

湖南省各城市农业基本情况、农业生产条件和农业生产总值 多元统计分析



|  |
| --- |
| **指导老师：汪炎汝** |
| **姓 名：汪俣充** |
| **班 级：统计1701** |
| **学 号：1303170104** |

摘要

“三农”问题是全面建成小康社会、实现“中国梦”所必须克服的关键性问题。农业生产总值和粮食产量关乎农民和农业。而湖南作为农业大省，省内各城市的农业发展相关研究具有较大意义。

本文搜集了2018年湖南省统计年鉴中14个城市的数据，从农业基本情况和农业生产条件这两个方面选取了10个指标，进行聚类分析和主成分分析，并通过回归分析研究其与粮食总产量和农业生产总值之间的关系。

本文首先进行了聚类分析，选择的聚类方法为离差平方和法，距离使用的是欧式距离。并根据聚类结果和实际情况，将14个城市分为了四类：良好类（邵阳、衡阳和常德），一般类（益阳、郴州、永州、怀化和岳阳），省会类（长沙），不佳类（株洲、湘潭、张家界、娄底和湘西）

然后对这10个指标进行主成分分析，针对中心化标准化后的数据，使用相关系数矩阵求主成分。根据方差贡献率和主成分碎石图，并结合实际情况，选择了3个主成分，累积方差贡献率达到92.11%。第一主成分为农村人员、耕地资源和生产条件的综合指标。第二主成分为土地资源和用电投入的综合指标。第三主成分为土地退化和生产条件的综合指标。

最后分别以粮食总产量和农业生产总值作为被解释变量，三个主成分作为解释变量，建立多元线性回归模型。并通过显著性检验和残差分析修正模型，剔除了第二主成分。农业生产总值的回归模型中剔除了长沙、常德和衡阳三个异常值点。根据研究结果为政府采取措施提高农业水平提出了相应的政策建议。

**关键字** 农业生产总值 农业基本情况 农业生产条件

聚类分析 主成分分析 回归分析

目录

[1.绪论 1](#_Toc18596339)

[1.1研究背景及意义 1](#_Toc18596340)

[1.2文献综述 2](#_Toc18596341)

[1.3 研究内容 3](#_Toc18596342)

[2理论知识及指标体系介绍 4](#_Toc18596343)

[2.1主要指标 4](#_Toc18596344)

[2.2理论基础 6](#_Toc18596345)

[2.2.1聚类分析 6](#_Toc18596346)

[2.2.2主成分分析 7](#_Toc18596347)

[2.2.3主成分回归 8](#_Toc18596348)

[3基于基本多元统计方法的分析 9](#_Toc18596349)

[3.1聚类分析 9](#_Toc18596350)

[3.1.1聚类过程 9](#_Toc18596351)

[3.1.2聚类数的选择 10](#_Toc18596352)

[3.1.3聚类结果 11](#_Toc18596353)

[3.2主成分分析 11](#_Toc18596354)

[3.2.1求解主成分及个数选择 11](#_Toc18596355)

[3.2.2主成分结果及解释 13](#_Toc18596356)

[3.3回归分析 13](#_Toc18596357)

[3.3.1粮食总产量模型的建立与优化 13](#_Toc18596358)

[3.3.2农业生产总值模型的建立与优化 16](#_Toc18596359)

[3.3.3针对模型结果提出的建议 21](#_Toc18596360)

[4总结与展望 22](#_Toc18596361)

[4.1本文优点 22](#_Toc18596362)

[4.2不足与展望 22](#_Toc18596363)

[参考文献 23](#_Toc18596364)

[附录 25](#_Toc18596365)

湖南省各城市农业基本情况、农业生产条件和农业生产总值 多元统计分析

# 1.绪论

## 1.1研究背景及意义

中国的“三农”发展历来都是关系党和人民事业发展的全局性和根本性任务，也是全面建成小康社会、实现“中国梦”所必须克服的关键性问题。研究“三农”问题就是为了促进农民增收、农业发展和农村稳定。改革开放40年以来，中国的经济状况突飞猛进，农业和农村经济发展取得了举世瞩目的重大成就。但是，我们也必须要注意到，长远来看，我国农业收入持续稳定增长的内在动力和后劲仍然不足，当前中国的城乡居民收入差距仍然处于高位徘徊的状态,城乡居民收入比持续下降的趋势不断弱化。因此，我们不得不承认，在眼下，甚至是之后较长的一段时期内，进一步解决“三农”问题、实现农业收入可持续增长和缩小城乡居民收入差距仍是中国经济持续发展必须面临的难题。只有解决好了农民增收的可持续性问题，不断缩小城乡居民收入差距，才能找到快速且精准地消除。

而湖南省是一个农业大省，农业人口数量众多，超四千多万，耕地资源也非常丰富，将近九百万公顷，素有“渔米之乡”的称号，农产品的种类十分丰富。而湖南省气候温和、光照充足、降雨量充沛，属于中亚热带季风湿润气候，非常适合各类农作物的生长。近年来，通过实施乡村振兴战略，湖南农业充分发挥传统优势和产业基础优势，着力培育优势产业，大力扶持发展农产品加工业，农业产业结构不断调整优化，农业产业化水平不断提高，农业产业的质量和效益显著提升。但是依旧存在着发展不平衡，要素配置不合理的问题。

农业基本情况和农业生产条件是农业发展的两个主要影响因素。而农业生产总值和粮食产量则是衡量农业发展水平的两个重要指标。探究他们之间的关系是解决“三农”问题的重要手段。而湖南省内各城市农业情况各不相同，农业发展水平并不均衡，根据农业基本情况和农业生产条件将它们进行分类，并针对不同的类提出不同的发展策略和政策性建议，有助于湖南省乡村振兴战略的实施。在农业基本情况和农业生产条件所包含的相关指标中，指标间相关性较强，反应的信息在一定程度上有所重叠，提取出其中的主要信息，将高维数据压缩，能给农业基本情况和农业生产条件的相关研究带来较大的简化。除此之外以这些指标作为解释变量，以粮食产量和农业生产总值作为被解释变量，建立多元回归模型，可以直观的反应各指标对于农业发展的影响，对于农业政策的调整、农业决策以及乡村振兴战略的实施有着较大的指导意义。

## 1.2文献综述

国家一号文件连续十五年关注三农问题，对于增加农民收入、缩小城乡差距的研究层出不穷。国内学者韩长赋、陈锡文、林毅夫等人已经把新阶段农业和农村问题的核心归结为农民收入的增长问题，进而有很多学者对农民增收问题进行了研究，其中多数是基于农民收入的影响因素分析，既有多因素分析，也有单因素分析。多因素分析例如：李雪艳、史少赞采用对数——线性模型探究农民农业收入影响的因素[1]，发现主要农产品单位面积产量和农业基础设施投入对农业收入正向影响显著；刘进宝、张延君利用辅助线性模型从宏观因素对农业收入影响因素进行计量分析[2]，发现农业劳动者人均农业收入和农业生产物质投入、国内人口总量以及所在地区国民经济发展水平具有正相关性，而技术进步引起农业收入下降；卢军静根据调查得到的500多份样本数据，采用单因素方差分析和多元线性回归[3]，总结出在当前，观光休闲农业是帮助农户实现农业增收的最有效途径，其次分别是养殖业和种植业；雷平、詹慧龙通过从生产、科技、制度、经营四个方面选取影响农业收入的14项指标构建模型对农民增收的影响因素进行分析[4]，得出结论：现阶段提高农民收入不仅在于促进非农就业，更在于提高农业效率和产能，加快农业机械化和信息化。单因素分析例如：例如从农业发展资金角度入手，探索农村贷款对促进农民收入的研究[5]，财政支农支出对农民收入的影响[6]；例如从市场的角度入手，探索农产品价格[7][8]与农业收入的关系；例如从农业生产条件角度入手，探索水利[9][10]、机械[11]、科技进步[12][13][14]等对农民农业收入的影响等等。

田珍在分析地区农业竞争力的基础上,依据全面性、可操作性、可比性、系统性和代表性原则,对地区农业竞争力从农业资源状况、农业经济增长与发展状况、农业产业化经营程度和农业现代化程度四个方面构建了一套农业竞争力评价指标体系。[15]

刁广付,王健,高卓以沧州市14个县区为研究对象,依据农业经济理论, 选择5个评价指标作为影响农业生产的主要因素, 通过聚类分析, 对沧州市14个县区农业生产条件的差异及其特征进行定量和定性分析, 并提出对策建议：农业生产条件好的地区, 要充分挖掘具有优势生产要素的潜力, 最大限度地提高农业生产效果。沧州市政府部门应根据各县区实际农业生产条件, 科学规划农业产区, 针对农业生产条件较好和一般的地区重点投资改造。[16]

张建平根据1991-2004年我国农业发展的统计资料,应用灰色关联分析方法,对农业生产条件与农业经济发展的关系进行了研究,结果显示农用固定资产资金额对农业经济发展的影响最大,农业物质生产条件对农业经济的影响明显,其中农机总动力对农业经济的影响近年来明显增强,而农村劳动力、耕地面积和灌溉面积对农业经济发展的影响较小[17]

王进,王礼力,庞新江通过对农村区域竞争力评估指标体系设计的探讨,, 设计出30个评估指标对我国各省进行了实证分析, 得出当前和将来很长一段时间, 农村区域竞争力非平衡性很难在短时间内得到扭转, 但对农村区域竞争力提升路径提出了新的要求, 应该有全新的思路努力提高农村地区的竞争力, 这需要各种力量的综合贡献。并建议中央和地方政府政策输出科学化和政策支持十分关键； 加快技术创新是必要手段； 加快推进农村区域工业化是现实选择。[18]

## 1.3 研究内容

本文选取了2018年湖南省统计年鉴的数据。首先，从农业基本情况和农业生产条件入手，选取10个指标变量，对湖南省内14个城市进行聚类分析。然后，针对这10个变量进行主成分分析，提取出较少的综合指标来表达原指标的信息。最后，分别以粮食产量和农业生产总值为被解释变量，进行主成分回归，量化研究这些农业基本情况和农业生产条件的相关指标对于城市农业发展水平的影响。并根据最终结果提出相应的政策性建议。

诚然，众多学者将新阶段农业和农村问题的核心归结为农民收入的增长问题，促进农民增收的因素的研究更普遍。但“三农”问题不仅是农民问题，还有农业问题。因此我们将农业生产总值和粮食产量这两个既能反映农民收入问题又能反映农业发展问题的对象作为研究主体，理由如下：

我们所研究的促进农民农业收入的因素——农业基本情况和农业生产条件主要与农民收入中的农业收入成分有关。农民收入结构在我国以前的统计口径中，农民收入按来源不同可分为：家庭经营收入 、工资性收入 、财产性收入 以及转移性收入 。2013年开始新的统计口径，农民可支配收入按来源不同分为：可支配工资性收入、可支配经营净收入、可支配财产净收入、可支配转移净收入。其中，工资性收入占有很大比例。如果研究农民收入与农业基本情况和农业生产条件的关系，则会导致随机误差项的因素过多，误差过大。

另外，广义的农业基本情况和农业生产条件包括自然条件、社会经济条件和技术装备条件等。农业生产自然条件又包括光热、水分、土地和生物条件等，社会经济条件评价包括社会经济发展历史、人口劳动力、交通运输、工业和城镇基础条件等；技术装备条件评价包括农业技术装备、技术水平和劳动素养等。这其中，自然条件和技术装备对农业生产的影响显然优于社会经济条件。例如，社会经济条件中的交通运输主要作用体现在农产品的市场流通环节；工业反哺农业可以体现在技术装备等方面；而由于中国自古以来人多地少，所以农业劳动力的投入对农业生产不能产生显著影响，在不考虑人均收入的情况下可以忽略。而水利、肥料、机械、耗材、种子等则直接作用于农业生产，因而选取这些因素作为本文研究的农业基本情况和农业生产条件。

# 2理论知识及指标体系介绍

## 2.1主要指标

**粮食产量（）**

粮食产量是全社会粮食作物的生产量。包括国有经济、集体经济和农民家庭经营的粮食产量以及工矿企业办的农场和其他生产单位的粮食产量。粮食除包括稻谷、小麦、玉米、高粱、谷子及其他杂粮外，还包括薯类和豆类。其产量计算方法，豆类按去豆荚后的干豆计算；薯类(包括甘薯和马铃薯，不包括芋头和木薯)1963年以前按每4公斤鲜薯折1公斤粮食计算，从1964年开始改为按5公斤鲜薯折1公斤粮食计算。城市郊区作为蔬菜的薯类(如马铃薯等)按鲜品计算，并且不作粮食统计。其他粮食一律按脱粒后的原粮计算。

**农业生产总值（）**

是一定时期（通常为一年）内以货币形式表现的农业全部产品的总量。反映农业生产总规模和总成果。计算方法是：产量分别乘以各自单位产品价格求得;少数生产周期较长，当年没有产品或产品产量不易统计的，采用间接方法匡算其产值。

**农业基本情况**

根据湖南省2018年统计年鉴的划分，农业基本情况包含乡村劳动力、年末实有耕地面积、当年减少耕地面积、造林面积、乡村户数和乡村从业人员数这几个指标。

**乡村劳动力（）**

乡村劳动力指乡村人口中经常参加合作经济组织（包括乡村办企业事业单位）和从事家庭经营生产劳动的整、半劳动力。

**耕地面积[年末实有耕地面积（）,当年减少耕地面积（）]**

耕地面积是可以用来种植农作物、经常进行耕锄的田地。包括：熟地、当年新开荒地、连续撂荒未满三年的耕地和当年的休闲地（轮歇地），以种植农作物为主并附带种植桑树、茶树、果树和其他林木的土地，沿海、沿湖地区已围垦利用的“海涂”、“湖田”等。

**造林面积（）**

造林面积指报告期内在荒山、荒地沙丘等一切可以造林的土地上，采用人工播种、植苗、飞机播种等方法种植成片乔木林和灌木林。

**乡村户数（）**

是指户口在农村的常住户数，包括参加乡村各级举办和各种行业的合作经济组织，并从中直接取得实物或货币收入家庭户数。但不包括乡村区域内的国有经济的机关、团体、学校、企业、事业单位的集体户。集体户口无论其人数多少，都以一户进行统计。

**乡村从业人员人数（）**

乡村从业人员指乡村人口中16岁以上实际参加生产经营活动并取得实物或货币收人的人员，既包括劳动年龄内实际参加劳动人员，也包括超过劳动年龄但实际参加劳动的人员，但不包括户口在家的在外学生、现役军人和丧失劳动能力的人，也不包括待业人员和家务劳动者。劳动者年龄为16岁以上。

**农业生产条件**

根据湖南省2018年统计年鉴的划分，农业生产条件包含农业机械总动力、有效灌溉面积、化肥施用量和农村用电量这几个指标。

**农业机械总动力（）**

农业机械总动力指主要用于农、林、牧、渔业的各种动力机械的动力总和。包括耕作机械、排灌机械、收获机械、农用运输机械、植物保护机械、牧业机械、林业机械、渔业机械和其他农业机械，按功率折成瓦计算。不包括专门用于乡、镇、村、组办工业、基本建设、非农业运输、科学试验和教学等非农业生产方面用的动力机械与作业机械。

**有效灌溉面积（）**

有效灌溉面积是地块比较平整，有一定水源、灌溉设施配套，在一般年景下当年能进行正常灌溉的农田面积。为能进行正常灌溉的水田和旱地中水浇地面积之和。有效灌溉面积的大小，是衡量农业生产单位和地区水利化程度和农业生产稳定程度的指标。

**化肥施用量（）**

指本年内实际用于农业生产的化肥数量，包括氮肥、磷肥、钾肥和复合肥。化肥施用量要求按折纯量计算数量。折纯量是指把氮肥、磷肥、钾肥分别按含氮、含五氧化二磷、含氧化钾的百分之一百成份进行折算后的数量。复合肥按其所含主要成分折算。

**农村用电量（）**

农村地区本年内消耗的总电量。

## 2.2理论基础

### 2.2.1聚类分析

从定义的角度来说，聚类分析是一组将研究对象分为相对同质的群组的统计分析技术。其基本思想是，我们所研究的样品或指标之间存在程度不同的相似性；于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间相似程度的统计量，以这些统计量为划分类型的依据；把一些相似程度较大的样品聚合为一类，把另外一些彼此之间相似程度较大的样品又聚合为另一类，直到把所有的样品聚合完毕。

为了将样品进行分类，就需要研究他们之间的关系。方法是将一个样品看作p维空间的一个点，并在空间定义距离，距离越近的点归为一类，距离较远的点归为不同的类。

欧氏距离（）

****

系统聚类的基本思想是，距离相近的样品先聚成类，距离相远的后聚成类，过程一直进行下去，每个样品总能聚到合适的类中。系统聚类过程是，假设总共有n个样品，第一步将每个样品独自聚成一类，共有n类；第二步根据所确定的样品 “距离”公式，把距离较近的两个样品聚合为一类，其它的样品仍各自聚为一类，共聚成n-1类；第三步将“距离”最近的两个类进一步聚成一类，共聚成n-2类；……，以上步骤一直进行下去，最后将所有的样品全聚成一类。为了直观地反映以上的系统聚类过程，可以把整个分类系统画成一张谱系图。

该方法的基本思想来自于方差分析，如果分类正确，同类样品的离差平方和应当较小，类与类的离差平方和较大。具体做法是先将n个样品各自成一类，然后每次缩小一类，每缩小一类，离差平方和就要增大，选择使方差增加最小的两类合并，直到所有的样品归为一类为止。

### 2.2.2主成分分析

在分析问题时，变量较多往往会增加分析问题的复杂性。而在实际问题中，变量之间可能存在一定的相关性，因此，多变量中可能存在信息的重叠。我们自然希望对这些相关的众多变量进行“降维”，通过克服相关性、重叠性，用较少的不相关的新变量反映原始变量所提供的绝大部分信息，通过对新变量的统计分析来解决问题。

利用线性变换，将原有的相关性较强的p个指标重新组合成p个新的相关性较弱的综合指标。通过选择适当的线性变换系数，使得新指标中的第一个指标方差最大(如有必要，还需使第二个指标方差尽可能地大)，从而最终可用较少的指标(如第1, 2个)表示原p个指标中足够多的信息，即降维，这便是主成分分析的基本思想。

值得一提的是，原始数据标准化可以克服量纲的影响，但是却会损失部分信息。一般而言，若变量计量单位不同或差异较大时，应该选基于相关系数矩阵的主成分分析，否则基于协方差的主成分分析效果可能更好。当原始数据标准化后，则样本协差阵和相关阵相同。

当一个变量只取一个数据时，这个变量（数据）提供的信息量是非常有限的，当这个变量取一系列不同数据时，我们可以从中读出最大值、最小值、平均数等信息。变量的变异性越大，说明它对各种场景的“遍历性”越强，提供的信息就更加充分，信息量就越大。主成分分析中的信息，就是指标的变异性，用标准差或方差表示它。 主成分分析的数学模型是，设p个变量构成的p维随机向量为X = （X1，…，Xp）′。对X作正交变换，令F = U′X，其中U为正交阵，要求F的各分量是不相关的，并且F的第一个分量的方差是最大的，第二个分量的方差次之，……，等等。为了保持信息不丢失，F的各分量方差和与X的各分量方差和相等。

### 2.2.3主成分回归

回归分析是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且自变量之间存在线性相关，则称为多重线性回归分析。

相关分析研究的是现象之间是否相关、相关的方向和密切程度，一般不区别自变量或因变量。而回归分析则要分析现象之间相关的具体形式，确定其因果关系，并用数学模型来表现其具体关系。

一般来说，回归分析是通过规定因变量和自变量来确定变量之间的因果关系，建立回归模型，并根据实测数据来求解模型的各个参数，然后评价回归模型是否能够很好的拟合实测数据；若能够很好的拟合，则可以根据自变量作进一步预测。

多元共线性是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。

而利用主成分作回归，一方面可以克服自变量间有较强的多重共线性的干扰，另一方面由于将变量降维处理，可以使回归方程的结构简化。

# 3基于基本多元统计方法的分析

## 3.1聚类分析

### 3.1.1聚类过程

首先为了克服量纲的影响，将数据中心化标准化。

指标体系中有10个变量，选用离差平方和法来进行聚类分析。在距离方面，选择欧式距离。算得各点间的欧式距离如下：

表1 各点之间的距离

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 长沙 | 株洲 | 湘潭 | 衡阳 | 邵阳 | 岳阳 | 常德 | 张家界 | 益阳 | 郴州 | 永州 | 怀化 | 娄底 |
| 株洲 | 4.42 |  |  |  |  |  |  |  |  |  |  |  |  |
| 湘潭 | 5.57 | 1.63 |  |  |  |  |  |  |  |  |  |  |  |
| 衡阳 | 4.11 | 5.23 | 6.71 |  |  |  |  |  |  |  |  |  |  |
| 邵阳 | 5.11 | 5.91 | 7.41 | 2.42 |  |  |  |  |  |  |  |  |  |
| 岳阳 | 3.43 | 3.36 | 4.6 | 3.35 | 3.99 |  |  |  |  |  |  |  |  |
| 常德 | 4.39 | 6.03 | 7.24 | 3.4 | 3.86 | 2.97 |  |  |  |  |  |  |  |
| 张家界 | 6.83 | 2.97 | 2.08 | 8 | 8.41 | 6.08 | 8.64 |  |  |  |  |  |  |
| 益阳 | 3.19 | 2.37 | 3.6 | 3.71 | 4.49 | 1.22 | 3.85 | 5.06 |  |  |  |  |  |
| 郴州 | 4.68 | 2.54 | 3.95 | 3.46 | 3.8 | 2.58 | 4.84 | 5.01 | 2.22 |  |  |  |  |
| 永州 | 4.56 | 3.95 | 5.35 | 2.54 | 2.9 | 1.91 | 3.47 | 6.6 | 2.43 | 1.85 |  |  |  |
| 怀化 | 3.58 | 3.16 | 4.62 | 4.05 | 3.87 | 2.88 | 4.72 | 5.23 | 2.61 | 2.63 | 3.09 |  |  |
| 娄底 | 4.69 | 1.13 | 2.03 | 5.31 | 5.79 | 3.7 | 6.43 | 3.2 | 2.78 | 2.53 | 4.07 | 3.33 |  |
| 湘西 | 5.88 | 1.88 | 1.54 | 6.82 | 7.12 | 4.71 | 7.25 | 1.67 | 3.84 | 3.81 | 5.29 | 4.07 | 2.29 |

其他的层次聚类方法也有尝试，但聚类结果不如离差平方和法理想，这也是意料之中的。

得到聚类结果图如下：

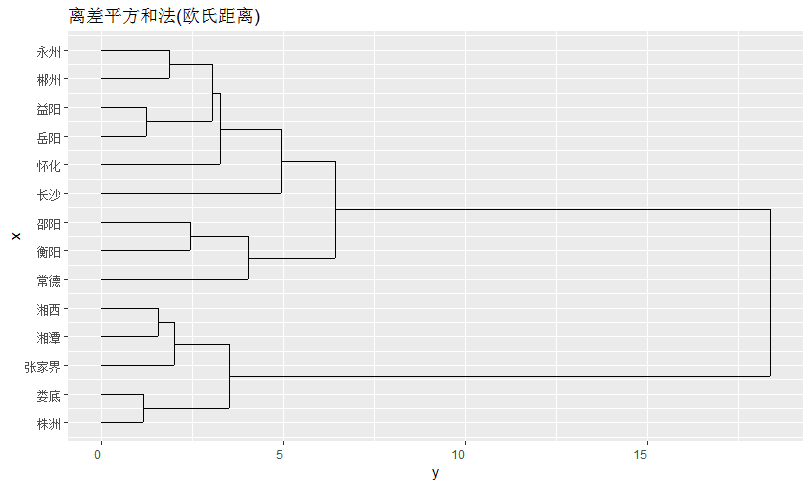


图1 聚类结果图

### 3.1.2聚类数的选择

从结果图来看，为保证各类中的元素不是太多，至少需要分成3类。而选取的研究对象为湖南省内14个市，根据研究目的，研究省内各市农业基本情况和农业生产条件的分类，类的个数不需要太多，将它们分成3-4类较为合适。如果只分为3类，长沙则会与永州和郴州等城市并为一类，而分为4类时，长沙则会分离出来单独成一类。长沙作为湖南的省会城市，在农业基本条件和农业生产投入这两个方面，相对于其他城市来说，有自己的特点。在农业基本情况方面，土地面积相关的指标具有明显的劣势，但是人口方面则不存在劣势；在农业生产条件方面，农业用电量和机械动力的指标有明显的优势，这是省会城市特有的优势。因此将长沙单独分出来，总共分为三类，会较为合适。

下面我们观察一下各类的中心点。

表2 各类的中心点

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 类别 | 乡村 劳动力 | | 年末实有耕地面积 | | 当年减少耕地面积 | 造林面积 | 乡村 户数 | 乡村从业人员 | 机械总动力 | 灌溉面积 | 化肥施用 | 用电量 |
| 1 | | 360.27 | | 44.92 | 0.08 | 6.32 | 166.6 | 330.03 | 561.77 | 347.7 | 28.04 | 142067 | |
| 2 | | 264.30 | | 33.17 | 0.06 | 4.87 | 120.3 | 241.92 | 531.89 | 247.5 | 19.91 | 78939 | |
| 3 | | 275.20 | | 27.42 | 0.13 | 0.98 | 124.5 | 244.20 | 602.04 | 231.6 | 18.56 | 200471 | |
| 4 | | 157.34 | | 17.42 | 0.03 | 2.23 | 67.7 | 139.62 | 262.26 | 126.6 | 8.60 | 52840 | |

从表格中，我们能较为直观的看出各类之间的差别。除第3类为省会类外，第1、2、4类各指标呈梯度下降，因此认为它们分别为农业基本情况和农业生产条件良好、一般和不佳类。

### 3.1.3聚类结果

综上所述，聚类结果如下：

**第一类：良好类**

邵阳 衡阳 常德

**第二类：一般类**

益阳 郴州 永州 怀化 岳阳

**第三类：省会类**

长沙

**第四类：不佳类**

株洲 湘潭 张家界 娄底 湘西

## 3.2主成分分析

### 3.2.1求解主成分及个数选择

本题选取的10个指标之间有一定的相关性，反应的信息在一定程度上有所重叠。

接下来使用主成分分析法将原来指标重新组合成一组新的互相无关的几个综合指标。

将标准化后的数据，基于相关系数矩阵进行主成分分析，获得主成分碎石图如下：

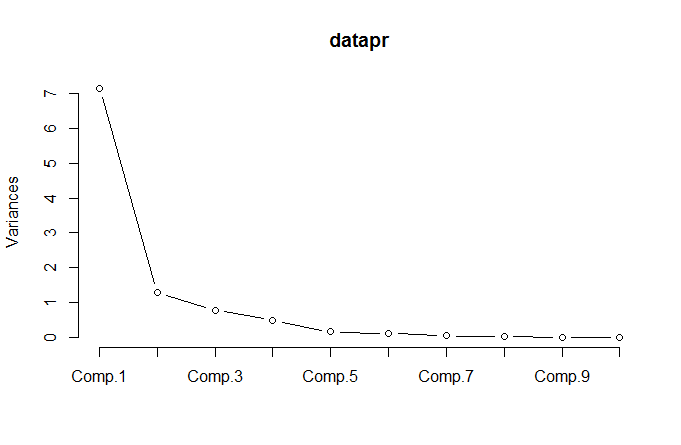
图2 主成分碎石图

表3 主成分方差贡献率表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 |
| 标准差 | 2.6732 | 1.1354 | 0.8807 | 0.6934 | 0.3898 | 0.3272 | 0.1880 |
| 方差贡献率 | 0.7146 | 0.1289 | 0.0775 | 0.048 | 0.0152 | 0.0107 | 0.0035 |
| 累积贡献率 | 0.7146 | 0.8435 | 0.9211 | 0.9692 | 0.9844 | 0.9951 | 0.9986 |

根据图表所示结果，选择前3个主成分较为合适，这三个主成分已经包含原指标超过92%的信息。

### 3.2.2主成分结果及解释

三个主成分的系数如下：

表4 主成分的系数表

|  |  |  |  |
| --- | --- | --- | --- |
| 原指标 | F1 | F2 | F3 |
| 乡村劳动力 | 0.349 | 0.214 | 0.241 |
| 年末实有耕地面积 | 0.358 | 0 | -0.174 |
| 当年减少耕地面积 | 0.211 | -0.47 | 0.517 |
| 造林面积 | 0.243 | 0.63 | 0 |
| 乡村户数 | 0.362 | 0.143 | 0.168 |
| 乡村从业人员 | 0.352 | 0.218 | 0.217 |
| 农业机械总动力 | 0.331 | -0.224 | -0.172 |
| 有效灌溉面积 | 0.325 | -0.157 | -0.465 |
| 化肥施用量 | 0.327 | -0.156 | -0.483 |
| 农村用电量 | 0.262 | -0.45 | 0.283 |

根据系数表，我们可以对主成分进行解释：

第一主成分F1中，各指标的系数较为平均，耕地面积和与人口有关的指标系数相对稍大一点，其次是三项农业生产条件。可以把第一主成分看成农村人员、耕地资源和生产条件的综合指标。

第二主成分F2中，减少的耕地面积、造林面积和农村用电量指标系数较大。可以把第二主成分看成土地资源和用电投入的综合指标。

第三主成分F3中，减少的耕地面积、有效灌溉面积和化肥施用量指标系数较大。可以把第三主成分看成土地退化和生产条件的综合指标。

## 3.3回归分析

### 3.3.1粮食总产量模型的建立与优化

农业基本情况和农业生产条件的10个指标间，具有较强的线性相关性。故而采用前文中提取的主成分作为解释变量来进行回归分析，以克服多重共线性的干扰。

首先以粮食总产量作为被解释变量，三个主成分作为解释变量，建立多元回归模型。由于前文已进行中心化和标准化处理，模型表达式中不含常数项。

模型中，F2的系数未通过t显著性检验，

表5 模型系数及参数检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 参数估计值 | 标准化残差 | t检验值 | Pr(>|t|) | 显著性 |
| F1 | 0.34809 | 0.02313 | 15.046 | 1.1e-08 | \*\*\* |
| F2 | -0.04801 | 0.05354 | -0.897 | 0.38907 |  |
| F3 | -0.26204 | 0.07038 | -3.723 | 0.00336 | \*\* |

于是接下来剔除F2重新建立二元回归模型，并进行对比。新模型各系数均通过了t显著性检验，调整可决系数由原模型的0.9472上升为0.9478。

表6 模型系数及参数检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 参数估计值 | 标准化残差 | t检验值 | Pr(>|t|) | 显著性 |
| F1 | 0.34997 | 0.02285 | 15.316 | 3.06e-09 | \*\*\* |
| F3 | -0.25709 | 0.06958 | -3.695 | 0.00307 | \*\* |

下面根据软件输出的图片，对模型进行残差分析。

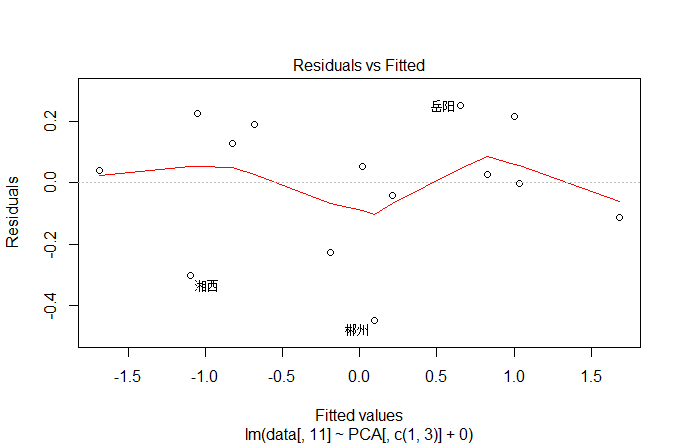


图3 残差图

残差分布随机，无明显趋势。

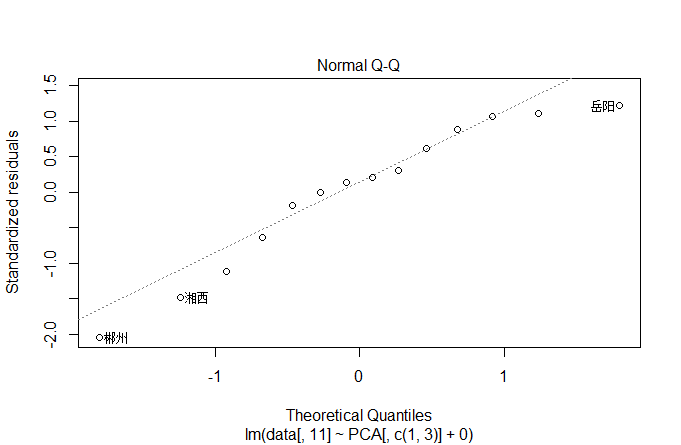


图4 Normal Q-Q图

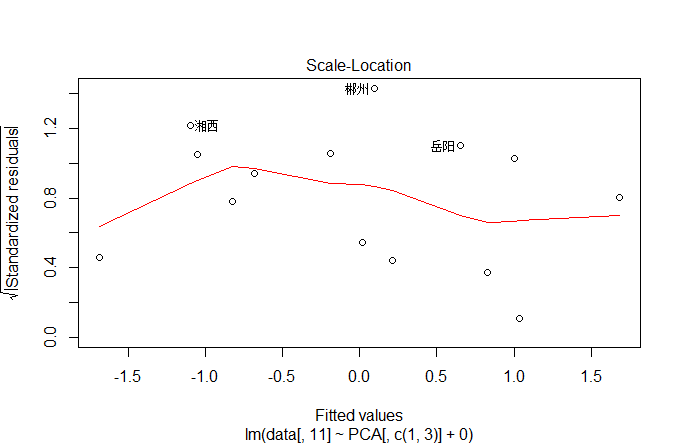
近似分布在一条直线上，残差满足正态假设。

图5 位置尺度图

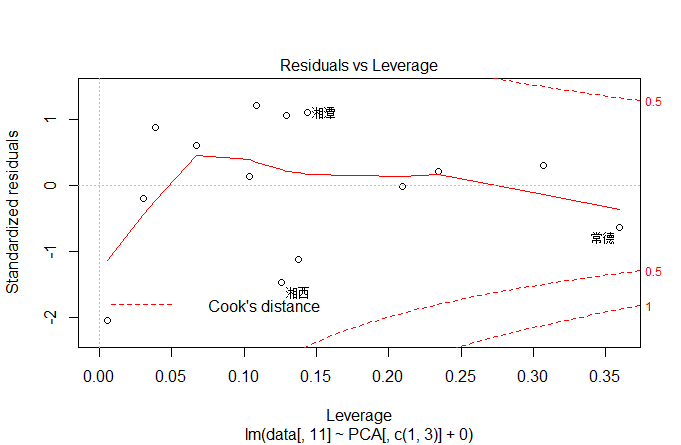
水平线附近的点分布随机，满足不变方差假设。 

图6 残差杠杆图

模型中无异常值点。

将主成分还原为原始的10个指标，

表7 原始变量标准化系数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 乡村劳动力 | 年末实有耕地面积 | 当年减少耕地面积 | 造林面积 | 乡村户数 |
| 0.0602 | 0.1700 | -0.0591 | 0.0850 | 0.0835 |
| 乡村从业人员 | 农业机械总动力 | 有效灌溉面积 | 化肥施用量 | 农村用电量 |
| 0.0674 | 0.1601 | 0.2333 | 0.2386 | 0.0189 |

再将中心化和标准化的变量还原为原始值。

得到最终的表达式为：

方程中各变量系数的正负性符合实际情况。

### 3.3.2农业生产总值模型的建立与优化

接下来，再以农业生产总值作为被解释变量，三个主成分作为解释变量，建立多元回归模型。该模型输出的残差杠杆图如下：

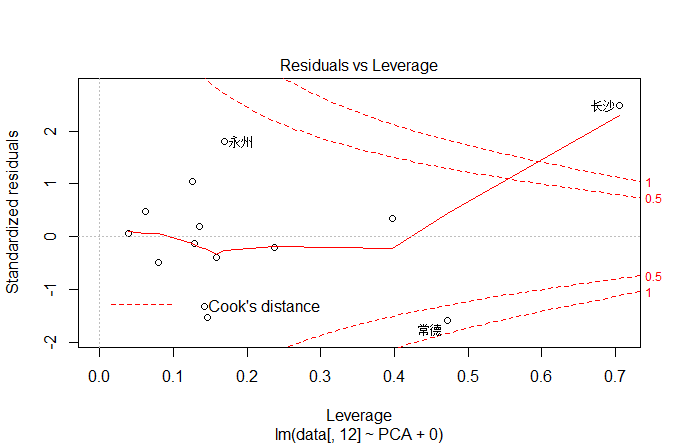


图7 残差杠杆图

图中可观察出，有两个明显的异常值点，即长沙和常德。于是我们将这两个点剔除，再重新建立多元回归模型。

新建立的模型中，F2的系数未通过t显著性检验，

表8 模型系数及参数检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 参数估计值 | 标准化残差 | t检验值 | Pr(>|t|) | 显著性 |
| F1 | 0.33842 | 0.02133 | 15.862 | 6.94e-08 | \*\*\* |
| F2 | -0.05479 | 0.07160 | -0.765 | 0.46378 |  |
| F3 | -0.28831 | 0.08683 | -3.320 | 0.00894 | \*\* |

于是剔除该自变量，重新建立二元回归模型。

表9 模型系数及参数检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 参数估计值 | 标准化残差 | t检验值 | Pr(>|t|) | 显著性 |
| F1 | 0.33506 | 0.02044 | 16.39 | 1.49e-08 | \*\*\* |
| F3 | -0.30619 | 0.08188 | -3.74 | 0.00385 | \*\* |

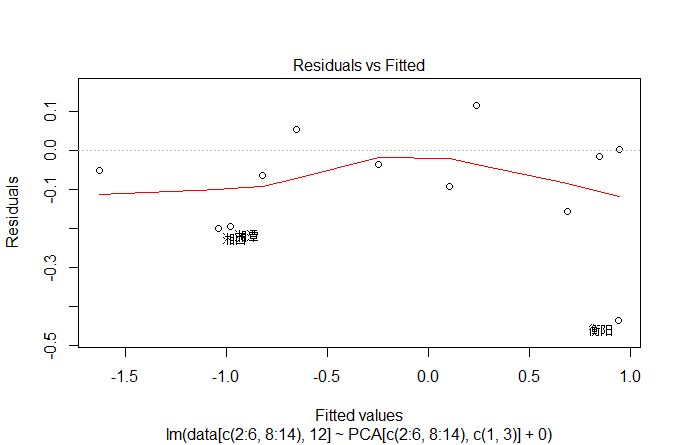
调整可决系数从0.9561提高到了到0.9579。但是输出的残差图和残差杠杆图中，仍存在一个明显的异常值点，衡阳。

图8 残差图

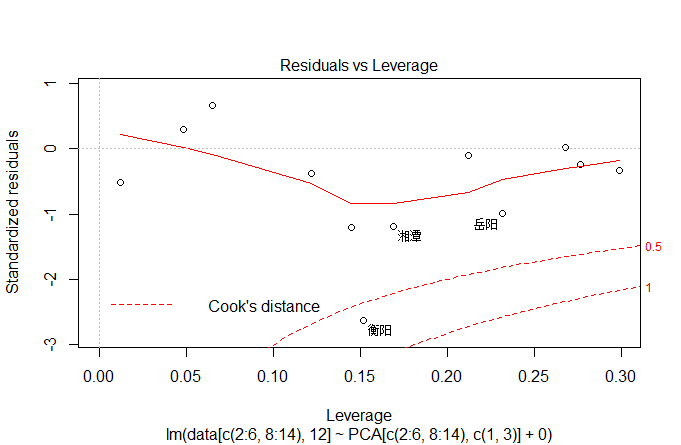


图9 位置尺度图

于是，剔除该点，重新建立二元回归模型。得到的最终结果为：

表10 模型系数及参数检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 参数估计值 | 标准化残差 | t检验值 | Pr(>|t|) | 显著性 |
| F1 | 0.35520 | 0.01288 | 27.58 | 5.25e-10 | \*\*\* |
| F3 | -0.27062 | 0.04894 | -5.53 | 0.000366 | \*\*\* |

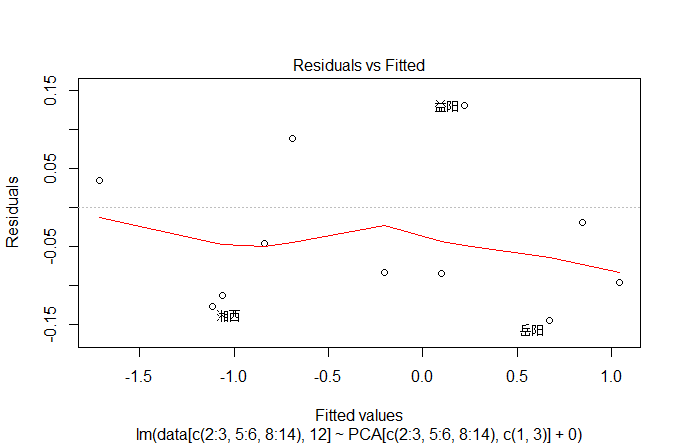


图10 残差图

残差分布随机，无明显趋势。

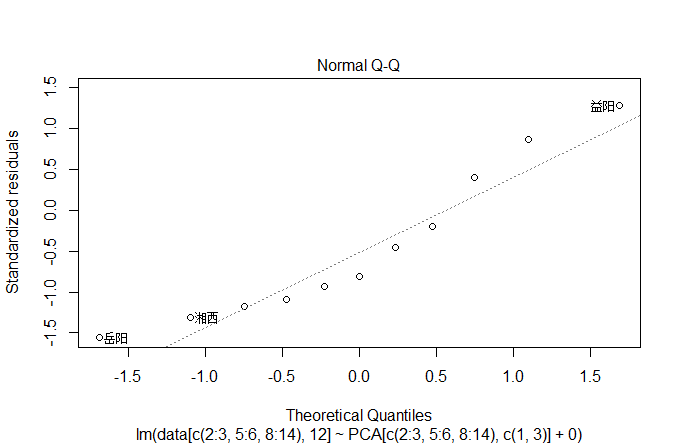


图11 Normal Q-Q图

近似分布在一条直线上，残差满足正态假设。

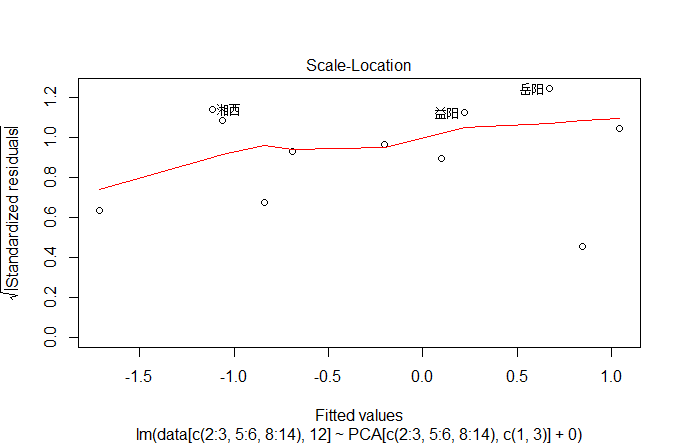


图12 位置尺度图

水平线附近的点分布随机，满足不变方差假设。

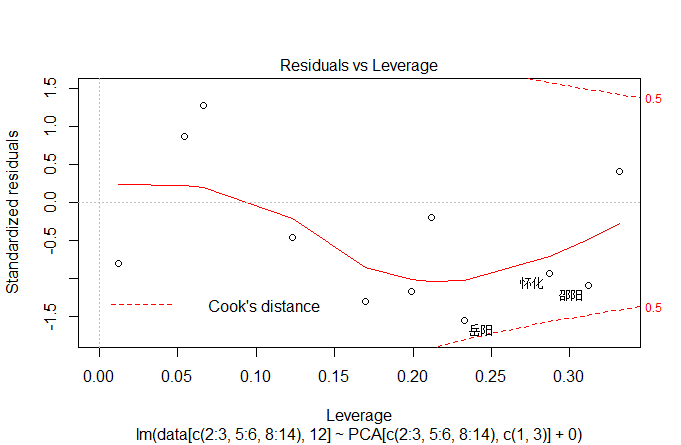


图13 残差杠杆图

模型中无异常值点。

将主成分还原为原始的10个指标，

表11 原始变量标准化系数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 乡村劳动力 | 年末实有耕地面积 | 当年减少耕地面积 | 造林面积 | 乡村户数 |
| 0.0587 | 0.1743 | -0.0650 | 0.0863 | 0.0831 |
| 乡村从业人员 | 农业机械总动力 | 有效灌溉面积 | 化肥施用量 | 农村用电量 |
| 0.0663 | 0.1641 | 0.2413 | 0.2469 | 0.0165 |

再将中心化和标准化的变量还原为原始值。

得到最终的表达式为：

方程中各变量系数的正负性符合实际情况。

### 3.3.3针对模型结果提出的建议

有效灌溉面积对农业收入的贡献率大于农业机械总动力。而灌溉面积大小则是主观和客观因素的统一。一方面取决于降水量和蒸发量的大小，另一方面也取决于人对自然条件的改善，诸如农田水利设施、中小型水库等等。现有耕地水平下，根据我国降水区域分布不均和季节分布不均的规律，东南部降水充沛，在水利措施做好的基础上，过犹不足，灌溉面积的增加对农业的贡献将不再明显。而西北部降水稀少，灌溉面积的增加主要依赖于人工水利设施的建设与升级。

在农作物有效灌溉得到保障的前提下，增加农业机械的投入使用是必不可少的。而且，在一定程度上要增加农业机械总动力对农业收入的贡献率。原因有三：一是农业机械投入主要取决于主观力量，受客观环境影响小；二是在灌溉达到一定水平，对农业的贡献不再明显以后，农业产量的增长将主要依赖于机械总动力的增加；三是在灌溉用水受到限制的农业地区，应当减少农业对水的依赖，使农业增长更多依赖机械化。可以预测，未来农业的发展必然要走机械化发展的道路。

除了模型体现出来的因素以外，其它随机因素的影响也是不容忽视的。其它农业生产条件因素，例如种子和化肥：种子革命曾经一度解决了我国的粮食问题，化肥的从无到有，也使粮食大幅度增产。这是技术进步带来的生产力发展。改良土壤、改善光热条件，这些都要靠技术进步。农业生产条件以外的因素例如调整农业生产关系，走集约化生产道路。

# 4总结与展望

## 4.1本文优点

本文的指标选择具有全面性与客观性。数据来源为湖南省统计局统计年鉴，权威而可靠。指标的选择较为全面，将与主题相关的指标全部收集了起来，囊括了农业基本情况和农业生产条件的全部信息，并对各指标的意义作出了充分的解释。

本文的模型具有严谨性。对于回归分析的结果，进行了全面的检验，并进行残差分析，并不断改进模型，最终得出了最佳模型。

本文注重了理论与实际的结合。在选择聚类个数时，将聚类结果，与湖南省的实际情况一起进行考虑。对于最终的回归分析结果，也进行了现实意义的检验，并从现实意义的角度对模型进行了补充说明并提出建议。

## 4.2不足与展望

本文在数据的选取方面，具有一定的局限性，只选取了2018年湖南省统计年鉴中的数据。然而农业发展水平和农业基本情况与农业生产条件间的关系可能会随着时间变化。因此，如果将连续几年的数据搜集起来，作为面板数据进行建模，结果可能会更加严谨，更加具有使用价值。

参考文献

|  |
| --- |
| [1]李雪艳,史少赞.农民农业收入影响因素分析[J].商业文化(学术版),2007(11):103-104. |
| [2]刘进宝,张延君.农业收入影响因素的计量经济分析[J].经济论坛,2004(05):117-118. |
| [3]卢军静,应银.我国农户农业收入影响因素深层探究[J].商业时代,2014(02):121-122. |
| [4]雷平,詹慧龙.新形势下农民收入增长影响因素研究——基于国家农业示范区面板数据[J].农林经济管理学报,2016,15(06):641-647. |
| [5]张召波.农业贷款对农民收入增长的影响[J].山西农经,2015(09):40+87. |
| [6]白志远,乐美媛.我国财政支农支出对农民收入的影响——基于1999-2013年省级面板数据的实证研究[J].财政监督,2018(09):86-93. |
| [7]陈王生.农产品价格因素对农民收入的影响相关性分析[J].南方,2018,49(03):102+106. |
| [8]张晓玲.粮食价格波动对农民收入的影响探析以贵州省为例[J].中国市场,2018(16):133-135 |
| [9]殷彩娟.农村水价对灌溉用水及农业收入的影响[J].现代农村科技,2018(12):103.郝悦,祁晓慧,乔光华. |
| [10]贾小虎,马恒运,朱玉春.考虑农户收入差异的农田水利设施供给效果的综合评价[J].农业工程学报,2018,34(17):1-8. |
| [11]农业机械投入水平影响因素实证分析——以内蒙古通辽市113户农户调查数据为例[J].内蒙古农业大学学报(社会科学版),2018,20(03):11-16. |
| [12]马轶群.农产品贸易、农业技术进步与中国区域间农民收入差距[J].国际贸易问题,2018(06):41-53. |
| [13]王雅芹,王震.农业科技进步对农民收入增长的影响分析[J].农业经济,2018(07):74-75. |
| [14]胡京蕊.农业技术对农业收入的影响分析[J].时代农机 |
| [15]田珍.地区农业竞争力评价指标系探讨[J].农业与技术,2002(06):33-35+38. |
| [16]刁广付,王健,高卓.沧州市农业生产条件聚类分析[J].合作经济与科技,2018(18):39-41 |
| [17]张建平.农业生产条件对农业经济发展影响的多元统计分析[J].农业与技术,2007(01):134-137. |
| [18]王进,王礼力,庞新江.我国农村区域竞争力评估指标体系及实证分析[J].北京理工大学学报(社会科学版),2009,11(02):43-48. |
| [19] 韩家炜.数据挖掘：概念与技术[M].第三版.机械工业出版社, 2012年 |
| [20] 任雪松,于秀林.多元统计分析[M].第二版.中国统计出版社, 2011年. |
| [21] 何晓群,刘文卿.应用回归分析[M].第四版.中国人民大学出版社, 2015年 |

附录

原始数据:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 城市 | 乡村劳动力（万人） | 年末实有耕地面积（万公顷） | | 当年减少耕地面积（万公顷） | | 造林面积  （万公顷） | 乡村户数（万户） | 乡村从业人员（万人） |
| 长沙 | 275.2 | 27.42 | 0.13 | | 0.98 | | 124.5 | 244.2 |
| 株洲 | 194.3 | 20.85 | 0.04 | | 2.93 | | 77.3 | 165.7 |
| 湘潭 | 129 | 15.11 | 0.02 | | 0.98 | | 58.3 | 127 |
| 衡阳 | 348.5 | 39.38 | 0.06 | | 7.35 | | 166.9 | 322.6 |
| 邵阳 | 428.8 | 44.78 | 0.08 | | 7.65 | | 185.3 | 391.3 |
| 岳阳 | 251.9 | 35.37 | 0.06 | | 3.19 | | 123.1 | 230.8 |
| 常德 | 303.5 | 50.59 | 0.09 | | 3.95 | | 147.7 | 276.2 |
| 张家界 | 88.4 | 11.91 | 0.04 | | 2.4 | | 43.8 | 80.7 |
| 益阳 | 227 | 29.92 | 0.06 | | 2.92 | | 108.1 | 209.3 |
| 郴州 | 276.3 | 29.91 | 0.03 | | 6.29 | | 118.2 | 247.3 |
| 永州 | 302.6 | 36.42 | 0.04 | | 6.75 | | 133.6 | 279.5 |
| 怀化 | 263.7 | 34.24 | 0.11 | | 5.2 | | 118.7 | 242.7 |
| 娄底 | 230 | 19.24 | 0.03 | | 2.48 | | 100.3 | 198 |
| 湘西 | 145 | 19.97 | 0.04 | | 2.35 | | 59.2 | 126.7 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 城市 | 农业机械总 动力  （万千瓦） | | | 有效灌溉面积  （千公顷） | 化肥施用量（万吨） | 农村用电量（万度） | 粮食合计（吨） | 农业总产值（万元） |
| 长沙 | | 602.04 | 231.62 | | 18.56 | 200471 | 2264435 | 3239403 |
| 株洲 | | 361.58 | 165.01 | | 10.94 | 84966 | 1716651 | 1428046 |
| 湘潭 | | 286.67 | 141.38 | | 10.84 | 67871 | 1390920 | 965149 |
| 衡阳 | | 573.28 | 287.66 | | 24.12 | 194440 | 3385827 | 2326289 |
| 邵阳 | | 478.16 | 286.56 | | 22.49 | 98423 | 3203751 | 2686765 |
| 岳阳 | | 639.47 | 317.08 | | 24.18 | 86829 | 3077099 | 2346220 |
| 常德 | | 633.88 | 469.15 | | 37.52 | 133340 | 3727699 | 2780822 |
| 张家界 | | 114.76 | 55.5 | | 5.72 | 19790 | 591179 | 555648 |
| 益阳 | | 538.09 | 237.78 | | 23.62 | 97290 | 2360081 | 2200818 |
| 郴州 | | 431.19 | 193.99 | | 17.12 | 72123 | 1850372 | 1927500 |
| 永州 | | 617.88 | 289.78 | | 23.41 | 78332 | 3029742 | 2589982 |
| 怀化 | | 432.81 | 199.13 | | 11.24 | 60125 | 1790818 | 1686969 |
| 娄底 | | 360.9 | 100.33 | | 8.51 | 72357 | 1519378 | 1198992 |
| 湘西 | | 187.41 | 170.9 | | 6.99 | 19219 | 829033 | 912718 |

聚类分析代码

setwd("C:/Users/user/Desktop")

data0=read.csv('Agriculture.csv')

data=scale(data0[,2:13])

rownames(data)=data0[,1]

data1=data[,1:10]##农业基本情况和农业生产条件的标准化矩阵

mdata1=as.matrix(data1)

Z1=dist(mdata1, method = "euclidean")#欧氏距离

library(ggdendro)

war=hclust(Z1,method="ward")#欧氏距离

war[["labels"]]=data0[,1]

ggdendrogram(war, rotate = TRUE, size = 4, theme\_dendro = FALSE, color ="tomato")+labs(title="离差平方和法(欧氏距离)")

主成分分析代码

datapr1=princomp(data1,cor=TRUE)

summary(datapr1,loadings=TRUE)

screeplot(datapr,type="lines")

回归分析代码

matrix1=cor(data)

pcoefficient=read.csv('主成分.csv')##主成分系数矩阵

PCA=as.matrix(data[,1:10])%\*%as.matrix(pcoefficient)##主成分的具体值

model1=lm(formula = data[,11]~PCA+0)

summary(model1)

model3=lm(formula = data[,11]~PCA[,c(1,3)]+0)

summary(model3)

plot(model3)

snum1=as.matrix(pcoefficient[,c(1,3)])%\*%as.matrix(coef(model3))#最终模型1

model2=lm(formula = data[,12]~PCA+0)

summary(model2)

plot(model2)##模型2第一次

model10=lm(formula = data[c(2:6,8:14),12]~PCA[c(2:6,8:14),]+0)

summary(model10)

plot(model10)##模型2第二次

model9=lm(formula = data[c(2:6,8:14),12]~PCA[c(2:6,8:14),c(1,3)]+0)

summary(model9)

plot(model9)#模型2第三次

model11=lm(formula = data[c(2:3,5:6,8:14),12]~PCA[c(2:3,5:6,8:14),c(1,3)]+0)

summary(model11)

plot(model11)##模型2

snum2=as.matrix(pcoefficient[,c(1,3)])%\*%as.matrix(coef(model11))

##将标准化系数还原

LJJ=1:10

for (i in 2:11) {

LJJ[i-1]=sum((data0[,i]-mean(data0[,i]))^2)

}

midx=1:10

for (i in 2:11) {

midx[i-1]=mean(data0[,i])

}

LYY1=sum((data0[,12]-mean(data0[,12]))^2)

LYY2=sum((data0[,13]-mean(data0[,13]))^2)

number1=sqrt(LYY1)/sqrt(LJJ)\*snum1

number1[11]=mean(data0[,12])/sqrt(LYY1)-sum(midx\*snum1/sqrt(LJJ))##还原常数项

number1

number2=sqrt(LYY2)/sqrt(LJJ)\*snum2

number2[11]=mean(data0[,12])/sqrt(LYY2)-sum(midx\*snum2/sqrt(LJJ))

number2