肥料登记数据分析

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。本文将以肥料产品为研究对象，利用肥料登记数据，根据养分百分比、有机质含量及生产原料细分肥料，分析产品数量的变化趋势及分布差异，并根据其技术指标设计处理流程，对肥料养分、有机质等含量进行多维度分析。

目录

1　问题分析

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 数据的基本处理

2.1 数据探索

本任务用于数据预处理的**“附件 1”**表格数据，利用 python 读取文件并将其命名为 task_1，使用 shape 函数打印数据形状，得到 task1 总共有 2925 行、12 列数据。

2.2 数据清洗

2.2.1 重复值处理

本次数据分析中，**重复值**是指该数据表中用于分析的各个字段数据均为一致。在数据清洗过程中，首先对数据进行重复值的删除处理。

2.2.2 缺失值处理

2.2.3 异常值处理

1. 异常值情况说明

完成重复值和缺失值的处理之后，对数据进行异常值处理，在本次数据分析中，**异常值**是指产品登记数据中**养分、有机质百分比为负值**或包含**特殊符号**的异常数据

2.3 规范化产品通用名称

2.3.1 规范标准说明

产品通用名称不规范，指的是**命名与通用名称不统一、包含特殊字符，如空格、换行符等无意义符号的产品名称数据。**

本文产品通用名称**规范化的要求为：**，按照复混肥料（掺混肥料归入此）、有机-无机复混肥料、有机肥料和床土调酸剂这 4 种类别，对其进行规范化处理；

2.4 总无机养分百分比统计

2.4.1 计算方法

总无机养分百分比，是指各肥料产品的氮、磷、钾养分百分比之和，结合附件 1 具体数据，可知总无机养分百分比的计算公式有：

3 肥料产品的数据分析

3.1 复混肥料产品分组登记数量分析

3.1.1 分组处理描述

3.1.2 产品登记数量分析——基于直方图

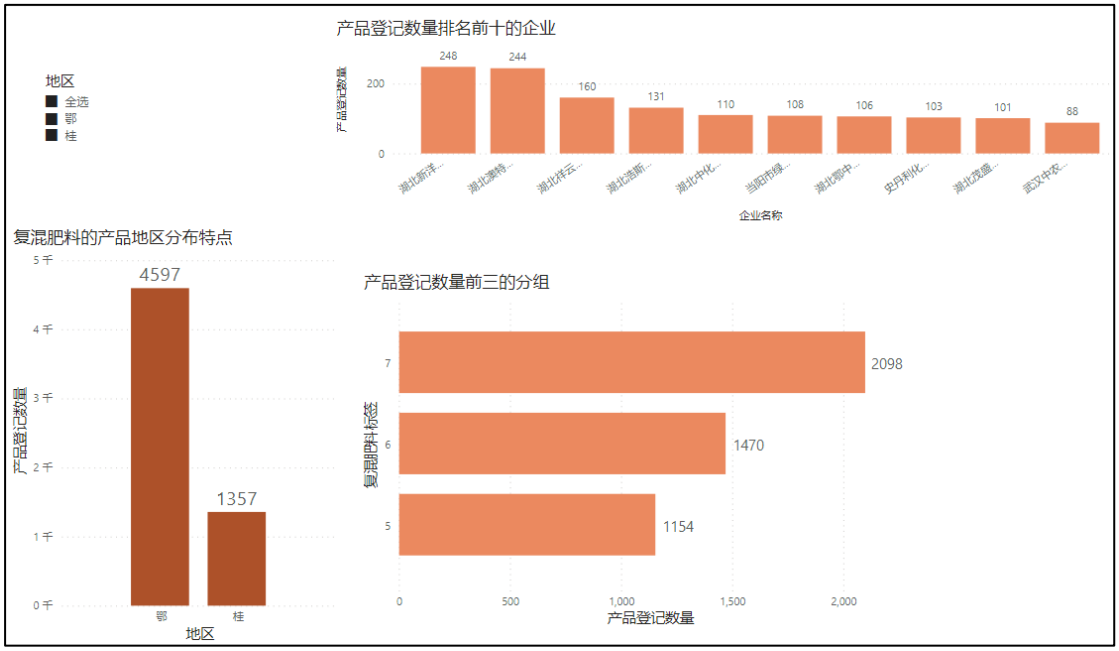

图 3-1 复混肥料产品登记数量分析

从上图按登记数量从大到小列出登记数量最大的前 3 个分组图可以看出，复混肥料第 7 组的复混肥料类的产品登记数量最多，达到了 2098 个复混肥料类的产品，其次是复混肥料第 6 组，达到了 1470 个复混肥料类的产品，最后是第 5 组，达到了 1470 个复混肥料类的产品；可以看出：复混肥料类的产品登记数量多的组别多集中在第 7 组、第 6 组和第 5 组这些总无机养分百分比较为适中的组别，由此可以推测：总无机养分百分比适中的的复混肥料更科学，也更容易被接受。

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

3.2 有机肥料产品分组登记数量分析

3.2.1 产品分布分析——基于热力图

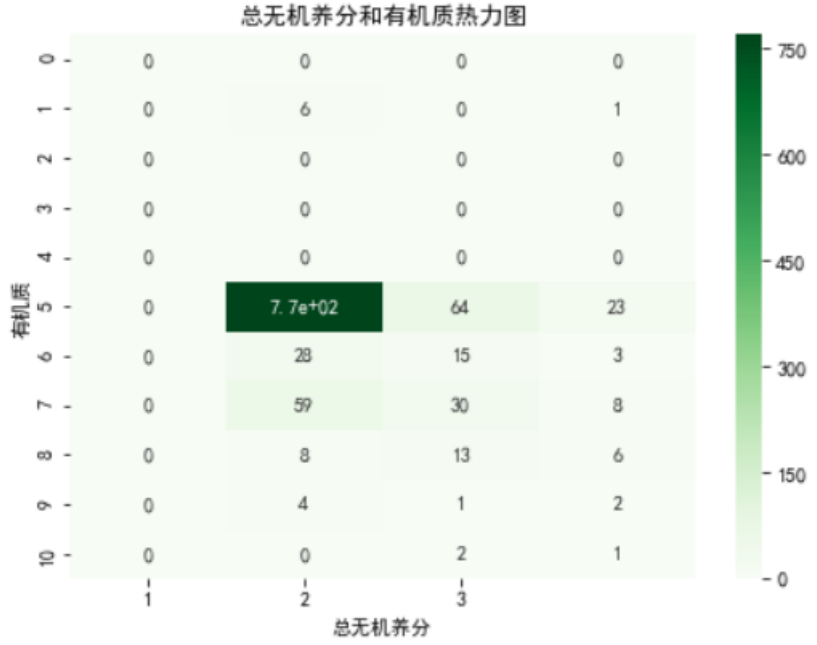

图 3-2 有机肥料产品含量分布热力图

3.3 基于 K-Means 复混肥料产品聚类分析

3.3.1 K-Means 聚类算法理论

K-means 算法[1]属于一种动态聚类算法，又称逐步聚类法,目的是将 n 个数据对象划分为 k 个簇类,每个类的对象具有高度的相似性。首先随机选取 k 个对象作为簇的均值点或中心点，然后计算每个对象与 k 个簇类均值或者中心点的距离，并将其指派到离它最近的簇类均值或者聚类中心所在的簇中，然后更新簇的均值或者中心点。

如此循环往复,直至均值或者中心点不再变化为止，即上式（3-2）收敛，本算法的基本流程如下:

输入：簇的数目 k 和包含 n 个对象的数据集；

输出：满足目标的 k 个簇集合。

①从数据集中任意选择 k 个对象作为初始的簇类中心;

②循环③到⑤,根据簇中对象的平均值,将每个对象赋予最类似的簇。直到目标函数 E 不再发生变化为止。

③计算更新簇的均值或者中心点,即有：

④计算每个对象：

⑤直至 E 不再明显变化。

3.3.2 聚类结果分析——基于三维散点图、散点图矩阵、雷达图

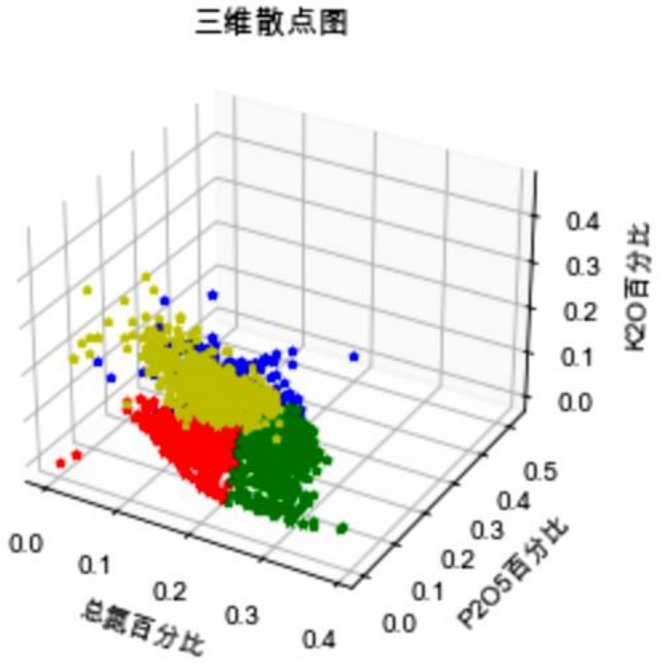

图 3-3 复混肥料聚类三维散点图

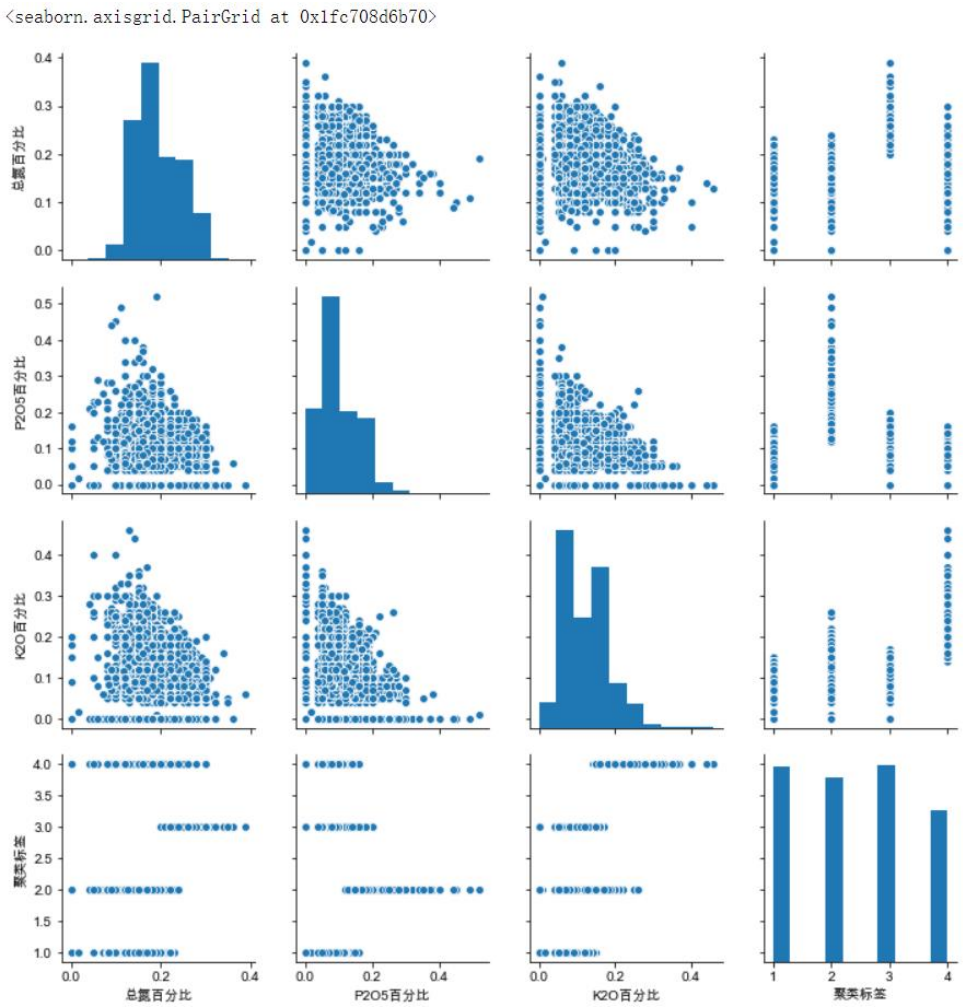

图 3-4 复混肥料聚类结果的散点图矩阵

4 品登记数量的多维度对比分析

4.1 复混肥料的登记数量变化趋势分析

4.1.1 年份提取描述

4.1.2 各组别不同年份产品登记数量可视化分析

4.2 有机肥料的分布差异分析

4.2.1 有效产品的提取描述

4.2.2 广西、湖北有效有机肥料产品分布差异可视化分析

4.3 肥料企业关于原料的杰卡德相似系数矩阵

4.3.1 杰卡德相似系数概念

杰卡德相似系数，在本次数据分析中，是指衡量两个企业所用的肥料产品原料集合相似度的指标。

企业所用原料集合 A 和集合 B 的交集元素在 A，B 并集中所占的比例，称为两个原料集合的杰卡德相似系数。

5 产品原料的多维度对比分析

5.1 技术指标规范化处理

5.1.1 计算肥料产品含量百分比

5.1.2 复混肥料有机质百分比替换

5.1.3 规范化含氯情况

5.2 基于原料与百分比提取各种原料的名称及其百分比

5.2.1 表格预处理

5.2.2 文本预处理

6 参考文献

[1] 李卫军.K-means 聚类算法的研究综述[J]研究与开发,2014:31.