

参赛承诺书

提交包含此承诺书的 pdf 文件，表明所有此文件的作者共同承诺：

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权北京理工大学数学建模竞赛组织方，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

2024. 4

基于二级模糊综合测评法的水质监测模型

摘要

本文针对北京地区水体水质分布和影响因素问题，考虑了**水温、pH值、溶解氧、电导率、浊度指标对高锰酸盐指数、氨氮、总磷、总氮**的影响，给出了从**居民、游客**角度出发对水质的新分类，采取**三次插值法、主成分分析法和二级模糊综合测评法**等多种建模方法对上述问题进行分析。结合客观因素和问题的结论，本文最终反映了北京地区水体水质受**多种**因素影响、水质空间分布差异大和普遍水质较差的问题。

关键词：三次插值法、主成分分析法、综合评价、相关性、二级模糊综合测评法

基于二级模糊综合测评法的水质监测模型

一、问题背景

随着北京地区城市化进程的持续推进、经济活动的不断繁荣以及人口密度的显著增长，人类活动对自然环境的影响日益凸显，特别是对水资源的污染和破坏问题日益严重^{[1][2]}。作为城市的生命线，水质的好坏直接关系到居民的生活质量和城市的可持续发展。因此，对北京地区水系的水质状况进行全面、深入的了解和分析，显得尤为重要。

我国水污染防治发展大致分为 4 个阶段。其中，第三阶段是在 2012 年 2020 年，主要是针对排放的氮(Nitrogen, N)、磷(Phosphorus, P) 等指标开展针对性治理，这一时期进一步提高了污水处理排放标准^[3]。近年来，随着环保意识的提高和科技手段的进步，水质监测和数据分析成为了水质管理和保护的重要手段。通过设立地表水水质自动监测站，我们可以利用站中的传感器技术、数据采集系统和数据传输系统，实时获取大量、准确的水质数据，这些数据不仅反映了水质的现状，也为水质变化趋势的预测和污染源的追溯提供了有力支持。

按照地理位置和水系分布，北京地区的水系被划分为若干个监测区域，每个区域包含若干个监测站点。每个监测站点都提供了详细的水质数据，这些数据反映了该站点所在区域的水质状况。

在此背景下，我们获得了 2024 年 3 月 22 日 16 时北京地区 19 个地表水水质自动监测站发布的数据。这些数据不仅涵盖了多个监测站点的水质指标，包括高锰酸盐指数、氨氮、总磷和总氮等关键参数，还提供了丰富的地理和时间信息，为我们全面了解北京地区水系的水质状况提供了宝贵的数据支持。

附件中列出了 19 个地表水水质自动监测站有关断面名称、水质类别、水温、pH 值、溶解氧、电导率、浊度、高锰酸盐指数、总磷、氨氮、总氮的水质监测数据。

基于附件中所提供的数据，本团队建立出多个数学模型，解决下列问题：

问题 1 分析北京市所在地区水系之中高锰酸盐的指数、总磷、氨氮、总氮的分布规律，得出其对应的总体分布特征。

问题 2 分析问题 1 中四种元素的高低，以及其分别与哪些影响因素具有何种相关性，并根据所给数据探究现有分布特征的形成原因。

问题 3 推翻国家对水质的分类标准，以居民和游客的生活体验为标准，给出所考虑的有关依据，将各水质监测站附近的水质进行重新分类。

二、问题分析

随着全球环境的日益变化和社会工业化水平的快速提升，水资源在社会发展中所占据的比重愈发加大。其中，水质的好坏更是直接影响着人民的日常生活和社会的持续发展。本文根据北京地区的地表水水质自动监测站所发布的数据，具体研究北京地区水系问题。

2.1 问题一的分析

对于问题一，经过对题目的理解和对数据的分析，我们采用数据分析，计算高锰酸盐指数、氨氮、总磷、总氮的平均值、中位数、标准差等统计量，以了解其在整个水系中的集中趋势和离散程度。其次，利用 python，采取三次插值法直观地展示出四个指标的分布情况。最后，结合统计量和可视化结果，分析这些指标在不同监测站之间的差异性，以及可能存在的空间分布规律。

2.2 问题二的分析

对于问题二，考虑到题目中仅提供了有限的数据项目，我们将主要分析这些指标与水温、pH 值、溶解氧、电导率、浊度等水质参数之间的相关性。首先使用主成分分析法对水域进行综合评价。通过主成分的主要元素因子和水域中高锰酸盐指数、氨氮、总磷、总氮的载荷矩阵和各个因子的累计贡献率以及百分比来具体分析各水质参数对元素浓度的主要原因。

2.3 问题三的分析

对于问题三，我们采取二级模糊综合测评法，将定量和定性分析结合起来。从居民、游客生活体验的角度出发分析选取指标时，先查阅大量文献资料，了解到在北京地区居民和游客更注重水质的具体方面，以及起受到什么因素的影响，得到一级指标和二级指标；再结合附件所给数据对因素的影响确定各因素的评语集。一般而言，各因素在综合评价中所起的作用是不同的，我们由此根据 Delohi 法和熵权法确定一个各因素之间的权重分配，然后确定模糊综合判定矩阵，最后得出对 19 个水体水质的综合新评判。

本文的研究路径如图 1 所示：



图 1 研究路径

三、模型假设

1. 假设所给的采样点的数据都是可靠准确的。
2. 不考虑元素间的相互作用的影响。
3. 假设短期内元素的物理、化学变化水域影响不大。
4. 不考虑元素历史沉积的影响。
5. 假设从网站上搜查的各地经纬度以及指标数据都是真实的。

四、模型构建与求解

4.1 问题1——北京地区水系内高锰酸盐的指数、总磷、氨氮、总氮的总体分布特征

4.1.1 研究四种元素在整个水系中的集中趋势和离散程度

经过对题目和数据的分析，我们采用数据分析计算高锰酸盐指数、氨氮、总磷、总氮的平均值、中位数、标准差等统计量，得出其在整个水系中的集中趋势和离散程度。

平均值公式：

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_i \quad (i=1, 2, \dots, n; j=1, 2, 3, 4; n=19) \quad (1)$$

标准差公式：

$$s_j^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (i=1, 2, \dots, n; j=1, 2, 3, 4; n=19) \quad (2)$$

方差公式：

$$s_j = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (i=1, 2, \dots, n; j=1, 2, 3, 4; n=19) \quad (3)$$

其中，j 分别对应高锰酸盐指数、氨氮、总磷、总氮浓度；i 分别对应断面。

所得表 1：

表 1 高锰酸盐指数、氨氮、总磷、总氮的统计量

	高锰酸盐指数 (mg/L)	氨氮 (mg/L)	总磷 (mg/L)	总氮 (mg/L)
总数	19	19	19	19
中位数	3.210	0.026	0.039	5.490
平均值	3.124	0.119	0.041	5.606
标准差	1.466	0.248	0.034	3.292
方差	1.210	0.498	0.184	1.814
最小值	0.900	0.025	0.005	1.190
0.250	2.145	0.025	0.005	2.770
0.500	3.210	0.026	0.039	5.490
0.750	4.215	0.070	0.059	7.375
最大值	5.830	1.053	0.108	12.120

根据表 1 可得，北京地区地表水中含量最多的为总氮，其余从多到少依次为高锰酸盐、总磷、氨氮。根据得出的方差和标准差可知，总氮和高锰酸盐含量数据较分散，说明这两种元素在北京地区水系中的分布差异较大；而氨氮和总磷含量数据较为集中，说明其在北京地区水系中分布差异相对较小。

4.1.2 研究四种元素各自分布情况

通过查询地表水水质监测系统，得到各个地区断面采点的经纬度，如下表所示：

表2 断面采点经纬度

断面名称	经度	纬度
鼓楼外大街	116.39537	39.956847
广北滨河路（桥）	116.44061	39.899404
沙窝	116.03895	39.757200
新八里桥	116.618333	39.907408
大红门闸上	116.425261	39.834406
南大荒桥	116.187123	39.890237
清河闸	116.349049	40.026901
花园路	116.365688	39.977668
白石桥	116.32528	39.941767
码头	116.305631	40.412399
罗庄	116.574886	39.654948
小屯桥	116.903395	39.735711
怀柔水库	116.592811	40.309662
东店	116.526519	39.673915
密云水库	116.913914	40.517502
大关桥	116.803076	40.563889
辛庄桥	116.298616	40.094418
谷家营	115.959848	40.430962
后城	116.633775	40.319538

接着，我们借助 python，采取三次插值法画出高锰酸盐指数、氨氮、总磷、总氮在北京地区水系的的空间分布图，由颜色浓度来反映各元素分布情况：

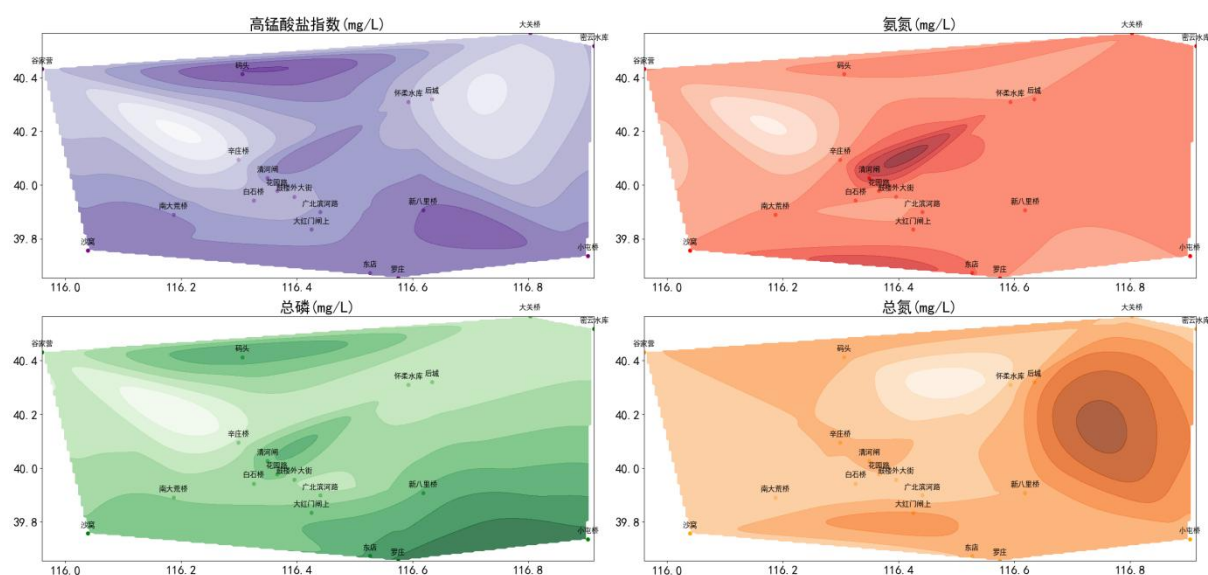


图2 四种元素分布情况等高线图

由上图可直观感受到北京地区水系中高锰酸盐指数、氨氮、总磷、总氮的总体分布特征：高锰酸盐指数南北高、东西低；氨氮含量中部高、四周低；总磷含量南北高、东西低；总氮含量东部高、中西部偏低。值得一提的是，我们发现高锰酸盐指数与总磷含量的空间分布有较强的相似性。高锰酸盐指数可以直接反映水中受还原性物质污染的过程，包括有机物、亚硝酸盐、亚铁盐、硫化物等；总磷是反映水体中总的磷化合物数量总和的指标，也导致水体富营养化和水质变差的重要因素，因此二者在分布上有相似度，符合客观事实规律。

4.2 问题 2——分析四种元素的含量高低及与影响因素的相关性

问题思路：使用主成分分析法对水域进行综合评价。通过计算主成分的主要元素因子和水域中高锰酸盐指数、氨氮、总磷、总氮的载荷矩阵，以及各个因子的累计贡献率和百分比，具体分析各水质参数对元素浓度产生影响的主要原因。

4.2.1 模型的建立

1. 对原始数据进行标准化处理

假设进行主成分分析的指标变量为水温、pH、溶解氧、电导率、浊度： x_1, x_2, x_3, x_4, x_5 ；共有 4 个评价对象：高锰酸盐指数、氨氮、总磷、总氮；第 i 个评价对象的第 j 个指标的取值为 a_{ij} 。将各指标值 a_{ij} 转换成标准化指标 \widetilde{a}_{ij} ，有式子：

$$\widetilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j}, i=1,2,3,4; j=1,2,3,4,5 \quad (4)$$

$$\mu_j = \frac{1}{n} \sum_{i=1}^n a_{ij} \quad (5)$$

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (a_{ij} - \mu_j)^2} \quad (6)$$

其中， $j=1, 2, 3, 4, 5$ ，即 μ_j 和 s_j 为第 j 个指标的样本均值和样本标准差。

对应的，称式（7）为标准化指标变量。

$$\widetilde{x}_j = \frac{x_j - \mu_j}{s_j}, j=1,2,3,4,5 \quad (7)$$

2. 计算相关系数矩阵 R

相关系数矩阵 $R=(r_{ij})_{5 \times 5}$ 有

$$r_{ij} = \frac{\sum_{k=1}^n \widetilde{a}_{ki} \cdot \widetilde{a}_{kj}}{n-1}, i, j=1,2,3,4,5 \quad (8)$$

其中， $r_{ii} = 1$ ， $r_{ij} = r_{ji}$ ， r_{ij} 为第 i 个指标与第 j 个指标的相关系数。

3. 计算特征值和特征向量

计算相关系数矩阵 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_5 \geq 0$, 及对应的标准正交特征向量 u_1, u_2, \dots, u_5 , 其中 $u_j = [u_{1j}, u_{2j}, \dots, u_{mj}]^T$, 由特征向量组成 m 个新的指标变量:

$$\begin{aligned} y_1 &= u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + \dots + u_{51}\tilde{x}_5, \\ y_2 &= u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \dots + u_{52}\tilde{x}_5, \\ &\vdots \\ y_5 &= u_{15}\tilde{x}_1 + u_{25}\tilde{x}_2 + \dots + u_{55}\tilde{x}_5, \end{aligned} \quad (9)$$

其中, y_1 为第一主成分, y_2 为第二主成分, \dots , y_5 为第五主成分。

4. 选择 p ($p < 5$) 个主成分, 计算综合评价值

1) 计算特征值 λ_j ($j = 1, 2, \dots, 5$) 的信息贡献率和累积贡献率。称式 (10) 为主成分 y_j 的信息贡献率,

$$b_j = \frac{\lambda_j}{\sum_{k=1}^5 \lambda_k}, j = 1, 2, \dots, 5 \quad (10)$$

同时, 有式 (11) 主成分 y_1, y_2, \dots, y_p 的累积贡献率:

$$\alpha_p = \frac{\sum_{k=1}^p \lambda_k}{\sum_{k=1}^5 \lambda_k} \quad (11)$$

当 α_p 接近于 1 (一般取 $\alpha_p = 0.85, 0.90, 0.95$) 时, 则选择前 p 个指标变量 y_1, y_2, \dots, y_p 作为 p 个主成分, 代替原来 5 个指标变量, 从而可对 p 个主成分进行综合分析。

2) 计算综合得分:

$$Z = \sum_{j=1}^p b_j y_j \quad (12)$$

其中, b_j 为第 j 个主成分的信息贡献率, 根据综合得分值就可进行评价。

4.2.2 模型的求解

经过对北京水域中水温、pH、溶解氧、电导率、浊度的数据标准化处理后利用 python 软件进行主成分分析, 可得出如下结果:

表 3 变量相关矩阵

	水温(°C)	pH(无量纲)	溶解氧(mg/L)	电导率(μ S/cm)	浊度(NTU)
水温(°C)	1.000	-0.147	0.120	0.599	0.062
pH(无量纲)	-0.147	1.000	0.688	0.274	-0.184
溶解氧(mg/L)	0.120	0.688	1.000	0.405	-0.122
电导率(μ S/cm)	0.599	0.274	0.405	1.000	0.331
浊度(NTU)	0.062	-0.184	-0.122	0.331	1.000

由此可得相关系数矩阵热图：

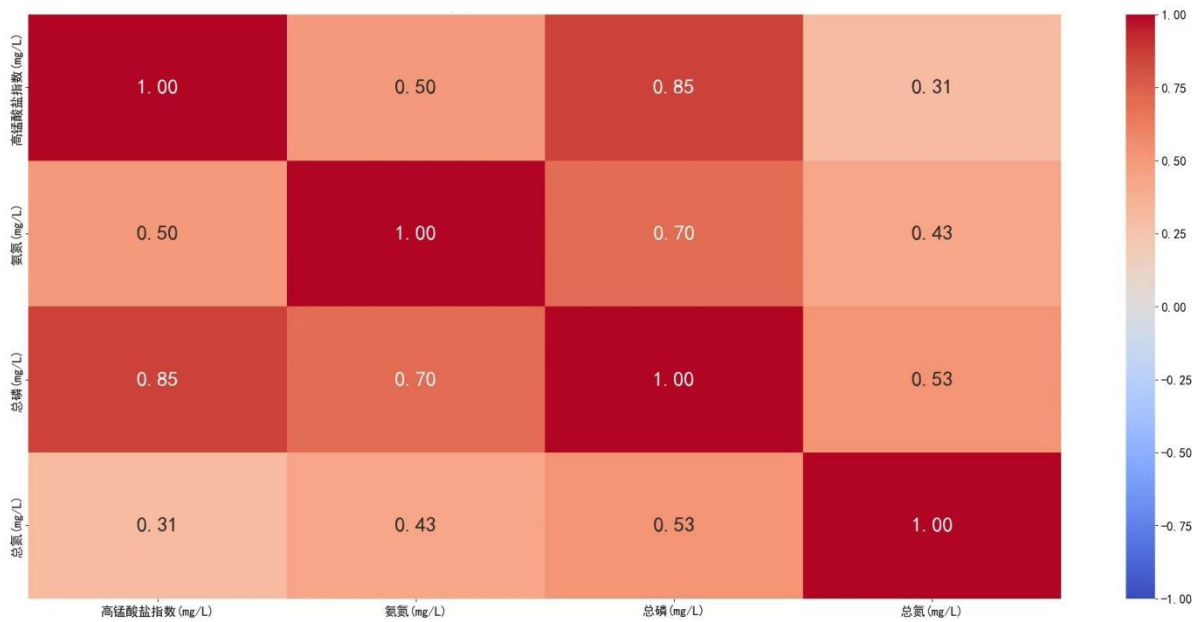


图 3 相关系数矩阵热图

再利用 python 软件统计可得前几个因子的累积贡献率和初始特征值，如表 4 所示：

表 4 解释的总方差

初始特征值				提取平方和载入		
主成分序号	特征值	方差贡献率	累计方差贡献率	特征值	方差贡献率	累计方差贡献率
0	2.028	0.406	0.406	2.028	0.406	0.406
1	1.570	0.314	0.720	1.570	0.314	0.720
2	0.916	0.183	0.903	0.916	0.183	0.903
3	0.300	0.060	0.963			
4	0.186	0.037	1.000			

由表 3 可见，前三个主成分分别累计提取了总方差的 90.3%，表明所提取的主成分能较好地代表源数据所蕴涵的信息，主成分分析效果很好。将前三个主成分进行综合评价，其对应的特征向量见表 5：

表 5 特征向量

主成分序号	x_1	x_2	x_3	x_4	x_5
0	-0.331	-0.483	-0.579	-0.565	-0.046
1	0.519	-0.484	-0.309	0.385	0.503
2	-0.571	0.210	0.048	0.040	0.791

以及得出的载荷矩阵如下表所示：

表 6 因子载荷系数表

	因子载荷系数			共同度（公因子方差）
	主成分 1	主成分 2	主成分 3	
水温($^{\circ}\text{C}$)	0.471	0.65	-0.547	0.944
pH(无量纲)	0.688	-0.606	0.201	0.881
溶解氧(mg/L)	0.825	-0.387	0.046	0.833
电导率($\mu\text{S}/\text{cm}$)	0.805	0.482	0.039	0.882
浊度(NTU)	0.066	0.63	0.757	0.974

由此可得 3 个主成分分别为：

$$\begin{aligned}
 y_1 &= -0.331\tilde{x}_1 - 0.483\tilde{x}_2 + \cdots - 0.046\tilde{x}_5, \\
 y_2 &= 0.519\tilde{x}_1 - 0.484\tilde{x}_2 + \cdots + 0.503\tilde{x}_5, \\
 y_3 &= -0.571\tilde{x}_1 + 0.210\tilde{x}_2 + \cdots + 0.791\tilde{x}_5,
 \end{aligned} \tag{13}$$

分别由三个主成分的贡献率为权重，构建主成分综合评价模型：

$$Z = 0.406y_1 + 0.314y_2 + 0.183y_3 \tag{14}$$

同时计算得出各成分回归系数如表 7：

表 7 主成分回归系数

	高锰酸盐	氨氮	总磷	总氮
水温($^{\circ}\text{C}$)	0.00133	0.10500	0.00282	0.04419
pH(无量纲)	-0.00840	-0.13977	0.00945	-0.03663
溶解氧(mg/L)	-0.00764	-0.10599	0.01000	-0.02316
电导率($\mu\text{S}/\text{cm}$)	0.00223	0.00077	0.00276	0.00163
浊度(NTU)	0.01234	-0.00866	-0.01020	-0.02722

并通过数据分析得到五个指标对四种元素浓度的影响程度，如图 4：

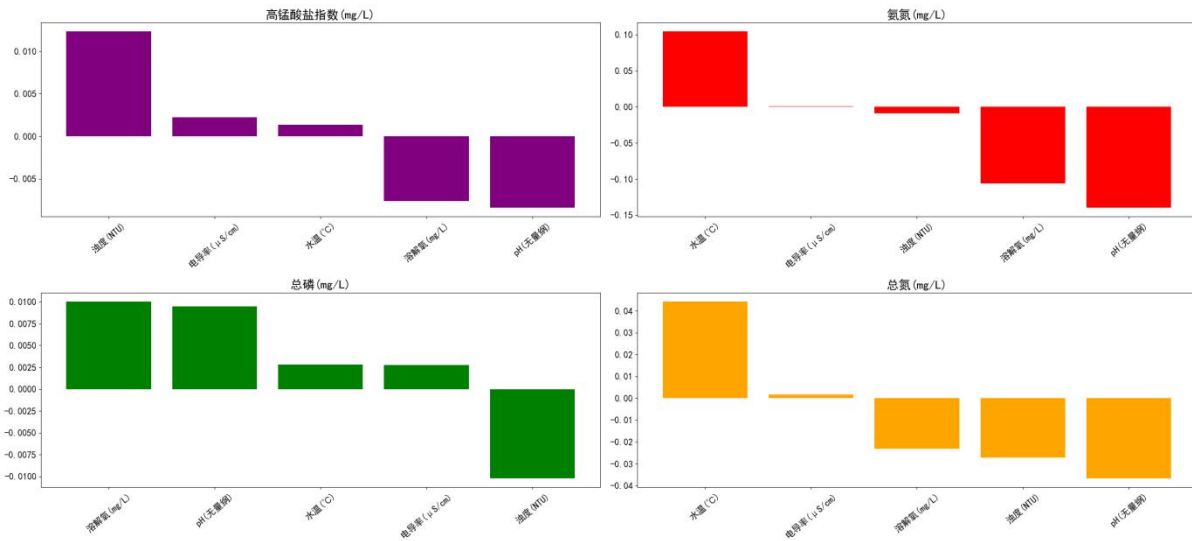


图 4 指标元素浓度的影响因素

综上结合客观事实并分析：北京地区水域中高锰酸盐指数受河流浊度影响最大，因为图 4 的左上图表“高锰酸盐指数”中，浊度明显高于其余四个指标；根据所查阅的文献，浊度可以使光散射或者吸收^[4]，且河流浊度增加所带来的颗粒物和有机物在河流中与水中的锰离子发生反应，导致高锰酸盐指数升高^[5]。氨氮和总氮含量受水温影响较大，因为图 4 的右方两个图表“氨氮”与“总氮”中，水温均明显高于其余四个指标；较高的水温增强了微生物的活性，从而加速氨氮和总氮的转化和分解过程^[6]，且水温升高也会促进水体中的硝化作用，使得氨氮转化为硝酸盐，从而影响总氮的含量^[7]。总磷浓度受溶解氧和 pH 值影响较大，因为图 4 的左下图表“总磷”中，溶解氧与 pH 值均明显高于其余三个指标；且据文献资料显示，溶解氧的含量会影响水体中微生物对磷的利用和转化过程，进而影响总磷的浓度^[8]；而 pH 值的变化则会改变磷的化学形态和存在状态，比如在较低的 pH 值下，磷酸盐更容易以固态形式沉淀下来，从而降低总磷浓度。

4.3 问题 3——对各个监测站的水质进行重新分类

问题思路：借助于二级模糊综合测评法，我们从居民、游客生活体验的角度出发，分析选取指标。我们首先查阅大量文献资料，了解到在北京地区居民和游客更注重水质的哪些具体方面，以及其受到什么因素的影响，得到一级指标和二级指标；再结合附件所给数据对因素的影响确定各因素的评语集。接着根据 Delphi 法和熵权法确定一个各因素之间的权重分配，然后确定模糊综合判定矩阵，最后得出对 19 个水体水质的综合新评判。

4.3.1 模型的建立

1. 确定因素集和子因素集

将因素集 $U = \{u_1, u_2, \dots, u_n\}$ 按某种属性分成 s 个子因素集 U_1, U_2, \dots, U_s ，其中 $U_i = \{u_{i1}, u_{i2}, \dots, u_{in_i}\}$, $i = 1, 2, \dots, s$ ，且满足：

1. $n_1 + n_2 + \dots + n_s = n$;
2. $U_1 \cup U_2 \cup \dots \cup U_s = U$;
3. 对任意的 $i \neq j$, $U_i \cap U_j = \phi$ 。

2. 确定各因素的权重

一般情况下，因素集中的各因素在综合评价中所起的作用是不相同的，综合评价结果不仅与各因素的评价有关，而且在很大程度上还依赖于各因素对综合评价所起的作用，这就需要确定一个各因素之间的权重分配，它是 U 上的一个模糊向量。

3. 综合评判

对于每一个因素集 U_i ，分别做出综合评判。设 $V = \{v_1, v_2, \dots, v_m\}$ 为评语集， U_i 中各因素相对于 V 的权重分配是 $A_i = [a_{i1}, a_{i2}, \dots, a_{in_i}]$ 。

若 \widetilde{R}_i 为单因素评判矩阵，则得到一级评判向量：

$$B_i = A_i \cdot \widetilde{R}_i = [b_{i1}, b_{i2}, \dots, b_{im}], i = 1, 2, \dots, s \quad (15)$$

将每个看作为一个因素，记为：

$$K = \{u_1, u_2, \dots, u_s\}. \quad (16)$$

这样， K 又是一个因素集。 K 的单因素评判矩阵为：

$$R = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_s \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{s1} & b_{s2} & \dots & b_{sm} \end{bmatrix} \quad (17)$$

每个 U_i 作为 U 的一部分，反映了 U 的某种属性，可以按它们的重要性给出权重分配 $A = [a_1, a_2, \dots, a_s]$ ，于是得到二级评判向量：

$$B = A \cdot R = [b_1, b_2, \dots, b_m] \quad (18)$$

4.3.2 模型的求解

经过查阅《居民生活用水舒适度调查研究》^[9]、《探险旅游中游客情绪满意度、情绪动态变化及其影响因素》^[10]、《北京市汛期降雨对清河水质影响及水环境保障对策研究》^[11]等文献，并结合附件中各要素的性质，我们将9个指标分为舒适度（ U_1 ）、安全性（ U_2 ）、清洁度（ U_3 ）这3个子因素集。根据各二级指标对一级指标的增益性，将其分为1-5五个等级，其中5表示增益性最好。为方便呈现，将19个地区名称简写为1, 2, ..., 19, 如表8所示：

表8 评价

一级指标	二级指标	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
舒适度	水温	3	2	3	4	5	2	5	4	3	4	4	3	2	3	1	1	2	2	2
	浊度	2	3	3	5	2	1	3	4	2	1	5	4	4	2	5	3	5	3	4
	溶解氧	4	4	4	3	3	3	5	4	4	2	2	1	3	3	3	4	3	2	3
安全性	高锰酸盐指数	3	3	4	5	3	4	3	3	2	5	4	4	2	4	3	2	1	2	1
	电导率	4	2	3	3	2	2	3	3	4	2	1	1	5	2	5	4	3	3	4
	pH 值	3	2	5	2	5	3	5	4	4	4	2	1	3	4	3	3	3	3	3
	氨氮	4	5	3	3	2	4	1	2	5	4	3	5	4	2	4	4	4	4	4
清洁度	氨氮	4	5	3	3	2	4	1	2	5	4	3	5	4	2	4	4	4	4	4
	溶解氧	4	4	4	3	3	3	5	4	4	2	2	1	3	3	3	4	3	2	3
	总磷	4	4	3	3	3	3	3	3	4	2	1	1	5	2	5	5	5	5	5
	总氮	4	5	3	2	1	5	3	3	4	3	2	3	5	3	5	3	3	3	1

由 Delphi 法设定一级权重指标为： $A = [0.2, 0.4, 0.4]$ 。

由熵权法设定二级权重指标：

表9 二级权重指标

权重类别	水质类别	水温	pH	溶解氧	电导率	浊度	高锰酸盐指数	氨氮	总磷	总氮
全部权重	0.079	0.039	0.027	0.052	0.073	0.12	0.055	0.374	0.113	0.068
舒适度权重		0.184		0.246		0.57				
安全性权重			0.051		0.138		0.105	0.706		
观赏性权重				0.085				0.617	0.186	0.112

$$A_1 = [0.184, 0.570, 0.246].$$

$$A_2 = [0.105, 0.138, 0.051, 0.706]. \quad (19)$$

$$A_3 = [0.617, 0.085, 0.186, 0.112].$$

对各个元素集进行模糊综合评判，再根据最大隶属度原则，可以得到从居民、游客生活体验的角度出发分析下的 19 个地区的水质综合评判数值，其中数值越高，表明该地区水质越好。通过 excel 排序得到水质的新分类，等级 1 表示该地区水质最好如下表 10：

表 10 水质新分类

	舒适度	安全性	清洁度	总分	地名	标准化	等级
2	3.062	4.222	4.729	4.193	广北滨河路	1	1
9	2.676	4.496	4.617	4.181	白石桥	0.9935618	1
15	3.772	3.982	4.011	3.952	密云水库	0.87235464	1
13	3.386	3.877	4.011	3.832	怀柔水库	0.809302695	1
12	3.078	4.137	3.691	3.747	小屯桥	0.764047896	2
17	3.956	3.496	3.787	3.705	辛庄桥	0.741645027	2
1	2.676	3.844	4.000	3.673	鼓楼外大街	0.724723677	2
16	2.878	3.739	3.973	3.661	大关桥	0.71831176	2
19	3.386	3.634	3.564	3.556	后城	0.663178016	2
18	2.57	3.601	3.601	3.395	谷家营	0.5776363	3
6	1.676	3.672	3.841	3.34	南大荒桥	0.54874035	3
4	4.324	3.158	2.888	3.283	新八里桥	0.518534718	3
10	1.798	3.828	3.346	3.229	码头	0.489928155	3
3	3.246	3.208	3.186	3.207	沙窝	0.477978033	3
11	4.078	2.777	2.532	2.939	罗庄	0.336342107	4
8	4.000	2.346	2.569	2.766	花园路	0.244642445	4
14	2.43	2.312	2.298	2.33	东店	0.013868496	5
5	2.798	2.259	2.159	2.327	大红门闸上	0.012239917	5
7	3.86	1.691	2.138	2.304	清河闸	0	5

由表中的舒适度、清洁度、安全性占各水体的权重进行综合打分并排序，分数越高的水体水质越好，最后将 19 个水体的水质按分数分成 5 类，1 类水体水质最好。

最后，我们用 Python 实现了自动化输出等级，对上述结果进行了验证。

综上所述，我们从居民、游客生活的角度，结合水温、pH、溶解氧、电导率、浊度、高锰酸盐指数、氨氮、总磷、总氮指标，以舒适度、清洁度、安全性为大分类依据将 19 个地区水质分为 5 类。其中，最好的一类水质水体有广北滨河路、白石桥、密云水库；二类水质水体有小屯桥、辛庄桥、鼓楼外大街、大关桥、后城；三类水质水体有谷家营、南大荒桥、新八里桥、码头、沙窝；四类水质水体有罗庄、花园路；五类水质水体有东店、大红门闸上、清河闸。根据附件中所给出的水质类别看，从居民、游客生活的角度分析与从水质本身对人和环境的影响角度分析整体上没有较大差别，仅在细小因素考虑权重中有不同，符合客观事实规律。

五、模型结果分析

对于问题一，我们采用数据分析，通过计算各元素的平均值、中位数、标准差等统计量，了解了其在整个水系中的集中趋势和离散程度，具有较直观和真实的现实意义。其次，使用三次插值法展示出四个指标的分布情况，将统计量和结果可视化，说明了北京地区水系中高锰酸盐指数、氨氮、总磷、总氮的总体分布特征，对现实中研究北京地区水系各元素含量分布变化提供参考指标

对于问题二，我们使用了主成分分析法对水域进行综合评价。通过主成分的主要元素因子和水域中高锰酸盐指数、氨氮、总磷、总氮的载荷矩阵和各个因子的累计贡献率以及百分比具体得出了五个指标对四种元素浓度的影响程度，在参考文献的辅助下我们发现结论符合实际情况，同时，这一研究分析结果对未来调节治理水域有重要参考价值

对于问题三，我们采取二级模糊综合测评法，将定量和定性分析结合起来，得出对 19 个水体水质的综合新评判。这一判定与“水质本身对人和环境的影响”角度相比仅存在微小差别，符合客观事实。同时，从居民和游客生活的角度出发，对水质进行的新分类，可以为当地房地产业和旅游业的发展提供参考指标。

六、总结

6.1 模型的优、缺点评价分析

6.1.1 模型的优点

三次插值法：

三次插值法具有较高的精度，可以通过拟合曲线来逼近原始数据。同时它具有广泛的适用性，可以适用于离散和连续的函数。并且它可通过调整插值节点的数量和位置，可以对插值结果进行灵活控制，以满足不同的需求，有很好的可控性。

主成分分析法：

主成分分析法可以将高维数据降低到低维空间，减少数据的维度。这有助于减少数据集的复杂性和存储空间，并提高计算效率。并且通过去除相关性较低的主成分，从而去除数据中的冗余信息。这有助于减少噪声和异常值对数据分析的影响，提高模型的鲁棒性。

综合评价：

综合评价法可以根据具体情况选择不同的评价指标，并赋予不同的权重，以适应不同的评价需求。这种灵活性使得综合评价方法可以应用于各种不同领域和问题的评价。同时，将不同的评价指标转化为统一的评分体系，从而使得不同对象之间的评价结果具有可比性。这样可以方便进行排名、比较和选择。

二级模糊综合测评法：

二级模糊综合测评法能够有效地处理评价对象的不确定性和模糊性。在实际问题中，往往存在着各种不确定因素，而二级模糊综合测评法可以通过引入模糊数学的概念，对这些不确定因素进行量化和处理。除此之外它可以同时考虑多个指标对评价结果的影响。在实际问题中，往往需要综合考虑多个指标来对一个对象进行评价，而二级模糊综合测评法可以将多个指标进行加权组合，得到一个综合的评价结果。

6.1.2 模型的缺点**三次插值法：**

三次插值法在边界处的表现可能不够理想。由于插值函数是通过数据点逼近得到的，边界处的数据点较少，可能导致插值函数在边界处的准确性下降。

主成分分析法：

对异常值比较敏感。由于 PCA 是基于数据的方差来进行变换，如果数据中存在异常值，它们可能会对结果产生较大的影响。而且它对数据的预处理要求较高，特别是对数据的缩放。如果数据的尺度差异较大，需要进行适当的缩放处理，否则可能会导致降维结果不准确。

综合评价：

在综合评价中，不同指标或者不同评价者的权重确定是一个关键问题。权重的确定往往需要依赖专家判断或者统计分析等方法，但这些方法都存在一定的主观性和不确定性。在不同评价者之间可能存在一致性问题，即不同评价者对同一被评价对象的评价结果可能存在较大差异。这可能是由于评价者的主观偏好、经验差异等因素导致的。

二级模糊综合测评法：

二级模糊综合测评法需要获取大量的评价数据，包括各个指标的评价值和权重值。然而，在实际应用中，获取这些数据可能会面临困难，特别是对于一些复杂的评价问题。需要进行大量的模糊数学运算，包括模糊数的加法、减法、乘法和除法等。这些运算往往会导致计算复杂度的增加，特别是在处理大规模评价问题时。

6.2 模型的改进与推广**6.2.1 模型的改进**

三次插值法：可以与其他插值方法进行组合，如线性插值、二次插值等。通过将不同插值方法的结果进行加权平均或者逐步逼近，可以得到更准确的指标并分析出在不同监测站之间的差异性。

主成分分析法：主成分分析默认假设数据是线性相关的，但实际数据中可能存在非线性关系。在分析题目中所给指标与水温、pH 值、溶解氧、电导率、浊度等水质参数之间的相关性时，可以采用核主成分分析（Kernel PCA）等方法，将数据映射到高

维空间中进行处理，以捕捉更多的非线性结构。

综合评价：为了进行综合评价，需要收集相关数据并进行分析。我们可以改进数据收集方法、提高数据质量，以及使用更先进的数据分析技术来获取更准确的评价结果。

二级模糊综合测评法：在某些决策问题中，时间因素可能对评价结果产生重要影响。因此，在进行二级模糊综合评价时，可以考虑引入时间因素，对不同时间段的评价结果进行比较和分析，以更好地反映决策对象的动态变化。

6.2.2 模型的推广

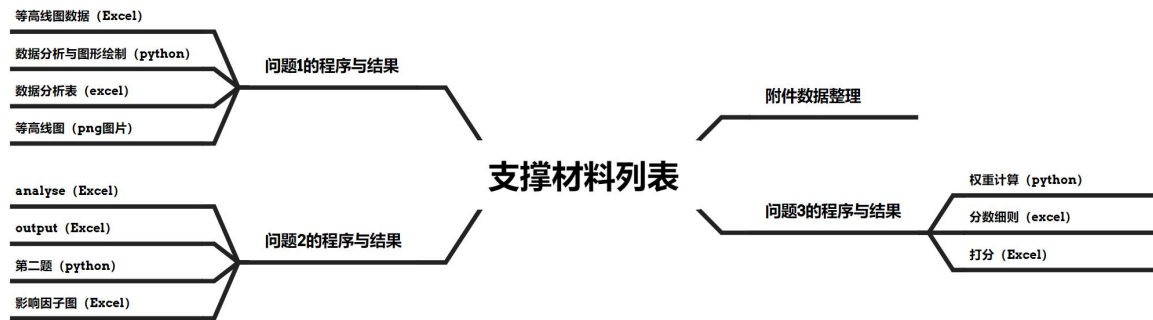
本文所采用的四个模型不仅具有广泛的适用性，能够灵活运用于不同地区的水系分析问题，而且在水污染防治方面发挥着精准有效的促进作用。通过精准判断特定地区水系的主要污染元素，我们可以实现“对症下药”，从而有针对性地制定防治措施。此外，这些模型对水体水质的深入分析，还能够为注重人文环境的产业，如旅游产业和房地产产业，提供有力的理论支持，促进这些产业的可持续发展。

七、参考文献

- [1] Pan D,Chen H, Zhang N,et al. Do livestock environmental regulations reduce water pollution in China[J]. Ecological Economics.2023,204:107637.
- [2] Bai M,Chen M,Zhang L,et al. Research on Evolutionary Game Analysis of Spatial Cooperation for Social Governance of Basin Water Pollution[J]. Water.2022,14(16):2564.
- [3] 付朝臣.城市河流水环境及水生态时空演变特征与模拟研究[D].河北工程大学,2023.
- [4] 侯顺婷.地表水中浊度与高锰酸盐指数的相关性分析[J].中国资源综合利用,2017,35(06):103-107.
- [5] 刘萍,翟崇治,余家燕,等.地表水浊度对高锰酸盐指数的影响[J].三峡环境与生态,2010,32(06):30-31+44.
- [6] 葛长宇,宋协法,毛玉泽,等.水温和体重对污损生物麦秆虫氨氮和磷排泄的影响[J].吉首大学学报(自然科学版),2008,(03):109-111+121.
- [7] 缪雨恒,俞天奇,茅学鹏.某城市地区水环境检测中总氮和氨氮的关系分析[J].山西化工,2023,43(07):206-208.
- [8] 徐闯,余香英,许泽婷,等.潭江干流溶解氧时空格局及其调控因素研究[J/OL].环境科学学报,1-9[2024-04-06].
- [9] 吴俊奇,乔晓峰,李庚,等.居民生活用水舒适度调查研究[J].给水排水,2016,52(08):84-88.
- [10] 王宇娟.探险旅游中游客情绪满意度、情绪动态变化及其影响因素[D].武汉大学,2019.
- [11] 杨莹,郝生超,张博.北京市汛期降雨对清河水质影响及水环境保障对策研究[J].水资源开发与管理,2024,10(02):46-51.

八、附录

8.1 支撑材料列表



8.2 主要程序代码

8.2.1 第一题

```

# %% [markdown]
# ## 调用

# %%
import pandas as pd
import matplotlib.pyplot as plt

excel = pd.read_excel("B 题数据.xlsx")
lon_lat = pd.read_excel('等高线图.xlsx')
x_label = excel['断面名称']
y_labels = {
    '高锰酸盐指数(mg/L)': 'purple',
    '氨氮(mg/L)': 'red',
    '总磷(mg/L)': 'green',
    '总氮(mg/L)': 'orange'
}

plt.rcParams['font.sans-serif'] = ['SimHei'] #显示中文
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号

# %% [markdown]

```

```

### 第一题
#
# 1、分析北京地区水系中高锰酸盐指数、氨氮、总磷、总氮的总体分布特征。
# *（分布特征分析）*
#
# 另外，为了绘图，我把广北滨河路（桥）的（桥）给去掉了，特此说明

# %% [markdown]
#### 数据分析

# %%
ana = excel[['高锰酸盐指数(mg/L)', '氨氮(mg/L)', '总磷(mg/L)', '总氮(mg/L)']]
print(ana.describe())

# %% [markdown]
#### 地区分布条形图

# %%
plt.figure(figsize=[27.50, 12.50])

# 遍历 y_labels 字典，为每个指标创建子图
for i, (ylabel, color) in enumerate(y_labels.items(), start=1):
    plt.subplot(2, 2, i) # 根据循环的索引选择子图位置
    plt.bar(x_label, excel[ylabel], color=color, label=ylabel)
    plt.xticks(rotation=40, fontsize=10)
    plt.title(ylabel, fontsize=20)
    plt.gca().xaxis.set_tick_params(labelrotation=40)

# 显示图例
plt.legend()

# 显示图形
plt.show()

# %% [markdown]
#### 等高线图

# %%

```

```

point_label = list(x_label)

# %%
from scipy.interpolate import griddata
# print(lon_lat['经度'])
# 定义一维数组
x = lon_lat['经度']
y = lon_lat['纬度']
point=np.vstack([x,y]).T
# print(list(point))
print(lon_lat.columns)
xx,yy=np.meshgrid(np.linspace(x.min(),x.max(),200),np.linspace(y.min(),y.max(),200))

plt.figure(figsize=[27.50, 12.50])
for i, (ylabel, color) in enumerate(y_labels.items(), start=1):
    index = lon_lat[ylabel]
    zz=griddata(points=point, values=index, xi=(xx,yy), method="cubic")
    plt.subplot(2, 2, i)
    plt.xticks(fontsize=20)
    plt.yticks(fontsize=20)
    plt.title(ylabel,fontsize=25)
    plt.scatter(x, y, c=color, s=30)
    plt.contourf(xx, yy, zz, alpha=0.75, cmap=f'{color.title()}s')
    for j, name in enumerate(x_label):
        plt.annotate(name, xy = list(point)[j], textcoords="offset points",
xytext=(0,10), ha='center', fontsize=10)
plt.tight_layout()
plt.show()

```

8.2.2 第二题

```

# %% [markdown]
# ## 调用

# %%
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

```

```

# from scipy.stats import spearmanr
import pingouin as pg
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

plt.rcParams['font.sans-serif'] = ['SimHei'] #显示中文
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号
plt.rcParams["figure.figsize"] = [27.50, 12.50]

excel = pd.read_excel("B 题数据.xlsx")

data = excel.iloc[0:,2:]
ana_ff1 = pd.DataFrame(data)

# %% [markdown]
# ## 第二题
#
# 2、分析高锰酸盐指数、氨氮、总磷、总氮的高低，与哪些因素具有相关性，仅
限给出的数据项目，并据此评估呈现当前分布特征的原因。
# *（相关性分析，评估原因）*

# %% [markdown]
# ### 主成分分析法

# %% [markdown]
# ##### 创建类

# %%
import numpy as np
from numpy import linalg

class PCA:
    """ dataset 形如 array([样本 1,样本 2,...,样本 m]),每个样本是一个 n 维的 ndarray"""

    def __init__(self, dataset):
        # 这里的参数跟上文是反着来的(每行是一个样本)，需要转置一下
        self.dataset = np.matrix(dataset, dtype='float64').T

```

```
# def standard2(self, dataset):
#     scaler = StandardScaler()
#     scaled_data = scaler.fit_transform(np.asarray(dataset))
#     # scaled_df = pd.DataFrame(scaled_data)
#     # return scaled_df
```

```
def standard(self, dataset):
```

```
    """标准化"""
```

```
    for (index, line) in enumerate(dataset):
```

```
        dataset[index] -= np.mean(line)
```

```
        # np.std(line, ddof = 1)即样本标准差(分母为 n - 1)
```

```
        dataset[index] /= np.std(line, ddof = 1)
```

```
def get_eig_vector(self, data, Cov):
```

```
    """求特征值和特征向量"""
```

```
    eigs, vectors = linalg.eig(Cov)
```

```
    # print(vectors)
```

```
    for i in range(len(eigs)):
```

```
        data.append((eigs[i], vectors[:, i]))
```

```
        # 按照特征值从大到小排序
```

```
        data.sort(key = lambda x: x[0], reverse = True)
```

```
    return eigs
```

```
def principal_comps(self, threshold = 0.85):
```

"""求主成分。threshold 可选参数表示方差累计达到 threshold 后就不再取后面的特征向量。

返回值是特征值、特征向量、方差贡献率、累计方差贡献率。"""

```
    ret = []
```

```
    data = []
```

```
    self.standard(self.dataset)
```

```
    Cov = np.cov(self.dataset)
```

```
    eigs = self.get_eig_vector(data, Cov)
```

```
    sum = 0
```

```
    for comp in data:
```

```
        sum += comp[0] / np.sum(eigs)
```

```
        ret.append(
```

```

tuple(
    map(lambda x: np.round(x, 5)
        # 特征值、特征向量、方差贡献率、累计方差贡
        献率
        , (comp[0], comp[1], comp[0] / np.sum(eigs), sum))
    )
)
if sum > threshold:
    return ret

return ret

# PMx = PCA(anal)
# op = PMx.principal_comps(threshold=0.85)

# %% [markdown]
# ##### 数据准备

# %%

#准备原始数据
anal = excel.drop(['高锰酸盐指数(mg/L)', '氨氮(mg/L)', '总磷(mg/L)', '断面名称', '总
氮(mg/L)', '水质类别'], axis=1)
index = ['水温(℃)', 'pH(无量纲)', '溶解氧(mg/L)', '电导率(μ S/cm)', '浊度(NTU)']
#输出用
output = [['高锰酸盐', '高锰酸盐指数(mg/L)'], ['氨氮', '氨氮(mg/L)'], ['总磷', '总磷
(mg/L)'], ['总氮', '总氮(mg/L)']]

# print(excel.columns)

#绘制影响因子图用
x_label = pd.read_excel('影响因子图.xlsx', sheet_name='Sheet5',
header=None).squeeze()
y_labels = {
    '高锰酸盐指数(mg/L)': 'purple',
    '氨氮(mg/L)': 'red',
    '总磷(mg/L)': 'green',
    '总氮(mg/L)': 'orange'
}

```

```

def standard(data):
    """标准化"""
    opt = (data - data.mean()) / data.std()

    return opt

# %% [markdown]
# ##### 分析

# %%
with pd.ExcelWriter('output.xlsx') as writer:
    for ii in output:
        PMx = PCA(anal)
        op = PMx.principal_comps(threshold=0.85)
        opd = pd.DataFrame(op, columns=['特征值', '特征向量', '方差贡献率', '累计
方差贡献率'])
        # print(opd)
        opd.to_excel('analyse.xlsx', index=True)
        matrix = pd.DataFrame(index=index)
        for i in range(len(op)):
            matrix[i] = op[i][1]
        #result 是主成分
        result = np.dot(anal, matrix)
        # print(result)
        model = LinearRegression()
        y = np.array(excel[ii[1]])
        x = np.array(result)
        model.fit(y=standard(y), X=x)
        result_mid = np.dot(matrix, model.coef_.T)
        pd.set_option('display.float_format', lambda x: '%.6f' % x)
        # print(result2)
        matrix['主成分回归系数'] = result_mid
        # print(matrix)
        matrix.to_excel(writer, sheet_name=ii[0], float_format='%.6f', index=True)
# print(result)

# 载荷系数=原始变量与主成分之间的相关系数
for i in range(3):

```



```

for j in index:
    load = pd.DataFrame([result[:, i], anal[j]])
    # print(load)
    print(load.iloc[0].corr(load.iloc[1]))
print("\n")

# %% [markdown]
# ##### 影响因子

# %%
for i, (ylabel, color) in enumerate(y_labels.items(), start=1):
    inf = pd.read_excel('影响因子图.xlsx', sheet_name=f'Sheet{i}',
header=None).squeeze()
    # print(list(inf))
    # print(list(x_label))
    plt.subplot(2, 2, i) # 根据循环的索引选择子图位置
    plt.bar(list(x_label[i-1]), list(inf), color=color, label=ylabel)
    plt.xticks(rotation=40, fontsize=15)
    plt.title(ylabel, fontsize=20)
    plt.gca().xaxis.set_tick_params(labelrotation=40)

plt.tight_layout()
plt.show()

# %% [markdown]

# ## 调用

# %%
import numpy as np
import pandas as pd

# plt.rcParams["figure.figsize"] = [27.50, 12.50]

excel = pd.read_excel("B 题数据.xlsx")

ana = pd.DataFrame(excel).iloc[:, 2:]
# print(ana)

```

8.2.3 第三题

```
# %% [markdown]
```

```
# ## 第三题
```

```
#
```

3、国家标准的 I-V 分类主要考虑水质本身对人和环境的影响，请从居民、游客生活体验的角度出发，对各个监测站的水质重新分类，并给出分类的依据。

```
# *（分类，分类依据制定）*
```

```
# %% [markdown]
```

```
# ### 熵权法计算原始数据打分到二级标准（实际是一级）的权重
```

```
#
```

注：不包括水质类别相对应的数据列。一方面是制定打分标准时忘记了，另一方面是后来反思也认为数据给出的水质类别与其他列是相关的，不必算进新的打分标准中

```
# %%
```

```
stdana = pd.DataFrame()
```

```
weight = []
```

```
k = 1 / np.log(len(ana.index))
```

```
for i, col in enumerate(ana):
```

```
    stdana[i] = ((ana[col]-min(ana[col]))/max(ana[col]-min(ana[col])))
```

```
    stdana[i] /= sum(stdana[i])
```

```
    for ji, j in enumerate(stdana[i]):
```

```
        if j != 0:
```

```
            stdana.iloc[ji, i] *= np.log(j)
```

```
        else:
```

```
            stdana.iloc[ji, i] = 0
```

```
    weight.append(1-(-k*sum(stdana[i])))
```

```
for k in range(len(weight)):
```

```
    weight[k] /= sum(weight)
```

```
weight = {k:v for k,v in zip(ana.columns, weight)}
```

```
# print(weight)
```

```
# %% [markdown]
```

```
# ### 分数计算
```

```
# %%
```

```
weight1 = [weight['水温(℃)'], weight['浊度(NTU)'], weight['溶解氧(mg/L)']]
```

```

weight2 = [weight['高锰酸盐指数(mg/L)'], weight['电导率(  $\mu$  S/cm)'], weight['pH(无量纲)'], weight['氨氮(mg/L)']]
weight3 = [weight['氨氮(mg/L)'], weight['溶解氧(mg/L)'], weight['总磷(mg/L)'], weight['总氮(mg/L)']]
weight1 /= sum(weight1)
weight2 /= sum(weight2)
weight3 /= sum(weight3)
weightn = [weight1, weight2, weight3]
s = pd.DataFrame(columns=['舒适度', '安全性', '清洁度'])
for i in range(3):
    score = pd.read_excel('分.xlsx', sheet_name=i)
    s.iloc[:, i] = np.dot(weightn[i], score.iloc[0:, 1:])
weightupper = [0.2, 0.4, 0.4]
s_last = pd.concat([excel['断面名称'], pd.Series(np.dot(s, weightupper))], axis=1)
s_last.columns = ['断面名称', '总分']
s_last_sorted = sorted(s_last.values, key=lambda x: x[1], reverse=True)
# print(excel['断面名称'])
# print(s_last_sorted)
total_scores = np.array([item[1] for item in s_last_sorted])
standardized_scores = (total_scores - min(total_scores)) / (max(total_scores) - min(total_scores))
levels = 6 - np.digitize(standardized_scores, np.linspace(-0.00001, 1.00001, 6), right=True)
s_last_sorted_with_scores = [np.append(item, [score, level]) for item, score, level in zip(s_last_sorted, standardized_scores, levels)]
for item in s_last_sorted_with_scores:
    print(item)
print(s_last_sorted)

```