

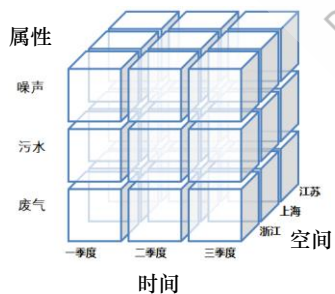
环境地学数据融合平台

ENVIFUSION

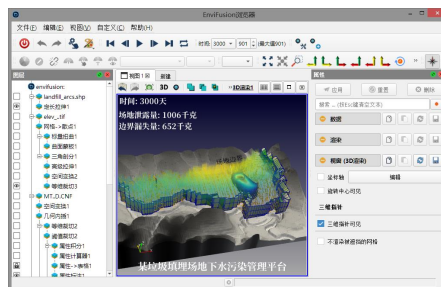
线上直播课

2021年7月

理论准备 工具准备



数据融合

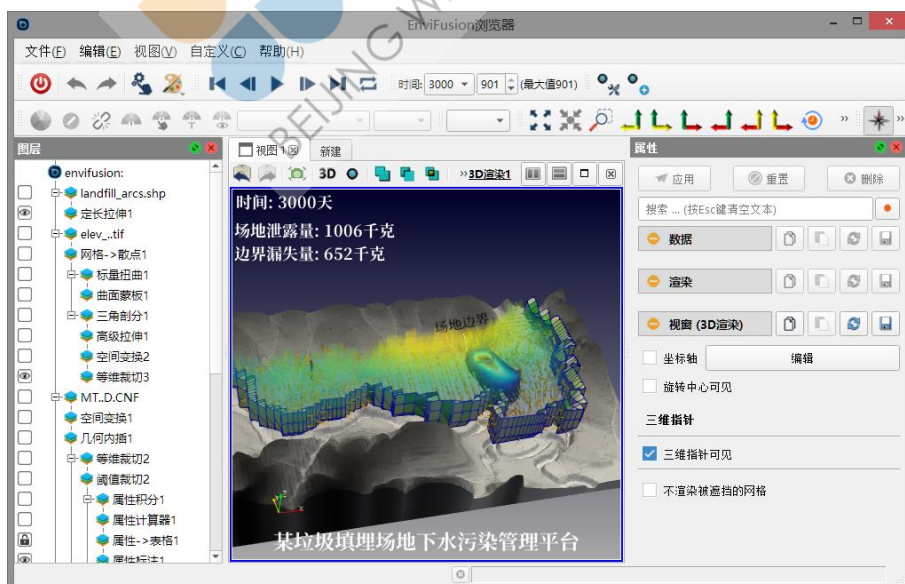


ENVIFUSION

EnviFusion软件发展

- 由中国地学科学家领导，在自然资源部支持下开发的环境地学计算平台。
- 在四维时空构建环境地学属性数据体。
- 帮助领域科学家对海量数据进行统计挖掘和可视化分析。
- 采用图层管道技术，允许设计复杂的数据模型。
- 提供二次开发接口，帮助生产单位为客户提供高水平咨询服务。

EnviFusion界面



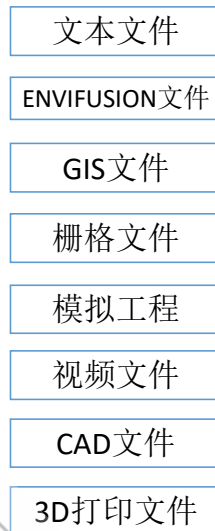
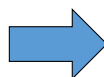
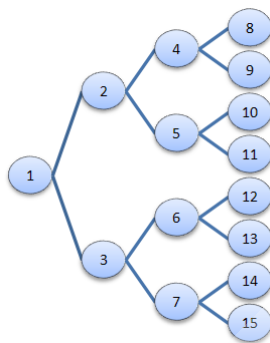
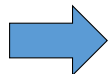
EnviFusion中的数据融合与变换工具

- 数据补齐 (时间/空间/属性等)
- 空间地理 (切片/扭曲/提取等)
- 地下水建模 (MODFLOW/MT3DMS/SEAWAT)
- 地下水数据挖掘 (TFPR)
- 地球化学 (等值/热点/聚类等)
- 文本分析 (自然语言处理)
- 拓扑变换 (点线面体间转换)
- 向量操作 (属性场生成与追踪)
- 时间变换 (时序差分/平滑等)
- 属性变换 (筛选/提取/插值/微分/积分等)

数值模拟引擎

- 梯度场引擎：势场->流场->追踪相关工具；
- MODFLOW引擎：边界+参数->势场->流场；
- MT3DMS引擎：流场->浓度场；
- SEAWAT引擎：变密度流场->浓度场；
- TFPR引擎：势场变化->含水层参数；
- 地球化学引擎：高维度地球化学数据->降维分析；
- 文本分析引擎：热词提取与语义关联；
- 数据补齐引擎。

数据流

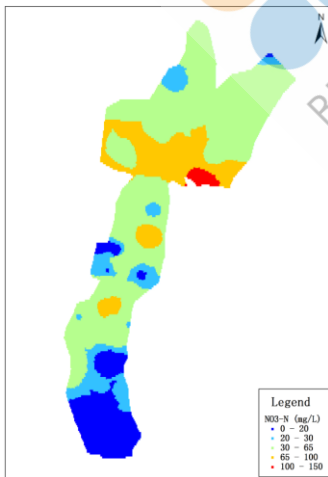


基于数据融合的地下水污染溯源
和缺失值补齐

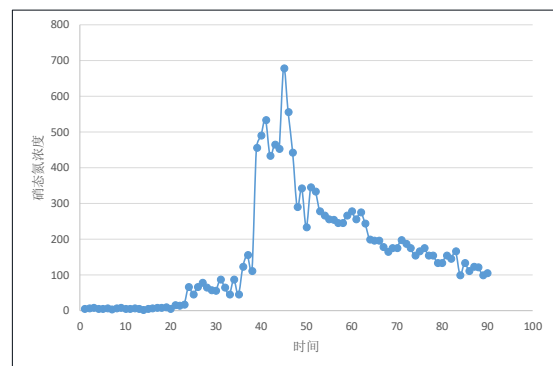
● 污染溯源

- 地下水污染指人类活动引起地下水化学成分、物理性质和生物学特性发生改变而使质量下降的现象。
- 当存在大量监测数据时，污染溯源很大程度是寻找离群数据的过程。

● 单一维度（属性/测试指标）污染溯源



空间分析

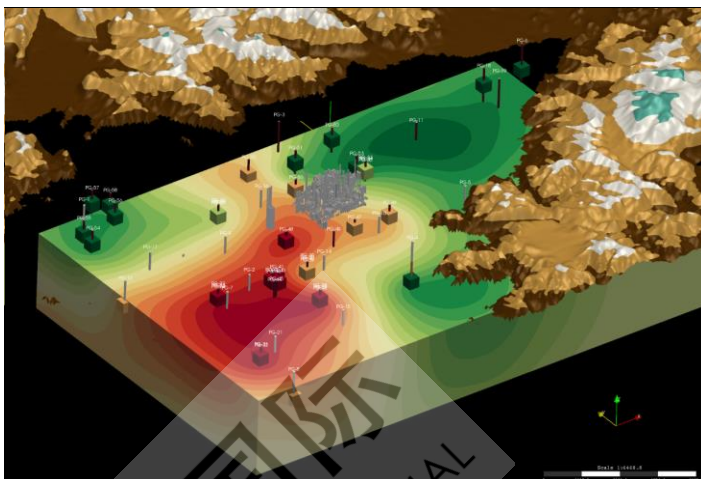


时间分析

高维度（多指标）污染溯源

50口监测井
20次采样
50个维度（属性/指标）
1000个样本
5万个数据值

山前农业区硝酸盐污染过程
平原腹地还原性增强后的铁锰超标
小区域污染场地
零星的硫酸盐污染
零星的pH值偏高
零星的重金属超标

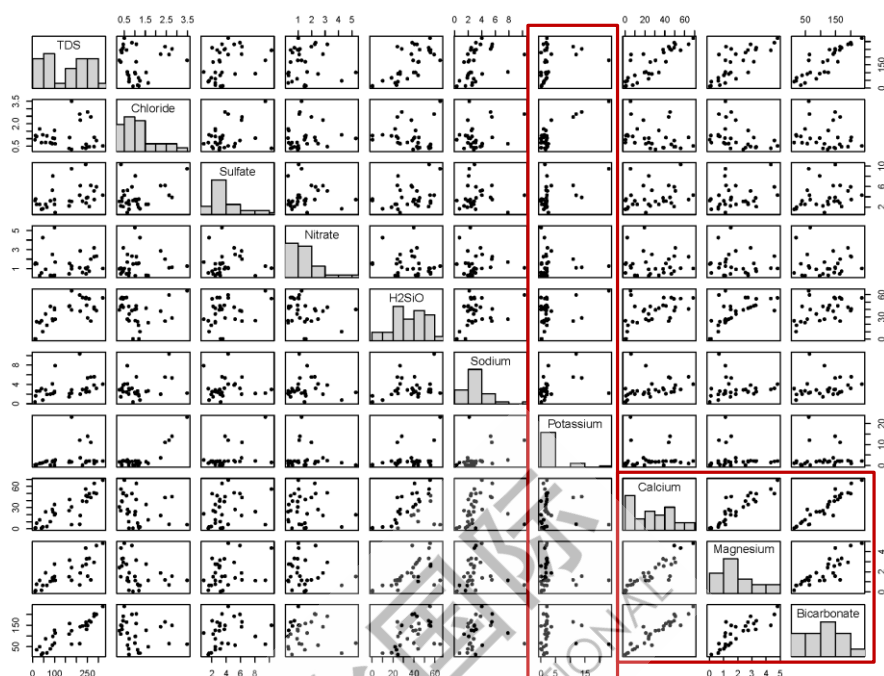


维度筛选

- 主要阳离子 Ca^{2+} , Mg^{2+} , Na^+ , K^+
- 主要阴离子 HCO_3^- , Cl^- , SO_4^{2-} , CO_3^{2-}
- 无机污染物 金属、放射性物质、营养物质
- 有机污染物 石油烃类、非氯化溶剂、氯化溶剂、杀虫剂等

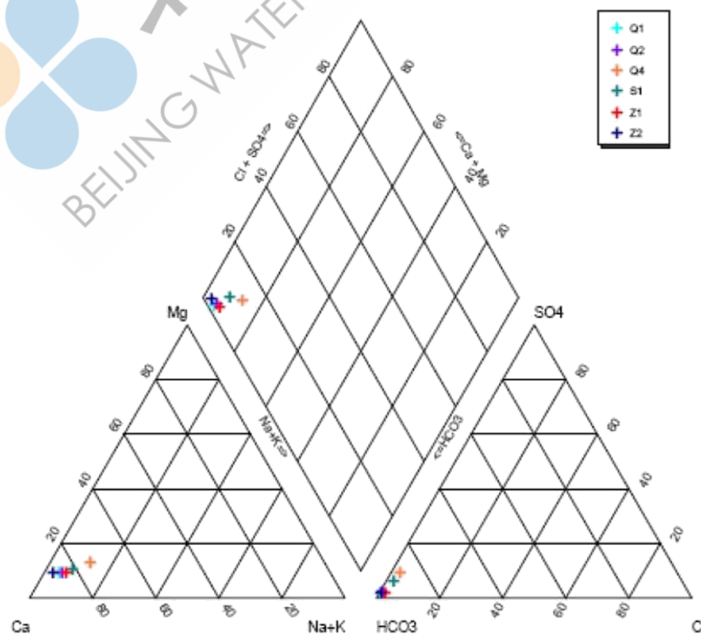
Pairs Plot分析

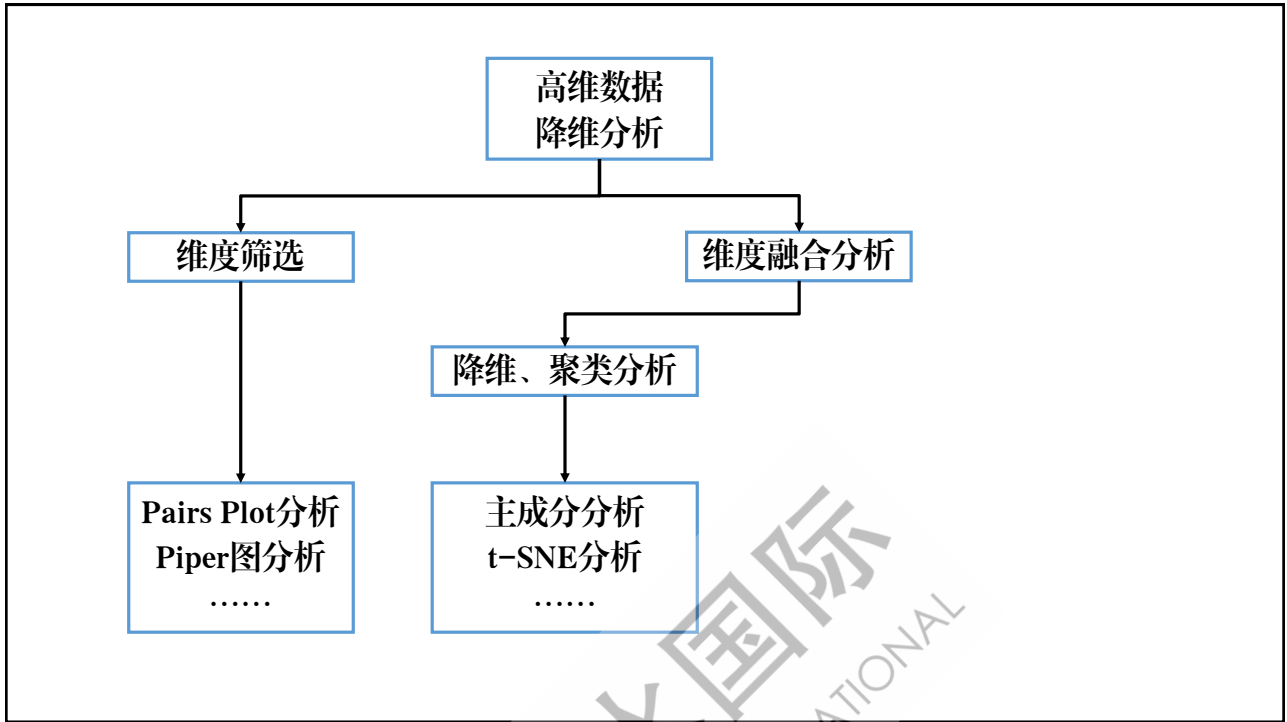
- 维度间两两关系分析
- 寻找离群超标数据



Piper图分析

- 各离子含量分布比例
- 水化学演化过程





降维分析——主成分分析

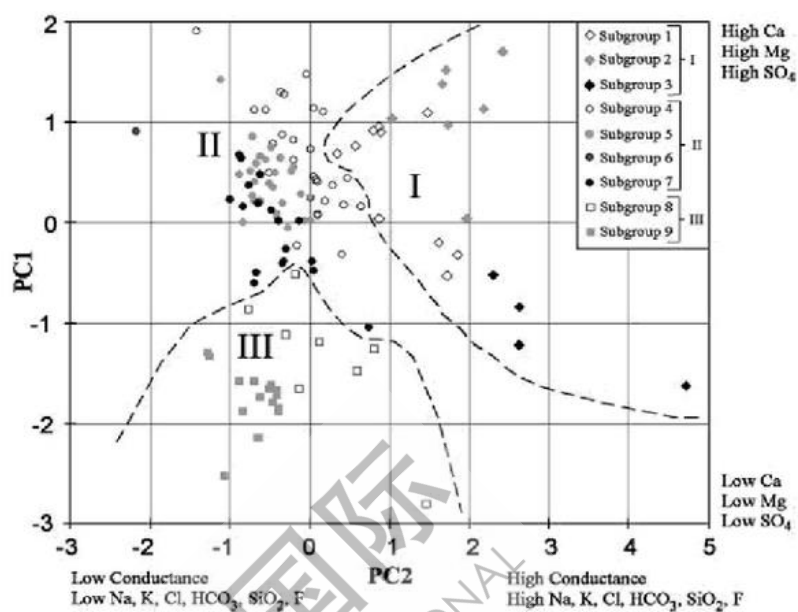
主成分分析（PCA）是一种高效的**线性**降维算法。

主成分分析的**基本方法**是通过构造原变量 $(X_1, X_2, \dots, X_n)$ 的适当的线性组合，以产生一系列互不相关的信息，从中选出**少数几个新变量** F_k

$(k < n)$ 并使它们含有尽可能多的原变量带有的信息，从而使得用这几个新变量代替原变量分析问题和解决问题成为可能。

主成分分析

- 十个原始维度线性转换为十个融合维度，选取前两个融合维度称为主成分PC1与PC2
- PC1-PC2图中，根据样本分布识别化学端元，并根据端元位置对样本进行分类。
- 污染溯源时，一个端元常常对应一类污染源。



降维分析——t-SNE

- t-SNE是一种非线性降维算法；
- 可将高维水质数据降维至新的2-3个维度；
- 便于识别发现化学演化规律；
- 使用聚类方法对水质数据进行污染溯源；

仅适用于维度完整的数据集

水质数据缺失，不利于降维聚类

时间缺失

空间缺失

站点缺失

属性缺失

原始数据

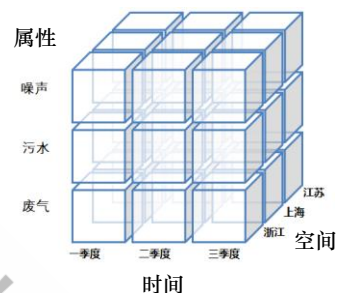
时间补齐

空间补齐

站点互采

属性补齐

补齐数据



维度融合分析

维度筛选

Pairs Plot分析
Piper图分析
.....

高维数据
降维分析

维度融合分析

降维、聚类分析

主成分分析
t-SNE分析
.....

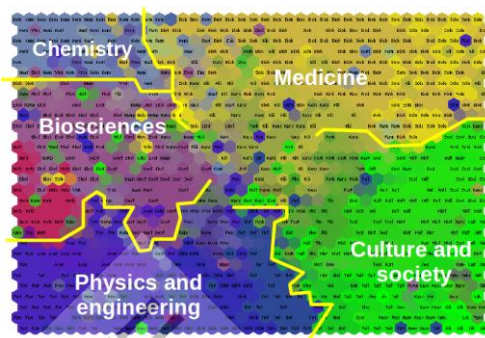
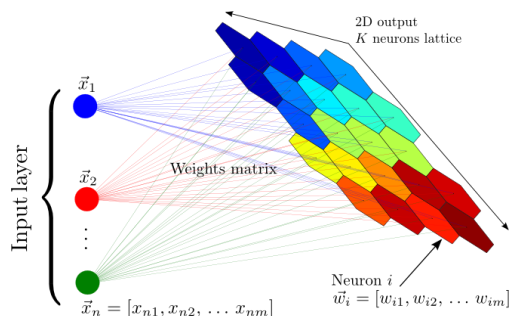
神经网络

数据补齐

维度映射分析
样本聚类分析



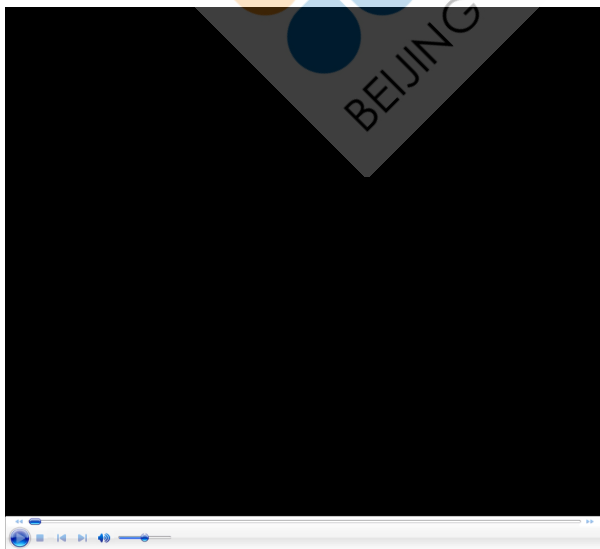
使用神经网络作为媒介，降低数据体维度



每个样本对应一个最优神经元



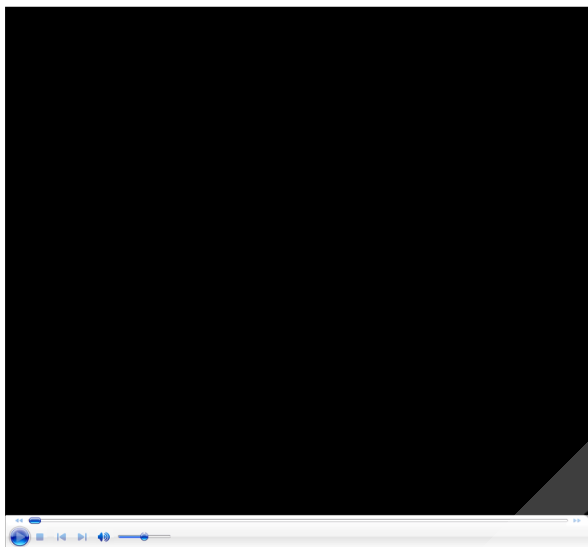
神经网络训练过程



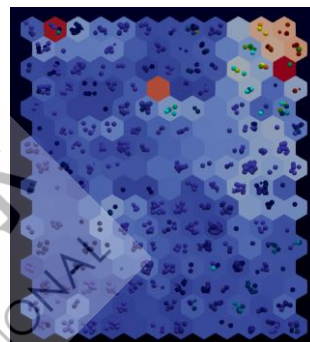
- 数据体应尽量连续，若有离群点应单独考虑；
- 样本量越大，训练结果越准确；
- 维度映射与样本聚类；
- 基于样本聚类结果进行数据补齐



预测缺失水质指标 (027-026)



- 根据水质指标之间的关联补齐缺失指标;
- 已知指标越多, 预测越准确。



华北示范区水质指标

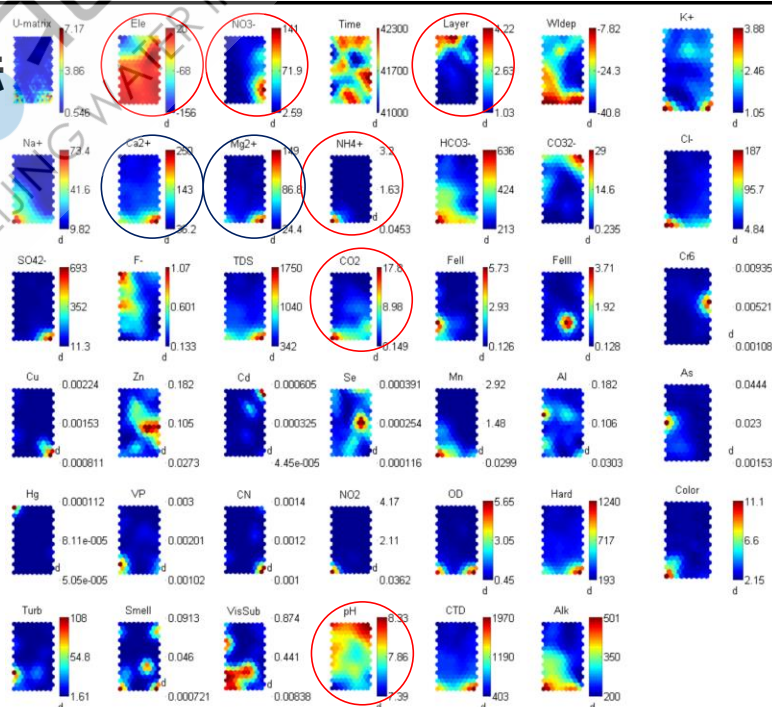
维度映射分析

埋深-层位

CO₂-pH

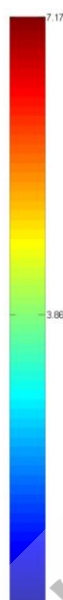
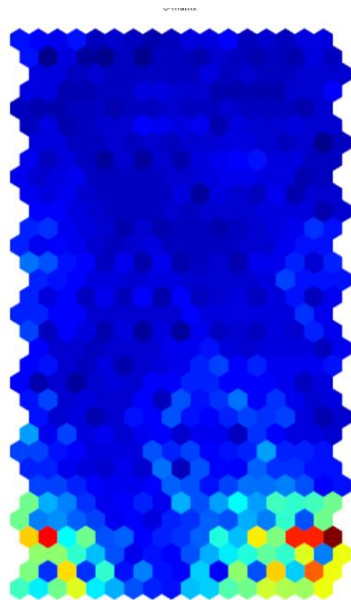
NO₃⁻-NH₄⁺

Ca²⁺-Mg²⁺





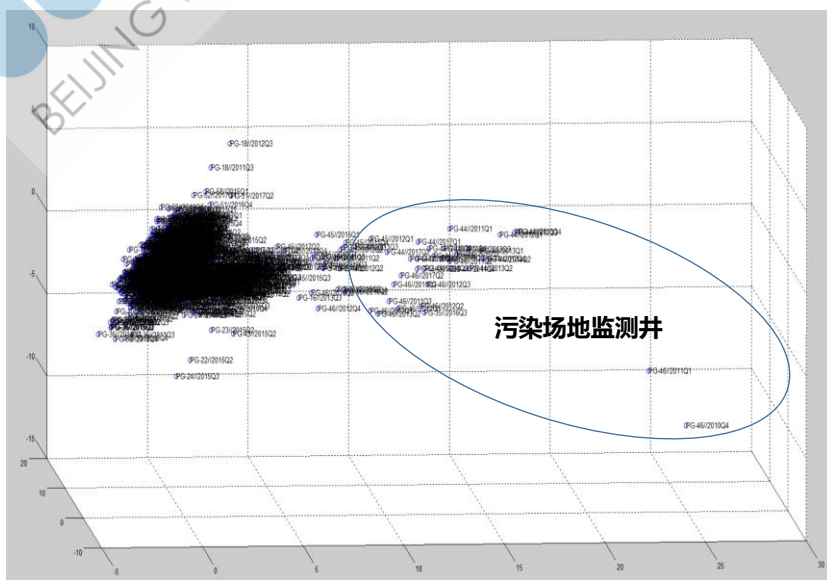
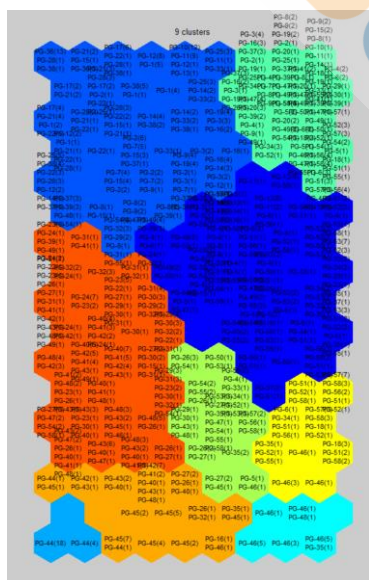
水质样本聚类分析 (污染溯源)



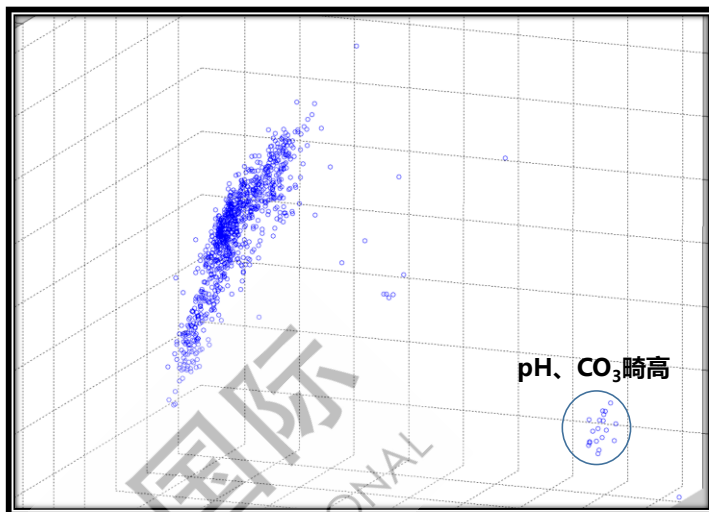
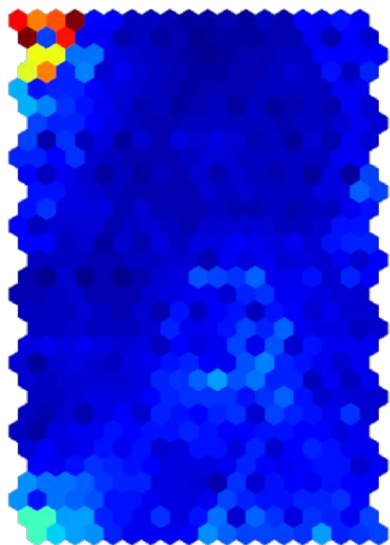
第一轮
地下水水质
异常



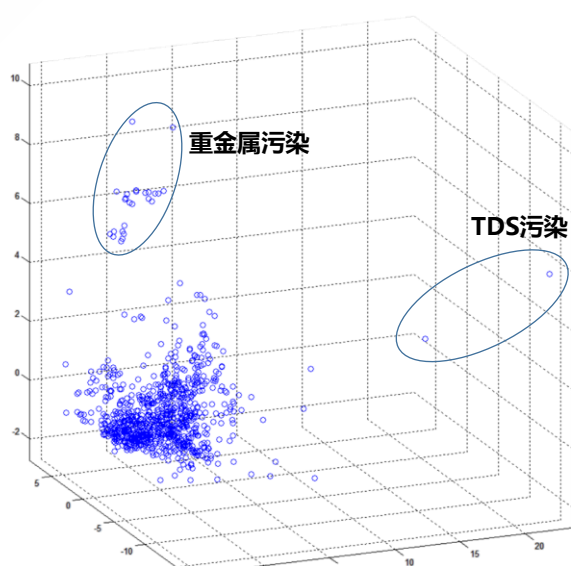
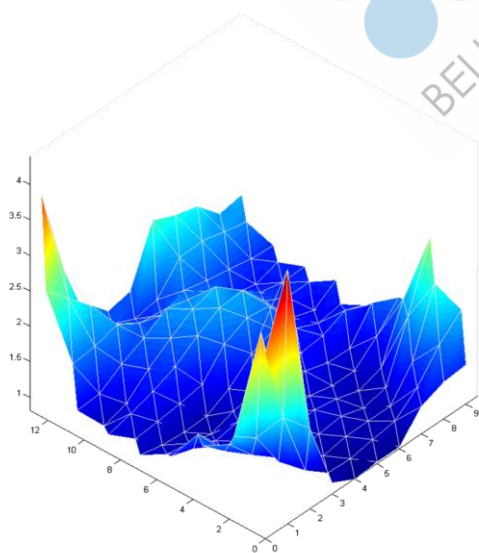
迭代聚类依次侦测地下水水质异常



● 第二轮水质异常检视

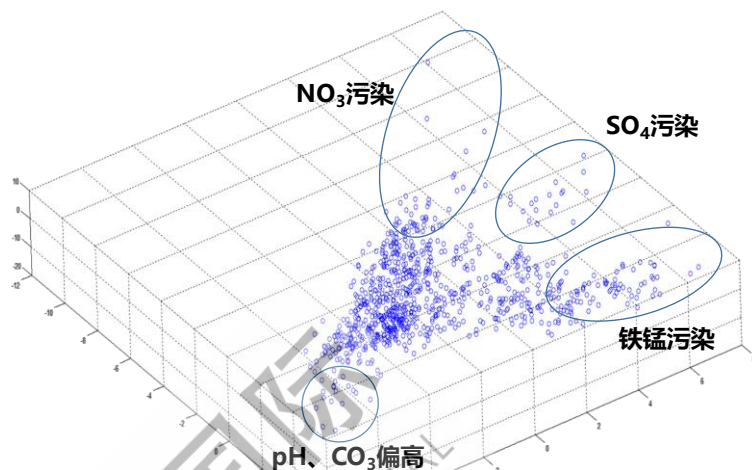
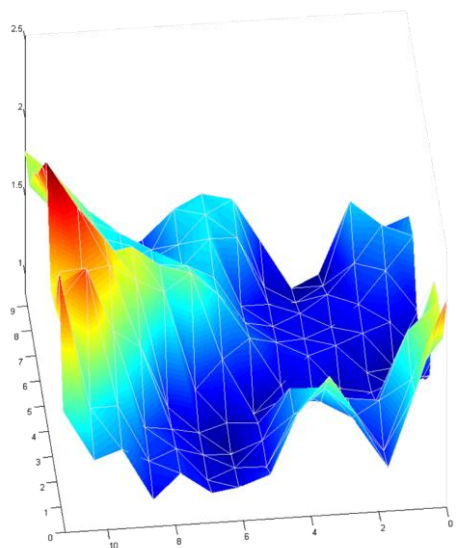


● 第三轮水质异常检视





第四轮水质异常检视



参数空间向物理空间的映射

