Hans Journal of Agricultural Sciences 农业科学, 2025, 15(4), 498-504
Published Online April 2025 in Hans. https://www.hanspub.org/journal/hjas
https://doi.org/10.12677/hjas.2025.154062

Hans汉斯

# 基于K-Means聚类算法的地区农业差异性分类研究

**杨　鹏**

四川建筑职业技术学院基础教学部，四川　德阳

收稿日期：2025年3月18日；录用日期：2025年4月17日；发布日期：2025年4月27日

## 摘　要

**农业是国民经济的基础，不仅为我们的生存和发展提供了基本的生活资料，还对社会经济的发展与稳定具有至关重要的作用。因此，从农业差异性的角度，对我国各地区进行分类，可以为制定更具针对性和科学性的政策提供数据支持，其具有重大意义。本文通过选取1999年至2019年我国31个地区的地区生产总值及农业总产值，建立了基于K-means聚类算法的地区农业差异性分类研究模型。首先，以农业经济占比率作为聚类基础，利用K-means聚类算法将31个地区分为农业为主型地区和非农业为主型地区；进一步地，为优化聚类结果，在K-means聚类算法中结合肘部法则，确定最优聚类数为4类，进而将31个地区分为农业高质量地区，农业中质量地区，农业发展一般地区及农业欠发展地区，且通过三个评价指标的优化率，即轮廓系数优化率为10.40%；DBI优化率为28.42%，CH优化率为104.72%，可以看出优化后的聚类效果得到了显著提升；最后，基于分类结果，对4类地区，提出与之相适应的农业发展建议，为相关地区在制定农业发展规划时提供针对性意见。**

## 关键词

**农业总产值，地区生产总值，K-Means聚类算法，肘部法则，优化**

# Research on the Classification of Regional Agricultural Differences Based on K-Means Clustering Algorithm

**Peng Yang**

Department of Basic Education, Sichuan College of Architectural Technology, Deyang Sichuan

Received: Mar. 18th, 2025; accepted: Apr. 17th, 2025; published: Apr. 27th, 2025

文章引用: 杨鹏. 基于 K-Means 聚类算法的地区农业差异性分类研究[J]. 农业科学, 2025, 15(4): 498-504.
DOI: 10.12677/hjas.2025.154062

## Abstract

**Agriculture is the foundation of the national economy. It not only provides us with the basic means of subsistence and development, but also plays a crucial role in the development and stability of the social economy. Therefore, classifying various regions in China from the perspective of agricultural differences can provide data support for formulating more targeted and scientific policies, which is of great significance. This paper selects the gross agricultural output value and regional gross domestic product of 31 regions in China from 1999 to 2019, and establishes a research model for classifying agricultural regions based on K-means clustering algorithm. Firstly, taking the proportion of agricultural economy as the basis of clustering, the K-means clustering algorithm is used to divide the 31 regions into regions mainly based on agriculture and regions not mainly based on agriculture. Furthermore, in order to optimize the clustering results, the elbow method is combined with the K-means algorithm, and the optimal number of clusters is determined to be 4. Then, the 31 regions are divided into four categories, namely regions with high-quality agriculture, regions with medium-quality agriculture, regions with general agricultural development, and regions with underdeveloped agriculture. And through the optimization rates of three evaluation indicators, that is, the silhouette coefficient optimization rate is 10.40%; the DBI optimization rate is 28.42%, and the CH optimization rate is 104.72%. It can be seen that the clustering effect after optimization has been significantly improved. Finally, the corresponding agricultural development suggestions are put forward for the four classified regions based on the classification results, which providing targeted opinions when formulating agricultural development plans.**

## Keywords

**Gross Agricultural Output Value, Regional Gross Domestic Product, K-Means Clustering Algorithm, Elbow Method, Optimization**



## 1. 引言

改革开放以来，我国对农业的重视程度日益加深，为实现我国由农业大国到农业强国的转化，推出了一系列的农业发展政策，使得我国农业总产值获得了极大的提高。但由于地区差异性，导致各地区的农业发展存在较大差异，为了给各地区制定更符合各自农业发展水平的农业政策，在政府领导下，1981年编制的《中国综合农业区划》依据人口、耕地、农、林、牧、渔业的分布将全国划分为10个一级农业区和38个二级农业区；2011年中国农业出版社出版的《中国农业功能区划研究》依据大的地理界线和大的区域发展水平，又将全国农业功能区划分为了10个一级区、45个二级区。通过农业区的划分，使得各地区在制定相应的农业政策时更具针对性和科学性。当前农业发展的新态势下，本文通过国家统计局官网获得我国31个地区的地区生产总值和农业总产值数据，基于此对31个地区农业差异性的分类问题进行了深入研究，可为相关部门在制定地区性的农业政策时提供数据支持，具有重大的意义。

## 2. 基于K-Means聚类算法的地区农业差异性分类模型建立与求解

### 2.1. 基于K-Means算法的农业地区分类模型的建立

为更好地分析我国31个地区的农业差异性，并基于该差异性对各地区实现分类。首先，本文所用于

分析的地区生产总值及农业总产值均为经济指标，而一个地区是否可划分为农业为主的地区，可通过该地区的农业总产值占据地区生产总值的比值来确定，即如果一个地区的农业总产值占地区生产总值的比值较大，那么这个地区可划为农业为主型的地区，反之，若一个地区的农业总产值占地区生产总值的比值较小，那么这个地区可划为非农业为主型的地区。因此，为便于实现 31 个地区的农业差异性分类，首先将各地区农业总产值占该地区生产总值的比值定义为农业经济占比率，即：

$$w_i = \frac{N_i}{G_i}(i = 1, \cdots, 31)$$

其中，$N_i$ 表示第 $i$ 个地区的平均农业总产值，$G_i$ 表示第 $i$ 个地区的平均地区生产总值。

因此，本文首先对 1999 年至~2019 年 31 个地区的地区生产总值及农业总产值进行平均值的求取，进而计算出相应的农业经济占比率如表 1 所示。

**Table 1.** The proportion of agricultural economy of 31 regions
**表 1.** 31 个地区的农业经济占比率

| 地区名 | 农业总产值(亿元) | 地区生产总值(亿元) | 农业经济占比率 | 地区名 | 农业总产值(亿元) | 地区生产总值(亿元) | 农业经济占比率 |
|---|---|---|---|---|---|---|---|
| 北京 | 123.4980 | 14220.8371 | 0.0087 | 湖北 | 1779.1043 | 17631.5313 | 0.1009 |
| 天津 | 151.2813 | 8998.9681 | 0.0168 | 湖南 | 1796.2552 | 16844.8830 | 0.1066 |
| 河北 | 2137.0532 | 19237.2805 | 0.1111 | 广东 | 1828.4877 | 45839.5005 | 0.0399 |
| 山西 | 573.0795 | 8308.9862 | 0.0690 | 广西 | 1382.7532 | 9815.8386 | 0.1409 |
| 内蒙古 | 873.4812 | 9832.3561 | 0.0888 | 海南 | 376.9590 | 2206.8157 | 0.1708 |
| 辽宁 | 1130.2112 | 15965.7927 | 0.0708 | 重庆 | 658.8440 | 8916.0167 | 0.0739 |
| 吉林 | 806.4498 | 8056.5438 | 0.1001 | 四川 | 2126.6853 | 18389.7433 | 0.1156 |
| 黑龙江 | 1704.1702 | 9578.3333 | 0.1779 | 贵州 | 892.7634 | 5870.6462 | 0.1521 |
| 上海 | 133.6886 | 16652.7467 | 0.0080 | 云南 | 1129.5464 | 8321.9948 | 0.1357 |
| 江苏 | 2276.8566 | 42325.0952 | 0.0538 | 西藏 | 47.2713 | 612.2096 | 0.0772 |
| 浙江 | 987.0119 | 27253.8171 | 0.0362 | 陕西 | 1125.3151 | 10555.5053 | 0.1066 |
| 安徽 | 1429.8441 | 13579.7681 | 0.1053 | 甘肃 | 708.1965 | 4149.8732 | 0.1707 |
| 福建 | 991.8406 | 16047.4295 | 0.0618 | 青海 | 87.0161 | 1371.1410 | 0.0635 |
| 江西 | 855.5382 | 10008.6700 | 0.0855 | 宁夏 | 175.8365 | 1673.9667 | 0.1050 |
| 山东 | 3150.7045 | 37860.9344 | 0.0832 | 新疆 | 1237.3900 | 5727.3076 | 0.2161 |
| 河南 | 3013.3776 | 23076.3357 | 0.1306 | | | | |

## 2.2. 基于 K-Means 算法的农业地区分类模型的求解

在获得各地区农业经济占比率后，以该地区农业经济占比率作为聚类基础，采用 K-means 算法对其进行聚类，从而实现各地区的农业差异性分类。

K-means 算法是典型的聚类算法，其具有可伸缩性强，收敛速度快等优点，是一种广泛使用的聚类算法。其算法基本原理为：首先，设 $k$ 为整个数据集需要划分的类别数，随机选择 $k$ 个样本数据作为 $k$ 个类别的初始聚类中心；然后，计算其他样本数据到每个聚类中心的距离，通过比较距离的大小，将该样本数据划分到与之距离最近的聚类中心所在的类别中；最后，计算每个类别的几何中心，将几何中心分

别作为 $k$ 个类别新的聚类中心，重新计算每个样本数据到新的聚类中心的距离，并将样本数据重新划分到与之距离最近的聚类中心所在的类别中，不断重复这个过程，直到每个类别的中心点收敛为止，此时每个类别内的数据相似性最大[1] [2]。

利用 K-means 聚类算法对 31 个地区进行基于农业经济占比率的划分步骤如算法 1 所示：

算法 1：

Step 1：选择表 1 中 31 个地区的农业占比率作为聚类样本集 $X$；

Step 2：设定初始聚类数为 $k=2$；

Step 3：在样本集 $X$ 中，随机选取 $k$ 个样本，将这 $k$ 个样本作为初始聚类中心，设为 $C_1, C_2, \cdots, C_k$；

Step 4：分别计算其它样本数据到每个聚类中心的欧氏距离，计算公式为：

$$d\left(X_i, C_i\right)=\sqrt{\left(x_i-c_j\right)^2},\ i=1,2,\cdots,31;\ j=1,2,\cdots,k$$

Step 5：通过比较各样本点到各个聚类中心的距离，将样本点划分到与其距离最小的类中；

Step 6：重新计算每个类的几何中心，作为新的聚类中心；

Step 7：重复进行第 Step 4~Step 6，直至聚类中心位置收敛，聚类结束。

基于算法 1，可得到将 31 个地区聚为 2 类的结果，其具体的地区分类结果如表 2 所示。

**Table 2.** Regional clustering results (number of clusters = 2)
**表 2.** 地区聚类结果(聚类数为 2)

| 聚类类别 | 地区 | 平均农业经济占比率 |
|---|---|---|
| 农业为主型(聚类类别 1) | 河北、吉林、黑龙江、安徽、河南、湖北、湖南、广西、海南、四川、贵州、云南、陕西、甘肃、宁夏、新疆 | 13.41% |
| 非农业为主型(聚类类别 2) | 北京、天津、山西、内蒙古、辽宁、上海、江苏、浙江、福建、江西、山东、广东、重庆、西藏、青海 | 6.24% |

由表 2 可以看出，聚类类别 1 中所包含的 16 个地区，其平均农业经济占比率为 13.41%，其中的新疆，黑龙江，甘肃等地区都是我国重要的粮食生产地区，因此，将聚类类别 1 定义为农业为主型地区；聚类类别 2 所包含的 15 个地区，其平均农业经济占比率为 6.24%，其中的上海，江苏，浙江等地区都是我国经济发达地区，其主要依赖第三第二产业，因此将其定义为非农业为主型地区。

## 3. 基于 K-Means 聚类算法的地区农业差异性分类模型结果分析及优化

### 3.1. 基于 K-Means 聚类算法的地区农业差异性分类模型结果分析

**Table 3.** Evaluation metrics for clustering results (number of clusters = 2)
**表 3.** 聚类结果评价指标(聚类数为 2)

| 轮廓系数 | DBI | CH |
|---|---|---|
| 0.519 | 0.651 | 48.014 |

为更好的评价上述聚类结果，本文通过轮廓系数，DBI 及 CH 三个指标对其进行评价，其中轮廓系数是一个样本集合中所有样本轮廓系数的平均值，其取值范围是[−1, 1]，当同类别样本距离越相近，而不同类别样本距离越远时，分数越高，此时聚类效果越好，即轮廓系数越大表示聚类效果越好；DBI (Davies-bouldin)：用来衡量任意两个簇的簇内距离与簇间距离之比，该指标越小表示聚类效果越好；CH (Calinski-Harbasz Score)是通过计算类内各点与类中心的距离平方和来度量类内的紧密度(分母)，通过计算类间中

心点与数据集中心点距离平方和来度量数据集的分离度(分子)，CH 指标由分离度与紧密度的比值得到，CH 越大表示聚类效果越好。为此，基于上述聚类结果，分别计算三个评价指标，其结果如表 3 所示。

由表 3 可得，将 31 个地区聚为 2 类时，其轮廓系数为 0.519，勉强超过一半，而 DBI 为 0.651，CH 为 48.014 这两个指标均未达到一半，可见该聚类结果总体来说并不理想，有待进一步提高。

### 3.2. 基于 K-Means 聚类算法的地区农业差异性分类模型结果优化

基于聚类结果三个指标的评价而言，可知其聚类结果并不理想，究其原因，是因为在利用 K-means 算法进行聚类求解时，由于事先并不知道类的特征，因此初始聚类数 $k$ 通常是人为设定的，其具有一定的偶然性，也就可能造成较差的聚类结果，为更好的确定初始聚类数，可利用肘部法则来确定最优聚类数[3]。在肘部法则的应用中需要用到误差平方和(SSE: Sum of the Squared Errors)，SSE 与聚类效果之间的关系为：若 SSE 越小，即在同一类别中的数据距离聚类中心越近，则聚类效果越好，反之，若 SSE 越大，即在同一类别中的数据距离聚类中心越远，则聚类效果越差[4]。

肘部法则的原理为：在 K-means 聚类算法中，随着 $k$ 的增大，SSE 会逐渐减小。比如，当 $k=1$ 时，所有的数据都在同一类中，则 SSE = 0，当 $k$ 取样本数时，每个样本独立为一类，则 SSE 为最大值，因此，随着 $k$ 的增大，SSE 会逐渐减小，直到达到一个临界点，此时再增加 $k$ 的值，SSE 就不会显著减小了，这个临界点就是图像中的“肘部”点，对应的 $k$ 值就是最优的 $k$ 值。

因此，结合肘部法则，可得到确定最优聚类数的步骤如算法 2 所示：

算法 2：

Step 1：设定聚类数 $k$ 的范围为 2 到 10；

Step 2：对于每个 $k$ 值，运行 K-means 聚类算法，并计算 SSE；

Step 3：绘制 $k$ 与 SSE 之间的关系图，找出 SSE 下降速度最快的点，即图像中的“肘部”点；

Step 4：根据“肘部”点确定最优的 $k$ 值。

基于算法 2，可绘制出 $k$ 与 SSE 之间的关系图如图 1 所示。

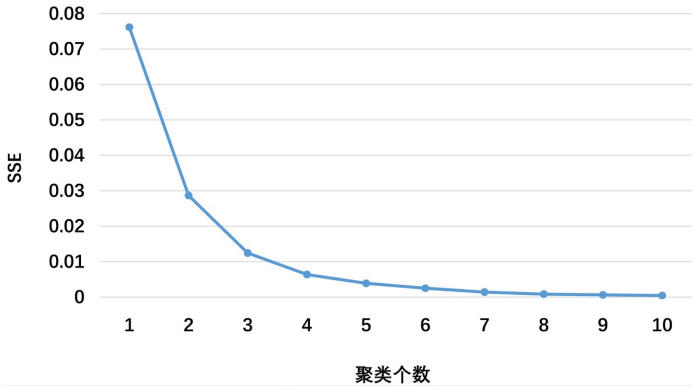


**Figure 1.** Trend plot of the number of clusters and SSE
**图 1.** 聚类数 $k$ 与 SSE 变化趋势图

由图 1 可看出，随着聚类数 $k$ 的增大，SSE 逐渐减小，且其减少趋势为先快后慢，最终趋于稳定。具体而言，当 $k<4$ 时，随着 $k$ 增大，SSE 减少较快，但当 $k>4$ 以后，随着 $k$ 的增大，SSE 减少便不那么显著了，由此可以得到，$k=4$ 即为“肘部”点，即是所求的最佳聚类数。

基于算法 2，可得最佳聚类数为 $k=4$，因此，重新设定算法 1 中的聚类数为 4，可得优化后的地区聚类结果如表 4 所示：

**Table 4.** Regional clustering results (number of clusters = 4)
**表 4.** 地区聚类结果(聚类数为 4)

| 聚类类别 | 地区 | 平均农业经济占比率 |
|---|---|---|
| 农业高质量地区(聚类类别 1) | 黑龙江、海南、贵州、甘肃、新疆 | 17.75% |
| 农业中质量地区(聚类类别 2) | 河北、吉林、安徽、河南、湖北、湖南、广西、四川、云南、陕西、宁夏 | 11.44% |
| 农业发展一般地区(聚类类别 3) | 山西、内蒙古、辽宁、江苏、福建、江西、山东、重庆、西藏、青海 | 7.27% |
| 农业欠发展地区(聚类类别 4) | 北京、天津、上海、浙江、广东 | 2.19% |

由表 4 可以看出，聚类类别 1 中所包含的 5 个地区，其平均农业经济占比率为 17.75%，其平均农业经济占比率排第一，可见这些地区农业发展质量高，因此将聚类类别 1 定义为农业高质量地区；聚类类别 2 中所包含的 11 个地区，其平均农业经济占比率为 11.44%，其平均农业经济占比率排第二，可见这些地区农业发展质量较高，因此，将聚类类别 2 定义为农业中质量地区；聚类类别 3 中所包含的 10 个地区，其平均农业经济占比率为 7.27%，其平均农业经济占比率排第三，可见这些地区农业发展质量一般，因此，将聚类类别 3 定义为农业发展一般地区；聚类类别 4 中所包含的 5 个地区，其平均农业经济占比率为 2.19%，其平均农业经济占比率排名最后，可见这些地区农业发展质量较弱，因此，将聚类类别 4 定义为农业欠发展地区。

为便于评价该优化后聚类效果，仍然选取轮廓系数，DBI 及 CH 三个评价指标，其结果如表 5 所示。

**Table 5.** Evaluation metrics for clustering results (number of clusters = 4)
**表 5.** 聚类结果评价指标(聚类数为 4)

| 轮廓系数 | DBI | CH |
|---|---|---|
| 0.573 | 0.466 | 98.296 |

由表 5 可得：相较于优化前聚类数为 2 时的聚类结果，该优化后聚类结果的三个指标中，轮廓系数由 0.519 提高到了 0.573，其优化率为 10.40%；DBI 由 0.651 下降到了 0.466，优化率为 28.42%；CH 由 48.014 提高到了 98.296，优化率为 104.72%。可见，结合肘部法则选取得到最优聚类数后，使得优化后的 K-means 聚类算法的聚类效果得到了显著的提升。

## 4. 结论与建议

本文建立了基于 K-means 聚类算法的地区农业差异性分类模型，对 1999 年至 2019 年期间 31 个地区的农业总产值及地区总产值进行了深入的研究，利用结合肘部法则的 K-means 聚类算法将 31 个地区分为农业高质量地区、农业中质量地区、农业发展一般地区及农业欠发展地区，从轮廓系数、DBI 及 CH 三个评价指标来看，优化后的聚类算法所分的四个类别，分类效果更好，更符合地区差异性。针对所聚类后的四类农业地区，分别对各类地区的农业发展提出如下建议。

### 4.1. 农业高质量地区

农业发展高质量地区包括黑龙江、海南、贵州、甘肃、新疆等 5 个地区，这些地区具有良好的自然条件、基础设施和技术水平，农业发展水平高。对于这些地区，应当在现有技术上继续推进农业机械化，进一步优化种植结构，减少化肥和农药的使用，保持农业发展的可持续性和稳定性。此外，可进一步发展现代农业，延伸产业链，增加附加值等。

### 4.2. 农业中质量地区

农业发展中质量地区包括河北、吉林、安徽、河南、湖北、湖南、广西、四川、云南、陕西、宁夏等11个地区，这类地区已初步形成符合各自地区特点的地区性农业，具有一定的发展潜力和优势，在进一步的农业发展中应继续结合当地优势，发展多元化农业；增强农业基础设施投资，提高农业生产条件和农民农业收入[5]；优化产业布局，发挥区域特色优势，培育和扩大特色农业产业。

### 4.3. 农业发展一般地区

农业发展一般地区包括山西、内蒙古、辽宁、江苏、福建、江西、山东、重庆、西藏、青海等10个地区，这10个地区农业生产水平相对较低，需在农业发展过程中加强政策支持，为农业生产提供优惠政策和财政支持；引进先进的农业生产技术和管理经验，提高农业生产效率和产品质量；加强农业产业链建设，促进农产品加工、物流、销售等环节的协调发展。

### 4.4. 农业欠发展地区

农业欠发展地区包括北京、天津、上海、浙江、广东，这5个地区均为我国经济靠前的城市或省份，这些地区有着人均耕地较少，生产规模较小，科技发展较高的特点。对此类地区，可以充分利用其科技优势，将传统农业与现代科技进行有效结合，创建科技主导的现代化农业；积极推进农业结构的调整，通过科学布局，发展规模化、集约化的生产基地；加强智慧农业建设，引进推广新的农业科技成果和优良品种及配套技术，推进农业标准化生产，提高农业园区的智能化和高端化水平。

## 参考文献

[1] 黄志敏, 梁承东. 基于K-means聚类算法的等级测评数据分析[J]. 电子质量, 2023(12): 40-44.

[2] 杨阳. 数据挖掘K-means聚类算法的研究[D]: [硕士学位论文]. 长沙: 湖南师范大学, 2015.

[3] 吴广建, 章剑林, 袁丁. 基于K-means的手肘法自动获取K值方法研究[J]. 软件, 2019, 40(5): 167-170.

[4] Cui, M. (2020) Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance*, **1**, 5-8.

[5] 郭翔宇. 推进农业高质量发展, 以农业强省支撑农业强国建设[J]. 农业经济与管理, 2022(6): 4-7.