Operations Research and Fuzziology 运筹与模糊学, 2024, 14(2), 138-146
Published Online April 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.142119

Hans汉斯

上海市规模以上工业企业的从业人数探讨

——基于聚类分析和主成分分析模型

马飞雅

上海工程技术大学管理学院，上海

收稿日期：2024年1月10日；录用日期：2024年1月31日；发布日期：2024年4月10日

摘　要

本文选取了2021年上海市规模以上的工业企业(包含33个行业)，不同职业类型的从业人员期末人数数据。通过系统聚类和K均值聚类方法进行聚类分析，对这33个行业进行分类，系统聚类将其分为4类，K均值聚类分别分为4类和5类两种情形。再通过主成分分析，确定不同职业类型中的第一主成分和第二主成分，然后计算得分，得到主成分综合评价。结果表明，上海市各行业里从事计算机、通信和其他电子设备制造业，汽车制造业和通用设备制造业的人数较多。随着大数据时代的发展，未来相关领域的从事人数是否会进继续增加，值得学者进一步探讨。

关键词

系统聚类，K均值聚类，主成分分析，工业企业，从业人数

Exploration of the Number of Employees in Industrial Enterprises above Designated Size in Shanghai

—Based on Clustering Analysis and Principal Component Analysis Models

Feiya Ma

Department of Business Administration, Shanghai University of Engineering Science, Shanghai

Received: Jan. 10th, 2024; accepted: Jan. 31st, 2024; published: Apr. 10th, 2024

Abstract

This article selects data on the number of employees from industrial enterprises above designated

文章引用: 马飞雅. 上海市规模以上工业企业的从业人数探讨[J]. 运筹与模糊学, 2024, 14(2): 138-146.
DOI: 10.12677/orf.2024.142119

size (including 33 industries) and different occupational types in Shanghai in 2021. Cluster analysis was conducted using systematic clustering and K-means clustering methods to classify these 33 industries. Systematic clustering divided them into 4 categories, while K-means clustering was divided into 2 categories: 4 and 5, respectively. Then, through principal component analysis, determine the first and second principal components in different occupational types, and calculate the scores to obtain a comprehensive principal component evaluation. The results indicate that there are more people engaged in computer, communication, and other electronic equipment manufacturing, automotive manufacturing, and general equipment manufacturing in various industries in Shanghai. With the development of the big data era, it is worth further exploration by scholars whether the number of professionals in related fields will continue to increase in the future.

Keywords

System Clustering, K-Means Clustering, Principal Component Analysis, Industrial Enterprises, Number of Employees

1. 引言

随着社会经济结构的日益复杂和新兴业态、产业的层出不穷，产业链管理和混合经营逐渐成为各类经济实体的主要运营模式，给行业归类判断造成了不少困难[1]。新兴产业的不断涌现，推动了制定这些新兴产业领域行业分类统计标准的发展。黄雯怡等通过全球工业企业的发展现状，进而提出江苏省工业企业存在的问题及对策建议[2]。范震等通过三阶段 DEA-Malmquist 模型测算我国“十三五”期间地区工业企业的创新效率[3]。孙兰芸等对河北省全部和主要行业规上工业企业技术创新的现状进行分析，归纳出其呈现的特点并提出加快河北省工业企业技术创新发展的对策建议[4]。本文选取上海市规模以上的工业企业的相关数据。采取《上海市统计年鉴》里划分行业的准则，按照行业分类，将上海市的行业划分成 33 个。从事在各行业中人员，按照职业类型，依次被划分为中层及以上管理人员，专业技术人员，办事人员和有关人员，社会生产服务和生活服务人员和生产制造及有关人员。而在这所有行业中，热门行业中从业人数的密度，鲜有人探讨。本文基于聚类分析和主成分分析的方法，探讨各类型从业人员在 33 个行业里的分布情况，以及从事的目的和原因。

2. 数据来源和方法

2.1. 数据来源

本文使用的数据来源于《上海市统计年鉴》，数据是关于上海市规模以上的工业企业，按 33 个行业类别划分，分析 2021 年不同职业类型的从业人员期末人数的分布情况。其中职业类型分为中层及以上管理人员，专业技术人员，办事人员和有关人员，社会生产服务和生活服务人员和生产制造及有关人员。

2.2. 研究方法

2.2.1. 系统聚类和 K 均值聚类

系统聚类法又称分层聚类法，是聚类分析的一种方法。其做法是开始时把每个样品作为一类，然后

把距离最小的样品首先聚为小类，再将已聚合的小类按其类间距离再合并，不断继续下去，最后把一切子类都聚合到一个大类[5]。K 均值聚类法，是根据给定的参数 k，先把 n 个对象粗略地分为 k 类，然后按照某种最优原则，通常表示为一个准则函数，修改不合理的分类，直到准则函数收敛为止，就得到了一个最终的分类[6]。实际运用中，系统聚类法每一步都要计算类间距离，计算类偏大。尤其当样本量很大时。所以一般学者采用较多的方法是 K 均值聚类法[7]。

2.2.2. 主成分分析

主成分分析也称主分量分析，是由 Hotelling 于 1933 年首先提出的。由于多元统计分析处理的是多变量问题，变量较多，维数较大，增加了分析问题的复杂性[8]。但在实际问题中，变量之间可能存在一定的相关性，因此，所讨论的全部变量中可能存在信息的重叠．为去除这些信息重叠，人们自然希望用个数较少但是保留了原始变量大部分信息的几个不相关的主成分来代替原来较多的变量[9]。主成分分析的本质就是“有效降维”，既要减少变量个数，又不能损失太多信息。换句话说，就是“降噪”或者“沉余消除”，将高维数据有效地转化为低维数据来处理，揭示变量之间的内在联系，进而分析解决实际问题。

3. 上海市规模以上工业企业的从业人数探讨的聚类分析

3.1. 数据处理

本文将要对上海市规模以上的工业企业(包含 33 个行业)进行聚类分析。将上海市规模以上的工业企业，33 个行业进行分类。为了方便进行聚类分析和主成分分析，将行业用变量 X 替代，整合到文本文件，之后导入 R 软件进行分析。具体替代见表 1。

Table 1. Introduction to industrial enterprises above designated size in Shanghai (including 33 industries)
表 1. 上海市规模以上的工业企业(包含 33 个行业)简介

行业名称	变量名	行业简称	行业名称	变量名	行业简称
石油和天然气开采业	*x*1	开采	橡胶和塑料制品业	*x*18	橡胶
农副食品加工业	*x*2	农副	非金属矿物制品业	*x*19	非金属矿物
食品制造业	*x*3	食品	黑色金属冶炼和压延加工业	*x*20	黑色金属
酒、饮料和精制茶制造业	*x*4	茶水	有色金属冶炼和压延加工业	*x*21	有色金属
烟草制品业	*x*5	烟草	金属制品业	*x*22	金属
纺织业	*x*6	纺织	通用设备制造业	*x*23	通用设备
纺织服装、服饰业	*x*7	服装	专用设备制造业	*x*24	专用设备
皮革、毛皮、羽毛及其制品和制鞋业	*x*8	皮革	汽车制造业	*x*25	汽车
木材加工和木、竹、藤、棕、草制品业	*x*9	木材	铁路、船舶、航空航天和其他运输设备制造业	*x*26	铁路
家具制造业	*x*10	家具	电气机械和器材制造业	*x*27	机械
造纸和纸制品业	*x*11	造纸	计算机、通信和其他电子设备制造业	*x*28	计算机
印刷和记录媒介复制业	*x*12	印刷	仪器仪表制造业	*x*29	仪器仪表
文教、工美、体育和娱乐用品制造业	*x*13	文教	其他制造业	*x*30	其他
石油、煤炭及其他燃料加工业	*x*14	燃料	废弃资源综合利用业	*x*31	废弃资源
化学原料和化学制品制造业	*x*15	化学原料	金属制品、机械和设备修理业	*x*32	设备修理
医药制造业	*x*16	医药	电力、热力、燃气及水生产和供应业	*x*33	电力
化学纤维制造业	*x*17	化学纤维			

3.2. 系统聚类

根据处理好的数据，在 R 软件中对上海市 33 个行业类别进行系统聚类分析。通过简单连接法、完全连接法和平均连接法生成系统树图，根据生成的三种系统树图选择最优的分类效果[10]。选用平均连接法进行系统聚类，如图 1 所示。通过平均连接法的系统聚类，将上海市 33 个行业类别分为 4 类：

第 1 类：铁路、设备修理、非金属矿物、仪器仪表、茶水、化学纤维、开采、烟草、木材、废弃资源、皮革、有色金属、其他、黑色金属、家具、印刷、纺织、造纸、农副、服装、文教。

第 2 类：橡胶、金属。

第 3 类：电力、食品、医药、机械、化学原料、专用设备。

第 4 类：通用设备、汽车、计算机。

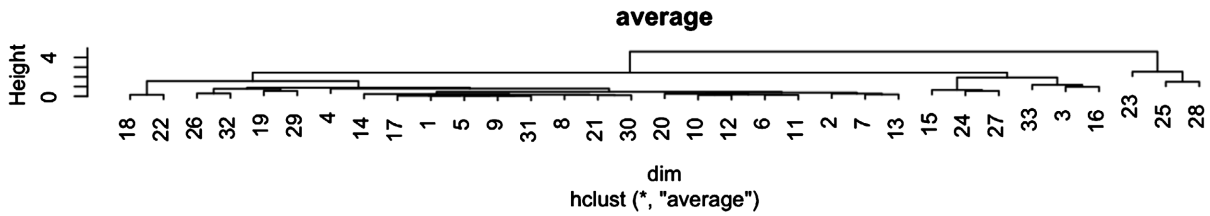

Figure 1. Complete connection method system tree diagram
图 1. 完全连接法系统树图

3.3. K 均值聚类

根据处理好的数据，在 R 软件中对上海市 33 个行业类别进行 K 均值聚类分析。碎石图能够直观地判断聚类的合适数目[11]。在 R 软件中输出了相应的碎石图，如图 2 所示。从碎石图来看，K 取 3 或者 4 时，能够较好地反应整体。

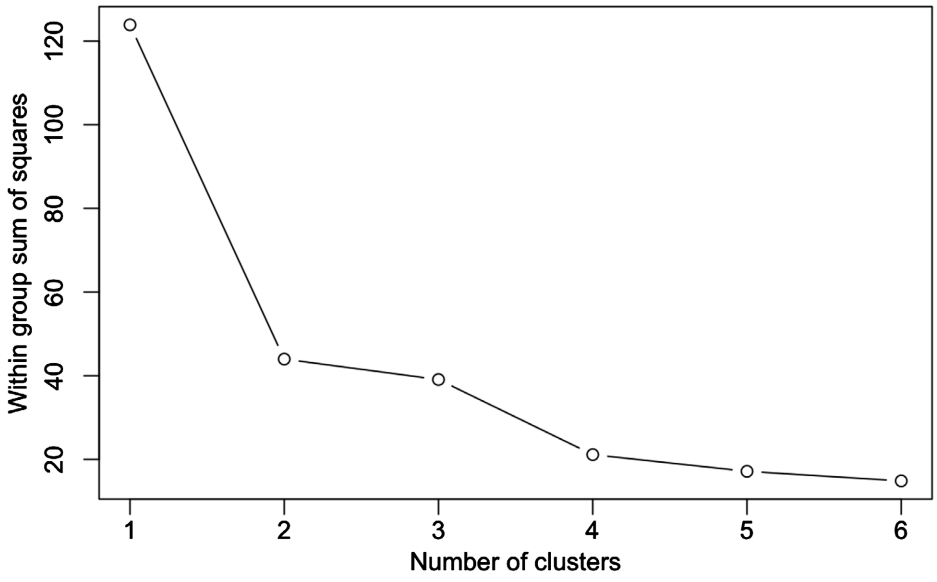

Figure 2. K-means clustering gravel map
图 2. K 均值聚类碎石图

当 K 均值聚类取 4 时，将 33 个行业分成四大类，如图 3 所示。K 均值聚类中类间平方和在总平方和

的占比为 83.0%。33 个行业分为四大类，第一类包含 7 个行业，第二类包含 3 个行业，第三类包含 16 个行业，第四类包含 7 个行业。

第一类：食品、化学原料、医药、橡胶、金属、专业设备、机械。

第二类：通用设备、汽车、计算机。

第三类：开采、农副、茶水、烟草、纺织、服装、皮革、木材、家具、造纸、印刷、文教、燃料、化学纤维、黑色金属、有色金属。

第四类：非金属矿物、铁路、仪器仪表、其他、废弃资源、设备维修、电力。

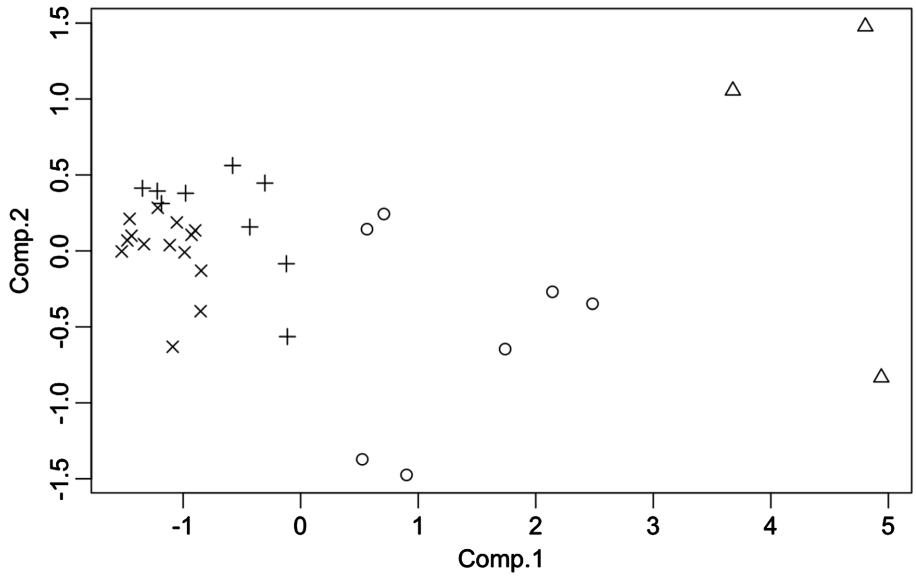

Figure 3. K-means clustering diagram divided into four categories
图 3. 分为四大类的 K 均值聚类图

当 K 均值聚类取 5 时，将 33 个行业分成五大类，如图 4 所示。K 均值聚类中类间平方和在总平方和的占比为 86.4%。33 个行业分为五大类，第一类包含 14 个行业，第二类包含 5 个行业，第三类包含 3 个行业，第四类包含 8 个行业，第五类包含 3 个行业。

第一类：开采、农副、茶水、烟草、纺织、服装、皮革、木材、家具、造纸、印刷、文教、燃料、化学纤维。

第二类：食品、医药、橡胶、金属、电力。

第三类：通用设备、汽车、计算机。

第四类：非金属矿物、黑色金属、有色金属、铁路、仪器仪表、其他、废弃资源、设备修理。

第五类：化学原料、专用设备、机械。

两种 K 均值聚类分析输出结果有所差异。在两种 K 均值聚类情形下，化学原料，专业设备和机械聚都为一类。通用设备，汽车和计算机都聚为一类。其他行业的聚类发生一定变化。其中，K 取 5 的聚类是将 K 取 4 聚类的第五类划分成一类。

4. 上海市规模以上工业企业的从业人数探讨的主成分分析

4.1. 相关系数矩阵

在进行主成分分析时，首先求出变量间的相关系数矩阵。通过相关系数矩阵，观察变量间的相关性

[12]。若变量间存在较强的相关性，则适合做主成分分析；反之，则不适合做主成分分析[13]。此次分析中包含 6 个变量，其中 $y1$ 为从业人员期末人数，$y2$ 为中层及以上管理人员，$y3$ 为专业技术人员，$y4$ 为办事人员和有关人员，$y5$ 为社会生产服务和生活服务人员，$y6$ 为生产制造及有关人员。在相关系数矩阵中，如表 2 所示，$y2$、$y3$、$y4$ 之间相关性较强。即中层及以上管理人员，专业技术人员，办事人员和有关人员和社会生产服务和生活服务人员相关性较强。所以该数据适合做主成分分析。下一步做主成分分析，求样本相关矩阵的特征值和主成分载荷。

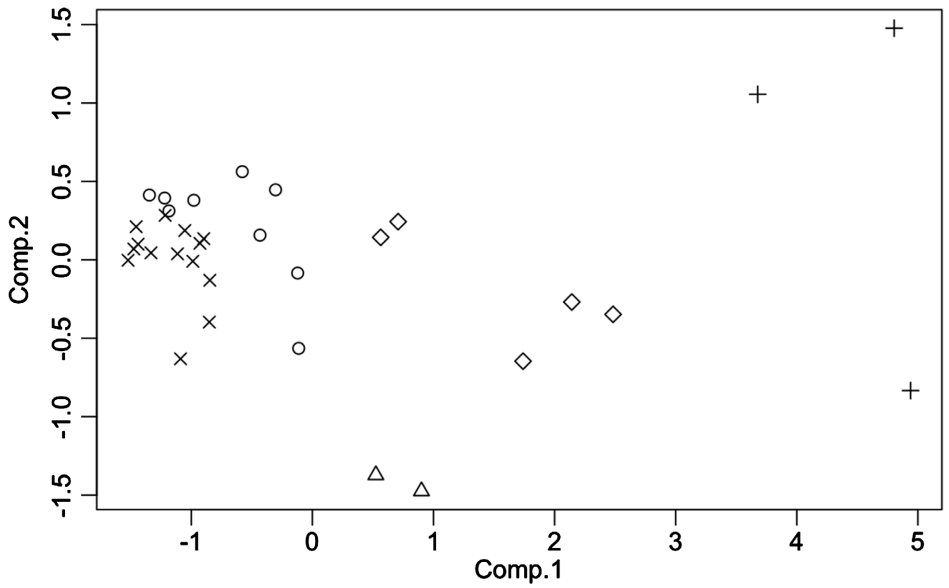

Figure 4. K-means clustering diagram divided into five categories
图 4. 分为五大类的 K 均值聚类图

Table 2. The correlation coefficient matrix of occupational type variables
表 2. 职业类型变量的相关系数矩阵

	$y1$	$y2$	$y3$	$y4$	$y5$	$y6$
$y1$	1	0.922	0.961	0.909	0.656	0.983
$y2$	0.922	1	0.909	0.975	0.765	0.848
$y3$	0.961	0.909	1	0.911	0.643	0.911
$y4$	0.909	0.975	0.911	1	0.807	0.822
$y5$	0.656	0.765	0.643	0.807	1	0.554
$y6$	0.983	0.848	0.911	0.822	0.554	1

4.2. 主成分分析

主成分分析能够有效降维，在 R 软件中先得到变量的相关系数矩阵，基于相关系数矩阵对中层及以上管理人员，专业技术人员，办事人员和有关人员，社会生产服务和生活服务人员和生产制造及有关人员五个变量进行降维。输出结果如表 3 所示。

在职业类型变量的相主成分分析中，前二个主成分的累积的贡献率为 95.73%，于是取前二个主成分，可以得出：

Table 3. Principal component analysis results of occupational type variables
表 3. 职业类型变量的主成分分析结果

Importance of components:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.0682218	0.7134335	0.35549412	0.25941948	0.140700308
Proportion of Variance	0.8555083	0.1017975	0.02527521	0.01345969	0.003959315
Cumulative Proportion	0.8555083	0.9573058	0.98258099	0.99604068	1
Loadings:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
$y2$	0.472		0.453	0.45	0.608
$y3$	0.46	−0.316		−0.804	0.195
$y4$	0.473	0.105	0.435		−0.754
$y5$	0.39	0.795	−0.443	−0.103	
$y6$	0.435	−0.506	−0.636	0.366	−0.125

$$Z1^* = 0.472y2^* + 0.46y3^* + 0.473y4^* + 0.39y5^* + 0.435y6^*$$

$$Z2^* = -0.316y3^* + 0.105y4^* + 0.795y5^* - 0.506y6^*$$

第一主成分对应的系数符号全为正数且 $y2^*$、$y3^*$和 $y4^*$对应的载荷值较大，可视为反应中层及以上管理人员，专业技术人员，和办事人员和有关人员的主成分，第二主成分对应的系数中 $y5$ 上的取值为负且载荷值特别大，可视为反应社会生产服务和生活服务人员。

再通过分析在 R 软件获得的基于协方差的主成分分析的碎石图，确定主成分。从碎石图，我们可以看出，前二个主成分的方差占了总方差变化的大部分，因此本文主成分的个数为 2 是适当的。如图 5 所示。

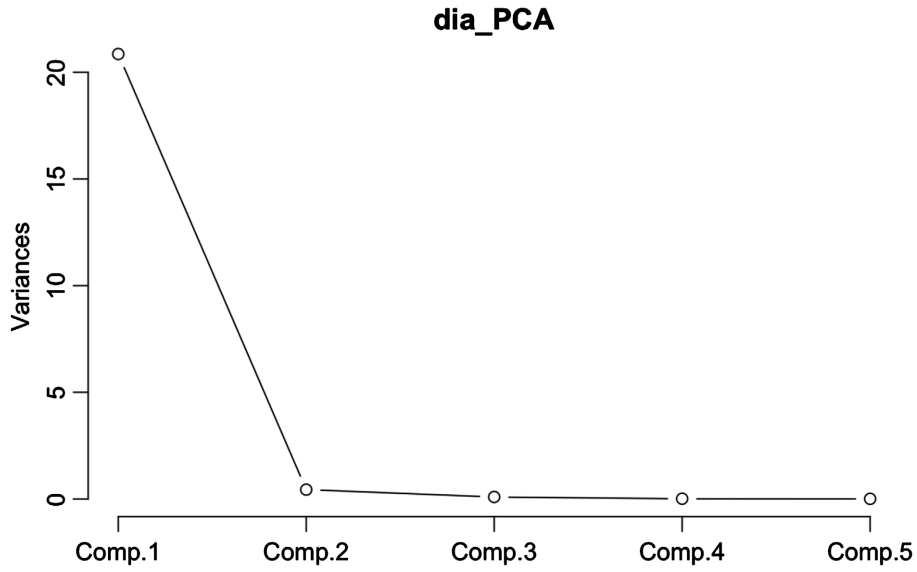

Figure 5. K-means clustering diagram divided into five categories
图 5. 基于协方差的主成分分析碎石图

4.3. 主成分综合评价

通过聚类分析和主成分分析，以所得分类结果为基准，对上海市 33 个行业的从业人员期末人数进行主成分综合评价。为便于分析，将上海市 33 个行业划分为 4 个梯队。各主成分的因子得分均值和主成分综合得分均值，结果如表 4 所示。

Table 4. Factor scores and comprehensive scores of principal components
表 4. 主成分的因子得分和综合得分

序号	行业	comp.1	comp.2	average
1	通用设备、汽车、计算机、机械	5.205	0.081	3.801
2	专用设备、化学原料、金属、医药、橡胶、电力、仪器仪表、铁路、设备修理	0.814	0.430	0.709
3	食品、非金属矿物、其他、废弃资源、黑色金属、有色金属、印刷	−0.984	0.415	−0.601
4	文教、家具、服装、造纸、燃料、化学纤维、农副、纺织、茶水、皮革、木材、烟草、开采	−1.635	−0.546	−1.337

结果如表 4 可知，第一主成分中得分最高是 5.20，包含的行业有通用设备、汽车、计算机和机械。第二主成分中得分最高的是 0.43，包含的行业有专用设备、化学原料、金属、医药、橡胶、电力、仪器仪表、铁路和设备修理。第一梯队的第一主成分的因子得分均值为 5.205，第二主成分为 0.081，主成分综合得分为 3.801。遥遥领先于其他梯队。可知在规模以上工业企业中，通用设备、汽车、计算机和机械热度不减，从业人员密度较高。互联网浪潮下，电子设备几乎是生活中的必备物品，如手机、电脑。近几年来，大数据技术，人工智能不断融入到现代生活中[14]。电子设备行业同时发展迅速[15]。行业的发展离不开人才的引入[16]。这使得从事电子设备的人员相较于其他行业是很高的[17]。汽车行业也是如此。近年来，新能源汽车进入人们视野，这种环境友好型，价格亲民的汽车备受人们亲睐。随着“双碳”政策的实行，未来新能源汽车行业将会吸引更多人才加入[18]。

5. 结论

本文基于《上海统计年鉴》的数据，探讨了上海市规模以上的工业企业，按 33 个行业类别划分，不同类型职业从业人数的分布情况。通过系统聚类模型选用平均连接法，将 33 个行业分成 4 类。通过 K 均值聚类模型，将 33 个行业划分成 4 和 5 类进行讨论。结果表明，通用设备、汽车、计算机行业的从业人员密度较大。行业的选择和时代大背景密切相关。之后进行主成分分析，在第一梯队的行业是计算机、电气机械和器材制造业，汽车制造业和通用设备制造业。近年来，互联网发展迅猛，电子行业同时发展起来，如“互联网大厂”等公司是很多从业人员的选择，计算机行业仍是热门就业选择。随着“双碳”政策的实行，汽车行业的新能源汽车备受人们亲睐[19]。汽车行业在未来可能会吸引更多人员加入。使得规模以上工业企业的发展效率大大增高。

参考文献

[1] 刘宇. 部门分类标准与中国供给使用表的编制[J]. 统计与决策, 2023, 39(3): 33-38.

[2] 黄雯怡, 张睿. 全球工业企业研发现状、经验及江苏启示[J]. 现代管理科学, 2023(4): 98-105.

[3] 范震, 姜顺婕. 创新价值链视角下我国“十三五”期间地区工业企业科技创新效率评价研究——基于三阶段 DEA-Malmquist 模型[J]. 统计与管理, 2023, 38(7): 47-57.

[4] 孙兰芸, 胡云红, 曹文海. 河北省规上工业企业技术创新现状分析及对策研究[J]. 统计与咨询, 2023(3): 34-37.

[5] 韩平平, 郭佳林, 董玮, 等. 基于系统聚类法的含新能源电力系统分区策略[J/OL]. 电力系统及其自动化学报: 1-8. https://doi.org/10.19635/j.cnki.csu-epsa.001292, 2024-04-03.

[6] 周湘贞, 李帅, 隋栋. 数据驱动下基于量子人工蜂群的 K 均值聚类算法优化[J]. 南京理工大学学报, 2023, 47(2): 199-206.

[7] 林伟杰, 王勇, 周林. 基于加权二分图的 K 均值最佳聚类数确定算法[J]. 计算机工程与设计, 2023, 44(4): 1104-1111.

[8] 徐婷婷, 胡摇, 郝群, 沈添天. 基于主成分分析重建图像的互信息视觉伺服[J]. 光学技术, 2023, 49(6): 736-742.

[9] 李月, 朱俊烨, 刘子涵, 等. 基于主成分分析与聚类分析评价茯砖茶滋味品质[J]. 食品安全质量检测学报, 2023, 14(21): 283-291.

[10] 徐業釗, 张晗, 冯贯昂, 韩立言, 庞华. 基于系统聚类法的基金资产配置策略研究[J]. 中国商论, 2021(17): 102-104.

[11] 陈玉明, 蔡国强, 卢俊文, 等. 一种邻域粒 K 均值聚类方法[J]. 控制与决策, 2023, 38(3): 857-864.

[12] 孟银凤, 李庆方. 基于多元函数主成分表示的识别学习[J]. 山东大学学报(工学版), 2022, 52(3): 1-8+17.

[13] 房汉国. 基于主成分分析法的宏观经济景气指数研究[J]. 当代经济, 2022, 39(1): 26-31.

[14] 赖晨华. 大数据时代事业单位人力资源管理的变革方法[J]. 人才资源开发, 2022(8): 33-34.

[15] 姜昊, 董直庆. 人工智能技术应用会存在选择性偏向吗?——行业属性与就业偏向[J]. 南方经济, 2023(12): 37-61.

[16] 鲁轶. 关于经济活动的行业归类判断方法的研究——行业分类标准的应用与实践[J]. 统计科学与实践, 2014(6): 26-29.

[17] 黄吉婷, 郭可歆, 齐佳音. 企业数字化转型对就业规模及结构影响的实证研究[J]. 智能科学与技术学报, 2023, 5(3): 352-365.

[18] 张涛, 唐僖, 吴君民, 等. 基于微分博弈的互联网企业与新能源汽车企业技术协同创新策略研究[J]. 运筹与管理, 2023, 32(10): 43-49.

[19] 刘宗巍, 宋昊坤, 赵福全. 中国新能源汽车产业人才需求预测研究[J]. 中国科技论坛, 2023(12): 137-148.