

澳 門 科 技 大 學

Macao University of Science and Technology

数据科学常用工具期末报告

Final Report on Spatial Database Systems

組 長： 吴百恒

組 員： 黄熙 李杰相 彭嘉男

學 院： 太空科學研究所

課 程： 数据科学常用工具

完成日期： 27/5/2025

目 录

[目 录 2](#_Toc197354514)

[1 前言 3](#_Toc197354515)

[**1.1** 内容简介 3](#_Toc197354516)

[**1.2** 完成环境 3](#_Toc197354517)

[2 文章复现及优化 4](#_Toc197354518)

[**2.1** 复现文章选择 4](#_Toc197354519)

[**2.2** 文章复现思路 4](#_Toc197354520)

[**2.3** 文章具体复现 5](#_Toc197354521)

[**2.3.1** 雷达数据预处理-去重复道以及数据拼接 5](#_Toc197354522)

[**2.3.2** 雷达数据后处理 7](#_Toc197354523)

[**2.4** 个人优化 8](#_Toc197354524)

[**2.4.1** 读取数据优化 8](#_Toc197354525)

[**2.4.2** 引入早期人工智能算法优化数据预处理 8](#_Toc197354526)

[**2.4.3** 使用更精准的零时矫正 9](#_Toc197354527)

[**2.4.4** 设置新型带通滤波 10](#_Toc197354528)

[**2.4.5** 额外使用新的方法进行背景噪声去除 10](#_Toc197354529)

[**2.5** 未来工作 11](#_Toc197354530)

[参 考 文 献 13](#_Toc197354531)

[总 结 14](#_Toc197354532)

1 前言

**1.1** 作业概述

使用包含445个斑岩岩石样品数据集（每个样品包含36个主、微量元素特征）构建一个端到端工作流，该流程需实现：区分未知样品为Cu-rich或Au-rich； 解释主要控矿地球化学特征；交付一个可离线运行的预测工具。

**1.2** 环境适配

PC：Microsoft Win10/11

运行环境：Python3.10.16

Server：Ubuntu 24.04.1 LTS

运行环境：Python3.11.11

**1.3** 报告撰写格式

标题：

字体使用黑体，字号由小三递减

1.5倍行距，大标题段后1行，其余段前0.5行

正文：

汉字：宋体小四 1.25倍行距

英文、数字：Times New Roman

参考文献：

APA格式，用方括号“[]”和数字按顺序以右上角标形式标注在引用处

2 任务实现流程

**2.1** 数据预处理

**2.1.1** 读取数据

构建本地模块data\_preprocessing.loader，借助os以及pandas两个库文件，实现根据文件后缀加载数据功能模块。功能见表2.1，试运行展示见图2.1。

|  |  |
| --- | --- |
| 功能 | 详细描述 |
| 1.多方式输入支持 | 支持文件流、本地路径 |
| 2.文件扩展名识别 | 提取扩展名并转为小写 |
| 3.主体读取、异常处理 | 根据扩展名选择读取方式，限制支持Excel和CSV文件 |
| 4.内容初判 | 判断文件是否为空，并设定状态码值 |
| 5.异常捕获 | 针对不同异常返回专属状态码 |

表2.1 本地模块data\_preprocessing.loader功能介绍

***亮点自述：***

1.支持本地路径和流式对象

2.设定多种异常分析以及专属状态码，方便问题查找

3.设定标准化返回，返回执行信息，成功或错误描述

日历

AI 生成的内容可能不正确。

图2.1 读数据模块试运行结果图

**2.1.2** 定义单位映射

设置专属映射字典，定义每个化学元素/氧化物的计量单位。试运行展示见图2.2、2.3。

***亮点自述：***

1. 为数据的每一列指定计量单位，便于后续数据解释和处理。
2. 指定后调用专属校验模块验证单位信息、数据类型等是否设置成功
3. 设定基础以及专属异常分析以及对应状态码
4. 设定标准化返回

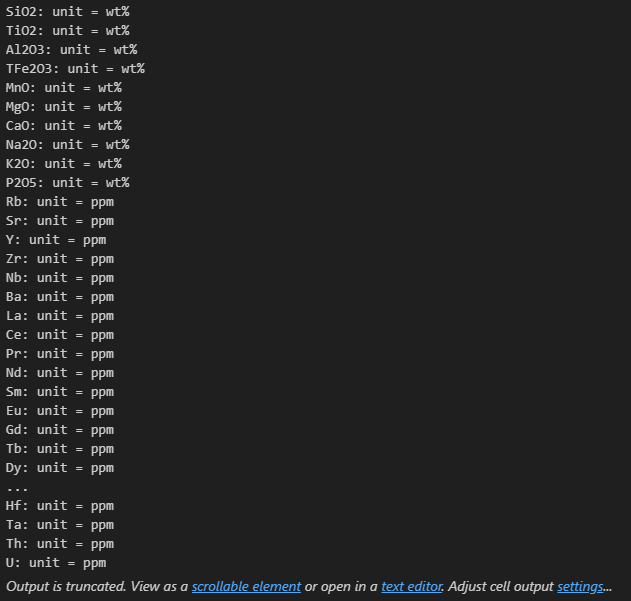


图2.2 单位映射试运行结果图

文本

AI 生成的内容可能不正确。

图2.3 校验验试运行结果图

**2.1.3** 缺失值处理

设置专属缺失值处理模块，自动检测所有数值列，并支持检测指定列；对缺失值支持“填0”“填中位数”处理。试运行展示见图2.4。

***亮点自述：***

1. 支持自动检测所有列以及指定列
2. 支持多种方式处理缺失值
3. 准确反馈状态和日志

文本

AI 生成的内容可能不正确。

图2.4 缺失值处理试运行结果图

**2.1.4** CLR变换

针对数据进行CLR变换，消除比例型数据的闭合效应。试运行展示见图2.5。

***亮点自述：***

1.在功能模块中实现单位自动统一、CLR变换、常规对数变换（任意底数）

2.不改变原始数据，可对针对专属列进行操作

3.支持元数据跟随转换，不会“换算了数值却忘了改单位”

4.增加非正值处理，防止脏数据报错（参数可兼容科研中零/极小异常值）

5. 同步索引，适合后续直接拼接

图形用户界面

AI 生成的内容可能不正确。

图2.5 CLR变换试运行结果图

**2.1.5** 异常值标记

针对数据进行检测标记，按Z分数阈值3.0检测异常值。试运行展示见图2.6。

***亮点自述：***

1. 检测双侧异常（取绝对值）
2. 定义专属列标记异常值

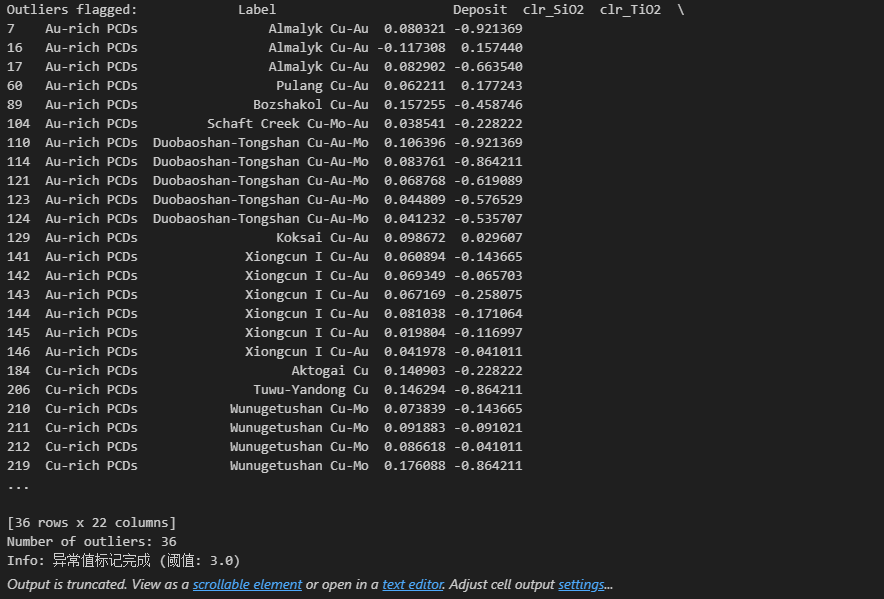


图2.6 异常值标记试运行结果图

**2.2** 探索性分析

**2.2.1** 散点矩阵图

为分析元素对之间的线性、非线性关系、不同类别样品分布规律、区分度，以及快速检查异常，选择了SiO2,TiO2,Al2O3,TFe2O3,MnO,MgO,CaO,Na2O,K2O,P2O5，十种元素生成散点矩阵，见图2.7。

由图可以看出：Au-rich和Cu-rich PCDs在部分组合上可以明显区分，MgO、CaO、SiO2与其它变量组合的散点关系较强，可能主控样品的分组，并且可以看到部分子图中有孤立的点。



图2.7 散点矩阵

**2.2.2** 相关性热力图

为分析元素之间整体相关性结构，尽可能找出可能的矿物组合、分带、迁移趋势，分别使用皮尔逊和斯皮尔曼法对上述元素生成相关性热力图，见图2.8、2.9。

根据两张相关性热力图，可以发现：大多数主量氧化物之间存在明显相关性，但相关性强弱、正负方向有差异。其中TiO2与MgO、TFe2O3，以及MgO与TFe2O3三者之间相关性最强（皮尔逊0.83/0.79，斯皮尔曼0.85/0.82），表明这些元素（及其组合）在样品中共变性强，可能源于同一地球化学过程、同源分带或矿物共生；SiO2与TiO2、TFe2O3、MgO等，都有很强的负相关（皮尔逊-0.78/-0.78/-0.75，斯皮尔曼-0.75/-0.81/-0.78），这显示SiO₂的增多通常伴随这些元素的减少，可能反映了地球化学分异、岩浆演化或不同矿床成因。

图表, 条形图, 树状图

AI 生成的内容可能不正确。

图2.8 皮尔逊法-相关性热力图

图表, 条形图, 树状图

AI 生成的内容可能不正确。

图2.9 斯皮尔曼法-相关性热力图

**2.2.3** PCA双标图

为查看样品是否有天然分群、查看哪些元素主导主要变异，分析可疑异常样品，生成PCA双标图，见图2.10。

由图可知，Au与Cu两类样品在主成分空间可见明显分布趋势，PC1方向上有判别能力。

图表, 散点图

AI 生成的内容可能不正确。

图2.10 PCA双标图

**2.2.4** 地球化学比值图

为判断分异趋势、成因路径，查看组分分布是否有分段、突变等情况，生成地球化学比值图，见图2.11。

左图可知：大部分数据点的K2O/Na2O分布在0到3之间，仅有少量离群值；SiO2含量从50%到75%不等，K2O/Na2O的变化主要集中在SiO2 60%到70%区间。说明这两类PCDs在K2O/Na2O与SiO2的关系上区分度有限，即K/Na比不能单独很好地区分这两类矿床类型。

右图可知：Sr/Y的值分布很广，大部分样品集中在0到120之间，也有少数极高离群值；样品点在SiO2 60%到70%范围内分布最密集，Sr/Y在这个区间内有较大的变异。总体来说，Sr/Y和SiO2的联合分布也未表现出特别明显的类型分带。

综上所述，两张比值图揭示了K2O/Na2O和Sr/Y这两个常用地球化学比值随SiO2含量的变化特征。

图表, 散点图

AI 生成的内容可能不正确。

图2.11 地球化学比值图

**2.3** 模型训练准备

**2.3.1** 单位均同

为保证所有后续数学分析、统计、机器学习和深度学习处理前提成立。使不同来源、数量级的元素数据可以公平、准确地参与模型训练和解释，必须要进行单位均同，运行展示见图2.12。

屏幕上有字

AI 生成的内容可能不正确。

图2.12 单位均同运行结果图

**2.3.2** 对数变换

为消除数量级影响，减小极差，避免训练时出现“强信号被噪声淹没”，使模型训练和特征评估更可靠，对数据进行对数变换，运行展示见图2.13。

***亮点自述：***

1.在变换过后，对异常值进行二次检测与去除

图形用户界面, 文本

AI 生成的内容可能不正确。

图2.13 对数变换运行结果图

**2.3.3** 去异常值前后对比

为消除数量级影响，减小极差，避免训练时出现“强信号被噪声淹没”，使模型训练和特征评估更可靠，对数据进行对数变换，运行展示见图2.13。

# 参 考 文 献

[1] Roncoroni, G., Forte, E., Santin, I., Černok, A., Rajšić, A., Frigeri, A. L. E. S. S. A. N. D. R. O., & Pipan, M. (2024). High frequency Lunar Penetrating Radar quality control, editing and processing of Chang’E-4 lunar mission. Scientific Data, 11(1), 118.

[2] Cao, H., Xu, Y., Xu, L., Zhang, L., Bugiolacchi, R., & Zhang, F. (2023). From Schrödinger to Von Kármán: An Intriguing New Geological Structure Revealed by the Chang'e‐4 Lunar Penetrating Radar. Geophysical Research Letters, 50(2), e2022GL101413.

# 总 结

通过本次期末报告，完成了雷达数据处理领域的入门及优化，在准备课上Pre过程也将许多雷达专业术语以及领域内不足的知识查明弄清，也养成了查文献读文献等相关习惯的培养。

感恩老师一学期的辛苦授课与课上的答疑解惑，祝老师心想事成、诸事顺利、论文顺利，也提前祝老师假期愉快！