肥料登记数据分析报告

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。基于我国是农业大国的情况，为了充分掌握肥料的相关登记信息，利用数据分析技术对各类肥料的原料、种类和省份等登记数据信息进行处理和分析则具有很大的现实意义。

目录

1. 任务一 数据的预处理

1.1 数据的分析和规范化处理——附件 1

1.2 总无机养分百分比的计算处理

2. 任务二 肥料产品的数据分析

2.1 复混肥料的处理与分析——附件 2

2.1.1 复混肥料的等距分组情况

2.1.2 复混肥料的产品登记数量直方图分析

根据分组结果在 Excel 表格中绘制直方图 3：

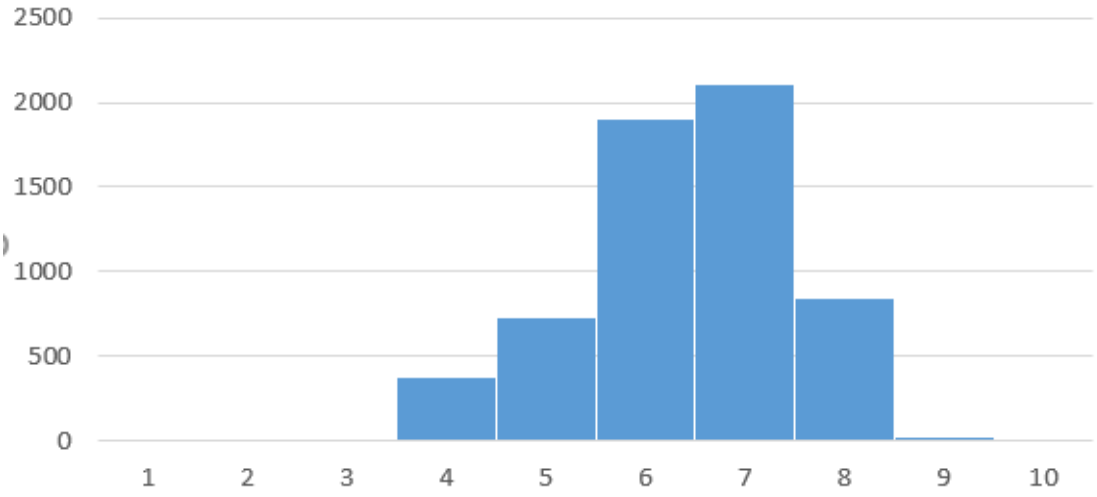


图 1 产品登记数量直方图

分析复混肥料产品的分布特点:

从图表可以看出：复混肥料产品分布比较集中，在总无机养分百分比 43.2%~50.4% 之间的肥料数量最多，35%~43.2% 次之。在其余期间数量成断崖式下降，甚至在 7.2%~14.4%、14.4%~21.6%组别 2、3 数量为 0，同时列出数量最大的前 3 个分组如下图 4：

排名	一	二	三
分组标签	7	6	8
产品登记数量	2098	1895	841

图 2 数量前 3 组图

2.2 复混肥料产品分类——聚类算法

算法介绍：

kmeans 算法又名 k 均值算法,K-means 算法中的 k 表示的是聚类为 k 个簇，means 代表取每一个聚类中数据值的均值作为该簇的中心，或者称为质心，即用每一个的类的质心对该簇进行描述。

其算法思想大致为：先从样本集中随机选取 k 个样本作为簇中心，并计算所有样本与这 k 个“簇中心”的距离，对于每一个样本，将其划分到与其距离最近的“簇中心”所在的簇中，对于新的簇计算各个簇的新的“簇中心”。

根据以上描述，我们大致可以猜测到实现 kmeans 算法的主要四点：

（1）簇个数 k 的选择

（2）各个样本点到“簇中心”的距离

（3）根据新划分的簇，更新“簇中心”

（4）重复上述 2、3 过程，直至"簇中心"没有移动

Step1.K 值的选择

k 的选择一般是按照实际需求进行决定，或在实现算法时直接给定 k 值。

说明：

A.质心数量由用户给出，记为 k，k-means 最终得到的簇数量也是 k

B.后来每次更新的质心的个数都和初始 k 值相等

C.k-means 最后聚类的簇个数和用户指定的质心个数相等，一个质心对应一个簇，每个样本只聚类到一个簇里面

D.初始簇为空

Step2.距离度量

将对象点分到距离聚类中心最近的那个簇中需要最近邻的度量策略，在欧式空间中采用的是欧式距离，在处理文档中采用的是余弦相似度函数，有时候也采用曼哈顿距离作为度量，不同的情况实用的度量公式是不同的。

1.欧式距离

2.曼哈顿距离

3.余弦相似度

A 与 B 表示向量(x1,y1)，(x2,y2)

分子为 A 与 B 的点乘，分母为二者各自的 L2 相乘，即将所有维度值的平方相加后开方。

说明：

A.经过 step2，得到 k 个新的簇，每个样本都被分到 k 个簇中的某一个簇

B.得到 k 个新的簇后，当前的质心就会失效，需要计算每个新簇的自己的新质心

Step3.新质心的计算

对于分类后的产生的 k 个簇，分别计算到簇内其他点距离均值最小的点作为质心（对于拥有坐标的簇可以计算每个簇坐标的均值作为质心）

说明：

A.比如一个新簇有 3 个样本：[[1,4], [2,5], [3,6]]，得到此簇的新质心 =[(1+2+3)/3, (4+5+6)/3]

B.经过 step3，会得到 k 个新的质心，作为 step2 中使用的质心

Step4.是否停止 K-means

质心不再改变，或给定 loop 最大次数 loopLimit

说明：

A 当每个簇的质心，不再改变时就可以停止 k-menas

B.当 loop 次数超过 looLimit 时，停止 k-means

C.只需要满足两者的其中一个条件，就可以停止 k-means

C.如果 Step4 没有结束 k-means，就再执行 step2-step3-step4

D.如果 Step4 结束了 k-means，则就打印(或绘制)簇以及质心

我们使用 jupyter notebook 进行 python 代码编写，得出分类图像和所有数据的分类结果。部分代码展示图如下：

三维图像展示：

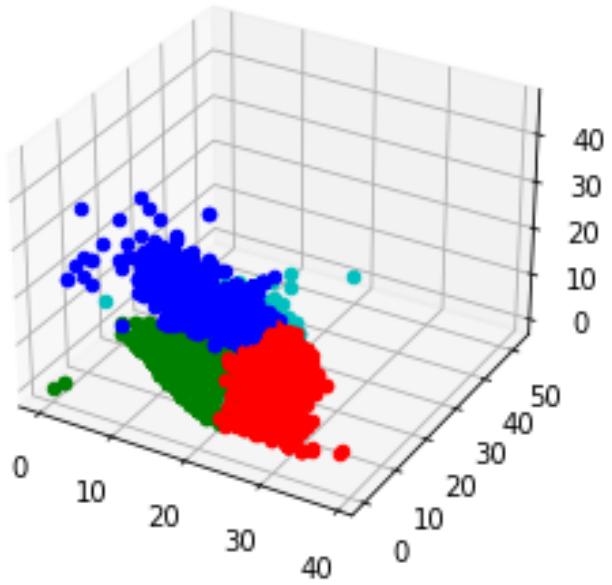


图 3　三维图像显示

使用 matlab 绘图工具绘制散点图矩阵如下：

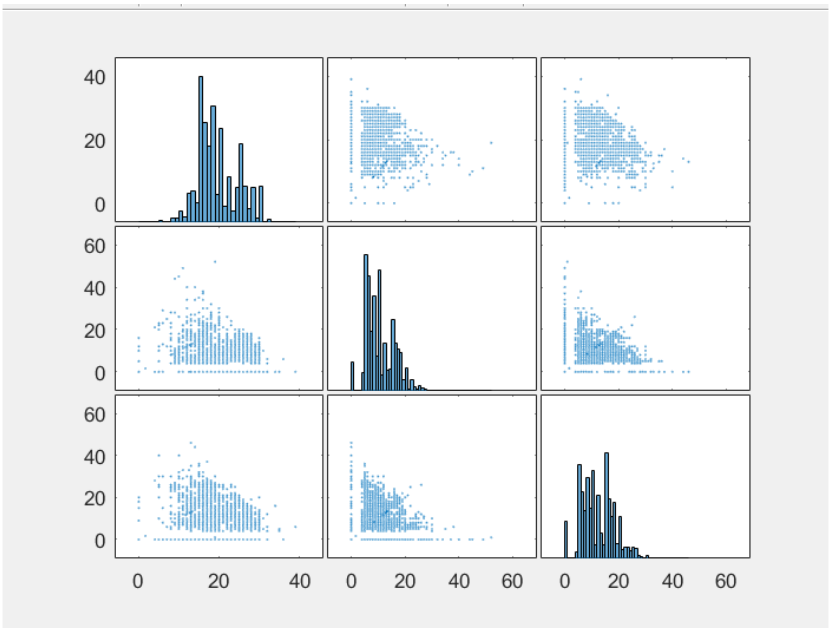
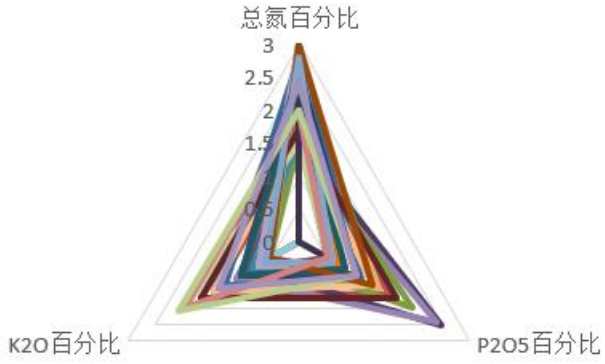


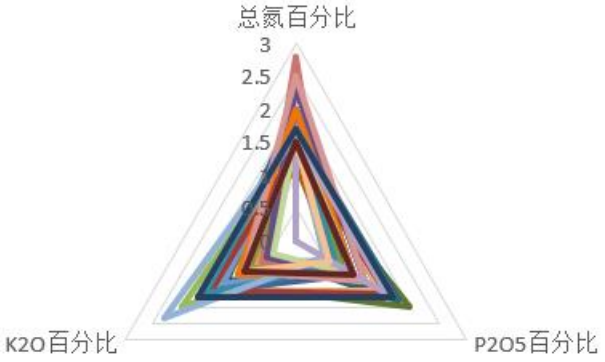
图 4　散点图矩阵

雷达图：
第一类分类雷达图：

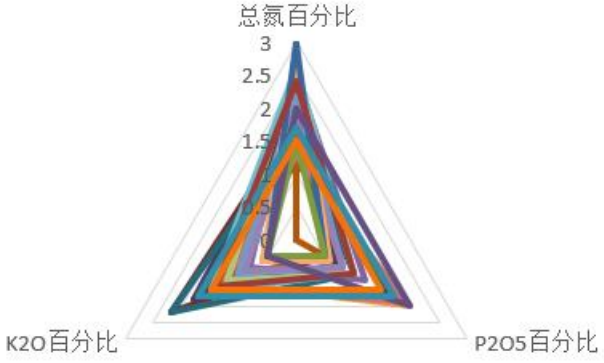
第一类雷达图

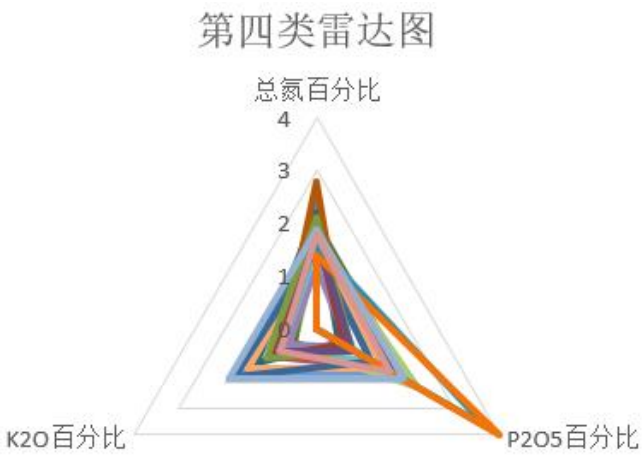


第二类雷达图



第三类雷达图





3. 任务三 肥料产品的多维度对比分析 1

3.1 复混肥料各组别不同年份产品登记数量的变化趋势

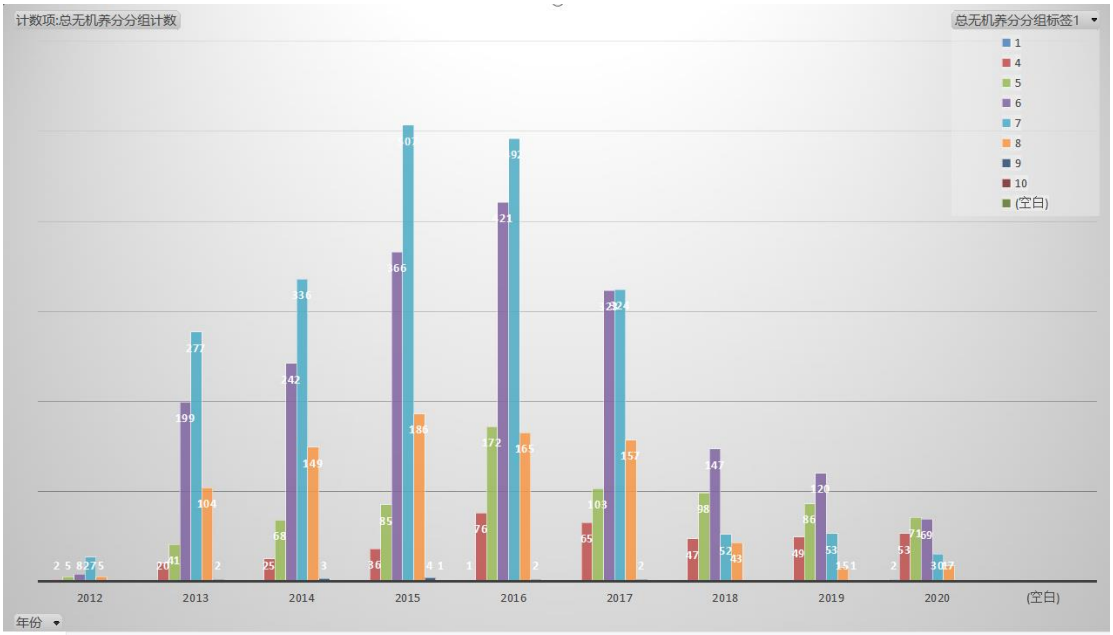


图 5 各组别不同年份数据透视图

3.2 杰卡德相似系数矩阵分析

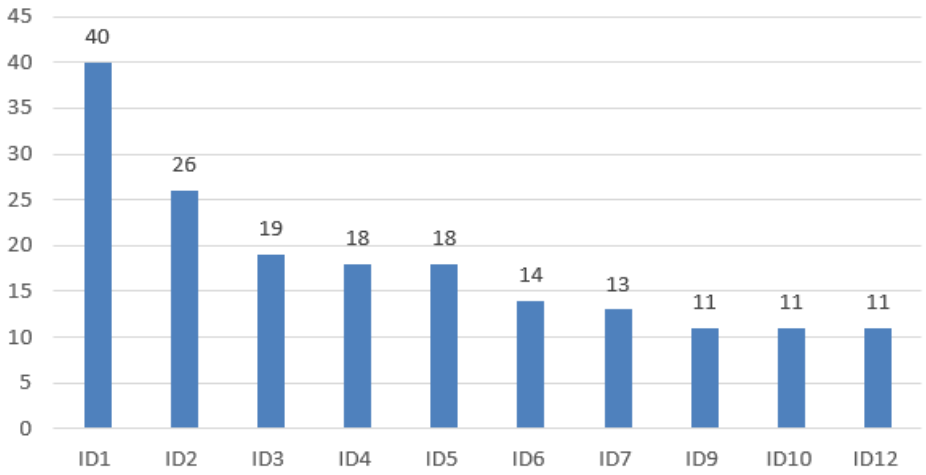


图 6　登记数量大于 10 的企业图

4. 任务四 肥料产品的多维度对比分析 2

4.1 各项百分比和含氯程度值的提取

4.2 原料名称和百分比的提取

5. 参考文献

[1] 肥料的分类与科学施用 刘玲 湖南农业 2021-06-01 期刊

[2] 分类中基于 k-means 的特征选择算法研究 陈晨 西安电子科技大学 2014 年

[3] 一种改进的散点图矩阵及其在 R 软件中的实现 金林 统计与决策