
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，现对肥料的成分和类别进行分析，从而加强肥料管理，保护生态环境，保障人畜安全，促进农业生产。

目录

1、问题分析	3
2、任务 1 数据的预处理	3
2.1 产品通用名称存在规范化处理	3
2.2 计算各肥料产品的氮、磷、钾养分百分比之和	3
3、任务 2 肥料产品的数据分析	3
3.1 分析复混肥料产品的分布特点	3
3.1.1 数据探索与预处理	3
3.1.2 绘制直方图复混肥料产品的分布和分析特点	4
3.2 分析有机肥料产品的分布特点	6
3.2.1 数据探索与预处理	6
3.2.2 绘制有机肥料产品的分布热力图和分析特点	6
3.3 使用聚类算法对产品进行分类	7
4、任务 3 肥料产品的多维度对比分析	8
4.1 复混肥料中各组别不同年份产品登记数量变化	8
4.2 分析广西、湖北省份有效期内有机肥料产品的分布差异	8
4.3	8
5、任务四 肥料产品的多维度对比分析	8
5.1	8
5.2 设计算法提取各种原料的名称及其百分比	8

1、任务目标

1. 对肥料登记数据进行预处理。
2. 根据养分的百分比对肥料产品进行细分。
3. 从省份、日期、生产商、肥料构成等维度对肥料登记数据进行对比分析。
4. 对非结构化数据进行结构化处理。

2、任务 1 数据的预处理

2.1 产品通用名称存在规范化处理

2.2 计算各肥料产品的氮、磷、钾养分百分比之和

3、任务 2 肥料产品的数据分析

3.1 分析复混肥料产品的分布特点

3.1.1 数据探索与预处理

针对同一列数据出现的不同情况，本次分析将会采用的处理方式如下：

1、针对缺失值情况，需要进行实际的分析，回归到原始数据中去，判断该数据为 0 时是否具有实际意义，如果没有就进行缺失值占比分析，发现缺失占比很小，可将其做删除处理，删除缺失值这一行数据。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7619 entries, 0 to 7618
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  ---                ---
0   序号                    7619 non-null   int64
1   企业名称                7619 non-null   object
2   产品通用名称            7619 non-null   object
3   产品形态                7618 non-null   object
4   总氮百分比              7619 non-null   float64
5   P205百分比              7619 non-null   float64
6   K20百分比              7619 non-null   float64
7   含氯情况                7619 non-null   object
8   有机质百分比            7619 non-null   float64
9   正式登记证号            7619 non-null   object
10  发证日期                7619 non-null   object
11  有效期                  7619 non-null   object
12  产品商品名称            1534 non-null   object
13  适用作物                1 non-null      object
14  总无机养分百分比        7619 non-null   float64
dtypes: float64(5), int64(1), object(9)
memory usage: 893.0+ KB
```

图 1 附件 2 数据表信息描述

2、针对异常值情况，在本次数据分析过程中，针对已经进行数据缺失值处理之后的产品形态这一列数据进行异常值处理。

。

3.1.2 绘制直方图复混肥料产品的分布和分析特点

由上图和产品登记数量直方图可知：总无机养分百分比的取值在(0.432, 0.504]范围内的产品数量最多，几乎占了 1/3，产品登记数量最大的前 3 个分组的总无机养分百分比的取值分别为：(0.432, 0.504]、(0.36, 0.432]、(0.288, 0.36]，因此可以了解到复混肥料产品主要分布在总无机养分百分比的范围是[0.30, 0.50]内，查询资料发现复混肥料产品的分布特点主要是分布在中高浓度的范围内，这个浓度的复混肥料养分齐全，可促进土壤养分平衡，明显增强作物抗逆性，适宜于生育期长的多年生、需肥量大、经济价值高的作物。

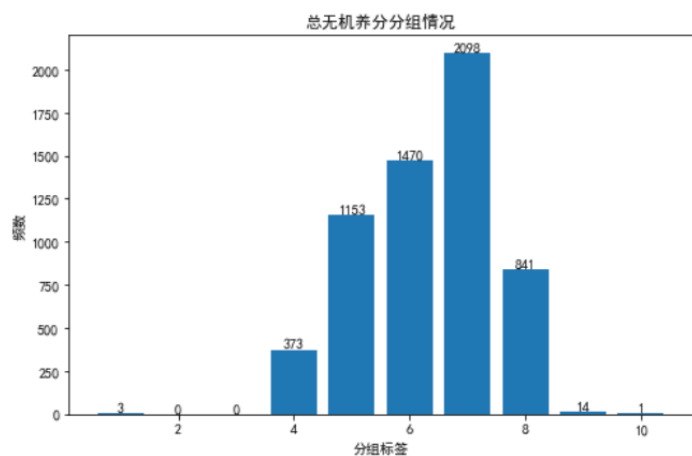


图 2 产品登记数量直方图

排名	一	二	三
分组标签	7	6	5
产品登记数量	2098	1470	1153

3.2 分析有机肥料产品的分布特点

3.2.1 数据探索与预处理

3.2.2 绘制有机肥料产品的分布热力图和分析特点

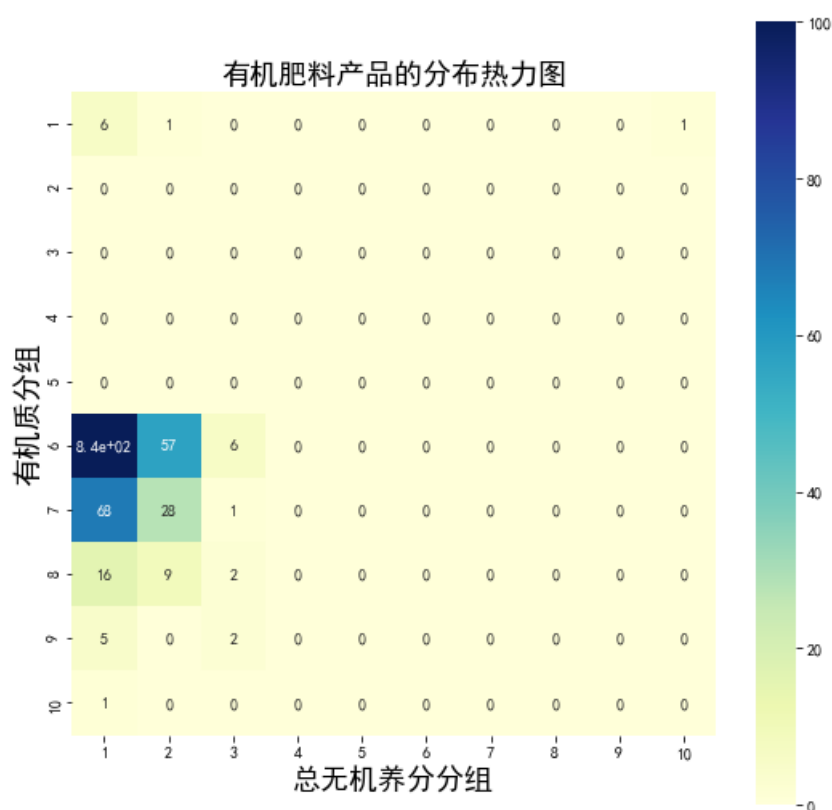


图 3 有机肥料产品分布热力图

由上图可知，在有机肥料中，按照总无机养分分组，标签越小表示总无机养分百分比越少，按照有机质分组，标签越小表示有机质百分比越小。通过观察热力图，可以发现：在有机肥料中，无机养分的含量是很少的，有机质的含量比较多的，而这也符合实际生活中的有机肥料特点，同时，也能观察到所有总无机养分分组会集中在前三个分类中，在有机肥料中，利用 `value_counts` 函数对有机质百分比这一列分组计数结果可得，有机质的百分比集中在（0.45，0.63）中。

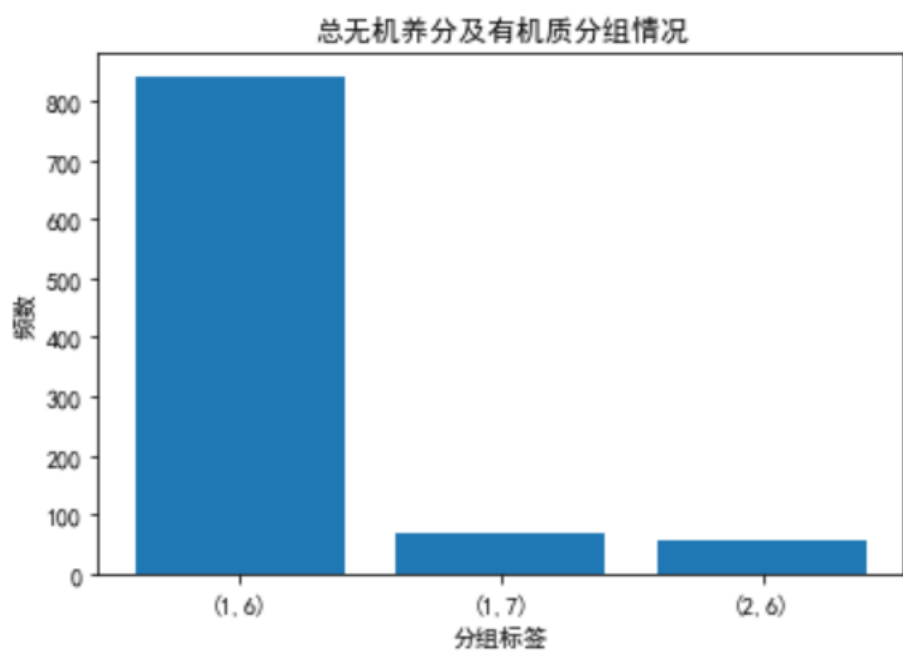


图 4 总无机养分及有机质分组数量登记

3.3 使用聚类算法对产品进行分类

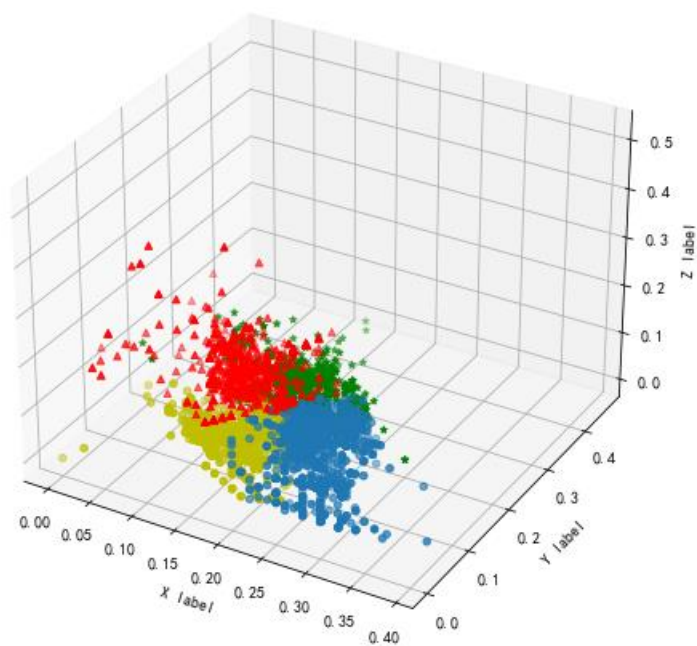


图 5 聚类标签三维散点图

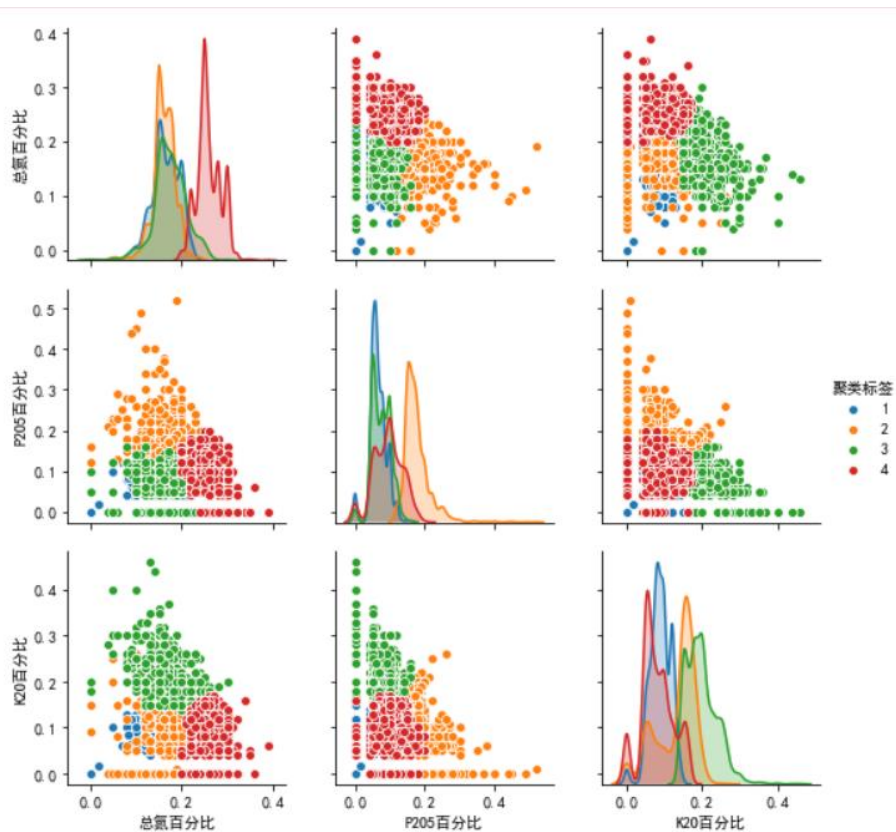


图 6 聚类标签散点图矩阵

4、任务 3 肥料产品的多维度对比分析

4.1 复混肥料中各组别不同年份产品登记数量变化

4.2 分析广西、湖北省份有效期内有机肥料产品的分布差异

4.3 计算产品登记数量大于 10 的企业间的杰卡德相似系数

5、任务四 肥料产品的多维度对比分析

5.1 提取氮、磷、钾养分、有机质的百分比和含氯程度

5.2 设计算法提取各种原料的名称及其百分比