(q,\varepsilon)\log p_i(\varepsilon)}{\log\varepsilon}$$

$$f(\alpha(q))=\lim_{\varepsilon\to 0}\frac{\sum_{i=1}^{N(\varepsilon)}u_i(q,\varepsilon)\log u_i(q,\varepsilon)}{\log\varepsilon}$$

#

当 $q$=0 时，$\alpha(0)$为奇异谱函数的均值，与分形结构上局部密集程度成反比。$\Delta\alpha=\alpha(q)_{max}-\alpha(q)_{min}$ 定义为多重分形奇异谱的谱宽，表征粒度分布分形结构上的差异性与不均匀性，$\Delta\alpha$越大，粒度分布越不均匀；反之，粒度分布越均匀。$\Delta f=f(\alpha(q)_{max})-f(\alpha(q)_{min})$代表了多重分形谱的形状特征，当$\Delta f<0$ 时，谱函数呈左钩状，指示粒度分布低频率组分在分形系统中占主导；当 $\Delta f>0$ 时，谱函数呈右钩状，指示粒度分布高频率组分在分形系统中占主导[18,53-56]（图 2b）。

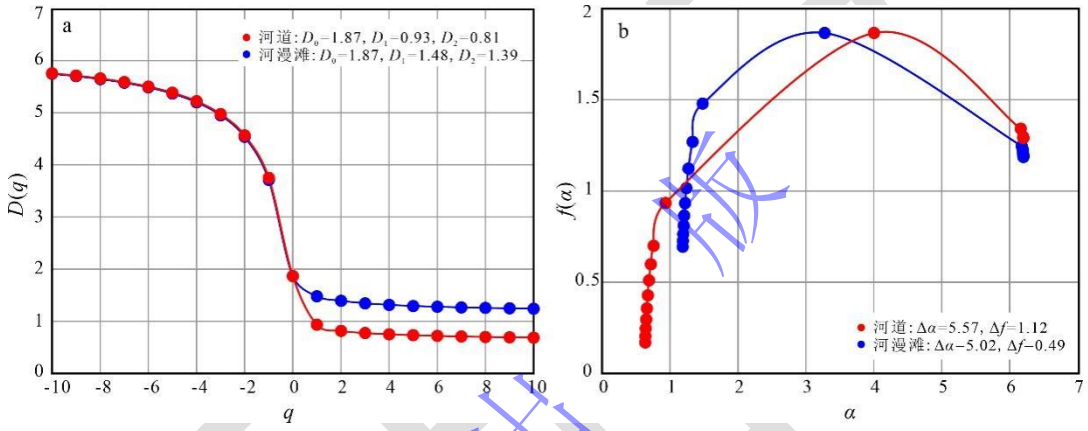


图 2　图 1a 中不同沉积环境中 2 个沉积物粒度分布频率数据多重分形结果

（a）广义维数谱；（b）多重分形奇异谱

Fig.2　Multifractal result of 2 GSDs from different depositional environments in Fig.1a

(a) generalized dimension spectra; (b) multifractal singularity spectra

Grout *et al.*[53]最早利用多重分形谱函数表示粒度分布频率数据的分形结构特征，有效描述了不同沉积环境中沉积物粒度分布的差异性和不均匀性。Miranda *et al.*[54]利用多重分形表征了巴西巴拉那河流域表层土壤再搬运后的腐泥粒度分布特征。常宏等[55]分析了黄河乌兰布和沙漠段两岸地表沉积物粒度分布的多重分形特征，证明了东岸地表沉积物粒度局部叠加强、沉积过程复杂、沉积环境多样。Qiao *et al.*[18]在研究黄土高原渗流带（地表以下 50~200 m）沉积物粒度分布的多重分形时，发现黄土粒度分布的分散程度、局部多变性和奇异谱在时间和空间上存在明显差异。Li *et al.*[56]解剖了三峡水库悬浮沉积物的粒度分布分形特征，证实了相比雨季的悬浮沉积物，旱季悬浮沉积物粒度分布分形结构更复杂、局部密集性和多变性更高、粒度非均质性更强。

## 3　基于粒度分布分解的挖掘方法

碎屑沉积物是在复杂的沉积环境中，由多种水动力机制共同作用不断搬运堆积形成的。

受不同水动力机制驱动的颗粒组成了粒度分布中不同的次总体[21-25]。多个次总体相互叠加组成的双峰或者多峰粒度分布频率曲线反映了沉积环境的总体特征[57]。这些粒度分布次总体是沉积物的最小结构单元，具有沉积“基因”的意义[3,27]。从粒度分布中分离次总体、探讨各组分的沉积学意义是粒度分布沉积环境研究的另一种思路，主要分为两种手段：（1）单个粒度分布的概率密度函数（probability density function, PDF）拟合法，通过曲线拟合技术（curve-fitting techniques, CFT）将单个粒度分布频率数据分解为若干个连续的单峰分布[23-25]；（2）粒度分布数据集的端元模型算法，通过主成分分析、因子旋转、非负最小二乘法等运算将批量粒度分布频率数据集分解为对应于不同沉积动力的若干个单峰或多峰端元[26-27,58-59]。

### 3.1　基于概率密度函数的单个粒度分布分解

基于概率密度函数的单个粒度分布分解方法认为单一沉积动力所搬运的粒度组分在频率上服从统计学某种概率密度函数，粒度分布本质上是由多个次总体凸组合而成的混合分布（mixture distribution），符合统计学有限混合模型（finite mixture model, FMM）[23-25]：

$$f = \sum_{i=1}^{m} c_i p_i$$

其中，$m$ 为次总体的个数；$p_i$ 为第 $i$ 个次总体服从的概率密度函数；$c_i$ 为第 $i$ 个次总体在粒度分布中的所占百分比，$0 \leqslant c_i \leqslant 1$ 且 $\sum c_i = 100\%$；$f$ 为沉积物粒度分布频率拟合值。最终采用曲线拟合最优化求解得出概率密度函数及 $c_i$，进而计算每个次总体的均值$\mu$、方差$\sigma$、峰度 $Sk$ 和偏度 $K_G$。在统计学中，随机变量服从的概率密度函数有十多种类型，粒度分布次总体分解中常用的分布模型包括正态分布（ϕ刻度下也称为对数正态分布）[22,31]、偏正态[24-25]和威布尔分布[23,60-62]。

1）正态分布 $N(\mu, \sigma^2)$

正态分布是最常见的统计学概率密度函数。若粒度分布中每个次总体服从正态分布，则次总体的概率密度函数为：

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

式中：$x$ 为粒径大小，单位ϕ；$\mu$为次总体均值，单位ϕ；$\sigma$ 为次总体标准差；每个次总体偏度为 0，峰度为 3。最优化求解可得出每个次总体在正态混合模型中的参数（$c_i, \mu_i, \sigma_i^2$）（图 3a）。

Spencer[31]认为所有的碎屑沉积物粒度分布是 3 个或 3 个以下服从正态分布的次总体组合，这 3 个正态分布次总体分别对应泥、砂、砾 3 种颗粒组分。Clark[63]利用正态分布分解

了粒度分布频率数据，并讨论了该方法的实用性和适用性。Ashley[64]利用正态分布从加拿大皮特湖流域 190 个沉积物粒度分布中分离了细砂、粉砂和黏土 3 个次总体。Xiao *et al*.[22]利用正态分布从呼伦湖现代沉积物粒度分布中分离出 6 个次总体，分别对应了不同的沉积环境和搬运方式。然而，沉积物的粒度分布次总体是受多种因素控制的离散随机分布，其频率曲线不一定完全服从正态分布，每个次总体也可能存在偏度和峰度[24-25]。

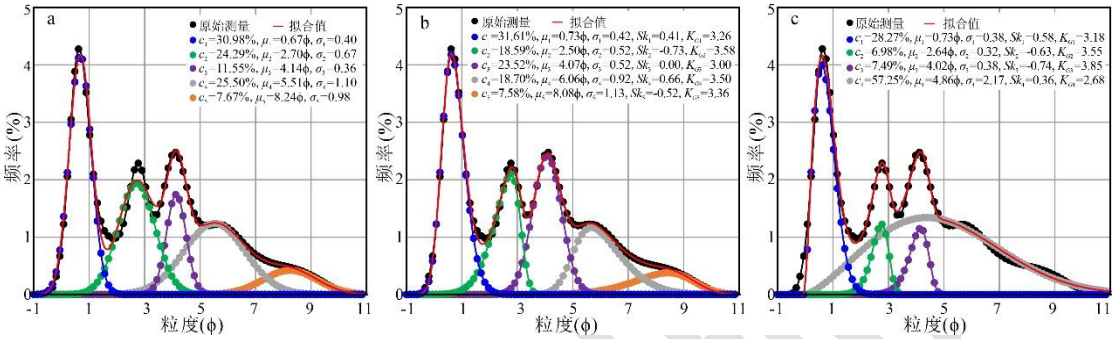


图 3　利用 QGrain[19]不同概率密度函数分解多峰粒度分布频率数据

（a）正态分布最优分解成 5 个次总体；（b）偏正态分布最优分解成 5 个次总体；（c）威布尔分布最优分解成 4 个次总体

Fig.3　Unmixing of GSD multimodal frequency curve by different probability density functions in QGrain[19]

(a) 5 subpopulations optimally unmixed by normal distribution; (b) 5 subpopulations optimally unmixed by skew normal distribution; (c) 4 subpopulations optimally unmixed by Weibull distribution

2）偏正态分布 $SN(\mu, \sigma^2, \lambda)$

Azzalini[65]在正态分布中添加了偏度参数λ（又称形状参数）形成了偏正态分布，可便捷计算偏度与峰度。若粒度分布中每个次总体服从偏正态分布，则次总体的概率密度函数为：

$$p(x)=\frac{1}{\pi\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\int_{-\infty}^{\lambda\frac{x-\mu}{\sigma}}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t$$

其中，$\mu$为位置参数，$\sigma\geqslant 0$为尺度参数，分别与正态分布均值和标准差含义一致。最优化求解可得出每个次总体在偏正态混合模型中的参数（$c_i, \mu_i, \sigma_i^2, \lambda_i$），进一步计算次总体均值、方差、峰度和偏度（图 3b)。

Gan *et al*.[25]首次利用偏正态分布分解了加纳博苏姆推湖 530 个粒度分布，分析了次总体参数与粒度分布参数和沉积年代的相互关系，认为与正态分布相比，偏正态分布的分解结果误差更小，携带偏度和峰度的次总体能更加定量化评价沉积属性。袁瑞等[24]从鄱阳湖流域不同沉积环境中 214 个粒度分布数据中分离得到了 977 个服从偏正态分布的次总体，统计了各个次总体的参数，分析了不同沉积环境中次总体的组合特征，认为结合粒度分布概率累积曲线和分离的次总体，可以定量计算不同搬运方式颗粒的所占百分比和确定相应沉积过程个数，为沉积过程的定量化研究提供参考。

3）威布尔分布 $W(\alpha, \beta)$

威布尔分布是产品可靠性分析和寿命检验最常用的分布模型，其概率密度函数曲线形状

可以右偏、左偏或者对称，具有较强的灵活性。若粒度分布中每个次总体服从威布尔分布，则次总体的概率密度函数为：

$$p(x)=\frac{\alpha}{\beta^{\alpha}}x^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$$

其中，$\alpha$为形状参数，决定概率密度曲线的形态（左偏、右偏或对称）；$\beta$为尺度参数，控制概率密度曲线众数的位置。最优化求解可得出每个次总体在威布尔混合模型中的参数（$c_i$, $\alpha_i$, $\beta_i$）以及次总体统计参数（图 3c）。

Carder *et al*.[60]证明威布尔分布可以更好地拟合太平洋赤道附近粒径大于 2.22 μm 的沉积物粒度分布。Kondolf *et al*.[61]对比分析了砾石河流中威布尔和正态分布拟合粒度分布的优缺点。孙东怀等[66]利用威布尔分布从呈双峰特征的黄土粒度分布频率数据中分离得到 2 个粒度组分，分别指示了黄土高原两种粉尘来源和沉积环境特征；他们也利用威布尔和正态分布对比分解了不同沉积环境中的多峰粒度分布，认为正态分布适用于河流和湖泊沉积环境，而威布尔分布适合研究河流、湖泊、风成沙丘、黄土和深海等多种环境中的沉积物[67-68]。Peng *et al*.[62]提出了一种基于威布尔分布更加高效和灵活的粒度分布分解数值方法，并结合 *k*-均值聚类指示了不同族类的沉积环境意义。

基于以上概率分布模型，国内外学者编制形成了一些操作简单、方便快捷的开源软件，例如基于偏正态分布的 SNDD[25]、基于威布尔分布的 CFLab[23]和多种分布模型可选的 QGrain[19]等。此外，概率密度函数 Rosin 分布[69]、对数双曲（log-hyperbolic）分布[47,70]、对数偏拉普拉斯（log-skew-laplace）分布[70]和伽马（gamma）分布[21]等也被应用于粒度分布频率数据的拟合，但是这些概率密度函数需要更多的参数，增加了粒度分布频率数据分解的难度。

### 3.2　基于端元模型的粒度分布数据集分解

Weltje[58]认为仅仅从单个粒度分布中分解的次总体及其参数在沉积属性解释中存在一定的不足，而相同或相似沉积环境中蕴含与沉积过程相关的若干个粒度区间，这些粒度区间可以用相互独立的固定端元表示[3]，最终形成了粒度分布端元模型算法。随着该方法的大量使用，涌现了许多改进的算法及对应的软件工具。Seidel *et al*.[71]免费公开了早期粒度分布端元模型分析算法的R语言函数包RECA（R-based end-member composition algorithm），提取了地中海钻孔岩心粒度分布的3个端元。Dietze *et al*.[72]提出了一种基于特征空间分析并考虑内在不确定性的柔性端元模型算法，并利用R语言开发了开源工具包EMMAgeo[73]，得到了青海冬给错纳湖表层沉积物的5个粒度端元。Paterson *et al*.[26]在原始端元模型算法中引入非负矩阵克服了最优化求解和耗时问题，形成了可为单峰或多峰粒度端元的分解方法，最终在

MATLAB平台下编制了易于操作的开源工具包AnalySize。Yu *et al*.[74]在贝叶斯框架下建立了分层贝叶斯端元模型分析（hierarchical Bayesian end-member modeling analysis, BEMMA）方法，分解了青海高台湖钻孔岩心粒度分布的3个端元。Zhang *et al*.[10]利用遗传算法模拟自然求解过程寻找基本端元，建立了基本端元模型算法（basic end-member model algorithm, BasEMMA）和确定最合适端元个数的方法，分离了东海多个钻孔岩心粒度分布的3个端元。Liu *et al*.[19]提出了基于神经网络的端元模型分析（neural network based end-member modeling analysis, NNEMMA）算法，利用Python语言开发了相应的开源软件QGrain；他们尝试综合单个粒度分布分解和粒度分布数据集端元模型分析的优点，提出了两者相结合的通用分解模型（universal decomposition model, UDM）[11]，解析了渭河盆地浅层钻孔黄土粒度分布的端元。

不同端元模型算法在应用过程中整体效果较好，但也存在各自的不足。例如，RECA在随机性较高的粒度分布中拟合误差较大；EMMAgeo能准确模拟粒度分布频率曲线主峰的位置，次峰拟合效果不佳；BEMMA在粒度分布频率曲线较复杂时拟合的端元组分不足[27]。相较于其他粒度分布端元模型分析算法，AnalySize具有较高的适应性和准确性[27]、容易确定最优端元个数和灵活设置端元曲线形态，是使用较多的粒度分布端元模型分析工具包之一。AnalySize要求输入至少10个粒度分布频率数据，自动模拟1~10个端元的结果。在确定最优端元个数时，需要综合考虑较大线性相关性（$R^2$）、较小端元相关度和较小角度偏差的原则。线性相关性是指粒度分布测量值与拟合值的相关性，线性相关性越高，端元拟合越好，一般要求大于0.9（图4a）。端元相关度用于衡量各个分离端元之间的相互独立性，端元相关度越小，各个端元之间重合区间越小（图4a）。角度偏差是端元在拟合粒度频率曲线时产生的形状偏差，偏差值越小，形状拟合越好，一般要求小于5°[26,75-76]（图4b）。AnalySize可利用非参数法和参数法分别实现多峰和单峰端元分析，提供丰富的端元形态（图4c，d）。

尽管利用不同的端元模型算法从不同沉积环境的粒度分布数据集中最终得到的端元个数不同，但是多数文献中3个粒度端元是最普遍的结果，尤其是沙漠和黄土等风成环境中[27]。河流、湖泊和冰川等环境沉积动力复杂，颗粒组合多样，粒度分布频率曲线呈多峰态，提取的粒度端元个数往往大于3个，但最多一般不超过6个[27]。基于端元模型分析的粒度分布分解更加适用于沉积环境相关联的大量样本，提取的粒度端元可提供更多的粒度组分细节，进而分析每个粒度端元对应的沉积意义，在河流、湖泊、海洋、沙漠和黄土等几乎所有碎屑沉积环境分析中得到了广泛应用[19,27,59,71-78]。

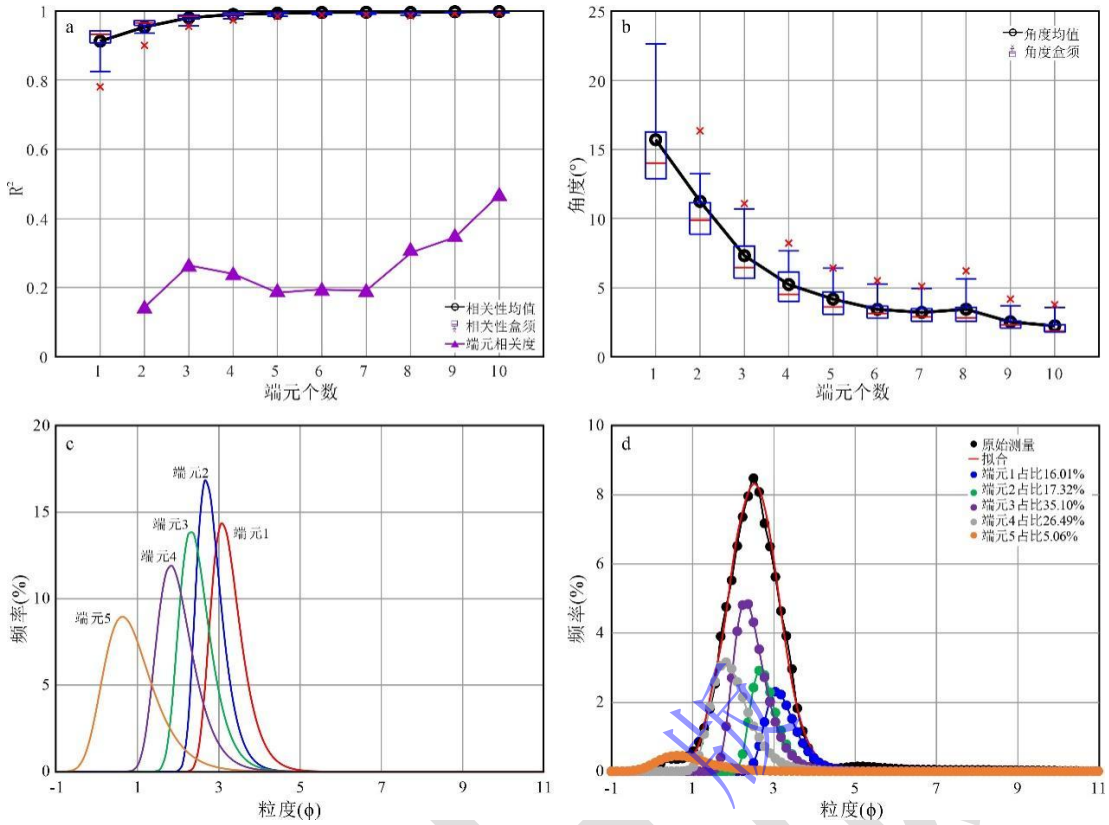


图 4　利用 AnalySize[26]参数法分析图 1a 中 15 份河道沉积物粒度端元

（a）线性相关性和端元相关度与端元个数的关系；（b）角度偏差与端元个数的关系；（c）优选的 5 个粒度端元；（d）1 个粒度分布分解成 5 个端元

Fig.4　GSD parametric end-member analysis result for 15 river channel samples in Fig. 1a using AnalySize[26]

(a) relationship between linear correlation, end-member correlation and end-member number; (b) relationship between angular difference and end-member number; (c) optimized 5 end-members; (d) 1 GSD unmixed to 5 end-members

## 4　大数据背景下的展望

当前，在大数据背景下，以数据密集型计算为基础的研究即将进入新的科研范式[79]，广泛依赖于各类地球数据的地球科学研究已经步入大数据时代[28]。粒度分布的沉积学研究已历经150年，尽管一些经典和非传统研究方法揭示了粒度分布中所蕴含的沉积信息，为现代和古代沉积环境分析和沉积过程重建提供了有力证据。为了顺应地质学大数据的前沿发展趋势，未来粒度分布的沉积学研究应重点开展沉积信息智能挖掘和大数据库构建两个方面。

### 4.1　沉积信息智能挖掘

粒度分布是原始沉积信息的载体，真实地记录了碎屑沉积物的形成演化过程[1-4]。沉积学期刊 *Sedimentary Geology* 于 2007 年出版了“From particle size to sediment dynamics”的专刊，Hartmann *et al.*[2]对粒度分布的沉积学研究提出了 8 点深层次问题：（1）从岩石到颗粒（物源问题）、（2）从源到汇（搬运问题）、（3）从野外到实验（取样分析问题）、（4）从实际到模拟（参数化问题）、（5）从点到面（空间趋势问题）、（6）从现今到以后（时

间趋势问题）、（7）从好到更好（实际验证问题）、（8）从一处到多处（适用性问题）。粒度分布的沉积学解释将更加注重“基于过程”（process-based）[2]。

在大数据的驱动下，基于机器学习、深度学习和人工智能等方法的地球科学大数据技术逐渐形成，已成功应用于油气勘探与开发、矿产资源定量评价和地质图件绘制等方面[29-30,80-81]。目前，尽管众多文献中形成了针对各自研究区沉积环境有效的粒度分布分析方法，但是非大量的数据体和非普适的研究方法仍然难以解决以上8个问题。以海量的碎屑沉积物粒度分布数据为对象，结合沉积学其他资料，利用大数据技术智能挖掘深度隐藏的粒度分布沉积学信息，包括监督学习粒度分布与沉积环境的耦合关系、构建不同沉积过程下粒度的组合模式、智能判别沉积环境、刻画沉积水动力条件、反演沉积动力学成因、表征沉积物粒度的时空分布模式等等，将成为粒度分布沉积学分析的必然发展趋势。

### 4.2 大数据库构建

地球表面约70%的面积被碎屑沉积物覆盖，全球积累的碎屑沉积物粒度分布数据已无法估量。综合大数据的常规特点，粒度分布大数据具有以下特征：（1）大量性（volume），沉积物样品易获取、粒度分布数据易整理，已积累了海量的数据体；（2）高速性（velocity），在标准的样品前处理和高性能的粒度分析测试仪器下，可快速获得可靠的粒度分布数据；（3）多样性（variety），粒度分布数据可来源于河流、湖泊、海洋、沙漠和黄土等多种沉积环境；（4）价值性（value），粒度分布蕴含了重要的沉积信息；（5）混合性（mixing），粒度分布是多种沉积作用叠加的结果；（6）因果性（causality），不同大小颗粒和水动力条件导致不同的粒度组合和粒度分布；（7）时空性（time-space），同一时间、不同位置的粒度差异表征了环境的不同，同一位置、不同时间的粒度变化刻画了气候和环境的变迁；（8）关联性（correlation），同一环境下的粒度分布数据具有一定的相似性；（9）差异性（difference），相同沉积体系内，不同相带上的粒度分布特征不同。

深时数字地球（deep-time digital earth, DDE）是由我国科学家发起的国际大科学计划，旨在整合地球演化全球数据、共享全球地学知识[82]。国外众多优秀的沉积学相关专业数据库已经兴起，例如全球沉积岩碎屑矿物年代学数据库GeoChron[83-84]、全球沉积物地球化学数据库EarthChem[84-85]、适用于深时研究的沉积学数据库Macrostrat[84,86]等等。然而，在大数据背景下，仍未见与沉积物粒度分布数据库相关的公开报道。根据粒度分布大数据的特点，未来可在Hadoop大数据生态系统下搭建开放、共享的粒度分布数据库，设计多个功能模块：（1）数据层，利用非关系型数据库和分布式文件系统存储粒度分布数据及其相关信息，包括取样位置、取样时间、深度、层位、样品岩性、沉积体系、照片和视频等等；（2）网络

层，通过互联网技术，注册用户可访问、浏览、绘图、下载和上传粒度分布相关信息；（3）方法层，嵌入粒度分布数据智能挖掘方法；（4）文献层，将文献导入数据库后，开发文本拾取技术自动获取发表时间、研究方法、取样时间、数据来源、粒度分布参数和沉积环境等信息，分门别类（图5）。搭建存储粒度分布相关资料、智能挖掘沉积信息的开放、共享大数据库，为未来的沉积物综合研究建立数据档案和存储库。若能融入到DDE计划，可与其他大数据库相互支撑，将极大推动地质学和沉积学大数据的发展与应用。

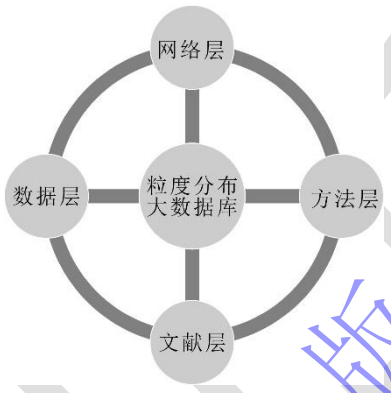


图5　粒度分布大数据库框架构想

Fig.5　Frame conception of big database for GSDs

## 5　结束语

（1）如果把沉积物类比为生物有机体、粒度分布次总体类比为沉积物的“基因”、则粒度分布次总体的组合模式类似于沉积物的“基因组”、不同沉积环境下沉积物的粒度分布次总体数据库类似于沉积物的“基因库”。获取高分辨率（纵向）和高覆盖率（横向）的粒度分布数据后，从粒度分布中分离、提取次总体对于理解沉积物搬运方式和沉积过程、自动定量解释沉积环境具有非常重要的意义。

（2）粒度分布的沉积学研究历程可总结为 4 个阶段：①起始阶段，19 世纪 70 年代—20 世纪 30 年代，粒度分布的沉积学价值被发现；②探索阶段，20 世纪 40 年代—20 世纪 80 年代，经典沉积学分析手段逐渐形成；③初期发展阶段，20 世纪 90 年代—21 世纪初，将粒度分布作为整体的半定量、定量研究方法逐渐发展；④快速发展阶段，进入 21 世纪以来，基于粒度分布次总体分解的定量分析方法逐渐成熟。

（3）在大数据背景下，粒度分布沉积学分析应重点开展 2 个方面的工作：①结合沉积学其他资料，建立粒度分布沉积信息深度挖掘的智能方法；②基于粒度分布大数据特点，搭建存储粒度分布相关资料、智能挖掘沉积信息的开放、共享大数据库。在不久的将来，粒度分布的沉积学研究将进入大数据技术阶段。

致谢　在文章撰写过程中得到了长江大学地球科学学院张昌民教授的大力指导和帮助；三位审稿专家和期刊编辑提出了针对性的修改意见，在此一并表达诚挚的谢忱。

## 参考文献（References）

[1] McManus J. Grain size determination and interpretation[M]//Tucker M. Techniques in sedimentology. Oxford: Backwell, 1988: 63-85.

[2] Hartmann D, Flemming B. From particle size to sediment dynamics: An introduction[J]. Sedimentary Geology, 2007, 202(3): 333-336.

[3] Weltje G J, Prins M A. Genetically meaningful decomposition of grain-size distributions[J]. Sedimentary Geology, 2007, 202(3): 409-424.

[4] Bright C, Mager S, Horton S. Response of nephelometric turbidity to hydrodynamic particle size of fine suspended sediment[J]. International Journal of Sediment Research, 2020, 35(5): 444-454.

[5] Sternberg H. Untersuchungen uber langen-und querprofil geschiebefuhrender flusse[J]. Zeitschrift fur Bauwesen, 1875, 25: 483-506.

[6] Udden J A. The mechanical composition of wind deposits[M]. Illinois: Augustana Lihrary Publications, 1898: 1-69.

[7] Krumbein W C. Size frequency distributions of sediments[J]. Journal of Sedimentary Petrology, 1934, 4(2): 65-77.

[8] Passega R. Grain size representation by CM patterns as a geological tool[J]. Journal of Sedimentary Petrology, 1964, 34(4): 830-847.

[9] IJmker J, Stauch G, Dietze E, et al. Characterisation of transport processes and sedimentary deposits by statistical end-member mixing analysis of terrestrial sediments in the Donggi Cona lake catchment, NE Tibetan Plateau[J]. Sedimentary Geology, 2012, 281: 166-179.

[10] Zhang X D, Wang H M, Xu S M, et al. A basic end-member model algorithm for grain-size data of marine sediments[J]. Estuarine, Coastal and Shelf Science, 2020, 236: 106656.

[11] Liu Y M, Wang T, Liu B, et al. Universal decomposition model: An efficient technique for palaeoenvironmental reconstruction from grain-size distribution[J]. Sedimentology, 2023, 70(7): 2127-2149.

[12] Román-Sánchez A, Temme A, Willgoose G, et al. The fingerprints of weathering: Grain size distribution changes along weathering sequences in different lithologies[J]. Geoderma, 2021, 383: 114753.

[13] Hjulstrom F. The load of the River Fyris in Central Sweden[J]. Bulletin Geology Institution University of Upsala, 1936, 25: 221-527.

[14] Sahu B K. Depositional mechanisms from the size analysis of clastic sediments[J]. Journal of Sedimentary Research, 1964, 34(1): 73-83.

[15] Doeglas D J. Grain-size indices, classification and environment[J]. Sedimentology, 1968, 10(2): 83-100.

[16] Visher G S. Grain size distributions and depositional processes[J]. Journal of Sedimentary Petrology, 1969, 39(3): 1074-1106.

[17] Nelson P A, Bellugi D, Dietrich W E. Delineation of river bed-surface patches by clustering high-resolution spatial grain size data[J]. Geomorphology, 2014, 205: 102-119.

[18] Qiao J B, Zhu Y J, Jia X X, et al. Multifractal characteristics of particle size distributions (50-200 m) in soils in the vadose zone on the Loess Plateau, China[J]. Soil and Tillage Research, 2021, 205: 104786.

[19] Liu Y M, Liu X X, Sun Y B. QGrain: An open-source and easy-to-use software for the comprehensive analysis of grain size distributions[J]. Sedimentary Geology, 2021, 423: 105980.

[20] McLaren P, Hill S H, Bowles D. Deriving transport pathways in a sediment trend analysis (STA)[J]. Sedimentary Geology, 2007, 202(3): 489-498.

[21] Risović D. Two-component model of sea particle size distribution[J]. Deep Sea Research Part I: Oceanographic Research Papers, 1993, 40(7): 1459-1473.

[22] Xiao J L, Chang Z G, Fan J W, et al. The link between grain-size components and depositional processes in a modern clastic lake[J]. Sedimentology, 2012, 59(3): 1050-1062.

[23] Wu L, Krijgsman W, Liu J, et al. CFLab: A MATLAB GUI program for decomposing sediment grain size distribution using Weibull

functions[J]. Sedimentary Geology, 2020, 398: 105590.

[24] 袁瑞，张昌民，赵芸，等. 基于偏正态概率分布的粒度分布次总体分离及其沉积环境指示意义[J]. 地质论评，2022，68（3）：1033-1047. [Yuan Rui, Zhang Changmin, Zhao Yun, et al. Decomposing subpopulations from grain-size distributions based on skew normal probability distribution and their significances for sedimentary environments[J]. Geological Review, 2022, 68(3): 1033-1047.]

[25] Gan S Q, Scholz C A. Skew normal distribution deconvolution of grain-size distribution and its application to 530 samples from Lake Bosumtwi, Ghana[J]. Journal of Sedimentary Research, 2017, 87(11): 1214-1225.

[26] Paterson G A, Heslop D. New methods for unmixing sediment grain size data[J]. Geochemistry, Geophysics, Geosystems, 2015, 16(12): 4494-4506.

[27] van Hateren J A, Prins M A, van Balen R T. On the genetically meaningful decomposition of grain-size distributions: A comparison of different end-member modelling algorithms[J]. Sedimentary Geology, 2018, 375: 49-71.

[28] 赵鹏大. 地质大数据特点及其合理开发利用[J]. 地学前缘，2019，26（4）：1-5. [Zhao Pengda. Characteristics and rational utilization of geological big data[J]. Earth Science Frontiers, 2019, 26(4): 1-5.]

[29] Karpatne A, Ebert-Uphoff I, Ravela S, et al. Machine learning for the geosciences: Challenges and opportunities[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(8): 1544-1554.

[30] MacLEOD N. Artificial intelligence & machine learning in the Earth sciences[J]. Acta Geologica Sinica (English Edition), 2019, 93(Supp.l): 48-51.

[31] Spencer D W. The interpretation of grain size distribution curves of clastic sediments[J]. Journal of Sedimentary Petrology, 1963, 33(1): 180-190.

[32] Udden J A. Mechanical composition of clastic sediments[J]. GSA Bulletin, 1914, 25(1): 655-744.

[33] Wentworth C K. A scale of grade and class terms for clastic sediments[J]. The Journal of Geology, 1922, 30(5): 377-392.

[34] Lane E W. Report of the subcommittee on sediment terminology[J]. Eos, Transactions American Geophysical Union, 1947, 28(6): 936-938.

[35] Friedman G M, Sanders J E. Principles of sedimentology[M]. New York: Wiley, 1978.

[36] Blott S J, Pye K. GRADISTAT: A grain size distribution and statistics package for the analysis of unconsolidated sediments[J]. Earth Surface Processes and Landforms, 2001, 26(11): 1237-1248.

[37] Blott S J, Pye K. Particle size scales and classification of sediment types based on particle size distributions: Review and recommended procedures[J]. Sedimentology, 2012, 59(7): 2071-2096.

[38] ISO 14688-1:2017. Geotechnical investigation and testing — Identification and classification of soil — Part 1: Identification and description[S]. International Organization for Standardization, 2017.

[39] 张昌民，王绪龙，朱锐，等. 准噶尔盆地玛湖凹陷百口泉组岩石相划分[J]. 新疆石油地质，2016，37（5）：606-614. [Zhang Changmin, Wang Xulong, Zhu Rui, et al. Litho-facies classification of Baikouquan Formation in Mahu Sag, Junggar Basin[J]. Xinjiang Petroleum Geology, 2016, 37(5): 606-614.]

[40] Folk R L. The distinction between grain size and mineral composition in sedimentary-rock nomenclature[J]. The Journal of Geology, 1954, 62(4): 344-359.

[41] Krumbein W C, Pettijohn F J. Manual of sedimentary petrography[M]. New York: Appleton-Century-Crofts, 1938.

[42] Inman D L. Measures for describing the size distribution of sediments[J]. Journal of Sedimentary Petrology, 1952, 22(3): 125-145.

[43] Folk R L, Ward W C. Brazos River bar: A study in the significance of grain size parameters[J]. Journal of Sedimentary Petrology, 1957, 27(1): 3-26.

[44] 贾建军，高抒，薛允传. 图解法与矩法沉积物粒度参数的对比[J]. 海洋与湖沼，2002，33（6）：577-582. [Jia Jianjun, Gao Shu, Xue Yunchuan. Grain-size parameters derived from graphic and moment methods: A comparative study[J]. Oceanologia et Limnologia Sinica, 2002, 33(6): 577-582.]

[45] Friedman G M. Distinction between dune, beach, and river sands from their textural characteristics[J]. Journal of Sedimentary Petrology, 1961, 31(4): 514-529.

[46] Klovan J E. The use of factor analysis in determining depositional environments from grain-size distributions[J]. Journal of Sedimentary Petrology, 1966, 36(1): 115-125.

[47] Bartholdy J, Christiansen C, Pedersen J B T. Comparing spatial grain-size trends inferred from textural parameters using percentile statistical parameters and those based on the log-hyperbolic method[J]. Sedimentary Geology, 2007, 202(3): 436-452.

[48] 高抒. 沉积物粒径趋势分析：原理与应用条件[J]. 沉积学报，2009，27(5)：826-836. [Gao Shu. Grain size trend analysis: Principle and applicability[J]. Acta Sedimentologica Sinica, 2009, 27(5): 826-836.]

[49] 李玉中，陈沈良. 系统聚类分析在现代沉积环境划分中的应用：以崎岖列岛海区为例[J]. 沉积学报，2003，21（3）：487-494. [Li Yuzhong, Chen Shenliang. Application of system cluster analysis to classification of modern sedimentary environment: A case study in Qiqu Archipelago area[J]. Acta Sedimentologica Sinica, 2003, 21(3): 487-494.]

[50] Ordóñez C, Ruiz-Barzola O, Sierra C. Sediment particle size distributions apportionment by means of functional cluster analysis (FCA)[J]. CATENA, 2016, 137: 31-36.

[51] 章婷曦，文莹亭，董丹萍，等. 太湖西北部表层沉积物粒度特征与沉积环境[J]. 湖泊科学，2018，30（3）：836-846. [Zhang Tingxi, Wen Yingting, Dong Danping, et al. Grain size features and sedimentary environment of surficial sediments in the northwest Lake Taihu[J]. Journal of Lake Sciences, 2018, 30(3): 836-846.]

[52] 刘祥奇，宋磊，吴奇龙，等. 基于粒度分布曲线的邻近传播聚类算法在沉积环境识别中的应用：以白洋淀地区为例[J]. 海洋地质与第四纪地质，2020，40（1）：198-209. [Liu Xiangqi, Song Lei, Wu Qilong, et al. Application of the affinity propagation clustering algorithm based on grain-size distribution curve to discrimination of sedimentary environment: A case study in Baiyangdian area[J]. Marine Geology & Quaternary Geology, 2020, 40(1): 198-209.]

[53] Grout H, Tarquis A M, Wiesner M R. Multifractal analysis of particle size distributions in soil[J]. Environmental Science & Technology, 1998, 32(9): 1176-1182.

[54] Miranda J G V, Montero E, Alves M C, et al. Multifractal characterization of saprolite particle-size distributions after topsoil removal[J]. Geoderma, 2006, 134(3/4): 373-385.

[55] 常宏，左合君，王海兵，等. 黄河乌兰布和沙漠段两岸地表沉积物多重分形特征及其指示意义[J]. 干旱区研究，2019，36（6）：1559-1567. [Chang Hong, Zuo Hejun, Wang Haibing, et al. Multi-fractal features and their significances of surface sediments along both banks of the Yellow River Reach in the Ulanbuh Desert[J]. Arid Zone Research, 2019, 36(6): 1559-1567.]

[56] Li J L, He X B, Wei J, et al. Multifractal features of the particle-size distribution of suspended sediment in the Three Gorges Reservoir, China[J]. International Journal of Sediment Research, 2021, 36(4): 489-500.

[57] Kuhnle R A. Fluvial transport of sand and gravel mixtures with bimodal size distributions[J]. Sedimentary Geology, 1993, 85(1/2/3/4): 17-24.

[58] Weltje G J. End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem[J]. Mathematical Geology, 1997, 29(4): 503-549.

[59] Weltje G J, Prins M A. Muddled or mixed? Inferring palaeoclimate from size distributions of deep-sea clastics[J]. Sedimentary Geology, 2003, 162(1/2): 39-62.

[60] Carder K L, Beardsley Jr G F, Pak H. Particle size distributions in the eastern Equatorial Pacific[J]. Journal of Geophysical Research, 1971, 76(21): 5070-5077.

[61] Kondolf G M, Adhikari A. Weibull vs. lognormal distributions for fluvial gravels[J]. Journal of Sedimentary Research, 2000, 70(3): 456-460.

[62] Peng J, Zhao H, Dong Z B, et al. Numerical methodologies and tools for efficient and flexible unmixing of single-sample grain-size distributions: Application to Late Quaternary aeolian sediments from the desert-loess transition zone of the Tengger Desert[J]. Sedimentary Geology, 2022, 438: 106211.

[63] Clark M W. Some methods for statistical analysis of multimodal distributions and their application to grain-size data[J]. Journal of the International Association for Mathematical Geology, 1976, 8(3): 267-282.

[64] Ashley G M. Interpretation of polymodal sediments[J]. The Journal of Geology, 1978, 86(4): 411-421.

[65] Azzalini A. A class of distributions which includes the normal ones[J]. Scandinavian Journal of Statistics, 1985, 12(2): 171-178.

[66] 孙东怀，鹿化煜，Rea D，等. 中国黄土粒度的双峰分布及其古气候意义[J]. 沉积学报，2000，18（3）：327-335. [Sun Donghuai, Lu Huayu, Rea D, et al. Bimode grain-size distribution of Chinese loess and its paleoclimate implication[J]. Acta Sedimentologica Sinica, 2000, 18(3): 327-335.]

[67] Sun D H, Bloemendal J, Rea D K, et al. Grain-size distribution function of polymodal sediments in hydraulic and Aeolian environments, and numerical partitioning of the sedimentary components[J]. Sedimentary Geology, 2002, 152(3/4): 263-277.

[68] Sun D H, Bloemendal J, Rea D K, et al. Bimodal grain-size distribution of Chinese loess, and its palaeoclimatic implications[J]. CATENA, 2004, 55(3): 325-340.

[69] Ibbeken H. Jointed source rock and fluvial gravels controlled by Rosin's law: A grain-size study in Calabria, South Italy[J]. Journal of Sedimentary Petrology, 1983, 53(4): 1213-1231.

[70] Purkait B. Patterns of grain-size distribution in some point bars of the Usri River, India[J]. Journal of Sedimentary Research, 2002, 72(3): 367-375.

[71] Seidel M, Hlawitschka M. An R-based function for modeling of end member compositions[J]. Mathematical Geosciences, 2015, 47(8): 995-1007.

[72] Dietze E, Hartmann K, Diekmann B, et al. An end-member algorithm for deciphering modern detrital processes from lake sediments of Lake Donggi Cona, NE Tibetan Plateau, China[J]. Sedimentary Geology, 2012, 243-244: 169-180.

[73] Dietze E, Dietze M. Grain-size distribution unmixing using the R package EMMAgeo[J]. E&G Quaternary Science Journal, 2019, 68(1): 29-46.

[74] Yu S Y, Colman S M, Li L X. BEMMA: A hierarchical Bayesian end-member modeling analysis of sediment grain-size distributions[J]. Mathematical Geosciences, 2016, 48(6): 723-741.

[75] 赵庆，郑祥民，周立旻，等. 末次冰期东海崎岛黄土粒度端元分析及其环境意义[J/OL]. 沉积学报. https://doi.org/10.14027/j.issn.1000-0550.2022.085. [Zhao Qing, Zheng Xiangmin, Zhou Limin, et al. Grain size end member characteristics and paleoclimatic significance of loess deposit in Shengshan Island during the Last Glacial Period[J/OL]. Acta Sedimentologica Sinica. https://doi.org/10.14027/j.issn.1000-0550.2022.085.]

[76] 袁瑞，冯文杰，张昌民，等. 长江武汉段天兴洲低滩沉积物粒度端元对河流—风成沙丘沉积环境的指示意义[J]. 地质论评，2023，69（5）：2023050025. [Yuan Rui, Feng Wenjie, Zhang Changmin, et al. Fluvial—Aeolian dune depositional environment significances from grain-size end-member in low-beach at the head of Tianxing Central-bar in Wuhan section of Yangtze River[J]. Geological Review, 2023, 69(5): 2023050025.]

[77] 朱海，张玉芬，李长安. 端元分析在长江武汉段古洪水识别中的应用[J]. 沉积学报，2020，38（2）：297-305. [Zhu Hai, Zhang Yufen, Li Chang'an. The application of end-member analysis in identification of paleo-floods in Wuhan Section of the Yangtze River[J]. Acta Sedimentologica Sinica, 2020, 38(2): 297-305.]

[78] 周声芳，刘秀铭，毛学刚，等. 美国Bryce峡谷Claron组粒度端元指示的风尘沉积及意义[J]. 沉积学报，2023，41（4）：1011-1024. [Zhou Shengfang, Liu Xiuming, Mao Xuegang, et al. Eolian deposition and its significance in the Claron Formation indicated by grain-size end members in the Bryce Canyon, Utah, USA[J]. Acta Sedimentologica Sinica, 2023, 41(4): 1011-1024.]

[79] Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery[M]. Washington: Microsoft Research, 2009.

[80] 左仁广. 基于数据科学的矿产资源定量预测的理论与方法探索[J]. 地学前缘，2021，28（3）：49-55. [Zuo Renguang. Data science-based theory and method of quantitative prediction of mineral resources[J]. Earth Science Frontiers, 2021, 28(3): 49-55.]

[81] 李灿锋，刘达，周德坤，等. 人工智能在地质领域的应用与展望[J]. 矿物岩石地球化学通报，2022，41（3）：668-677. [Li Canfeng, Liu Da, Zhou Dekun, et al. Application and prospect of artificial intelligence in the field of geology[J]. Bulletin of Mineralogy, Petrology and Geochemistry, 2022, 41(3): 668-677.]

[82] Wang C S, Hazen R M, Cheng Q M, et al. The Deep-Time Digital Earth program: Data-driven discovery in geosciences[J]. National Science Review, 2021, 8(9): nwab027.

[83] 李秋立，李扬，刘春茹，等. 地质年代学主要数据库现状分析与展望[J]. 高校地质学报，2020，26（1）：44-63. [Li Qiuli, Li Yang, Liu Chunru, et al. Analyses of current main geochronological databases and future perspectives[J]. Geological Journal of China Universities, 2020, 26(1): 44-63.]

[84] 蒋璟鑫，李超，胡修棉. 沉积学数据库建设与沉积大数据科学研究进展：以Macrostrat数据库为例[J]. 高校地质学报，2020，26（1）：27-43. [Jiang Jingxin, Li Chao, Hu Xiumian. Advances on sedimentary database building and related research: Macrostrat as an example[J]. Geological Journal of China Universities, 2020, 26(1): 27-43.]

[85] 张颖慧，王涛，焦守涛，等. 国内外岩浆岩数据库现状与应用前景[J]. 高校地质学报，2020，26（1）：11-26. [Zhang Yinghui, Wang Tao, Jiao Shoutao, et al. Review of igneous rock databases and their application prospect[J]. Geological Journal of China Universities, 2020, 26(1): 11-26.]

[86] Peters S E, Husson J M, Czaplewski J. Macrostrat: A platform for geological data integration and Deep-Time Earth crust research[J]. Geochemistry, Geophysics, Geosystems, 2018, 19(4): 1393-1409.

# Progress on Mining Methods of Sedimentological Information from Grain-size Distribution from the Background of Big Data

YUAN Rui

**School of Geophysics and Petroleum Resources, Yangtze University, Wuhan 430100, China**

**Abstract: [Significance]** The grain sizes of sediments contain information on multiple factors: transport path, depositional process, and environment. Grain-size distribution (GSD) is defined in sedimentology and geology as the frequency of occurrence of different-diameter particles. GSD is a record of the original sedimentological information. It is one aspect of the basic data used to reveal modern and ancient depositional environments in rivers, lakes, oceans, deserts, loess, etc. The traditional GSD analytical methods adopted to describe the overall features of depositional processes and environments, either qualitatively or semi-quantitatively, may not overcome problems of quantification and multiple solutions. **[Progress]** This study summarizes the range of different classification standards of grain-size scale, and compares moment and graphical frequency-curve methods of describing GSDs with morphological description standards. The applicability and usage of traditional methods of sedimentary environment analysis by GSD are reviewed, and some unconventional approaches are developed using mathematical methodology to tackle the entire range of GSD. Unsupervised clustering algorithms calculate the similarity of GSDs using their frequency, cumulative frequency or statistical parameters, then depositional environments are sorted according to the classes of clustering. Multifractal analysis is used to extract fractal parameters that represent the complexity of GSD frequency data. The different fractal structures reveal different depositional properties. When applied to multiple sedimentary processes in different sedimentary environments and dynamics and the GSD is superposed by multi-subpopulations, the corresponding frequency curve is found to be bimodal or multimodal. This implies that an inverse unmixing model of the sediments is ideally suited for obtaining genetically meaningful interpretations of these subpopulations. Two techniques are used to separate the grain-size component from GSD frequency data. To apply the statistical finite-mixture model, single-sample unmixing (SSU) uses a probability density function (normal, skew normal or Weibull distribution) to unmix the GSD by curve-fitting techniques. Each grain-size component is distributed in a unimodal fashion such that its statistical parameters (mean, sorting, skewness, kurtosis and percentage) may be calculated. The end-member modeling algorithm (EMMA) decomposes grain-size end-members from a GSD dataset. These unimodal or multimodal grain-size end-members are linearly independent and fixed within a single GSD dataset. Many

improved EMMAs are available in different open-source tools. To introduce examples of the application of these unconventional methods, in this study 27 GSDs from the central bar of the Kangshan River in the Poyang Lake drainage are processed by clustering, multifractal, SSU and EMMA. **[Conclusions and Prospects]** Problems of sedimentation analysis and the big-data properties of GSDs are solved. The trend of development of the depositional significance of GSDs is proposed based on analytical methods. With the advent of various modern grain-size analysis techniques and more sophisticated artificial intelligence procedures in earth sciences, new increasingly intelligent mining methods for GSDs are emerging for understanding the spatio-temporal grain-size patterns in sediments. Some excellent sedimentological related databases have been constructed. Accordingly, an open-access database will be established for GSDs to include various kinds of data, intelligent methods and a literature of reported research. From the background of big data, GSD big-data technology will provide a new driver for mining depositional properties intensively, and integrate them into sedimentological big data. Four phases-initial, exploratory, early development and rapid development – describe the history of GSD research. The future must hold a big-data phase for intelligent mining using sedimentological GSD information.

**Key words:** big data; grain-size distribution; sedimentological information; intelligent mining