**Advances in Geosciences** 地球科学前沿, 2022, 12(5), 668-678
Published Online May 2022 in Hans. http://www.hanspub.org/journal/ag
https://doi.org/10.12677/ag.2022.125067

Hans汉斯

# 基于多重主成分分析的地球化学异常提取

**喻姝研[1,2]，邓　浩[1,2]**

[1]中南大学有色金属成矿预测与地质环境监测教育部重点实验室，湖南　长沙
[2]中南大学地球科学与信息物理学院，湖南　长沙

收稿日期：2022年4月24日；录用日期：2022年5月23日；发布日期：2022年5月31日

## 摘　要

**地球化学异常识别是勘察地球化学工作中的中重要环节。充分利用地球化学元素的频率分布和空间分布规律，对识别多变量地球化学异常的方法有重要意义。本文以山东省胶东半岛西北部地区为例，提出一种基于多重主成分分析的地球化学异常提取新方法。从高维数据的角度出发，通过多重主成分分析提取地球化学数据空间维度和元素维度的主要信息。捕捉地球化学数据的元素相关性和空间结构，重建出地球化学数据，进而通过原始值与重建值之间的距离计算地球化学异常得分。已知金矿点分布在异常得分分数高的区域；使用AUC指标评价异常得分与已知金矿点之间的空间关联度(AUC = 0.856)。结果显示，本文提取的地球化学异常分数与已知金矿床之间存在密切的空间相关性，基于多重主成分分析的方法能够有效提取地球异常。**

## 关键词

**多重主成分分析，地球化学异常提取，空间分布**

# Geochemical Anomaly Extraction Based on Multiple Principal Component Analysis

**Shuyan Yu[1,2], Hao Deng[1,2]**

[1]Key Laboratory of Metallogenic Prediction of Nonferrous Metals, Ministry of Education, Central South University, Changsha Hunan
[2]School of Geosciences and Info-Pysics, Central South University, Changsha Hunan

Received: Apr. 24th, 2022; accepted: May 23rd, 2022; published: May 31st, 2022

文章引用: 喻姝研, 邓浩. 基于多重主成分分析的地球化学异常提取[J]. 地球科学前沿, 2022, 12(5): 668-678.
DOI: 10.12677/ag.2022.125067

## Abstract

**Geochemical anomaly identification is an essential part of geochemical exploration. Making full use of the distribution of frequency and spatial distribution patterns of geochemical elements is important for identifying multivariate geochemical anomalies. In this paper, a new method of geochemical anomaly extraction based on multiple principal component analysis is proposed for the northwestern part of Jiaodong Peninsula in Shandong Province as an example. From the perspective of high-dimensional data, the main information of spatial dimension and elemental dimension of geochemical data is extracted by multiple principal component analysis. The elemental correlation and spatial structure of the geochemical data are captured, and the geochemical data are reconstructed. Subsequently, the geochemical anomaly score is calculated by the distance between the original value and the reconstruction value. The known gold deposits are distributed in the areas with high anomaly scores; the spatial correlation between the anomaly score and the known gold deposits is evaluated using the AUC index (AUC = 0.856). The results show that there is a close spatial correlation between the extracted geochemical anomaly scores and the known gold deposits, and the multiple principal component analysis method can effectively extract the geochemical anomalies.**

## Keywords

**Multiple Principal Component Analysis, Identifying Geochemical Anomalies, Spatial Distribution**



## 1. 引言

矿产资源是国家经济可持续发展的重要物资基础和支持力量。地球化学异常识别是找矿工作的重要一环。在成矿作用的影响下，元素浓度的地表分布与成矿特征量相关，通过元素在地表的分布与地下空间矿体的相关性进行找矿预测[1]。地质过程的复杂性使得地球化学元素含量的分布具有复杂的空间结构特征，准确有效地刻画地球化学元素分布模式对于提取地球化学异常也有积极的意义。因此，地球化学异常信息提取方法和模型研究是勘查地球化学领域的一个研究热点和前沿。

为了识别地球化学异常，以前的研究通过探索地球化学数据的频率分布或空间分布规律，在地球化学异常提取工作上取得了可喜的进展。单变量统计方法如标准差、概率图等是最简单直接的传统统计方法，在对单一元素进行概率分布的研究基础之上，研究元素的统计分布规律，确定地球化学背景和下限。多变量数据处理方法如主成分分析[2]、聚类分析[3]、判别分析[4]、因子分析[5]等方法，以多元素或多元素组合为基本特征探索地球化学数据的规律，捕捉地球化学元素的频率分布特征并分离地球化学异常。

地球化学数据是一个典型的具有高维元素属性的地理空间数据。因此，许多方法通过分析地球化学数据中的基本空间结构和空间自相关的特点，在地球化学异常识别上取得较好的成果[6] [7]。如克里格法、地理加权回归法[8]、趋势面分析法等方法[9]、局部差距统计学[10]、局部邻域统计[11]、空间因素分析[12]、移动平均技术[13] [14]等方法被用来探索地球化学数据的空间结构特征。基于分形与多重分形理论的信息提取方法根据地球化学数据中的空间尺度不变量特性、几何特性和与空间面积有关的幂律关系来分离地球化学异常现象。如浓度－面积分形模型，频谱－面积模型(spectrum-area)，局部奇异性分析(LSA) [15] [16]

等方法在地球化学识别上都获得了较好的效果。

最近，将机器学习方法引入到地球化学异常识别中也是一个研究热点。包括单类支持向量机[17]，连续限制波尔兹曼机[18]，孤立森林[19]，高斯混合模型[20]，深度自动编码器网络[21]，深度变异自动编码器[22]，GANomaly 网络[23]等方法已被利用来学习地球化学元素的复杂模式和非线性关联。多卷积自动编码器[24]空间约束多自动编码器、空间约束多自动编码器[25]，深度卷积神经网络[26]堆叠卷积去噪自动编码器[27]等方法能够保留空间特征来分离地球化学异常现象。

上述地球化学异常识别方法都能有效提取异常，取得显著效果。然而，专注于从地球化学元素的频率分布特征或者元素相关特征中识别异常的这类方法，往往忽略了地球化学元素的空间结构，掩盖地球化学数据中一些与空间分布相关的潜在规律[7]。而由于处理高维变量的局限性，关注地球化学元素空间分布的方法大多依赖于对高维地球化学数据的降维方法。这样的地球化学数据处理可能会导致地球化学特征的缺失。

为了综合考虑地球化学数据的元素相关性和空间分布规律，本文提出了一种基于多重主成分分析的方法识别地球化学异常，并选取山东省胶东半岛西北部为案例进行分析。使用多重主成分分析，将地球化学数据视为张量，并从张量的空间维度和元素维度上分别使用主成分分析方法提取信息，既考虑了地球化学数据的空间分布又顾及了元素之间的相关性。

常规的主成分分析方法在地球异常识别工作中一般作为多元统计方法出现，使用主成分分析提取地球化学元素组合信息，将多个元素变量降维，再针对单独的一个或者几个主成分，进一步使用数学方法或理论模型识别地球化学异常。作为多元统计方法的主成分分析仅在元素维度进行降维或者信息融合。而本文利用了主成分分析方法的信息提取能力，不仅对元素维度进行主成分提取，同时还对空间维度进行主成分信息提取，捕捉到地球化学数据的元素关联信息和空间分布信息。在主成分分析方法的转换下，大部分地球化学样本的信息都被保留下来，而异常代表的小部分信息未包含在主成分中。因此，通过多重主成分分析提取研究区水系沉积物地球化学数据的空间－元素联合信息，重建地球化学数据。并计算重建数据与原始数据之间的距离，提取地球化学异常。

## 2. 研究区地质概况

胶东地区做为我国最大的金矿集区，拥有 200 多个金矿点，已查明的黄金储量超过 5000 吨[28] [29]。胶东地区的大地构造位置位于华北克拉通东南缘和大别－苏鲁超高压变质带东北端。研究区位于山东省胶东半岛西北部地区(见图 1)。主要岩性是前寒武纪元古代火山岩和褐铁矿－闪长岩片麻岩及中生代花岗岩，局部覆盖有古新世火山岩、碎屑沉积物和第四纪沉积物。

区域内重要的控矿构造主要由三条控矿断裂自西向东构成：三山岛断裂、焦家断裂、招远－平度断裂。这几条北北东向和北东向断裂构成了胶西北地区最突出的线性构造，是与金矿密切相关的断裂构造。近 90%的金矿资源都与北北东－北东向断裂有关，分布于主要断裂成矿带之间的区域。

区内金矿床的成矿条件十分复杂，具有明显的多期性、多因素控矿性。区域内大型－超大型金矿的形成主要受前寒武纪变质基底岩系、中生代燕山期构造－岩浆活动、北东、北北东向韧－脆性构造三大因素复合控制。主要金矿类型有焦家式的破碎带蚀变岩型金矿、玲珑式的含金石英脉型金矿和河西式的网脉型金矿。不同的金矿类型在伸展构造中出现的位置不同。其中，焦家式金矿分布于靠近主断面的黄铁绢英岩化碎裂岩带和黄铁绢英岩化花岗质碎裂岩带中，代表性矿床为焦家金矿、三山岛金矿和大尹格庄金矿。河西式金矿出现于断裂下盘焦家式金矿外围的黄铁绢英岩化花岗质碎裂带和黄铁绢英岩化花岗岩带中，玲珑式金矿则赋存于伸展构造下盘远离主构造带的黄铁绢英岩化花岗岩带和正常花岗岩内，比如玲珑金矿[30]。

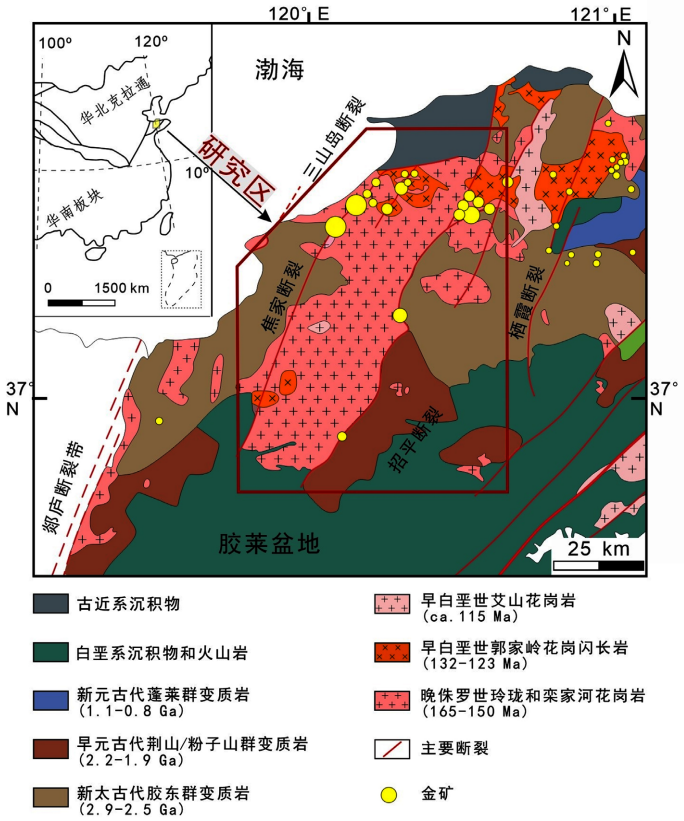



**Figure 1.** Geological map of the Jiaodong peninsula (modified after Liu *et al*., 2021 [31])
**图 1.** 胶西北地质特征图(改自 Liu *et al*., 2021 [31]))

## 3. 多重主成分分析方法框架

我们将 39 种元素的地球化学数据视为一个整体张量。为了从背景中分离出地球化学异常，使用多重主成分分析，对地球化学元素数据张量从空间维度 $x$–、空间维度 $y$–和元素维度这三个维度提取其信息。然后通过三个维度的主成分信息来重新构建地球化学张量。根据重建值和原始值之间的距离来计算异常值，识别地球化学异常。

### 3.1. 多重主成分分析

首先，对于高维地球化学数据数据 $\mathcal{X}\in\mathbb{R}^{M\times N\times L}$，将其分成空间上重叠的长方体块，我们可以将这种从完整的数据中分出来的长方体块称之为“斑块”。斑块大小为 $m\times n\times L$。每个斑块就是一个小型张量，表示为 $\mathcal{X}^{(i)}\in\mathbb{R}^{mnL}$，其中上标$(i)$表示的是第 $i$ 个斑块。

对于一个三阶张量 $\mathcal{M}$，其 $n$-模式展开矩阵由 $\mathcal{M}$ 的所有 $n$-模式展开向量进行排列而成$(n=1,2,3)$。我们将地球化学数据三阶张量的 1-模式、2-模式、3-模式展开矩阵称为 $x$-模式、$y$-模式、元素模式展开矩阵，它们分别表示沿 $x$、$y$ 和元素维度的展开矩阵。因此，斑块 $\mathcal{X}^{(i)}$ 可以被展开为三种模式的矩阵，即 $x$-模式展开矩阵 $\mathbf{X}_{(x)}^{(i)}\in\mathbb{R}^{m\times(nL)}$、$y$-模式展开矩阵 $\mathbf{X}_{(y)}^{(i)}\in\mathbb{R}^{n\times(mL)}$、元素模式展开矩阵 $\mathbf{X}_{(E)}^{(i)}\in\mathbb{R}^{L\times(mn)}$。

对于每个模式下的展开矩阵，我们分别利用主成分分析方法从中提取信息和特征，形成主成分信息矩阵。主成分分析(principal component analysis, PCA)是将多个变量指标综合为少数几个综合性变量指标的一种统计分析方法。主成分分析可以通过正交变换，将一组可能存在相关性的变量转换为一组线性不

相关的变量，转换后的这组变量叫主成分。因此，主成分能够保留原始变量的绝大部分信息，而且主成分之间是互不相关的。

对于斑块 $\mathcal{X}^{(i)}$ 的 $x$-模式展开矩阵 $\mathbf{X}_{(x)}^{(i)}$，我们可以对其进行主成分分析，则 $\mathbf{X}_{(x)}^{(i)}$ 可以表示成下列分解形式：

$$\mathbf{X}_{(x)}^{(i)} = \mathbf{T}_x \mathbf{U}_x^{(i)^{\mathrm{T}}}, \tag{3-1}$$

其中，主成分数据矩阵 $\mathbf{T}_x = [t_1, t_2, \cdots, t_t]$，$t_i$ 表示第 $i$ 个主成分组成的向量。$\mathbf{U}_x^{(i)}$ 是矩阵 $\mathbf{X}_{(x)}^{(i)}$ 的协方差阵的特征值，也叫做载荷(loading)矩阵。载荷矩阵是正交的，即：$\mathbf{U}_x^{(i)}\mathbf{U}_x^{(i)^{\mathrm{T}}} = \mathbf{I}$。由此，空间 $x$-维度的主成分矩阵 $\mathbf{T}_x$ 可以由 $x$-模式展开矩阵与载荷矩阵的矩阵乘法得到：

$$\mathbf{T}_x = \mathbf{X}_{(x)}^{(i)} \mathbf{U}_x^{(i)}. \tag{3-2}$$

我们取 PCA 结果的前 $p$ 个主成分( $p \le t$ )，也就是 $\mathbf{T}_x$ 的前 $p$ 列，形成主成分信息矩阵 $\mathbf{T}_x^* \in \mathbb{R}^{m \times p}$。即：$\mathbf{T}_x^* = [t_1, t_2, \cdots, t_p]$。$\mathbf{T}_x^*$ 可以代表空间 $x$-维度上的重要信息。

接下来，我们对斑块 $\mathcal{X}^{(i)}$ 的 $y$-模式展开矩阵进行主成分分析，并且得到空间 $y$-维度的主成分矩阵 $\mathbf{T}_y$：

$$\mathbf{T}_y = \mathbf{X}_{(y)}^{(i)} \mathbf{U}_y^{(i)} \tag{3-3}$$

其中，主成分数据矩阵 $\mathbf{T}_y$ 的每一个列是一个主成分向量，$\mathbf{U}_y^{(i)}$ 表示载荷矩阵。与 $\mathbf{T}_x^*$ 类似，空间 $y$ 维度信息矩阵 $\mathbf{T}_y^* \in \mathbb{R}^{n \times p}$ 可以通过取前 $p$ 个主成分( $p \le t$ )组成，$\mathbf{T}_y^* = [t_1, t_2, \cdots, t_p]$，$t_i$ 表示第 $i$ 个主成分向量。由此得到代表空间 $y$ 维度重要信息的矩阵 $\mathbf{T}_y^*$。

与空间维度上提取主成分信息矩阵过程相同，元素维度信息矩阵可以通过 PCA 结果中前 $q$ 个主成分获得。元素维度主成分矩阵 $\mathbf{T}_E$ 表示为：

$$\mathbf{T}_E = \mathbf{X}_{(E)}^{(i)} \mathbf{U}_E^{(i)} \tag{3-4}$$

其中，主成分数据矩阵 $\mathbf{T}_E$ 的一列表示元素维度上的一个主成分向量，$\mathbf{U}_y^{(i)}$ 表示载荷矩阵。元素维度信息矩阵 $\mathbf{T}_E^* \in \mathbb{R}^{L \times q}$ 由 PCA 结果中的前 $q$ 个主成分排列组成，包含着元素维度上的主要信息。

## 3.2. 多维信息重建

得到了空间维度上的信息矩阵 $\mathbf{T}_x^*$，$\mathbf{T}_y^*$ 和元素维度信息矩阵 $\mathbf{T}_E^*$ 后，我们可以得到一个联合的信息矩阵：

$$\mathbf{T} = \mathbf{T}_E^* \otimes \mathbf{T}_y^* \otimes \mathbf{T}_x^* \tag{3-5}$$

其中符号 $\otimes$ 表示的是 Kronecker 乘积运算[32]。联合的空间－元素维度信息矩阵能表示张量 $\mathcal{X}^{(i)}$ 的 $x$、$y$ 和元素维度的联合特征。我们能够根据空间－元素维度联合特征，还原重建出张量的主要信息。我们将张量 $\mathcal{X}^{(i)}$ 的向量化形式写为 $\mathbf{x}^{(i)}$，则有 $\mathbf{x}^{(i)} = \mathbf{T}\mathbf{c}^{(i)}$，其中 $\mathbf{c}^{(i)}$ 表示主成分信息矩阵 $\mathbf{T}$ 对应的重建核心系数，包含着矩阵 $\mathbf{T}$ 中各个主成分之间的关系。即，存在一组线性变换，能够将主成分信息矩阵 $\mathbf{T}$ 的各列(主成分)通过变换得到 $\mathbf{x}^{(i)}$。核心系数 $\mathbf{c}^{(i)}$ 可以通过最小化下列公式估计得到：

$$\min_{\mathbf{c}^{(i)}} \left\| \mathbf{x}^{(i)} - \mathbf{T}\mathbf{c}^{(i)} \right\|^2 + \lambda_c \left\| \mathbf{c}^{(i)} \right\|_1, \tag{3-6}$$

公式中，符号 $\|\cdot\|_1$ 表示的是 $\ell_1$-norm。这里我们使用使用交替乘法(ADMM) [33]求解出 $\mathbf{c}^{(i)}$。

主成分信息矩阵 $\mathbf{T}$ 包含着元素相关性和空间结构的信息。因此，核心系数 $\mathbf{c}^{(i)}$ 可以对斑块 $\mathcal{X}^{(i)}$ 的元素特征和空间特征进行编码，来重建地球化学背景。重建的过程中，我们需要使重建的数据 $\hat{\mathcal{X}}$ 接近原始的

数据 $\mathcal{X}$。我们先分别给定原始张量 $\mathcal{X}$ 和重建张量 $\hat{\mathcal{X}}$ 的向量化形式 $\hat{\mathbf{x}}=\mathrm{vec}\left(\hat{\mathcal{X}}\right)\in\mathbb{R}^{MNL}$ 和 $\mathbf{x}=\mathrm{vec}\left(\mathcal{X}\right)\in\mathbb{R}^{MNL}$。则求上述问题可以表述为：

$$\hat{\mathbf{x}}^{*}=\arg\min_{\hat{\mathbf{x}}}\left(\sum_{i=1}^{I}\left\|\mathbf{T}\mathbf{c}^{(i)}-\mathbf{R}^{(i)}\hat{\mathbf{x}}\right\|^{2}+\lambda\left\|\mathbf{x}-\hat{\mathbf{x}}\right\|^{2}\right) \tag{3-7}$$

式中，$\mathbf{R}^{(i)}\in\mathbb{R}^{nmL\times NML}$ 是一个由 0 和 1 组成的对角矩阵，用于从 $\mathcal{X}$ 中提取斑块 $\mathcal{X}^{(i)}$。在公式中，第一项确保关于 $\hat{\mathbf{x}}$ 的信息来自于空间元素信息矩阵 $\mathbf{T}$ 和核心系数 $\mathbf{c}^{(i)}$ 的信息，公式第二项可以约束重建数据的向量 $\hat{\mathbf{x}}$ 逼近原始数据向量 $\mathbf{x}$。$\lambda$ 是加权系数，控制 $\hat{\mathbf{x}}$ 与 $\mathbf{x}$ 之间的接近程度，在我们的工作中，$\lambda=1$。

公式(3-7)中的优化是一个简单的线性最小二乘法问题，对于 $\hat{\mathbf{x}}^{*}$ 的封闭式公式为：

$$\hat{\mathbf{x}}^{*}=\left(\lambda\mathbf{I}+\sum_{i=1}^{I}\mathbf{R}^{(i)^{\mathrm{T}}}\mathbf{R}^{(i)}\right)^{-1}\left(\lambda\hat{\mathbf{x}}+\sum_{i=1}^{I}\mathbf{R}^{(i)^{\mathrm{T}}}\mathbf{T}\mathbf{c}^{(i)}\right) \tag{3-8}$$

$\mathbf{I}\in\mathbb{R}^{MNL\times MNL}$ 是单位阵。最后，重建的地球化学数据 $\hat{\mathbf{x}}^{*}$ 被转换称张量形式，即：$\hat{\mathcal{X}}^{*}=\mathrm{vec}^{-1}\left(\hat{\mathbf{x}}^{*}\right)$。

得到重建数据之后，我们为每一个样本点计算重建值与原始值之间的距离，作为其异常得分，计算公式如下：[21] [27] [34]

$$\mathrm{Err}\left(i,j\right)=\sqrt{\sum_{k=1}^{L}\left(\mathcal{X}_{ijk}-\hat{\mathcal{X}}_{ijk}^{*}\right)^{2}} \tag{3-9}$$

其中，$\mathcal{X}_{ijk}$ 和 $\hat{\mathcal{X}}_{ijk}^{*}$ 分别表示 $\mathcal{X}$ 和 $\hat{\mathcal{X}}$ 在第 $k$ 个元素位置 $\left(i,j\right)$ 上的值。

## 4. 地球化学异常识别结果

我们以研究区 39 种元素地球化学数据为原始数据。地球化学元素数据是一种成分数据，我们使用等距对数比变换(isometric-logratio transformation, ilr)打开成分数据，剔除掉元素间的伪相关关系，消除成分数据结构中的“闭合效应”。然后再对 ilr 变换后的数据进行网格化，形成地球化学数据张量。

ilr 变换后的地球化学张量被分割成 16 × 16 × 38 大小的 97 个斑块。使用 PCA 分析中的主成分构成空间信息矩阵和元素信息矩阵，重建地球化学张量，得到地球化学异常得分。本文用接收者操作特征(receiver operating characteristic, ROC)曲线和 ROC 下曲线面积(area under the receiver operating characteristic curve, AUC)来量化异常得分与已知金矿之间的空间相关性[35] [36]，评价异常提取结果。

### 4.1. 参数的选择

主成分分析后得到的主成分拥有者最大的信息量。在使用多重 PCA 分析来提取地球化学张量的空间信息矩阵和元素信息矩阵时，我们需要确定所取得主成分的数量，即确定空间信息矩阵和元素信息矩阵的列数。因此，我们用 GridSearch 策略寻求最佳的列数。空间特征的主成分个数搜索范围是从 2 到 16，元素特征的主成分个数搜索范围是从 2 到 20，使用 AUC 对结果对模型结果进行评价。从图 2 中可以看到，元素信息矩阵列数的大小为 3 时，AUC 的值普遍比较高。而固定元素信息矩阵大小时，空间信息矩阵的列数对 AUC 的影响不大。

在分别学习 *x*-与 *y*-信息矩阵和元素信息矩阵的基础上，为了找到合适的列数，我们进一步探索空间信息矩阵和元素信息矩阵列数(即主成分个数)的影响，选取不同列数的空间信息矩阵提取元素和空间的信息，然后评估模型的结果。首先我们固定元素信息矩阵的大小，选取不同大小的空间信息矩阵进行主成分分析实验。根据 Grid Search 的结果(图 2)，我们选取列数为 3 的元素信息矩阵，空间信息矩阵的列数范围从 2 到 8。如图显示，当空间信息矩阵列数为 3 时，基于多重 PCA 信息矩阵的方法能够产生最大的 AUC 值 0.794。空间信息矩阵列数大于 3 时，增加主成分个数对结果并没有正向的影响。

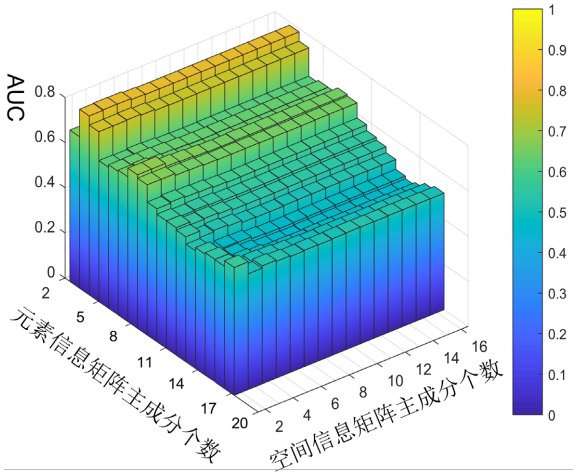
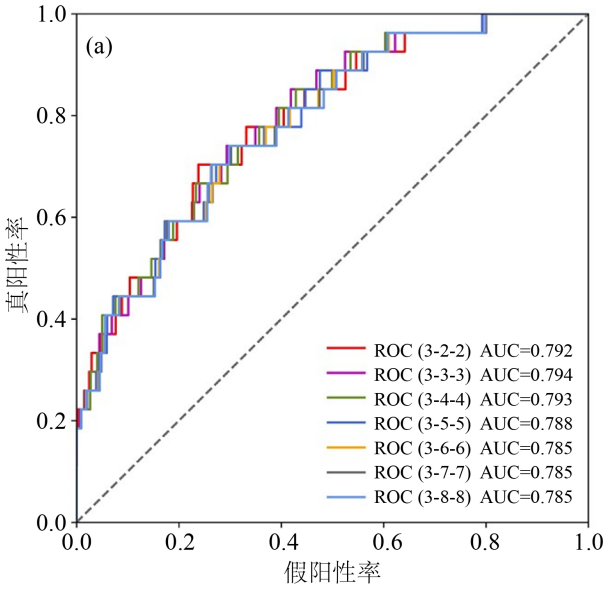



**Figure 2.** Grid-search results of the optimized atomic sizes for elemental informationmatrix and spatial informationmatrix

**图 2.** 空间信息矩阵和元素信息矩阵的主成分个数网格搜索结果

我们再次固定空间信息矩阵的大小，选取不同大小的元素信息矩阵进行实验，使用 AUC 评估基多重 PCA 方法的结果。根据不同空间信息矩阵列数的实验结果(图 3(a))，我们固定空间信息矩阵的列数为 3，采用范围从 2 到 8 的主成分元素信息矩阵列数进行实验。如图 3(b)显示，当使用 3 个主成分构成元素信息矩阵时，基于多重 PCA 方法的 AUC 达到最高值。说明在元素域 3 个主成分包含的信息量最适合提取地球化学异常。

下面我们采用元素信息矩阵列数(主成分个数)为 3，空间信息矩阵列数(主成分个数)也为 3 的设置进行多重主成分分析，计算地球化学异常得分。即：$p = 3$，$q = 3$。按照这种列数大小的设置，空间－元素联合信息矩阵大小为 9,728 × 27，其中 9,728 = 16 × 16 × 38，对应于“斑块”的尺寸，27 = 3 × 3 × 3 对应每个维度的主成分个数的乘积。即，由多重主成分分析提取的空间－元素联合信息矩阵的列数为 27。



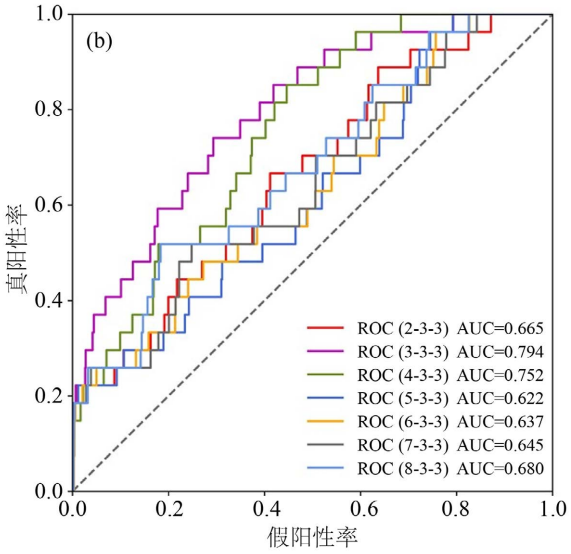



**Figure 3.** The ROC curves from spatial information matrix (a) and the elemental information matrix (b) with different numbers of principal components
**图 3.** 不同主成分个数的空间信息矩阵(a)和元素信息矩阵(b)产生的 ROC 曲线

PCA 方法能够变换出含有最大信息的变量。为了提升模型的性能，使主成分能充分代表性每个斑块的特征，我们用 Kmeans 算法，提前将 97 个斑块进行聚类分析，然后再对每一类的斑块进行信息矩阵学习。那么，同一个类的斑块将共享空间信息矩阵。Kmeans 算法中，聚类的类别个数也是至关重要的，这里我们画出了聚类的类别个数 $K$ 从 1 到 10 范围下模型结果的 AUC 值。从图 4 可以看到，当聚类数 $K = 5$ 的时候，AUC 值最大，ROC 曲线更偏向于左上角。并且聚类的操作对 AUC 的提升比较大，对结果有明显优化作用。

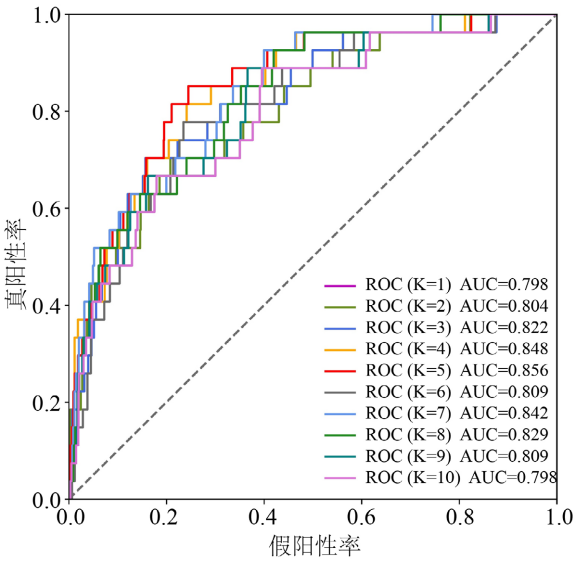



**Figure 4.** ROC curves of multiple PCA methods with different number of clusters
**图 4.** 多重 PCA 方法在不同聚类数下的 ROC 曲线图

### 4.2. 地球化学异常得分及评价

为了得到最优效果，最终我们使用 $p = 3$，$q = 3$ 的列数设置进行了地球化学异常识别，并且提前将 97 个“斑块”按照类别数 $K = 5$ 进行聚类，分别对每一类进行信息矩阵提取。

我们根据公式(3-9)估计了异常得分。基于多重主成分分析的方法得出的异常得分显示在图 5 中，大部分的金矿点都处在地球化学异常得分高的区域。为了评估已确定的地球化学异常和矿化之间的空间联系，我们绘制 ROC 曲线和一个显示累积异常区域与累积金矿的关系的图(如图 6)。

从异常分数图(图 5)可以看到，识别出的异常得分与金矿点大部分是吻合的。图 6(a)中的 ROC 曲线靠近左上角，并且 ROC 曲线下面积 AUC 的值为 0.856。从图 6(b)中看到，近 80%的已知金矿床位于异常得分较高的前 35%的区域。因此我们可以认为，基于多重主成分分析计算出的异常得分，与研究区的已知金矿点密切相关。

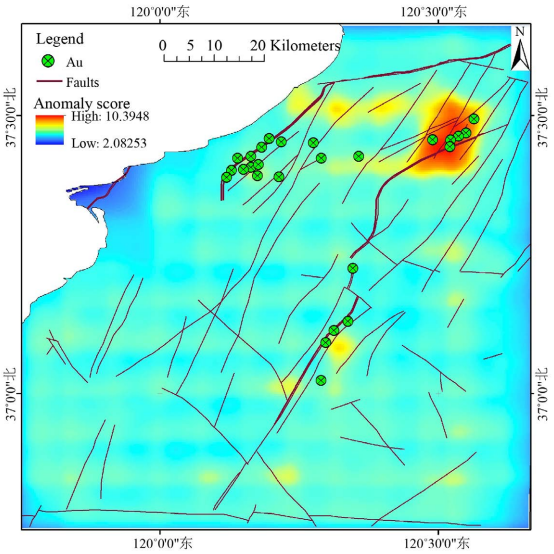



**Figure 5.** Geochemical anomaly scores based on multiple principal component analysis
**图 5.** 基于多重主成分分析提取的地球化学异常分数

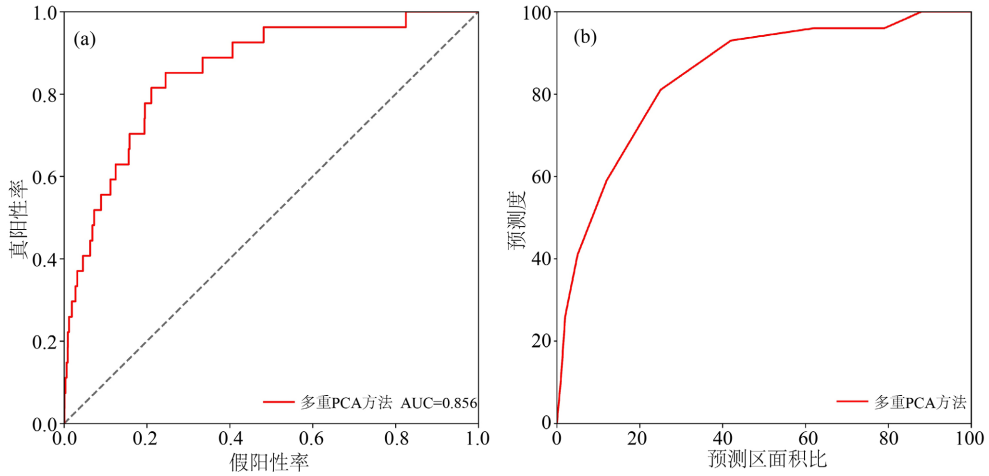



**Figure 6.** ROC curve (a), and predictive probability curve (b) based on multiple principal component analysis
**图 6.** 基于多重主成分分析的 ROC 曲线(a)，以及预测度曲线(b)

## 5. 结论

1) 本研究以中国山东省胶东半岛西北部地区为案例，对该区域的地球化学元素数据进行分析，基于多重 PCA 方法提取地球化学数据的空间－元素结构联合信息，由此重建地球化学高维数据。根据地球化学原始数据与高维数据之间的欧氏距离计算异常得分，实现研究区地球化学异常提取。

2) 基于多重 PCA 的方法从张量的角度出发，兼顾地球化学元素的空间维度和元素维度。将 39 种元素的地球化学数据视为张量，利用 PCA 提取空间维度和元素维度的主要信息，同时考虑地球化学数据的空间结构以及 39 种元素关联性。

3) 本研究使用 AUC 指标量化地球化学异常得分与已知金矿点之间的空间关联度，实现对多重 PCA 方法提取结果的评价。模型提取的地球化学异常得分分数高的区域，与金矿点分布相吻合；地球化学异常得分的 AUC 值为 0.856。评价结果表明：基于多重主成分分析计算的异常点得分与已知金矿床之间存在密切的空间相关性。因此我们可以认为，基于多重主成分分析的方法能够有效提取出地球化学异常。

## 参考文献


[1] 刘艳鹏, 朱立新, 周永章. 大数据挖掘与智能预测找矿靶区实验研究——卷积神经网络模型的应用[J]. 大地构造与成矿学, 2020, 44(2): 192-202.

[2] Jimenez-Espinosa, R., Sousa, A.J. and Chica-Olmo, M. (1993) Identification of Geochemical Anomalies Using Principal Component Analysis and Factorial Kriging Analysis. *Journal of Geochemical Exploration*, **46**, 245-256. https://doi.org/10.1016/0375-6742(93)90024-G

[3] Zhou, S., Zhou, K., Yang, G., *et al*. (2017) Application of Cluster Analysis to Geochemical Compositional Data for Identifying Ore-Related Geochemical Anomalies. *Frontiers of Earth Science*, **12**, 491-505. https://doi.org/10.1007/s11707-017-0682-8

[4] Sinclair, A.J. (1974) Selection of Threshold Values in Geochemical Data Using Probability Graphs. *Journal of Geochemical Exploration*, **3**, 129-149. https://doi.org/10.1016/0375-6742(74)90030-2

[5] 石文杰, 魏俊浩, 张德才, 赵少卿, 陈冲, 高翔, 翟亚峰, 易建. 基于数字高程模型因子分析的地球化学异常提取[J]. 物探与化探, 2012, 36(1): 103-108.

[6] Carranza, E.J.M. (2009) Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. *Handbook of Exploration and Environmental Geochemistry*, **11**, 351 p.

[7] Zuo, R., Carranza, E.J.M. and Wang, J. (2016) Spatial Analysis and Visualization of Exploration Geochemical Data. *Earth-Science Reviews*, **158**, 9-18. https://doi.org/10.1016/j.earscirev.2016.04.006

[8] Tian, M., Wang, X., Nie, L., *et al*. (2018) Recognition of Geochemical Anomalies Based on Geographically Weighted Regression: A Case Study across the Boundary Areas of China and Mongolia. *Journal of Geochemical Exploration*, **190**, 381-389. https://doi.org/10.1016/j.gexplo.2018.04.003

[9] Wang, H. and Zuo, R. (2015) A Comparative Study of Trend Surface Analysis and Spectrum-Area Multifractal Model to Identify Geochemical Anomalies. *Journal of Geochemical Exploration*, **155**, 84-90. https://doi.org/10.1016/j.gexplo.2015.04.013

[10] Wang, J. and Zuo, R. (2016) An Extended Local Gap Statistic for Identifying Geochemical Anomalies. *Journal of Geochemical Exploration*, **164**, 86-93. https://doi.org/10.1016/j.gexplo.2016.01.002

[11] Zuo, R. (2014) Identification of Weak Geochemical Anomalies Using Robust Neighborhood Statistics Coupled with GIS in Covered Areas. *Journal of Geochemical Exploration*, **136**, 93-101. https://doi.org/10.1016/j.gexplo.2013.10.011

[12] Grunsky, E.C. and Agterberg, F.P. (1988) Spatial and Multivariate Analysis of Geochemical Data from Metavolcanic Rocks in the Ben Nevis Area, Ontario. *Mathematical Geology*, **20**, 825-861. https://doi.org/10.1007/BF00890195

[13] Cheng, Q., Agterberg, F.P. and Bonham-Carter, G.F. (1996) A Spatial Analysis Method for Geochemical Anomaly Separation. *Journal of Geochemical Exploration*, **56**, 183-195. https://doi.org/10.1016/S0375-6742(96)00035-0

[14] Cheng, Q. (1999) Spatial and Scaling Modelling for Geochemical Anomaly Separation. *Journal of Geochemical Exploration*, **65**, 175-194. https://doi.org/10.1016/S0375-6742(99)00028-X

[15] Wang, J. and Zuo, R. (2018) Identification of Geochemical Anomalies through Combined Sequential Gaussian Simulation and Grid-Based Local Singularity Analysis. *Computers & Geosciences*, **118**, 52-64. https://doi.org/10.1016/j.cageo.2018.05.010

[16] Wang, J. and Zuo, R. (2019) Recognizing Geochemical Anomalies via Stochastic Simulation-Based Local Singularity Analysis. *Journal of Geochemical Exploration*, **198**, 29-40. https://doi.org/10.1016/j.gexplo.2018.12.012

[17] Chen, Y. and Wu, W. (2017) Mapping Mineral Prospectivity by Using One-Class Support Vector Machine to Identify Multivariate Geological Anomalies from Digital Geological Survey Data. *Australian Journal of Earth Sciences*, **64**, 639-651. https://doi.org/10.1080/08120099.2017.1328705

[18] Chen, Y., Lu, L. and Li, X. (2014) Application of Continuous Restricted Boltzmann Machine to Identify Multivariate Geochemical Anomaly. *Journal of Geochemical Exploration*, **140**, 56-63. https://doi.org/10.1016/j.gexplo.2014.02.013

[19] 吕岩. 基于机器学习系列方法的铁矿化地球化学异常识别[D]: [博士学位论文]. 长春: 吉林大学, 2021.

[20] Chen, Y. and Wu, W. (2019) Separation of Geochemical Anomalies from the Sample Data of Unknown Distribution Population Using Gaussian Mixture Model. *Computers & Geosciences*, **125**, 9-18. https://doi.org/10.1016/j.cageo.2019.01.010

[21] Xiong, Y. and Zuo, R. (2016) Recognition of Geochemical Anomalies Using a Deep Autoencoder Network. *Computers & Geosciences*, **86**, 75-82. https://doi.org/10.1016/j.cageo.2015.10.006

[22] Luo, Z., Xiong, Y. and Zuo, R. (2020) Recognition of Geochemical Anomalies Using a Deep Variational Autoencoder Network. *Applied Geochemistry*, **122**, Article ID: 104710. https://doi.org/10.1016/j.apgeochem.2020.104710

[23] Luo, Z., Zuo, R., Xiong, Y., *et al*. (2021) Detection of Geochemical Anomalies Related to Mineralization Using the GANomaly Network. *Applied Geochemistry*, **131**, Article ID: 105043. https://doi.org/10.1016/j.apgeochem.2021.105043

[24] Chen, L., Guan, Q., Feng, B., *et al*. (2019) A Multi-Convolutional Autoencoder Approach to Multivariate Geochemical Anomaly Recognition. *Minerals*, **9**, Article No. 270. https://doi.org/10.3390/min9050270

[25] Chen, L., Guan, Q., Xiong, Y., *et al*. (2019) A Spatially Constrained Multi-Autoencoder Approach for Multivariate Geochemical Anomaly Recognition. *Computers & Geosciences*, **125**, 43-54. https://doi.org/10.1016/j.cageo.2019.01.016

[26] Zhang, C., Zuo, R. and Xiong, Y. (2021) Detection of the Multivariate Geochemical Anomalies Associated with Mineralization Using a Deep Convolutional Neural Network and a Pixel-Pair Feature Method. *Applied Geochemistry*, **130**, Article ID: 104994. https://doi.org/10.1016/j.apgeochem.2021.104994

[27] Xiong, Y. and Zuo, R. (2021) Robust Feature Extraction for Geochemical Anomaly Recognition Using a Stacked Convolutional Denoising Autoencoder. *Mathematical Geosciences*, **54**, 623-644.

[28] 宋明春, 宋英昕, 丁正江, 李世勇. 胶东金矿床: 基本特征和主要争议[J]. 黄金科学技术, 2018, 26(4): 406-422.

[29] 宋英昕, 宋明春, 丁正江, 等. 胶东金矿集区深部找矿重要进展及成矿特征[J]. 黄金科学技术, 2017, 25(3): 4-18.

[30] 宋明春, 伊丕厚, 徐军祥, 崔书学, 沈昆, 姜洪利, 袁文花, 王化江. 胶西北金矿阶梯式成矿模式[J]. 中国科学: 地球科学, 2012, 42(7): 992-1000.

[31] Liu, Z., Hollings, P., Mao, X., *et al*. (2021) Metal Remobilization from Country Rocks into the Jiaodong-Type Orogenic Gold Systems, Eastern China: New Constraints from Scheelite and Galena Isotope Results at the Xiadian and Majiayao Gold Deposits. *Ore Geology Reviews*, **134**, Article ID: 104126. https://doi.org/10.1016/j.oregeorev.2021.104126

[32] Neudecker, H. (1969) A Note on Kronecker Matrix Products and Matrix Equation Systems. *SIAM Journal on Applied Mathematics*, **17**, 603-606. https://doi.org/10.1137/0117057

[33] Boyd, S., Parikh, N., Chu, E., *et al*. (2011) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[34] Zuo, R. and Xiong, Y. (2018) Big Data Analytics of Identifying Geochemical Anomalies Supported by Machine Learning Methods. *Natural Resources Research*, **27**, 5-13. https://doi.org/10.1007/s11053-017-9357-0

[35] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[36] Bergmann, R., Ludbrook, J. and Spooren, W.P.J.M. (2000) Different Outcomes of the Wilcoxon-Mann-Whitney Test from Different Statistics Packages. *The American Statistician*, **54**, 72-77. https://doi.org/10.1080/00031305.2000.10474513