一维到三维密度分布函数及其可视化在大数据分析中的应用
——以苦橄质玄武岩等为例

葛 粲[1,2,3],张 旗[4],李修钰[5],孙 贺[1,2,3],顾海欧[1,2,3],李伟伟[1,2,3],袁 峰[1,2,3]
GE Can[1,2,3], ZHANG Qi[4], LI Xiuyu[5], SUN He[1,2,3], GU Hai´ou[1,2,3], LI Weiwei[1,2,3], YUAN Feng[1,2,3]

1. 合肥工业大学资源与环境工程学院,安徽 合肥 230009;
2. 合肥工业大学矿集区立体探测实验室,安徽 合肥 230009;
3. 合肥工业大学安徽省矿产资源与矿山环境工程技术研究中心,安徽 合肥 230009;
4. 中国科学院地质与地球物理研究所,北京 100029;
5. 安徽省地质调查院,安徽 合肥 230001

1. *School of Resources and Environmental Engineering, Hefei University of Technology, Hefei* 230009, *Anhui, China;*
2. *Laboratory of Three-Dimension Exploration for Mineral District, Hefei University of Technology, Hefei* 230009, *Anhui, China;*
3. *Anhui Provincial Engineering Research Center for Mineral Resources and Mine Environments, Hefei University of Technology, Hefei* 230009, *Anhui, China;*
4. *Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing* 100029, *China;*
5. *Geological Survey of Anhui Province, Hefei* 230001, *Anhui, China*

摘要:提出不同维度的密度分布函数的计算方法和可视化方案,以解决不同数量级和不同测量误差的岩石样本数据分析对比困难的问题。通过SiO_2、全碱和MgO指标的三维密度分布函数和t-分布随机邻域嵌入可视化方法对GEOROC和PETDB数据库进行发掘,发现大洋岩(oceanite)和富辉橄玄岩(ankaramite)与苦橄质玄武岩(basalt, picritic)成分相近,而铁质苦橄岩(picrite,ferro)与侵入的橄榄辉长岩和苦橄岩(picrite)成分相似。利用二维密度分布函数和可视化技术,对比分析了不同岩石在TAS图解和硅镁图上的数据分布状态和数据集中核心区域。发现总体分布上,更富镁的苦橄岩的SiO_2含量高于苦橄质玄武岩,超基性的苦橄岩(picrate)核心区域主要分布在TAS图解的B区,这与以SiO_2=45%划分基性岩和超基性岩界线的观点矛盾。

关键词:密度分布函数;可视化;大数据;苦橄岩;苦橄质玄武岩
中图分类号:P588.14;P628 **文献标志码**:A **文章编号**:1671-2552(2019)12-2043-10

Ge C, Zhang Q, Li X Y, Sun H, Gu H O, Li W W, Yuan F. One-dimensional to three-dimensional density distribution functions and their applications in visualized big data analysis: Exemplified by picritic basalt and some other rocks. *Geological Bulletin of China*, 2019, 38(12):2043-2052

Abstract: In this paper, the calculation methods and visualization schemes of density distribution functions of different dimensions are proposed to solve the problem of difficulties in analysis and comparison of rock sample data with different orders of magnitude and different measurement errors. Data mining based on the GEOROC and PETDB databases by using the three-dimensional density distribution function of SiO_2, total alkali and MgO index as well as the t-distribution random neighborhood embedding visualization

收稿日期:2019-04-17;**修订日期**:2019-07-16
资助项目:国家青年科学基金项目《地震和重力数据联合约束下的苏鲁皖地区壳幔结构反演研究》(批准号:41504042)、《大别山双河超高压变质大理岩及其包裹榴辉岩的Li同位素地球化学研究》(批准号:41603005)和中国地质调查局项目《资源环境重大问题综合区划与开发保护策略研究》(编号:DD20190463)
作者简介:葛粲(1986-),男,博士,副教授,从事地球科学大数据分析和可视化研究。E-mail:gecan2008@gmail.com

method revealed that picritic basalt is similar to oceanite and ankaramite, while picrate is similar to intrusive olivine gabbro and ferropicrate. Comparisons between two-dimensional density distribution function and cumulative density contour visualization were used to analyze the data distribution of different rocks on TAS and Si-Mg maps and the core area of data concentration. It is found that the SiO_2 content of magnesium-rich picrite is higher than that of picrite basalt in general distribution. The core area of picrite is mainly located in the B area of TAS diagram, which is contrary to the traditional view that SiO_2=45% is used as the boundary between basic and ultramafic rocks.

Key words: density distribution function; visualization; big data; picrite; picritic basalt

随着信息时代的来临,地球科学数据已经呈现出爆炸式的增长趋势,如何从海量数据中发掘蕴含在数据中的关联知识是当前地球科学发展亟待解决的难题[1]。大数据分析需要利用科学有效的分析方法,在实际的数据分析中,遇到很多难题。①数据分布不均衡是大数据分析中常见的现象。在收集到的地球化学数据库中,不同岩石数据量差距极大,可以达到2~3个数量级,上万条数据如何与几百条数据对比分析?②数据库中不同来源(年代、仪器)的数据测量误差不同。而每个科研数据获取的代价都是昂贵的,笔者不希望丢弃任何有价值的数据。因此,如何去综合利用不同误差的数据?③数据量大的数据往往存在数据相互遮蔽的现象,散点图难以了解数据的疏密情况和集中的核心区域。④离群数据是不可避免的,有的离群数据是由于录入错误或其他原因导致的,而有的离群数据则具有研究价值和意义。只是受制于当前的认知水平,可能还难以被理解,对于离群数据的研究也可能成为学科新的增长点。离群数据数量少,但是往往会对均值、标准差等特征量统计产生很大的影响。如何能够在保留离群数据的情况下不影响分析数据的总体特征?针对以上大数据分析中遇到的问题,本文提出了基于密度分布函数的研究方法。

由于密度分布函数展现了数据分布的状态,不同类别的数据,虽然数据量有很大差别,但是转换成密度分布函数后可以对比,不受数据量多少的影响,基本解决了第一个难题。相比原先统计每个子区域样本数量再构造数据密度函数的方法[2],本文提出的方法可以针对不同样本使用不同的高斯函数。在年代久远的时期,用低精度的测量仪器和方法获得的测量数据,可以使用较大的标准差构建高斯函数;在当代用高精度测量手段获得的测量数据,可以使用较小的标准差构建高斯函数。这样就解决了第二个难题,可以融合不同测量误差的数据。在以散点图显示时,可通过累积密度值对数据点进行涂色,解决数据相互遮蔽的问题,方便了解数据的疏密情况和集中的核心区域。利用基于密度分布函数的分析方法,不需要假设计算传统统计量(均值、方差等),能够在保留离群数据的情况下不影响分析数据的总体特征。

本文利用该方法研究TAS图解中苦橄质玄武岩的问题,并提出一系列的可视化方案用于辅助研究。Pc分区是TAS图解14个分区SiO_2和全碱含量最低的一个分区,根据国际地科联的规定,该分区岩石名称定为苦橄质玄武岩(picrobasalt)[3-4]。在GEOROC和PETDB数据库中,笔者发现了名称类似的岩石,分别是苦橄岩(picrite)和苦橄质玄武岩(basalt, picritic)。苦橄岩是一种超镁铁质熔岩,由于接近原始岩浆成分,常被用来反演源区的成分及熔融的物理化学条件[6-7]。苦橄质玄武岩可能形成于地幔热柱的中心区域[8],对研究深部地质作用过程具有重要意义[9]。苦橄岩与苦橄质玄武岩如何区分?它们存在什么差异?其他名称的岩石是否也可能是苦橄质玄武岩?为了回答以上问题,笔者利用密度分布函数对数据库中样本量大于100个的142种岩石进行了系统研究,希望可以抛砖引玉,推进地质学的大数据研究。

1 密度分布函数构建与应用方法

数据库中的数据测量值有测量误差,可能测量值越远离真值其出现概率越小。假设测量的标准差已知,其分布是正态分布,将每个测量值转换成高斯函数,然后将所有测量值转换后的高斯函数累加起来,并除以测量样本数量,即得到连续的密度分布函数。根据研究内容的不同,可以构造一维、二维、三维甚至更高维度的密度分布函数。

一维连续密度分布函数是一条曲线,可视为统计直方图的变体。但无需定义统计区间,降低人为

因素影响。一维连续密度分布函数构建公式如下:

$$f(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{dx}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad (1)$$

其中,μ_i是测量数据,σ_i是该数据的测量误差标准差,N是参与计算的样本数量。计算时,x为均匀离散的网格点,dx是计算时网格点距。

二维连续密度分布函数是一个曲面,可视为二维统计直方图的变体。二维连续密度分布函数构建公式如下:

$$f(x,y)=\frac{1}{N}\sum_{i=1}^{N}\frac{dxdy}{2\pi\sigma_{xi}\sigma_{yi}}e^{-\frac{(x-\mu_{xi})^2}{2\sigma_{xi}^2}}e^{-\frac{(x-\mu_{yi})^2}{2\sigma_{yi}^2}} \quad (2)$$

其中,μ_{xi},μ_{yi}是测量数据测的2个指标,σ_{xi},σ_{yi}是该数据2个指标各自的测量误差标准差,N是参与计算的样本数量。计算时,x,y为均匀离散的网格点,dx和dy是计算时的二维网格点距。

三维连续密度分布函数为三维不均匀密度体,构建公式如下:

$$f(x,y,z)=\frac{1}{N}\sum_{i=1}^{N}\frac{dxdydz}{(2\pi)^{\frac{3}{2}}\sigma_{xi}\sigma_{yi}\sigma_{zi}}e^{-\frac{(x-\mu_{xi})^2}{2\sigma_{xi}^2}}e^{-\frac{(x-\mu_{yi})^2}{2\sigma_{yi}^2}}e^{-\frac{(x-\mu_{zi})^2}{2\sigma_{zi}^2}} \quad (3)$$

其中,μ_{xi},μ_{yi},μ_{zi}是测量数据3个指标,σ_{xi},σ_{yi},σ_{zi}是该数据各个指标的测量误差标准差,N是参与计算的样本数量。计算时,x,y,z为均匀离散的网格点,dx,dy和dz是计算时的三维网格点距。

当网格范围足够大时,理论上该方法计算出的密度分布函数满足累积密度为100%。

$$\sum_{k=1}^{M}f_k=100\% \quad (4)$$

其中,M为总的网格点数。当全部计算网格的总体累积密度与100%有较大差异时,需要考虑计算网格是否覆盖全部数据点。当无需考虑计算网格以外的数据时,可以利用总体累积密度对全体计算网格进行归一化。

密度分布函数f值较大的区域,数据集中程度更高。因此,可以根据密度分布函数的等高线、等值面确定数据集中分布的区域和空间范围,进而进行可视化和数据筛选。例如,需要确定80%数据集中分布区域的等高线和等值面,并找到对应等值线或等值面的密度阈值f^*,可通过以下步骤实现。

(1)需要建立累积密度函数G,该函数可以通过数值方法拟合构建。首先计算密度分布函数f的最大值和最小值,然后从最小值到最大值,取一系列具体的密度值f_i,将网格中累积密度大于等于f_i值的所有f_k求和获得累积密度值g_i:

$$gi=\sum_{f_k\geqslant f_i}f_k \quad (5)$$

拟合散点(f_i,g_i)可以构建累积密度函数G。

(2)通过拟合散点(g_i,f_i)可以构建累积密度函数的反函数G^{-1}。

(3)根据累积密度反函数G^{-1},可以求得累积密度阈值g^*=80%的密度阈值$f^*=G^{-1}(g^*)$。

(4)由密度分布函数$f(x,f,\cdots)=f^*$计算等高线和等值面,对数据分布区域进行圈定和可视化,以区分数据分布的核心区域和边缘区域,并在此基础上进一步分析。

一维到三维的密度分布函数构建和可视化的主要流程如图1所示。

2　数据来源

GEOROC(Geochemistry of Rocks of the Oceans and Continents)数据库由德国美因茨马克斯普朗克化学研究所建立并维护(http://GEOROC.mpch-mainz.gwdg.de/GEOROC/Start.asp)。该数据库综合收集、汇总了1.7万篇已发表论文、11个不同的地质背景、52万个火成岩和地幔捕虏体岩石样品、140.8万个分析结果的分析资料。样品信息包含地理位置、纬度和经度、岩石类别和岩石类型、分析方法、实验室、参考资料和引用来源,以及主量和微量元素含量、放射性和非放射性同位素比值,还包含岩石、玻璃、矿物和包含物的分析年龄。

PETDB是关于岩石、矿物和熔融包裹体的化学、同位素和矿物学数据的全球综合数据库(http://www.earthchem.org/petdb)。PETDB目前的内容集中在海底火成岩和变质岩的数据,特别是大洋中脊玄武岩和深海橄榄岩,以及来自地幔和下地壳的捕虏体样品[10],目前由Earthchem组织进行维护和持续更新。

本文从GEOROC和PETDB数据库中提取了岩石样本的岩石名称标签、SiO_2、Na_2O、K_2O和MgO含量数据作为分析数据。

3　基于密度分布函数的快速数据挖掘

在GEOROC和PETDB数据库中,样本数超过

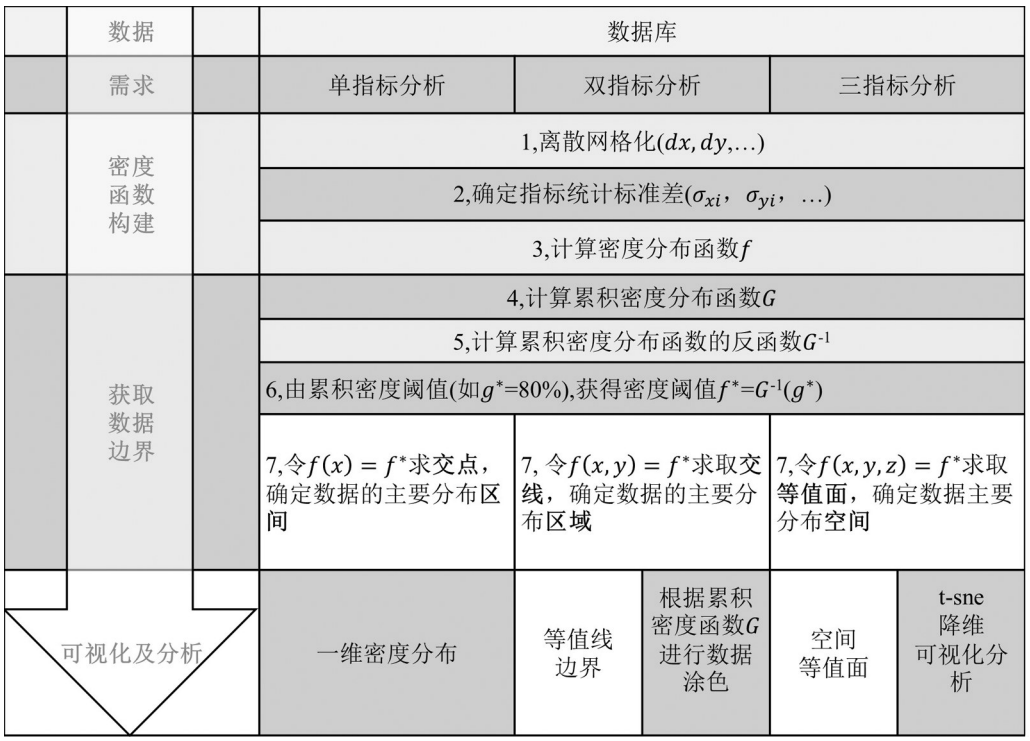

图1 一维到三维密度分布分析及其可视化流程图

Fig. 1 1-D to 3-D density distribution analysis and visualization flow chart

100个的岩石子类有142个，它们之间的关系如何？哪些岩石的化学成分相近？为了回答这些问题，在进行三维密度分布函数估计时，本文使用粗网格进行快速计算。由于TAS图解和硅镁图是常用的岩石分类图解，因此，本文使用SiO_2、全碱和MgO指标进行分析。没有对数据库中不同样本的测量方式及其精度进行统计，但是根据数据的分布区间进行了假定。当数据分布较分散时，可以利用标准差较大的高斯函数；当数据分布较集中时，利用标准差较小的高斯函数。以SiO_2为三维数据的x指标，取值范围为0%～85%，假设测量标准差为1%；以全碱(Na_2O+K_2O)为三维数据的y指标，取值范围为0%～21%，假设测量标准差为0.3%；以MgO含量为三维数据的z指标，取值范围为0%～50%，假设测量标准差为1%；三维网格以0.5%为网格点距，网格点数超过77万个。本文利用这3种指标，计算了142种岩石子类的各个网格点上三维密度分布函数值。142种岩石子类的密度分布函数的对比，可以视为每个网络点上密度分布函数值的对比，运用t-sne(t-distributed stochastic neighbor embedding，t-分布随机邻域嵌入)技术将高维数据压缩到二维空间进行可视化[11]。表1显示了本文从GEOROC和PETDB数据库中提取的各类岩石的样本数量和t-sne技术投影后的坐标。图2显示了各种岩石的相近程度，三维密度函数越接近的岩石在二维图中距离越接近。

通过图2和表1分析发现，与苦橄质玄武岩(basalt,picritic)分布接近的有大洋岩(oceanite)和富辉橄玄岩(ankaramite)，与苦橄岩(picrite)分布接近的有铁质苦橄岩(picrate,ferro)和橄榄辉长岩(gabbro,olivine)，而它们与玄武岩(basalt)保持了一定的距离，表明成分分布有明显差别。

大洋岩是一类含有大量橄榄石斑晶和略少辉石的岩石，由于其不仅见于大洋岛屿，也见于大陆地区，因此该命名已不常使用，目前维基百科已将其并入了苦橄质玄武岩(picritebasalt)条目。富辉橄玄岩也是具有镁铁质成分的火山岩，含丰富的辉石和橄榄石斑晶，它被认为是难熔的二辉橄榄岩地幔的原始熔融物[12]，维基百科将其认为是一种暗色斑状碧玄岩(basanite)。然而，实验研究表明，碧玄岩岩浆熔出于石榴橄榄岩[13]，且碧玄岩应该落在U1分区，而富辉橄玄岩只有33%的数据点落在U1分

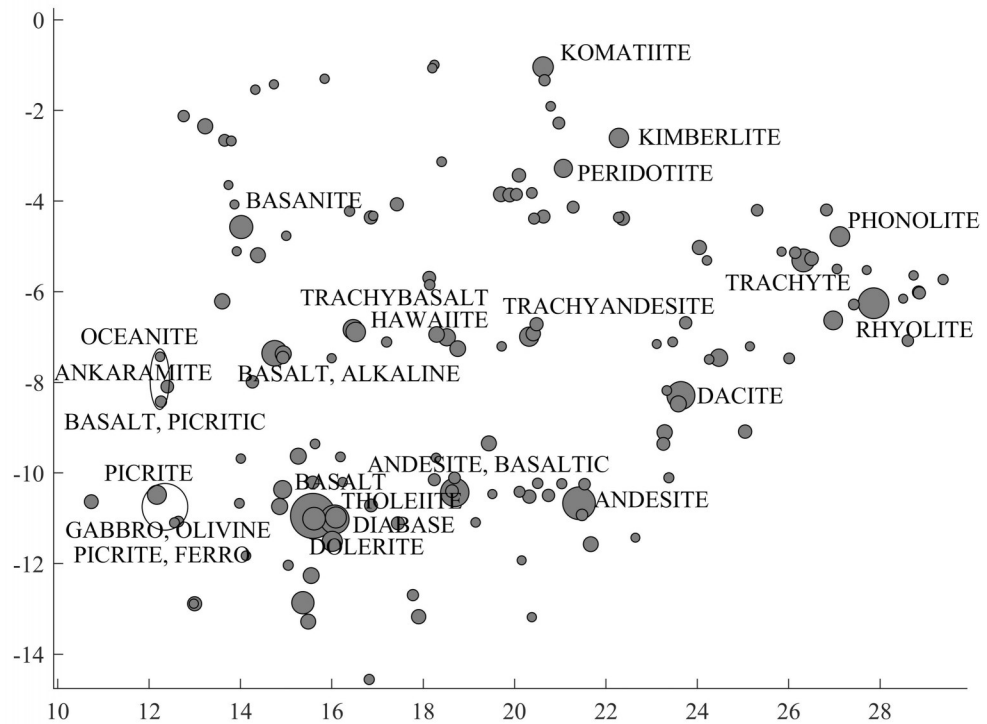

图2　SiO_2−全碱−MgO 三维密度分布函数的t−sne可视化

(圆圈的半径代表数据量的大小,圆圈的距离与密度分布函数的相似程度有关,三维密度函数接近的岩石在二维图中距离越接近)

Fig. 2　T−sne visualization of three−dimensional density distribution function of SiO_2−Na_2O+K_2O−MgO

区[2]。图2显示富辉橄玄岩与碧玄岩距离较远,表明其元素成分的高维密度分布函数有较大差异。因此,富辉橄玄岩并不属于碧玄岩,而与苦橄质玄武岩有较强的关联。橄榄辉长岩是深成侵入岩,不属于火山岩,但在成分上与苦橄岩类似,可能存在一定的联系。

4　苦橄质玄武岩、苦橄岩与玄武岩对比

为了更直观地对比分析苦橄质玄武岩、苦橄岩、玄武岩,以及与之相近的8类岩石之间的关系,笔者对数据进行了SiO_2−全碱和SiO_2−MgO投图,并根据各类岩石的二维累积密度分布函数进行了涂色和可视化对比,如图3、图4所示。

4.1　SiO_2−全碱指标的二维密度分布函数对比

图3中苦橄质玄武岩密度核心位于TAS图解Pc,B和U1分区交接处,苦橄岩密度核心主要分布在B区,小部分位于Pc分区,其硅和全碱含量略低于玄武岩分布的核心区域。苦橄质玄武岩的SiO_2总量低于橄榄石更富的苦橄岩,该结果与前人的认识大相径庭。

4.2　SiO_2−MgO指标的二维密度分布函数对比

图4中苦橄质玄武岩和苦橄岩均表现出富镁特征,SiO_2含量低于玄武岩和拉斑玄武岩而MgO含量高于这2类岩石。苦橄岩核心区域的SiO_2和MgO含量均略高于苦橄质玄武岩。

5　讨　论

5.1　对苦橄质玄武岩研究的启示

本文运用密度分布函数,对苦橄岩、苦橄质玄武岩、玄武岩等进行了研究,发现有2个问题值得注意。

(1)苦橄质玄武岩(包括富辉橄玄岩)比苦橄岩SiO_2含量低,这与前人的认识不同。苦橄岩的MgO含量高于苦橄质玄武岩(图4),由于橄榄石贫硅,苦橄岩的硅含量应比苦橄质玄武岩(为玄武岩的亚类)低才合理,出现相反的情况令人费解。笔者推测,可能是苦橄质玄武岩和苦橄岩的矿物组成不一样所致:苦橄质玄武岩除含橄榄石外,还含一定数量的斜方辉石。因为在暗色矿物中唯有斜方辉石既富镁又富硅(SiO_2含量为54%~57%),而苦橄岩主要的暗色矿物可能是橄榄石和单斜辉石(单斜辉石

表1 各类岩石样本数量和t-sne投影的坐标统计

Table 1 Statistics of the number of rock samples and the coordinates of t-sne projection

岩石名称	样本数量	*x*	*y*	岩石名称	样本数量	*x*	*y*
BASALT	64352	15.58721	-10.9474	SERPENTINITE	308	20.96866	-2.27882
ANDESITE	17916	21.4111	-10.6686	SYENITE, NEPHELINE	304	26.82648	-4.19666
RHYOLITE	14007	27.85793	-6.25479	NEPHELINITE, OLIVINE	300	13.65233	-2.65991
THOLEIITE	10955	16.06146	-11.0261	ANDESITE, THOLEIITIC	291	18.6852	-10.1085
ANDESITE, BASALTIC	10280	18.69086	-10.4295	DIORITE, QUARTZ	282	21.47555	-10.9256
DACITE	9632	23.64114	-8.29206	DUNITE	278	25.30984	-4.2026
BASALT, ALKALINE	6740	14.7485	-7.35955	SANDSTONE	275	28.60756	-7.08053
BASANITE	4503	14.01691	-4.57348	ANDESITE, HORNBLENDE	274	21.53134	-10.2483
TRACHYTE	4281	26.31998	-5.30974	TRACHYTE, ALKALINE	274	26.14345	-5.14133
DOLERITE	4150	15.61033	-11.0084	FOIDITE	260	12.75653	-2.12454
GABBRO	3911	15.36645	-12.8639	BASALT, PICRITIC	258	12.25989	-8.42502
BASALT, THOLEIITIC	3086	16.08896	-10.9825	KOMATIITE, PERIDOTITIC	253	20.65545	-1.33371
KOMATIITE	2775	20.62374	-1.04221	BRECCIA	250	17.77428	-12.6955
TRACHYBASALT	2616	16.46776	-6.83638	HARZBURGITE, SPINEL, XENOLITH	233	20.42683	-4.38752
DIABASE	2577	16.00959	-11.5096	ANDESITE, HYPERSTHENE-AUGITE	224	20.10768	-10.4155
PHONOLITE	2360	27.12145	-4.78326	GRANOPHYRE	212	26.01027	-7.47253
HAWAIITE	2284	16.52253	-6.8893	GRANITE, BIOTITE	211	27.42636	-6.28287
KIMBERLITE	2183	22.28322	-2.60571	LHERZOLITE, GARNET, XENOLITH	208	20.37776	-3.82075
TRACHYANDESITE	2159	20.31939	-6.99264	GABBRO, OLIVINE	203	12.6403	-11.0718
PICRITE	2041	12.17082	-10.4823	MINETTE	189	18.14181	-5.84798
GRANITE	1950	26.97359	-6.63372	DUNITE, XENOLITH	187	22.27814	-4.36065
PERIDOTITE	1692	21.06825	-3.27859	ANDESITE, 2-PYROXENE	186	20.50536	-10.2287
BASALT, OLIVINE	1628	14.923	-10.3652	LEUCITITE	179	16.38811	-4.22292
RHYODACITE	1420	24.47356	-7.45667	BASALT, SHOSHONITIC	171	17.19932	-7.11009
TRACHYANDESITE, BASALTIC	1371	18.51673	-7.00667	SCHIST	170	16.81794	-14.5571
THOLEIITE, OLIVINE	1062	14.85753	-10.7395	RHYOLITE, PERALKALINE	168	29.3796	-5.73169
GRANODIORITE	1052	23.58413	-8.47572	ANDESITE, CALC-ALKALINE	167	21.03326	-10.2375
BASALT, ALKALINE, OLIVINE	1043	14.93627	-7.37283	GREYWACKE	163	23.37688	-10.1065
BASALT, TRANSITIONAL	1028	15.26546	-9.62936	GABBRO, HORNBLENDE	161	15.04572	-12.0369
AMPHIBOLITE	994	15.54768	-12.2647	PICRITE, FERRO	154	12.55557	-11.0981
SHOSHONITE	964	18.75574	-7.2569	KAMAFUGITE	149	13.79827	-2.67388
MUGEARITE	929	18.29329	-6.94383	META-BASALT	141	13.97634	-10.6699
LAMPROPHYRE	860	13.60072	-6.21256	BASALT, OLIVINE-AUGITE	140	16.18754	-9.64572
ADAKITE	851	23.28493	-9.10343	LATITE, QUARTZ	140	23.46039	-7.10849
NEPHELINITE	828	13.22807	-2.34986	CARBONATITE, NATRO	139	29.91534	-5.4673
TEPHRA	793	19.43718	-9.34711	GRANODIORITE, HORNBLENDE- BIOTITE	138	23.33243	-8.18094
TEPHRITE	782	14.38127	-5.19469	IJOLITE	137	18.40342	-3.13221
GRANULITE	779	15.48294	-13.2801	BASALT, PLAGIOCLASE	134	15.63237	-9.35873
DIORITE	771	21.66719	-11.5719	GRANODIORITE, BIOTITE	133	24.2588	-7.49518
LHERZOLITE, SPINEL, XENOLITH	770	19.70123	-3.84926	SYENITE, QUARTZ	133	27.05632	-5.49837
LATITE	676	20.40854	-6.92619	WEHRLITE	132	20.79014	-1.90888
BONINITE	651	17.89986	-13.172	OCEANITE	130	12.24017	-7.43482

续表1-1

岩石名称	样本数量	x	y	岩石名称	样本数量	x	y
CARBONATITE	630	24.04427	-5.02497	ANDESITE, OLIVINE	129	19.14709	-11.091
BASALT, KOMATIITIC	625	12.99641	-12.8862	GABBRONORITE	129	14.12052	-11.8272
PYROXENITE	576	10.73488	-10.634	CARBONATITE, CALCITE	125	24.21048	-5.31015
PERIDOTITE, XENOLITH	567	22.3647	-4.38247	PYROXENITE, XENOLITH	125	18.24421	-0.99496
LHERZOLITE, XENOLITH	561	19.89378	-3.86592	MONZONITE	124	19.71715	-7.20648
SYENITE	529	26.50009	-5.27382	CLINOPYROXENITE	123	18.19991	-1.06767
LHERZOLITE	512	20.09564	-3.43263	NEPHELINITE, MELILITE	123	14.32507	-1.5421
GNEISS	502	25.04601	-9.08702	ANDESITE, BASALTIC, CALC-ALKALINE	122	18.2815	-9.66621
ANDESITE, AUGITE- HYPERSTHENE	492	20.32324	-10.5224	ECLOGITE	122	14.00808	-9.68333
BENMOREITE	475	20.48017	-6.71672	ASH	121	28.73349	-5.64124
PHONOTEPHRITE	473	16.8542	-4.36147	BASALT, AUGITE-OLIVINE	121	16.23734	-10.202
PHONOLITE, TEPHRI	469	17.4231	-4.07109	BASANITE, LEUCITE	121	15.00156	-4.76666
HARZBURGITE, XENOLITH	454	20.6322	-4.34245	MELILITITE, OLIVINE	119	15.84212	-1.30137
BASALT, CALC-ALKALINE	447	17.44508	-11.11	TEPHRITE, LEUCITE	119	16.90621	-4.32697
TONALITE	430	23.25654	-9.35807	KOMATIITE, BASALTIC	115	12.97848	-12.8844
THOLEIITE, QUARTZ	427	16.85698	-10.7168	LAMPROPHYRE, SHOSHONITIC	114	20.37735	-13.1823
LAMPROITE	425	18.13194	-5.69225	ABSAROKITE	113	15.99625	-7.46797
BASALTIC ANDESITE	414	18.62987	-10.4017	MELILITITE	111	14.73207	-1.42496
BASALT, SUBALKALINE	406	15.58202	-10.2094	ANDESITE, OLIVINE- AUGITE- HYPERSTHENE	109	19.51345	-10.4657
ANKARAMITE	404	12.39872	-8.09435	TRACHYPHONOLITE	109	25.84436	-5.11671
PANTELLERITE	401	28.84587	-6.0132	TRONDHJEMITE	109	25.15069	-7.20645
ALKALI BASALT	376	14.92808	-7.45134	CAMPTONITE	107	13.91801	-5.10844
GREENSTONE	375	14.26073	-7.98944	BASANITE, NEPHELINE	103	13.86844	-4.0752
ANDESITE, PYROXENE	364	20.74207	-10.4955	RHYOLITE, ALKALINE	103	28.50523	-6.15424
TRACHYDACITE	358	23.74495	-6.6863	TURBIDITE	102	20.15457	-11.9278
COMENDITE	335	28.86098	-6.02774	PELITE	101	22.64466	-11.4284
BASALT, ANDESITIC	324	18.24308	-10.1489	LIMBURGITE	100	13.73771	-3.64633
HARZBURGITE	317	21.28159	-4.13447	MONZONITE, QUARTZ	100	23.11145	-7.15484
PERIDOTITE, SPINEL, XENOLITH	314	20.04048	-3.85302	TRACHYTE, COMENDITIC	100	27.70746	-5.52258

的SiO_2含量为51%~54%)。

(2)TAS图解的界线是30年前根据归纳法提出的岩石分类界线,其分界方案与数据库中的真实数据样本大致吻合,但准确性有待商榷[2]。本次研究发现,苦橄质玄武岩数据的核心区在TAS图解中Pc区靠近B区的部位,苦橄岩的核心区在B区,略靠近Pc区;玄武岩则跨了Pc区、B区和A区。前人在新疆北准噶尔的研究也表明,苦橄岩分布于苦橄玄武岩和玄武岩区界线下部,苦橄质玄武岩分布于玄武岩区上部[14]。TAS图解确定的SiO_2=45%的界线没有起什么作用,而前人研究认为该界线是一条重要的界线,是划分基性岩和超基性岩的界线[15]。令人惊讶的是,真正属于超基性火山岩的苦橄岩主要分布在B区。最初TAS图解中苦橄质玄武岩、玄武岩、玄武质安山岩、安山岩和英安岩垂直界线的确定[3],与更加古老的分类标准一致[16-18],而早期界线的确定不可避免地受到样本量和检测水平的影响。

本次的研究方法是从数据库数据研究引入的,目的是为了反映数据的真实情况。研究得出的一些认识与早先的认识不同,是可以理解的。这里的启示值得学术界深思。可以肯定的是,早先的研究采用了抽样的方法,而本次研究是采用的是全数据模式,因此

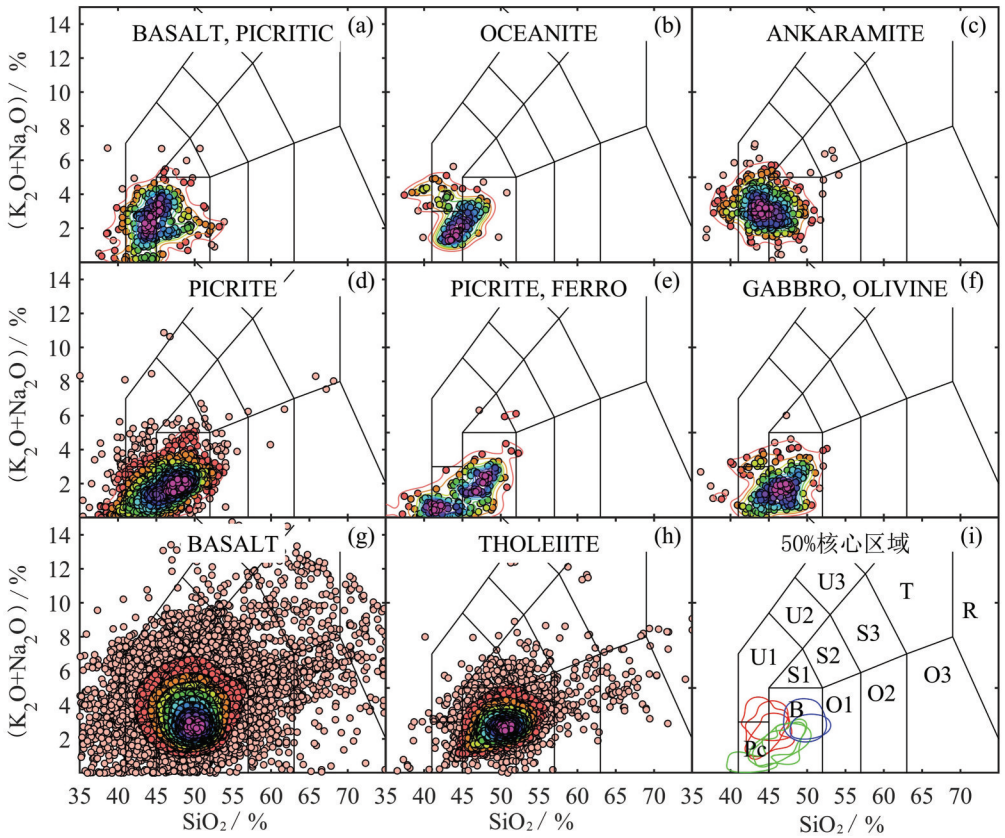

图3 8类岩石数据的SiO_2-全碱投图和二维累积密度分布函数可视化

(a~h分别为8类岩石数据的SiO_2-全碱投图和二维密度分布函数等高线,紫色-淡红色等不同颜色代表累积密度从10%递增到100%;样本数据点同样按照累积密度等高线的数值进行了涂色;i是8类岩石50%核心区域的等高线,其中红色曲线代表苦橄质玄武岩、大洋岩、富辉橄玄岩;绿色曲线代表苦橄岩、铁质苦橄岩、橄榄辉长岩;蓝色曲线代表玄武岩和拉斑玄武岩)

Fig. 3 SiO_2-(Na_2O+K_2O) maps and visualization of two-dimensional cumulative density distribution function for 8 rock data

Pc—苦橄玄武岩;B—玄武岩;O1—玄武安山岩;O2—安山岩;O3—英安岩;R—流纹岩;S1—粗面玄武岩;S2—玄武质粗面安山岩;S3—粗面安山岩;T—粗面岩、粗面英安岩;F—似长石岩;U1—碱玄岩、碧玄岩;U2—响岩质碱玄岩;U3—碱玄质响岩;Ph—响岩

全数据与抽样数据的结果显然是有差异的。

5.2 方法的局限性

本次提出的方法仍有一定的局限性,其中一些计算参数仍然需要靠人工主观确定。

(1)本次使用高斯函数作为核函数进行密度分布估计。高斯分布是一个在数学、物理、工程等领域都非常重要的概率分布,在统计学的许多方面有重大的影响力。然而,高斯分布并不一定适用于各种数据。例如,主量元素含量值的值域是有界的,其下界是0%,当数据集中分布在0%附近时,会导致密度分布函数0以下区域并非完全为0%,尤其是标准差取较大值时尤为明显。为避免这种不合理现象,一种可选的方法是,对有效区域内的累积密度函数再次进行归一化;另一种可选的方法是,选择其他合理的分布函数作为核函数进行计算。

(2)理论上,计算网格划分的越细,计算结果的精度越高,相应地计算量也越大。本文在计算二维和三维密度函数时,利用GPU设备进行了辅助计算。研究时可以从较粗的网格开始尝试,逐渐加密,观察密度分布函数是否有明显改变,当密度分布函数不再有明显改进时,说明网格密度已经足够。

(3)各个指标高斯函数的标准差,需要结合不同数据具体的分布情况、数据测量误差等因素确定。该参数目前还没有定量化的选择方法,需要进一步研究。

6 结 论

(1)密度分布函数是一种十分方便的技术,利

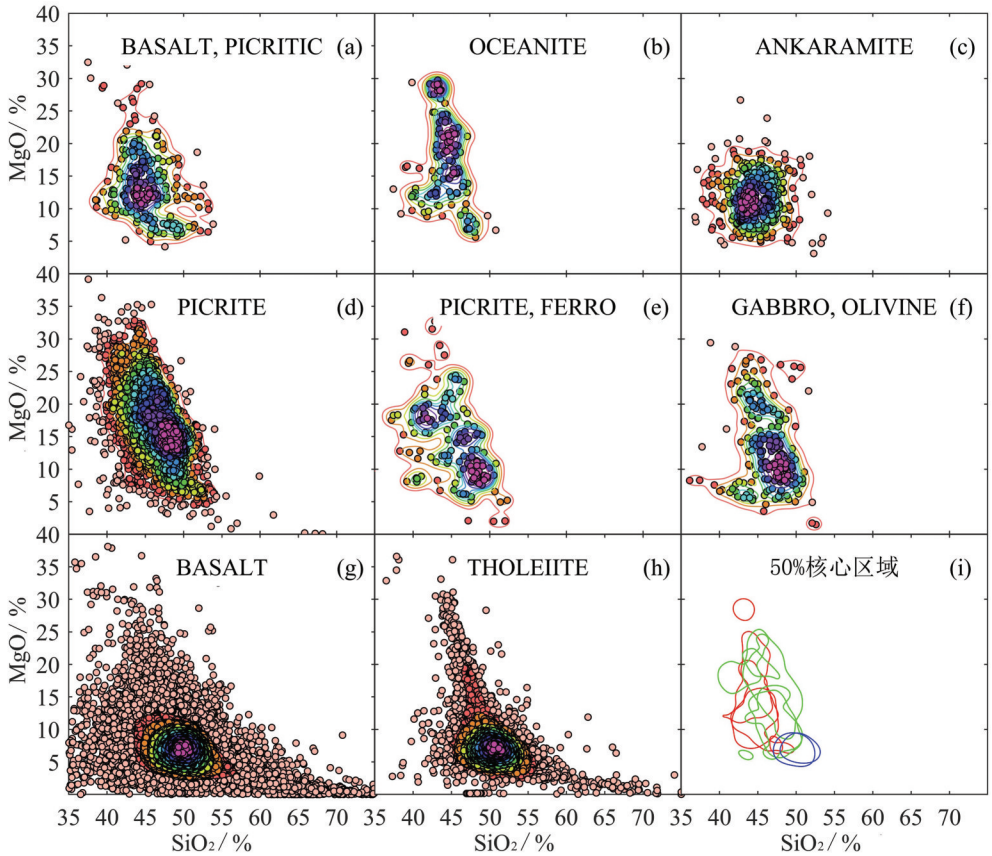

图 4　8类岩石数据的SiO_2−MgO投图和二维累积密度分布函数可视化

Fig. 4　SiO_2−MgO maps and visualization of two−dimensional cumulative density distribution function for 8 rock data

(a ~ h分别是8类岩石数据的SiO_2−MgO投图和二维密度分布函数等高线,其中图例同图3)

用密度分布函数计算获得的累积密度函数可以对散点数据进行可视化,以区分数据集中分布的核心区域与边缘区域,可用于多维度、全数据的分析、可视化和对比工作。

(2)通过SiO_2−全碱−MgO三维密度分布函数和可视化揭示了苦橄质玄武岩、大洋岩和富辉橄玄岩之间的相似性,以及苦橄岩、铁质苦橄岩和橄榄辉长岩之间的相似性。SiO_2−全碱和SiO_2−MgO二维密度分布函数和可视化揭示了苦橄岩、铁质苦橄岩、橄榄辉长岩、苦橄质玄武岩、大洋岩、富辉橄玄岩、玄武岩、拉斑玄武岩8类岩石之间的联系与区别。

(3)苦橄岩和苦橄质玄武岩作为超基性岩,与作为基性岩代表的玄武岩和拉斑玄武岩之间,并非是以SiO_2=45%为界线的,以此值为界线区分基性岩和超基性岩受到了地质大数据证据的挑战。

参考文献

[1]吴永亮, 陈建平, 贾志杰, 等. 地质数据本体构建及其在数据检索中的应用[J]. 地质通报, 2018, 37(5):945−953.

[2]葛粲, 顾海欧, 汪方跃, 等. 基于数据密度确定分布区域的方法:以TAS图解分析为例[J]. 地质科学, 2018, 53(4): 1240−1253.

[3]Le Maitre R W. A proposal by the IUGS Subcommission on the Systematics of Igneous Rocks for a chemical classification of volcanic rocks based on the total alkali silica (TAS) diagram[J]. Australian Journal of Earth Sciences, 1984, 31(2): 243−255.

[4]Le Maitre R W. A Classification of Igneous Rocks and Glossary of Terms: Recommendations of the InternationalUnion of Geological Sciences Subcommission on the Systematics of Igneous Rocks[M]. Blackwell. 1989: 1-193.

[5]Le Maitre R W. Igneous Rocks: A Classification and Glossary of Terms: Recommendations of the InternationalUnion of Geological Sciences Subcommission on the Systematics of Igneous Rocks[M]. Cambridge University Press, 2002 : 1-236.

[6]张招崇,郝艳丽, 王福生. 大火成岩省中苦橄岩的研究意义[J]. 地学前缘, 2003,(3): 105−114.

[7]张招崇,王福生, 郝艳丽. 峨眉山大火成岩省中的苦橄岩:地幔柱活动证据[J]. 矿物岩石地球化学通报, 2005, 24(1): 17−22.

[8]邓晋福, 赵海玲, 赖绍聪, 等. 中国北方大陆下的地幔热柱与岩石圈运动[J]. 现代地质, 1992, 6(3): 267−274.

[9]赖绍聪, 秦江锋, 赵少伟,等. 青藏高原东北缘柳坪新生代苦橄玄武岩地球化学及其大陆动力学意义[J]. 岩石学报, 2013, 30(2): 361−370.

[10]Lehnert K, Su Y, Langmuir C, et al. A global geochemical database structure for rocks[J]. Geochem. Geophys. Geosyst., 2000, 1(1):1−14.

[11]Der Maaten L V, Hinton G E. Visualizing Data using t−SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579−2605.

[12]Green D H, Schmidt M W, Hibberson W O. Island− arc Ankaramites: Primitive Melts from Fluxed Refractory Lherzolitic Mantle[J]. Journal of Petrology, 2004, 45(2): 391−403.

[13]Green D H. Conditions of melting of basanite magma from garnet peridotite[J]. Earth & Planetary Science Letters, 1973, 17(2): 456−465.

[14]陈毓川, 刘德权, 王登红, 等. 新疆北准噶尔苦橄岩的发现及其地质意义[J]. 地质通报, 2004, 23(11): 1059−1065.

[15]Johannsen A. A Descriptive Petrography of the Igneous Rocks[J]. Nature, 1938, 142(3594): 495−496.

[16]Frolova T I, Petrova M A. The classification diagram of effusive rocks[C]//IUGS Subcommission, 18th Circular, Contrib., 1974, 39: 25−30.

[17]Peccerillo A, Taylor S R. Geochemistry of Eocene calc− alkaline volcanic rocks from the Kastamonu area, northern Turkey[J]. Contrib. Mineral. Petrol., 1976, 58: 63−81.

[18]Streckeisen A. Classification and nomenclature of volcanic rocks, lamprophyres, carbonatites and melilitic rocks[C]//IUGS Subcommission on the Systematics of Igneous Rocks, GeolRundsch, 1980, 69(1): 194−207.