1、数据的预处理、清晰

1.1 字段的规范化

1.2 化肥养分占比分析

2、对于肥料产品的相关数据分析

2.1 复混肥料产品分组情况（任务 2.1）

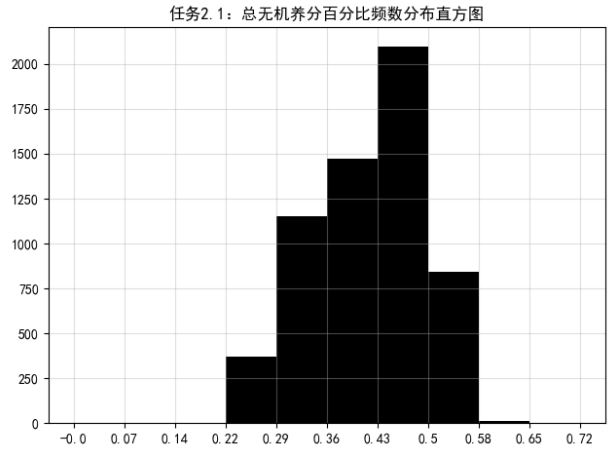


图 4.总无机养分百分比频数分布直方图

我们对分组详情画出登记数量由高到低的分组条形图，可以看出登记数量前 3 的分组为(0.43,0.5]、(0.36,0.43]、(0.29,0.36]分别为 2098 种、1470 种、1154 种。

排名	一	二	三
分组标签	7	6	5
产品登记数量	2098	1470	1154

2.2 有机肥料产品分组情况（任务 2.2）

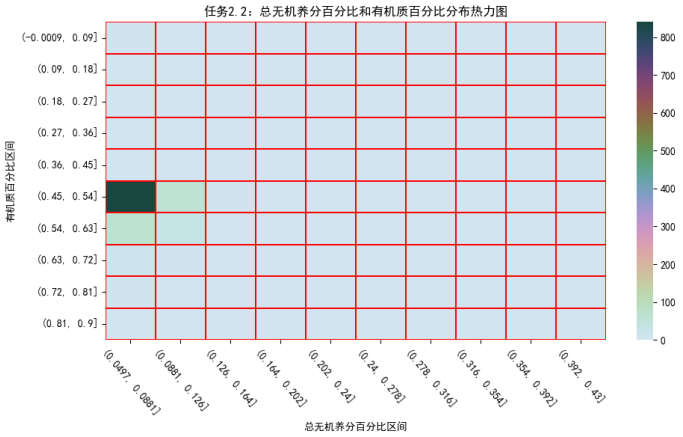


图 6.总无机养分百分比-有机质百分比分布热力图

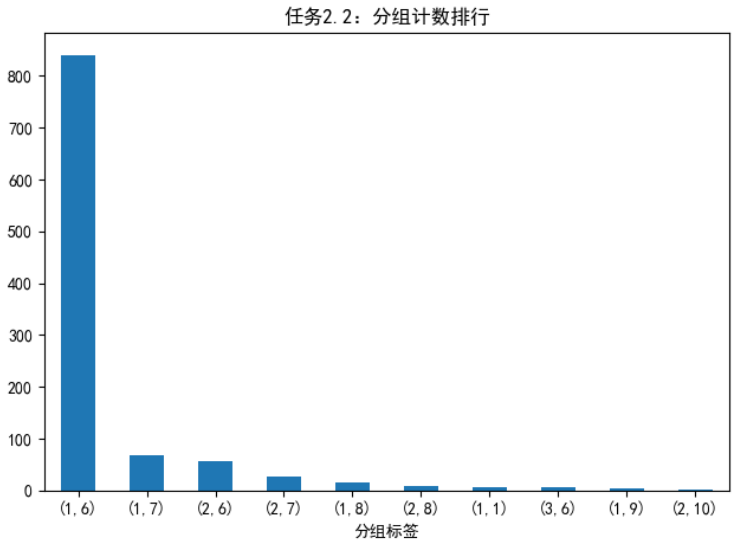


图 7. 有机肥料分组计数排行条形图

2.3 复混肥料的产品聚类分析（任务2.3）

2.3.1聚类定义

聚类分析根据一批样品的许多观测指标，按照一定的数学公式具体地计算一些样品或一些参数(指标)的相似程度，把相似的样品或指标归为一类，把不相似的归为一类。

2.3.2系统聚类法(本题所用):

(1) 基本思想: 一开始将n个样品各自自成一类,这时类间的距离与样品间的距离是等价的; 然后将距离最近的两类合并, 并计算新类与其他类的类间距离, 再按最小距离并类。这样每次缩小一类,直到所有的样品都成一类为止。这个并类过程可以用谱系聚类图形象地表达出来。

(2) 基本步骤:

(2.1) 数据变换:可以使用上节介绍的方法对数据进行变换.数据变换目的是为了便于比较、计算上的方便或改变数据的结构。选择度量样品间距离的定义(如欧氏距离)及度量类间距离的定义(如最短距离法,见下面“系统聚类分析的方法”中的介绍)。

(2.2) 计算n个样品(个体)两两间的距离,得初始的距离矩阵D(1) 。

(2.3) 一开始(第一步:i=1)n个样品各自构成一类,得类的个数k=n个类: Gt={X(t) } (t=1,…,n).此时类间的距离就是样品间的距离. 对步骤i=2,…,n执行并类过程的步骤③和④ .

(2.4) 步骤i得到的D(i-1) ,每次合并类间距离最小的两类为一新类.此时类的总个数k减少1类, 即k=n-i+1.

(2.5) 计算新类与其他类的距离,得新的距离矩阵D(i) .若此时类的总个数k大于1类,重复③和④步;直到类的总个数为1时止.

(2.6) 画谱系聚类图;

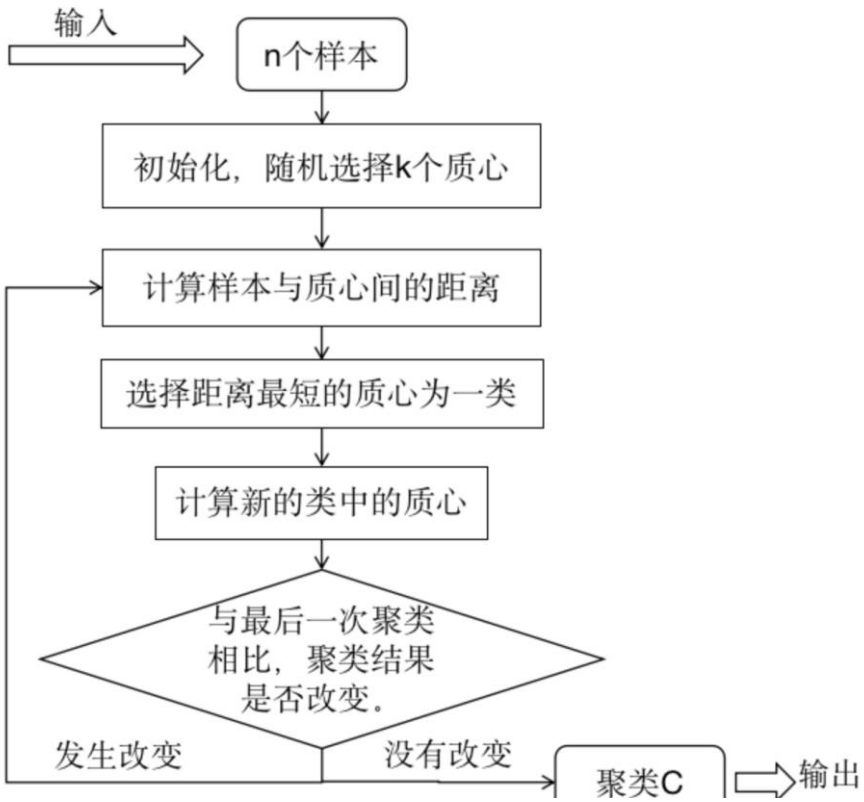
(2.7)决定分类的个数及各类的成员.

2.3.3 K-均值聚类算法 (本题所用):

(1) 基本思想: K 均值聚类 (k-means) 是基于样本集合划分的聚类算法。K 均值聚类将样本集合划分为 k 个子集, 构成 k 个类, 将 n 个样本分到 k 个类

中，每个样本到其所属类的中心距离最小，每个样本仅属于一个类，这就是 k 均值聚类，同时根据一个样本仅属于一个类，也表示了 k 均值聚类是一种硬聚类算法。

（2）基本步骤:



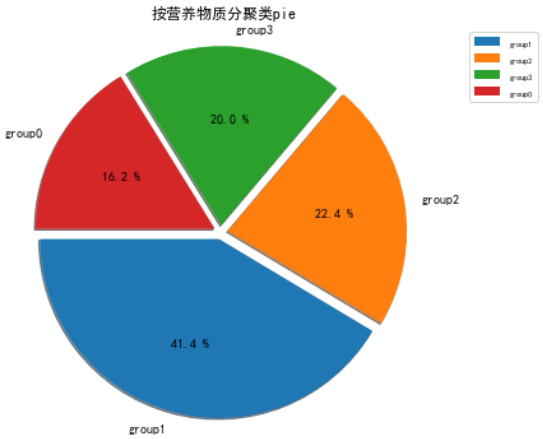


图 8.聚类输出结果样本数占比

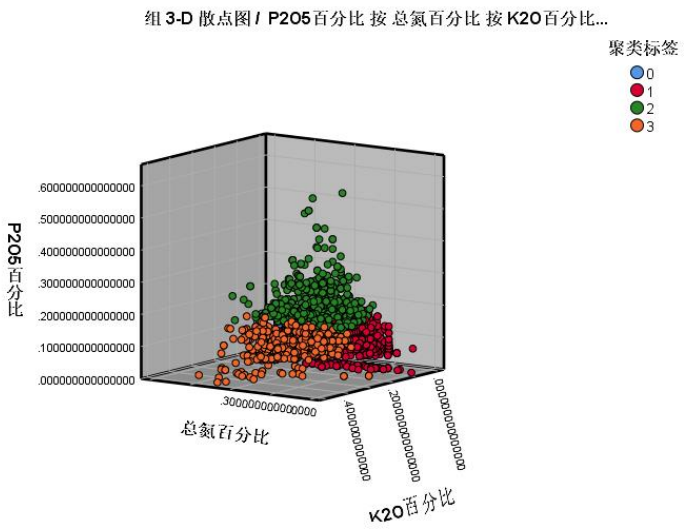


图 9.肥料产品 3 维散点图

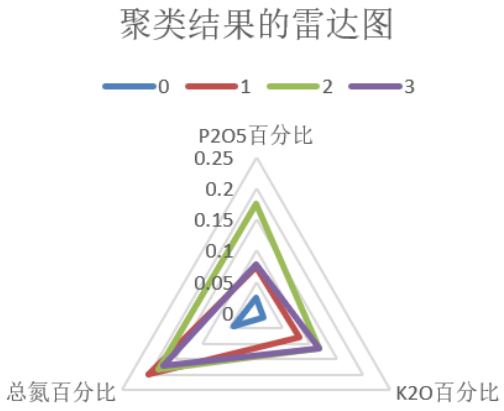


图 11.修正数据绘制的聚类结果的雷达图

3、肥料产品的多维度分析

3.1 各组别登记数量变化分析

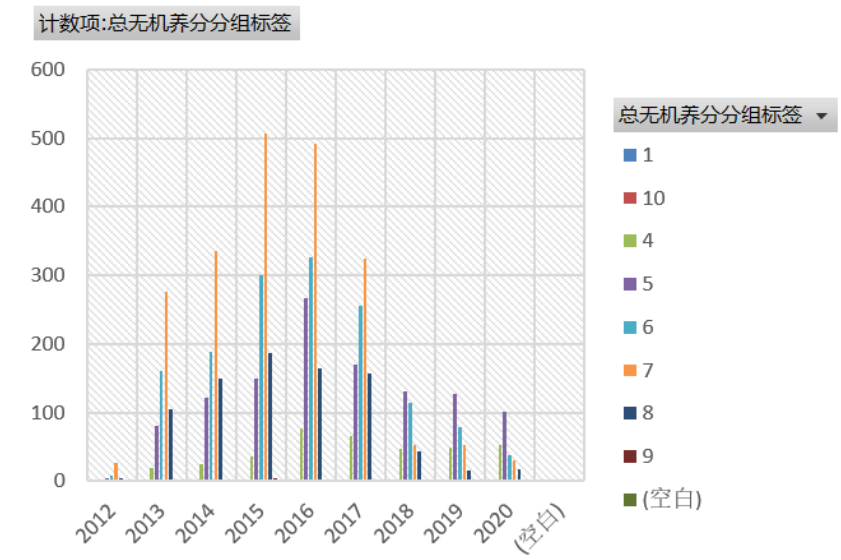


图 13.各组别分年份登记数分布条形图

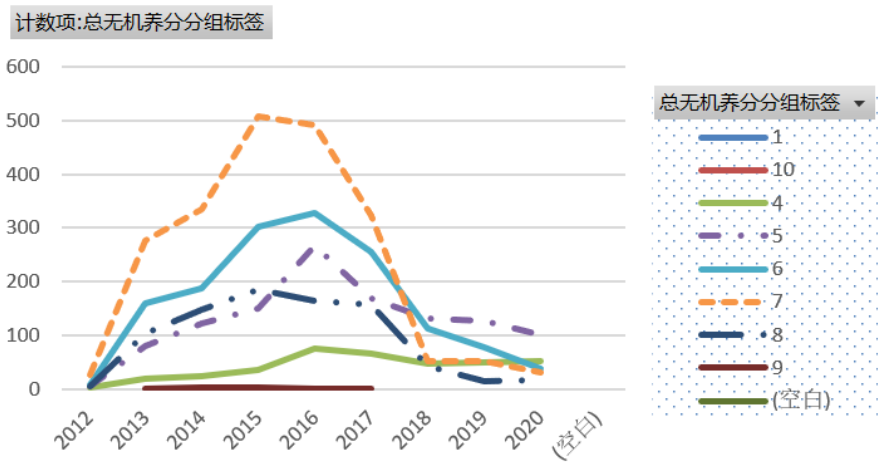


图 14.各组别分年份登记数变化曲线图

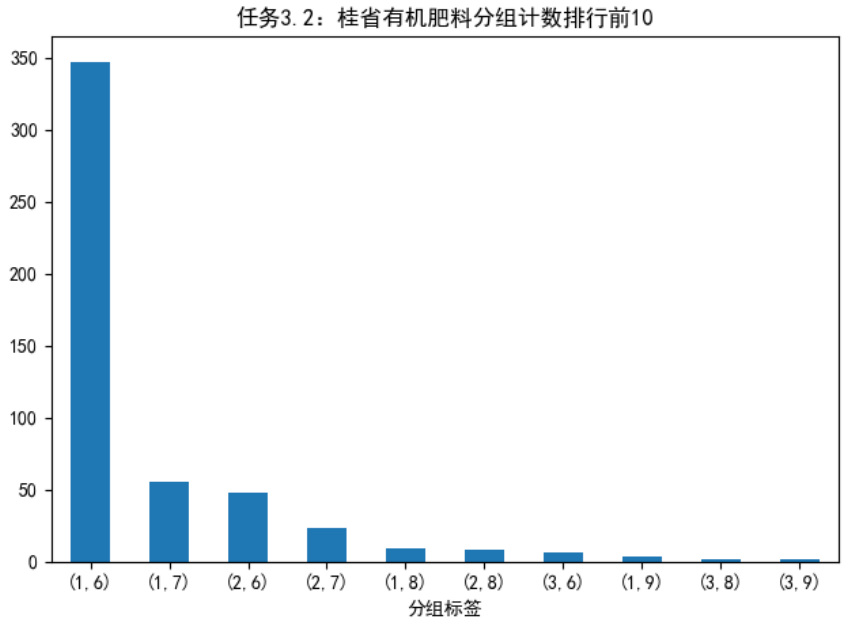


图 15.桂省有机肥料分组排行条形图

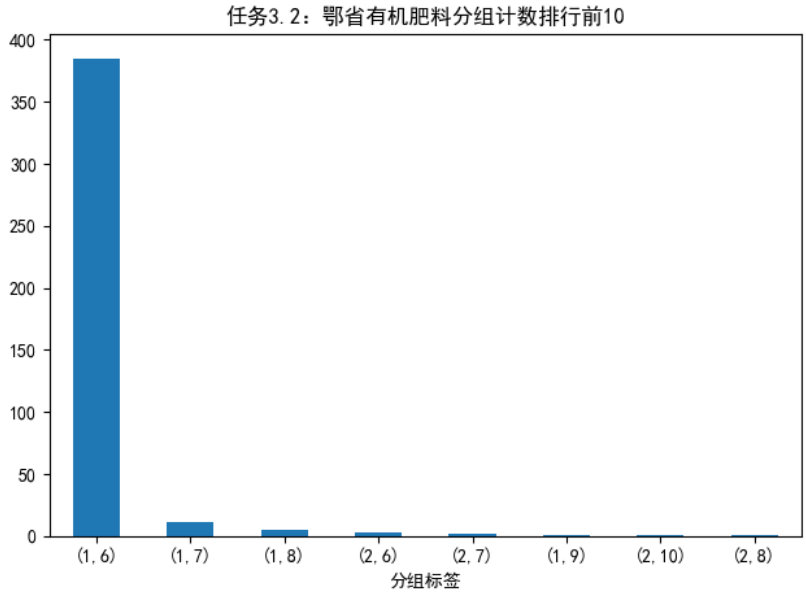


图 16.鄂省有机肥料分组排行条形图

3.3 杰卡德相似系数矩阵（任务 3.3）

3.3.1 定义（集合 A 与 B 的杰卡德相似系数）

杰卡德相似度算法没有考虑向量中潜在数值的大小，而是简单的处理为0和1，不过，做了这样的处理之后，杰卡德方法的计算效率肯定是比较高的，毕竟只

需要做集合操作。Jaccard系数主要用于计算符号度量或布尔值度量的个体间的相似度，无法衡量差异具体值的大小，只能获得“是否相同”这个结果，所以

4、肥料产品的多维度分析 2

4.1 数据处理（任务 4.1）