

目录

摘要.....	2
1、 任务一：数据预处理	2
1.1 数据规范化处理	2
1.1.1 方法包括以下内容：	2
1.1.2 实际规范化处理.....	3
1.2 总无机养分百分比分析.....	3
2、 任务二：肥料产品的数据分析	4
2.1 复混肥料产品的标签分组	4
2.2 有机肥料产品的标签分组与分布热力图	5
2.3 利用聚类标签绘图	5
2.3.1 变量与迭代设置.....	5
2.3.2 保存和选项.....	5
2.3.3 结果输出.....	5
2.3.4 绘制三维散点图.....	5
3、 任务三 肥料产品的多维度对比分析.....	6
3.1 复混肥料变化趋势	6
3.2 有效有机肥料产品分析	6
3.3 杰卡德相似系数矩阵	7
3.3.1 杰卡德相似系数.....	7
3.3.2 计算企业之间的杰卡德相似系数矩阵.....	7
4、 任务四 肥料产品的多维度对比分析.....	7
4.1 氮、磷、钾养分和有机质的百分比的提取.....	7
4.2 提取各种原料的名称及其百分比.....	8

摘要

肥料是农业生产中一种重要的生产资料，其生产销售必须遵循《肥料登记管理办法》，依法在农业行政管理部门进行登记。各省、自治区、直辖市人民政府农业行政主管部门主要负责本行政区域内销售的肥料登记工作，相关数据可从政府网站上自由下载。

1、任务一：数据预处理

1.1 数据规范化处理

数据规范化处理是数据挖掘的一项基础工作。不同的属性变量往往具有不同的取值范围，数值间的差别可能很大，不进行处理可能会影响到数据分析的结果。为了消除指标之间由于取值范围带来的差异，需要进行标准化处理。将数据按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析。

1.1.1 方法包括以下内容：

(1) 最小-最大规范化

假定 \min 和 \max 分别为属性 A 的最小值和最大值，则通过下面公

式将属性 A 上的值 v 映射到区间 $[new_min, new_max]$ 中的 v' :

(2) 零-均值规范化

将属性 A 的值根据其平均值 $mean$ 和标准差 std 进行规范化:

(3)小数定标规范化

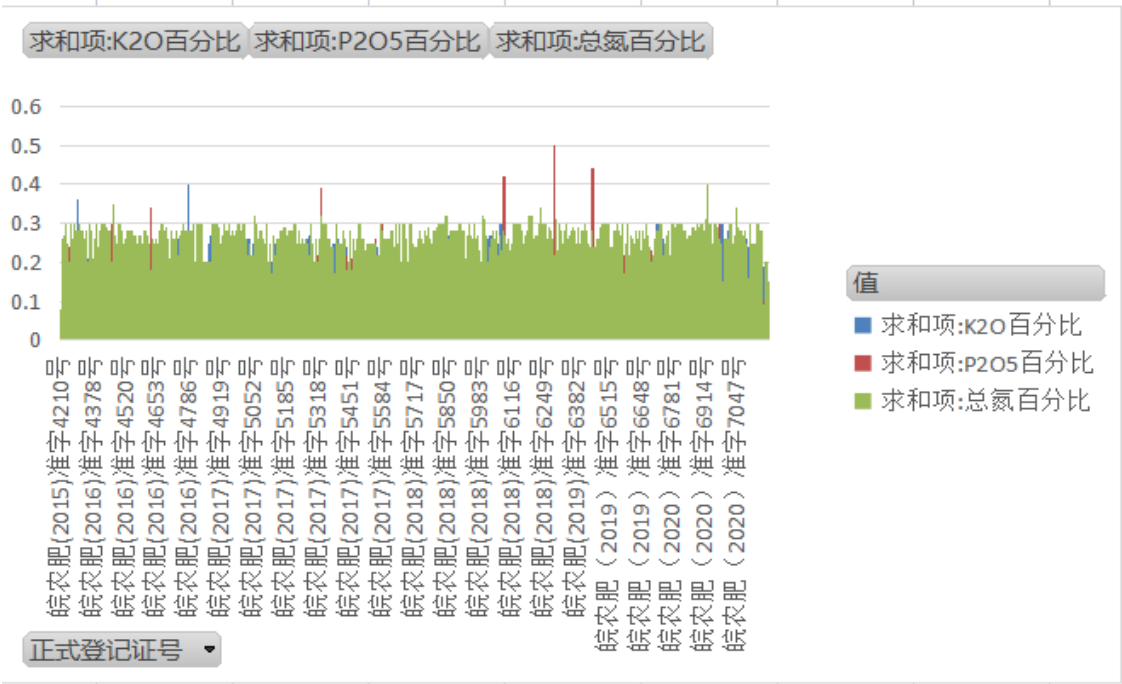
通过移动属性 A 的小数点位置进行规范化, 小数点的移动依赖于 A 的最大绝对值:

其中, j 是使 $Max(|v'|) < 1$ 的最小整数。

1.1.2 实际规范化处理

因所给题目中附件 1 的产品通用名称存在不规范的情况。且按照复混肥料（掺混肥料归入这一类）、有机-无机复混肥料、有机肥料和床土调酸剂这 4 种类别对附件 1 进行规范化处理。

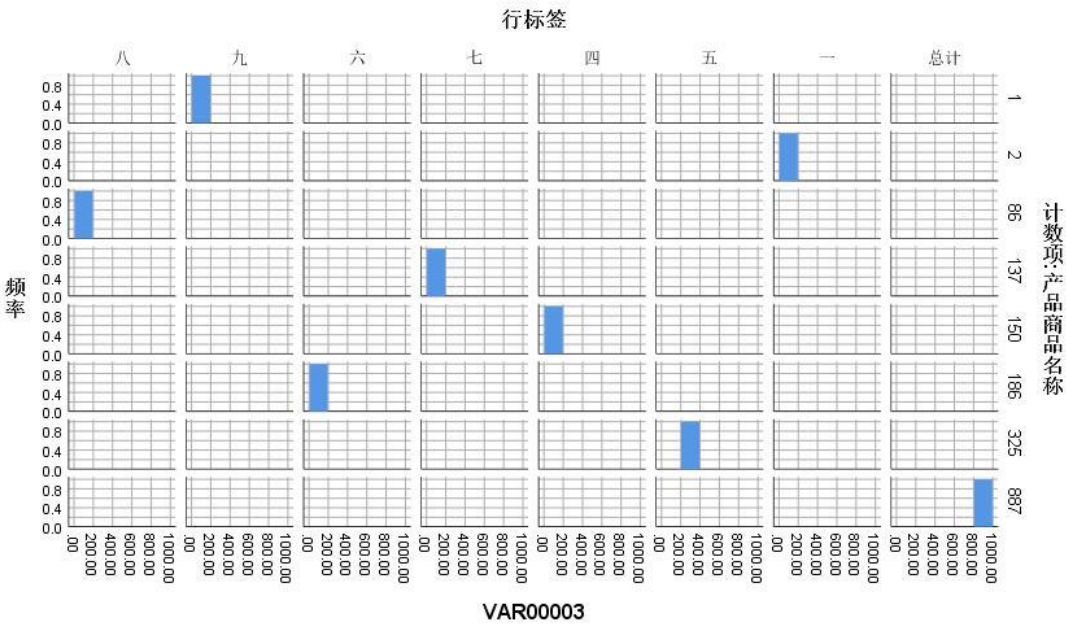
1.2 总无机养分百分比分析



2、任务二：肥料产品的数据分析

2.1 复混肥料产品的标签分组

我们利用 Excel 进行筛选将所有复混肥料按照总无机养分百分比的取值等距分为 10 组，且用 1-10 为其打上分组标签后制成一份新的表格，得到以下结果：



由该图我们可以观察到复混肥料多分布于标签为四、五、六的地方，少数分布标签为一、九的地方，即复混肥料总无机养分百分比多数为[29.00, 36.00]%, [24.99, 28.00]%次之，[58.00, 63.00]%与[0.00, 5.01]%占少数，而标签为二、三、十的为空。

行标签	计数项:产品商品名称
一	2
二	
三	
四	150
五	325
六	186
八	86
九	1
七	137
十	
(空白)	
总计	887

将行标签（总无机养分百分比标签）与产品商品名称用更加直观的图表表示出来后，我们可以清晰地看出登记数量最大的前 3 个分组及相应的产品登记数量即为标签为五、六、四的产品，登记数量分别为 325，186，150。

排名	三	一	二
分组标签	4	5	6
产品登记数量	150	325	186

2.2 有机肥料产品的标签分组与分布热力图

2.3 利用聚类标签绘图

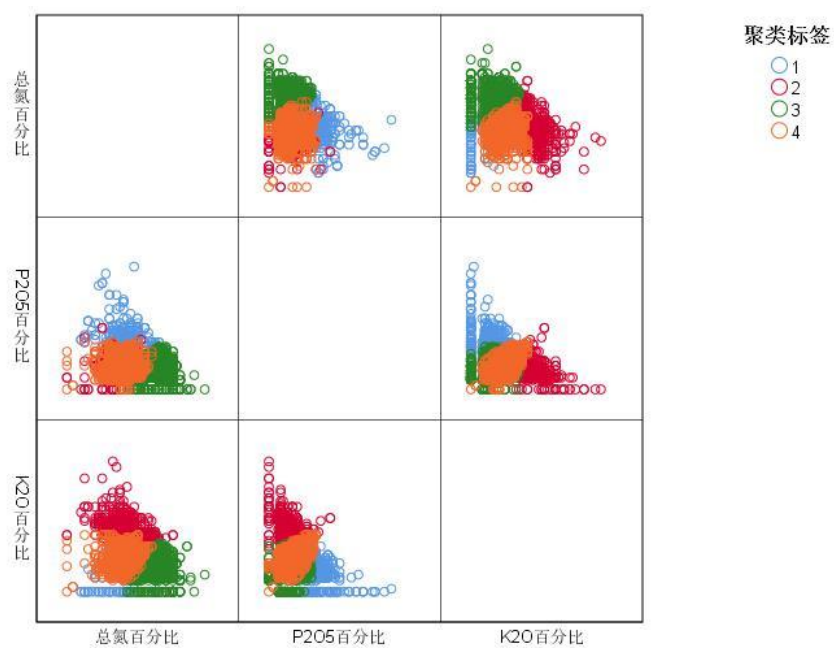
聚类分析是将研究对象按照一定的标准进行分类的方法，分类结果是每一组的对象都具有较高的相似度，组间的对象具有较大的差异。

2.3.1 变量与迭代设置

2.3.2 保存和选项

2.3.3 结果输出

2.3.4 绘制三维散点图

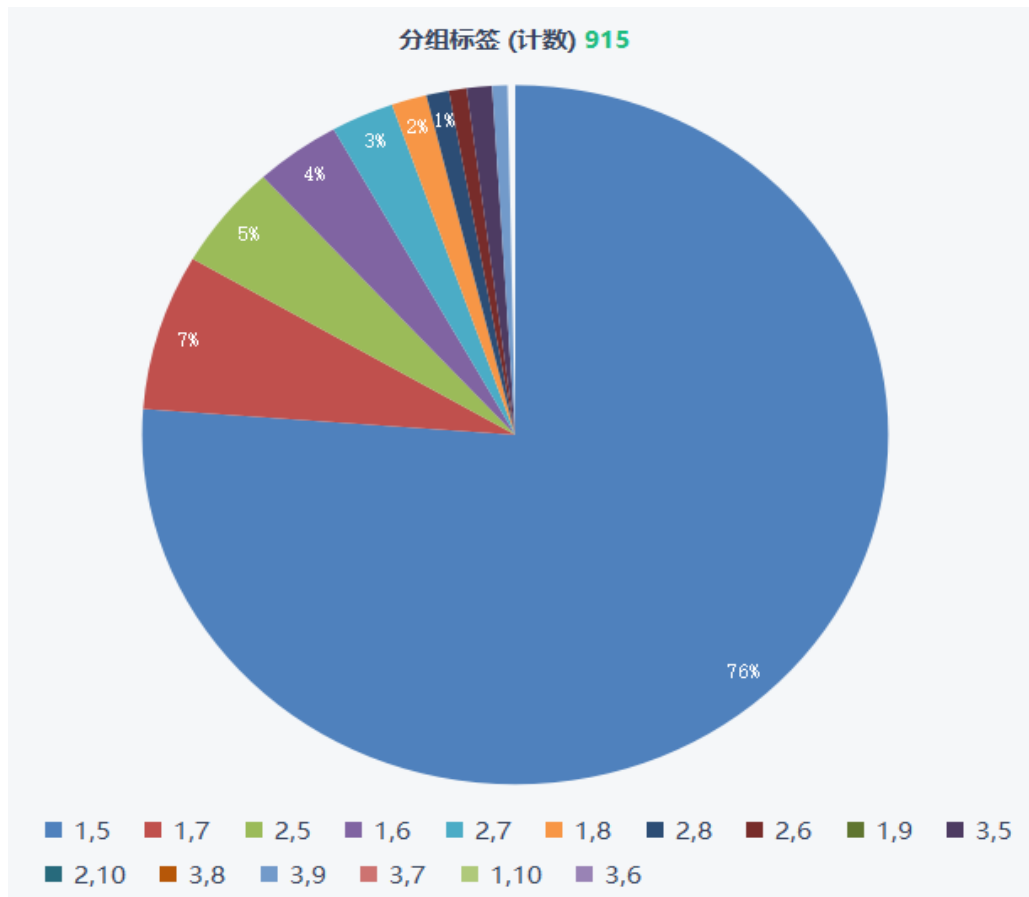


3、任务三 肥料产品的多维度对比分析

3.1 复混肥料变化趋势

3.2 有效有机肥料产品分析

利用数据透视图以分组标签与正式登记号为变量进行分析，可得下图：



3.3 杰卡德相似系数矩阵

3.3.1 杰卡德相似系数

两个集合 A 和 B 交集元素的个数在 A、B 并集中所占的比例，称为这两个集合的杰卡德系数，用符号 $J(A, B)$ 表示。杰卡德相似系数是衡量两个集合相似度的一种指标（余弦距离也可以用来衡量两个集合的相似度）。

jaccard 值越大说明相似度越高。

3.3.2 计算企业之间的杰卡德相似系数矩阵

4、任务四 肥料产品的多维度对比分析

4.1 氮、磷、钾养分和有机质的百分比的提取

4.2 提取各种原料的名称及其百分比