Statistics and Application 统计学与应用, 2020, 9(2), 277-287
Published Online April 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.92030

Hans汉斯

Factor Analysis and Cluster Analysis of Province Population under Multiple Development Indexes

Yuchen Jia, Jinhui Wen

School of Economics and Management, Beihang University, Beijing
Email: jyc9911@163.com, jhwen@buaa.edu.cn

Received: Mar. 29th, 2020; accepted: Apr. 10th, 2020; published: Apr. 17th, 2020

Abstract

With the increasing concern of environmental issues, people begin to think about the "population size and structure change in each region". This paper uses factor analysis and cluster analysis to classify the population of each province and municipality directly under the Central Government. In this paper, the factor analysis method is used to build a comprehensive scoring model, including data standardization, factor extraction, naming and building a comprehensive model. This paper will select 31 provinces (autonomous regions and municipalities) under the Central Government data for factor analysis, through factor extraction combined with the cumulative contribution rate to determine five factors. Then the factor score formula (the relationship between the extracted factor F and each index X) and the comprehensive model score formula (the relationship between the comprehensive score Y and each index X) are obtained by calculation. Then the data are substituted into the constructed comprehensive scoring model to get the comprehensive scoring of each province and municipality directly under the Central Government and rank the comprehensive scoring. Then, the K-means clustering method is used to cluster the comprehensive scores of each city in one dimension, and the final result is that 31 provinces (autonomous regions and municipalities) under the Central Government are divided into six categories.

Keywords

Factor Analysis, Cluster Analysis

多发展指标下的各省人口因子与聚类分析

贾宇尘，温锦辉

北京航空航天大学经济管理学院，北京
Email: jyc9911@163.com, jhwen@buaa.edu.cn

收稿日期：2020年3月29日；录用日期：2020年4月10日；发布日期：2020年4月17日

文章引用: 贾宇尘, 温锦辉. 多发展指标下的各省人口因子与聚类分析[J]. 统计学与应用, 2020, 9(2): 277-287.
DOI: 10.12677/sa.2020.92030

摘　要

随着环境问题日益受到关注，人们开始对“每个地区的人口规模与结构变化”进行思索，本文利用因子分析和聚类分析的方法对各个省及直辖市的人口进行一个综合排名分类。本文采用了因子分析的方法来构建综合得分模型，具体包括数据标准化处理、因子提取、命名以及构建综合模型。本文选取了31个省(自治区、直辖市)数据进行因子分析，通过因子提取结合得到的累计贡献率确定出5个因子。随后通过计算得出了因子得分公式(提取出来的因子F和各指标X之间的关系)和综合模型得分公式(综合得分Y和各指标X之间的关系)。然后将数据代入构建好的综合得分模型，得出每个省及直辖市的综合得分，进行综合得分排名。再利用K-means聚类的方法对各个城市的综合得分进行一维聚类，得到最终结果将31个省(自治区、直辖市)分为了六类。

关键词

因子分析，聚类分析

1. 研究意义

中国自 20 世纪 70 年代以来，先是实施计划生育政策，使得当今的人口结构相比计划生育之初，有了一个翻天覆地的变化，但是政策的实施也意味着多年后的今天，中国再次面临了与之前完全不同的问题——人口老龄化，导致了去年又再次实施单独二胎政策。回看这几十年的变动，中国的人口结构起起伏伏，始终无法保持在一个较为稳定的状态，同时，多年来各省因种种原因，也是体现出了各不相同的人口以及人口带来的其他各种问题。因此，对于各个省份及直辖市面临的不同问题，已有人口规模和未来的变动趋势的研究是至关重要的，因为这些关系到省、市政府对于自己这个地区的经济文化、教育水平、综合素质的未来发展认知和规划。而我国城镇未来经济增长的整体动力和水平在很大程度上又取决于城镇中各个经济要素的健康与完备[1]。因此，为了确保各个省份处于一个合理健康的人口规模和结构之下稳步发展，需要对其现状进行研究。

2. 人口发展水平因子分析

2.1. 数据标准化处理

本文选取了 31 个省(自治区、直辖市)包括消费指数、常住人口、道路面积等在内的 19 个指标[2]，由于各个指标不是同一个量纲和量纲单位，不能直接对不同量纲的指标进行比较。为了对整体指标进行综合评价，需要先对评价指标进行数学变换来消除指标量纲的影响，即进行无量纲化处理。标准化法是目前使用较为广泛的一种无量纲化方法，公式如下所示：

$$Z = \frac{X - \bar{X}}{\delta}$$

其中，Z 为原始变量 X 标准化后的数值，$\bar{X}$ 为 X 的期望值，δ 为 X 的标准差。

2.2. 提取因子

利用特征值作为保留主成分标准，可得主成分方差贡献率表，见表 1。

Table 1. Explained total variance
表 1. 解释的总方差

成份	初始特征值			提取平方和载入		
	合计	方差的%	累积%	合计	方差的%	累积%
1	6.978	36.724	36.724	6.978	36.724	36.724
2	4.365	22.973	59.698	4.365	22.973	59.698
3	1.675	8.814	68.512	1.675	8.814	68.512
4	1.451	7.639	76.151	1.451	7.639	76.151
5	1.044	5.497	81.649	1.044	5.497	81.649
6	0.850	4.472	86.120			
7	0.772	4.064	90.184			
8	0.450	2.369	92.553			
9	0.421	2.214	94.767			
10	0.310	1.632	96.399			
11	0.226	1.191	97.590			
12	0.175	0.922	98.512			
13	0.099	0.522	99.034			
14	0.083	0.434	99.468			
15	0.051	0.271	99.739			
16	0.025	0.132	99.872			
17	0.013	0.068	99.939			
18	0.011	0.055	99.994			
19	0.001	0.006	100.000			

提取方法：主成份分析。

根据上表可知，选取特征根大于 1 的因子有 5 个，累计方差贡献率达到 81.649%。相应的碎石图见图 1。

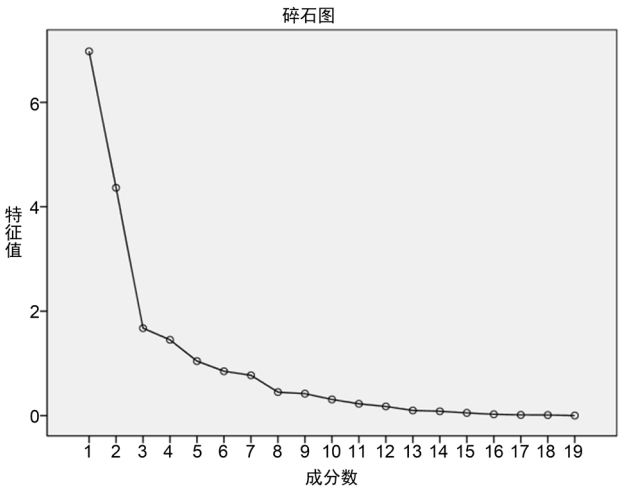

Figure 1. Scree plot
图 1. 碎石图

由上图可以直观看出，因子 4 与因子 5 的特征值比较相近，因子数大于 5 之后特征值呈现缓慢降低的趋势，因此选择 5 个因子是科学可行的。

2.3. 因子命名

5 个因子成分矩阵见表 2。

Table 2. Composition matrix
表 2. 成分矩阵

	成份				
	1	2	3	4	5
Zscore：居民消费水平指数(上年 = 100) (X1)	0.106	0.403	−0.342	0.522	0.064
Zscore：年末常住人口(万人)	0.970	0.084	−0.059	−0.084	−0.126
Zscore：全社会固定资产投资(亿元)	0.911	0.045	−0.147	−0.110	0.132
Zscore：地方财政一般预算支出(亿元)	0.926	−0.260	0.112	0.089	−0.028
Zscore：地方财政一般预算收入(亿元)	0.772	−0.437	0.217	0.210	0.156
Zscore：固定资产投资价格指数(上年 = 100)	0.044	−0.010	0.655	0.026	−0.568
Zscore：人均日生活用水量(升)	0.006	0.232	0.803	0.259	0.222
Zscore：每万人拥有公共交通车辆(标台)	0.239	−0.635	−0.057	0.352	−0.017
Zscore：人均城市道路面积(平方米)	0.237	0.368	−0.083	−0.388	0.575
Zscore：每十万人口初中阶段平均在校生数(人)	0.011	0.931	−0.055	0.229	−0.064
Zscore：每十万人口小学平均在校生数(人)	0.109	0.880	0.106	0.320	−0.009
Zscore：每十万人口高中阶段平均在校生数(人)	0.041	0.841	−0.168	0.252	−0.069
Zscore：每十万人口高等学校平均在校生数(人)	0.014	−0.840	−0.079	0.167	0.048
Zscore：卫生机构床位数(万张)	0.939	0.077	−0.119	−0.139	−0.170
Zscore：医院床位数(万张)	0.953	0.014	−0.122	−0.126	−0.142
Zscore：社区服务机构数(个)	0.817	−0.041	0.155	0.283	0.200
Zscore：文化娱乐类居民消费价格指数(上年 = 100)	0.112	0.380	0.503	−0.495	0.193
Zscore：公共图书馆业机构数(个)	0.703	0.370	−0.142	−0.349	−0.334
Zscore：交通事故发生数总计(起)	0.805	−0.009	0.119	0.195	0.175

根据成分矩阵我们可以看到，第 1 主因子与所有变量均为正相关，且年末常住人口(万人)、全社会固定资产投资(亿元)、地方财政一般预算支出(亿元)、卫生机构床位数(万张)、医院床位数(万张)、社区服务机构数(个)及交通事故发生数总计(起)与第一主成分有较强的正相关性，因此，第 1 主因子主要代表的是“社会总体基础设施水平”。

第 2 主因子与每十万人口初中阶段平均在校生数(人)、每十万人口小学平均在校生数(人)及每十万人口高中阶段平均在校生数(人)呈较强的正相关，与每十万人口高等学校平均在校生数(人)呈较强负相关性，说明第 2 主因子反映的是“初等教育普及指数”。

第 3 主因子与人均日生活用水量(升)、固定资产投资价格指数(上年 = 100)及文化娱乐类居民消费价格指数(上年 = 100)这三个变量有较强的正相关性，因此第 3 主因子主要代表的是“人口总体生活水平指数”。

第 4 主因子主要与居民消费水平指数(上年 = 100)、每万人拥有公共交通车辆(标台)呈正相关，说明第 4 主因子可以代表“居民总体经济水平”。

第 5 主因子与人均城市道路面积(平方米)呈较强正相关性，因此第 5 主因子主要反映的是“人均交通发展水平指数”。

3. 各省份人口发展水平分析

3.1. 人口发展水平指标

因子分析结果中，因子得分系数矩阵如下(见表 3)：

Table 3. Factor score coefficient matrix
表 3. 因子得分系数矩阵

	成份				
	1	2	3	4	5
Zscore：居民消费水平指数(上年 = 100)	0.015	0.092	−0.204	0.359	0.061
Zscore：年末常住人口(万人)	0.139	0.019	−0.035	−0.058	−0.121
Zscore：全社会固定资产投资(亿元)	0.131	0.010	−0.088	−0.076	0.127
Zscore：地方财政一般预算支出(亿元)	0.133	−0.060	0.067	0.061	−0.026
Zscore：地方财政一般预算收入(亿元)	0.111	−0.100	0.129	0.144	0.149
Zscore：固定资产投资价格指数(上年 = 100)	0.006	−0.002	0.391	0.018	−0.544
Zscore：人均日生活用水量(升)	0.001	0.053	0.479	0.178	0.213
Zscore：每万人拥有公共交通车辆(标台)	0.034	−0.145	−0.034	0.243	−0.016
Zscore：人均城市道路面积(平方米)	0.034	0.084	−0.049	−0.267	0.550
Zscore：每十万人口初中阶段平均在校生数(人)	0.002	0.213	−0.033	0.158	−0.061
Zscore：每十万人口小学平均在校生数(人)	0.016	0.202	0.063	0.220	−0.008
Zscore：每十万人口高中阶段平均在校生数(人)	0.006	0.193	−0.101	0.174	−0.066
Zscore：每十万人口高等学校平均在校生数(人)	0.002	−0.192	−0.047	0.115	0.046
Zscore：卫生机构床位数(万张)	0.135	0.018	−0.071	−0.095	−0.163
Zscore：医院床位数(万张)	0.137	0.003	−0.073	−0.087	−0.136
Zscore：社区服务机构数(个)	0.117	−0.009	0.093	0.195	0.191
Zscore：文化娱乐类居民消费价格指数(上年 = 100)	0.016	0.087	0.300	−0.341	0.185
Zscore：公共图书馆业机构数(个)	0.101	0.085	−0.085	−0.240	−0.320
Zscore：交通事故发生数总计(起)	0.115	−0.002	0.071	0.135	0.168

3.2. 各省份各因子得分

各省各因子得分矩阵如下(见表 4)：

Table 4. Factor score matrix
表 4. 因子得分矩阵

地区	因子 1	因子 2	因子 3	因子 4	因子 5
北京市	−0.53378	−3.07595	−0.06804	1.95064	−0.57757
天津市	−0.83191	−1.77601	−0.32528	−0.12363	0.85638
河北省	0.88933	0.45285	−0.99481	0.04845	−0.10508
山西省	−0.43616	0.06739	0.14063	−1.46393	−1.2815

Continued

内蒙古自治区	−0.51039	−0.09906	−0.55142	−1.81878	0.48267
辽宁省	−0.15251	−0.64568	−1.05877	−0.15436	−0.59993
吉林省	−0.73837	−0.95075	−0.66925	−1.56342	0.83825
黑龙江省	−0.45249	−0.75134	−0.43006	−1.46862	−0.62018
上海市	−0.62333	−1.9073	1.36007	−0.07874	−0.25773
江苏省	1.83407	−0.61031	−0.31225	0.23245	2.6933
浙江省	0.84	−0.48113	0.64029	−0.15566	0.85918
安徽省	0.42904	0.45708	−0.24919	−0.71298	1.07125
福建省	−0.15079	0.11772	0.35344	0.82343	−0.25461
江西省	−0.2151	0.75431	0.03359	0.17454	−0.66571
山东省	1.91521	0.08213	−0.90563	−0.93577	1.21179
河南省	1.10791	0.99235	−1.51566	−0.07485	−1.46992
湖北省	0.70037	−0.42697	0.52473	0.11567	−0.22374
湖南省	0.69321	0.22826	0.96667	−0.22573	−1.46642
广东省	2.55466	−0.24007	2.01139	1.68194	−0.37662
广西壮族自治区	−0.17059	1.10674	0.89924	−0.15695	0.2291
海南省	−1.28577	0.78394	1.95144	−0.17403	0.68481
重庆市	−0.562	−0.06112	−1.40432	1.33949	0.30739
四川省	1.24116	0.35924	0.82424	−1.32708	−1.22562
贵州省	−0.04545	1.74647	−1.1542	2.12722	0.72634
云南省	−0.0527	0.48501	−0.26893	−0.30875	−1.74509
西藏自治区	−1.40639	1.13383	2.15585	−0.10161	1.089
陕西省	−0.15577	−0.53094	−0.34974	0.31167	−0.81461
甘肃省	−0.73648	0.49649	−1.39573	−0.12236	0.48376
青海省	−1.37771	0.49362	−0.64492	1.58263	−0.85905
宁夏回族自治区	−1.30584	0.77985	0.18794	0.28929	1.1727
新疆维吾尔自治区	−0.46141	1.01937	0.24867	0.28983	−0.16254

经过对样本在每个因子上的得分进行从大到小排序，可得到各个省份在每个因子上的名次，见表 5：

Table 5. Ranking of various factors
表 5. 各因子得分排序

地区	因子 1	因子 2	因子 3	因子 4	因子 5
北京市	22	31	15	2	21
天津市	27	29	19	18	7
河北省	6	13	26	13	15
山西省	18	18	13	28	28
内蒙古自治区	21	20	22	31	12
辽宁省	14	26	27	19	22
吉林省	26	28	24	30	8

Continued

黑龙江省	19	27	21	29	23
上海市	24	30	4	15	19
江苏省	3	25	18	10	1
浙江省	7	23	8	20	6
安徽省	10	12	16	25	5
福建省	13	16	10	6	18
江西省	17	8	14	11	24
山东省	2	17	25	26	2
河南省	5	5	31	14	30
湖北省	8	22	9	12	17
湖南省	9	15	5	23	29
广东省	1	21	2	3	20
广西壮族自治区	16	3	6	21	14
海南省	28	6	3	22	10
重庆市	23	19	30	5	13
四川省	4	14	7	27	27
贵州省	11	1	28	1	9
云南省	12	11	17	24	31
西藏自治区	31	2	1	16	4
陕西省	15	24	20	7	25
甘肃省	25	9	29	17	11
青海省	30	10	23	4	26
宁夏回族自治区	29	7	12	9	3
新疆维吾尔自治区	20	4	11	8	16

公共因子 F_1 (社会总体基础设施水平)排名前五的城市分别是：广东省、山东省、江苏省、四川省和河南省。这几个省份除了省份本身人口规模十分庞大以为以外，政府预算支出也较高，累计基础设施建设也较高，因此人均占有资源指数比较高。北京和上海的名次分别为 22 和 24，说明这两个城市和其他省份相比人口规模不够。

公共因子 F_2 (初等教育普及指数)排名前五的城市分别是：贵州省、西藏自治区、广西壮族自治区、新疆维吾尔族自治区、河南省。这或许和我们的日常观念相悖，但是由于这前四个省人口基数很小，加上政府多年来的教育投资，也是能够解释的，而河南省是教育大省这一观念还是符合常理的。北京在此排名上为第 31 名，多少有些让人意外。

公共因子 F_3 (人口总体生活水平指数)排名前五的城市分别为：西藏自治区、广东省、海南省、上海市、湖南省。说明这几个城市的在一定程度上总体生活水平较高，在娱乐方面的消费相对较多。

公共因子 F_4 (居民总体经济水平)排名靠前城市的为：贵州、北京、广东、青海省、重庆省。贵州和青海的入围主要是由于居民消费指数增速较快，剩下的省份由于本身经济水平较高，且均为各地区的政治文化中心，高等学府聚集，因此在此项指标上得分较高。

公共因子 F_5 (人均交通发展水平指数)排名靠前的城市分别为：江苏省、山东省、宁夏省、西藏自治

区、安徽省。可以发现有三个省份交通发展水平本身较高，而宁夏和西藏省人口稀少，但国家大力投入交通建设，因此排名也比较靠前。

根据表可以得出：

$$\lambda_1 = 6.978\theta_1 = 36.724\%$$

$$\lambda_2 = 4.365\theta_2 = 22.973\%$$

$$\lambda_3 = 1.675\theta_3 = 8.814\%$$

$$\lambda_4 = 1.451\theta_4 = 7.639\%$$

$$\lambda_5 = 1.044\theta_5 = 5.497\%$$

可以得到 5 个主因子的表达式如下：

$$\begin{aligned}F_1 = {} & 0.015X_1 + 0.139X_2 + 0.131X_3 + 0.133X_4 + 0.111X_5 \\ & + 0.006X_6 + 0.001X_7 + 0.034X_8 + 0.034X_9 + 0.002X_{10} \\ & + 0.016X_{11} + 0.006X_{12} + 0.002X_{13} + 0.135X_{14} + 0.137X_{15} \\ & + 0.117X_{16} + 0.016X_{17} + 0.101X_{18} + 0.115X_{19}\end{aligned}$$

$$\begin{aligned}F_2 = {} & 0.092X_1 + 0.019X_2 + 0.010X_3 - 0.060X_4 - 0.100X_5 \\ & - 0.002X_6 + 0.053X_7 - 0.145X_8 + 0.084X_9 + 0.213X_{10} \\ & + 0.202X_{11} + 0.193X_{12} - 0.192X_{13} + 0.018X_{14} + 0.003X_{15} \\ & - 0.009X_{16} + 0.087X_{17} + 0.085X_{18} - 0.002X_{19}\end{aligned}$$

$$\begin{aligned}F_3 = {} & -0.204X_1 - 0.035X_2 - 0.088X_3 + 0.067X_4 + 0.129X_5 \\ & + 0.391X_6 + 0.479X_7 - 0.034X_8 - 0.049X_9 + -0.033X_{10} \\ & + 0.063X_{11} - 0.101X_{12} - 0.047X_{13} - 0.071X_{14} - 0.073X_{15} \\ & + 0.093X_{16} + 0.300X_{17} - 0.085X_{18} + 0.071X_{19}\end{aligned}$$

$$\begin{aligned}F_4 = {} & 0.359X_1 - 0.058X_2 - 0.076X_3 + 0.061X_4 + 0.144X_5 \\ & + 0.018X_6 + 0.178X_7 + 0.243X_8 - 0.267X_9 + 0.158X_{10} \\ & + 0.220X_{11} + 0.174X_{12} + 0.115X_{13} - 0.095X_{14} - 0.087X_{15} \\ & + 0.195X_{16} - 0.341X_{17} - 0.240X_{18} + 0.135X_{19}\end{aligned}$$

$$\begin{aligned}F_5 = {} & 0.061X_1 - 0.121X_2 + 0.127X_3 - 0.026X_4 + 0.149X_5 \\ & - 0.544X_6 + 0.213X_7 - 0.016X_8 + 0.550X_9 - 0.061X_{10} \\ & - 0.008X_{11} - 0.066X_{12} + 0.046X_{13} - 0.163X_{14} - 0.136X_{15} \\ & + 0.191X_{16} + 0.185X_{17} - 0.320X_{18} + 0.168X_{19}\end{aligned}$$

则各省份人口发展水平指标 Y 的表达式如下：

$$Y = 0.36724F_1 + 0.22973F_2 + 0.08814F_3 + 0.07639F_4 + 0.05497F_5$$

根据上文得出的公式对数据进行 Y 值计算，可得结果见表 6。

Table 6. Province Y value
表 6. 各省份 Y 值

地区	Y 值	排名	地区	Y 值	排名
广东省	1.168087	1	云南省	−0.05115	17
江苏省	0.671623	2	海南省	−0.09574	18

Continued

山东省	0.637516	3	宁夏回族自治区	−0.19727	19
贵州省	0.48522	4	重庆市	−0.22499	20
四川省	0.442232	5	陕西省	−0.23097	21
河南省	0.414732	6	甘肃省	−0.26218	22
河北省	0.340873	7	山西省	−0.31457	23
湖南省	0.294362	8	辽宁省	−0.34243	24
浙江省	0.289725	9	内蒙古自治区	−0.3712	25
广西壮族自治区	0.271467	10	青海省	−0.37572	26
安徽省	0.245024	11	黑龙江省	−0.52296	27
湖北省	0.201903	12	上海市	−0.56738	28
新疆维吾尔自治区	0.099855	13	吉林省	−0.62191	29
江西省	0.073994	14	天津市	−0.70455	30
福建省	0.051726	15	北京市	−0.7914	31
西藏自治区	−0.01389	16			

4. 聚类分析

所谓聚类问题，就是给定一个元素集合 D，其中每个元素具有 n 个可观察属性，使用某种算法将 D 划分成 k 个子集，要求每个子集内部的元素之间相异度尽可能低，而不同子集的元素相异度尽可能高。其中每个子集叫做一个簇。

K-means 算法是很典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。该算法认为簇是由距离靠近的对象组成的，因此把得到紧凑且独立的簇作为最终目标。

考虑到当 k 值确定时，聚类效果达到类间距比较大，组内元素聚集紧密时，k 值为较优值。于是我们构建了一个比值 μ 描述聚类效果：

$$\mu_{i1}=\frac{y_{i\min}-y_{i-1\max}}{y_{i\max}-y_{i\min}},\quad \mu_{i2}=\frac{y_{i\min}-y_{i-1\max}}{y_{i-1\max}-y_{i-1\min}},\quad i=2,3,\cdots$$

当 μ 越大时，我们可以判断此次聚类的效果越好。求出所有 μ 值后，若是有的 μ 太小，则说明此次分类不合理，分类的结果当中存在较大的类间距和较分散的聚类。最终我们得到的结果见表 7：

Table 7. Province clustering results
表 7. 各省份聚类结果

地区	分类	地区	分类	地区	分类
北京市	6	安徽省	5	重庆市	1
天津市	6	福建省	6	四川省	3
河北省	3	江西省	3	贵州省	1
山西省	3	山东省	5	云南省	3
内蒙古自治区	6	河南省	3	西藏自治区	4
辽宁省	6	湖北省	6	陕西省	6
吉林省	6	湖南省	3	甘肃省	1

Continued

黑龙江省	6	广东省	2	青海省	1
上海市	6	广西壮族自治区	4	宁夏回族自治区	4
江苏省	5	海南省	4	新疆维吾尔自治区	4
浙江省	5				

根据上述结果，分析我国 31 个省份自治区直辖市的人口结构和规模预测：

第一类为人口数量一般但是增速较大的地区，主要代表省份为重庆市、甘肃省、青海省和贵州省。这些省份在之前的经济水平方向增长也位于中上水准，因此容易吸引一些人来这些省份定居，预测这些省份接下来的数年人口会有小幅的增速。

第二类为人口数量较大而且增速较快的地区，主要代表省份为广东，我们知道，广东作为中国人口最多的省份，同时又是南部经济中心，拥有很多沿海城市和经济特区，一方面人口基数很大，另一方面也吸引了大量的外来人口，预测广东省接下来的数年人口增速同样会较快。

第三类为人口较大但是增速一般的省份，主要代表省份有河北省、山西省、江西省、河南省、湖南省、四川省和云南省。这些省份大部分的因子得分除了第一项以外都较为一般，可见这些省份大部分本身人口基数比较大，但是由于区位因素，人口增速较慢，对于人才的吸引不如其他省份强，预测这几个省接下来的数年人口增速会比较稳定，可能会有小幅增长。

第四类为人口数量较小但是增速较快的省份，主要代表省份为广西、海南、西藏、宁夏和新疆。这些省份大部分位于我国西部的边境线上，由于历史原因可能人口基数较小，但是由于近年来的发展和国家政策的支持，人口数量发生了大幅度的增长，预测这些省份接下来的数年人口会有大幅的增速。

第五类为人口数量一般而且增速也一般的地区，主要代表省份为江苏省、浙江省、安徽省和山东省。这些省份你大部分位于我国的东部沿海地区，发展较为平均，人口结构较为合理，因此人口数量和增速都在一定的区间范围，预测这些省份接下来人口的增速会较为稳定。

最后一类第六类为人口数量一般同时增速很慢的地区，主要代表省份为北京市、天津市、内蒙古、辽宁省、吉林省、黑龙江省、上海市、福建省、湖北省和陕西省。造成这些省份人口数量一般同时增速很慢的原因主要有两类，其中北京、上海和天津三个直辖市受限于城市的规模，无法使得城市在长期有较为稳定的增速；剩下的省份大部分处于我国的内陆，经济发展相对于其他省份较慢，没有办法吸引更多的人，因此预测这些省份在接下来数年人口的增速会很慢。

5. 结论

本文的主要结论：基于国内各个省份不同维度的指标，本文对各个省份的人口结构和规模预测进行了一定的探讨，从社会总体基础设施水平、初等教育普及指数、人口总体生活水平指数、居民总体经济水平、人均交通发展水平指数等 5 个方面构建了符合各个省份的人口结构和规模预测的指标体系。对于每个方面都根据科学性、全面性、代表性、可操作性等原则选取指标。接着通过基于因子分析的综合得分模型以及聚类分析对省份进行排名和分类。总得来说，各个省份的人口结构和规模预测由很多方面决定的，因此衡量一个省份的人口规模与结构预测也需要很多数据共同的支撑，各个省份的人口结构和规模预测在某些程度上也代表了这些省份的综合水平，今后各个省份应更注重多方面的综合发展。这也启示不同省份的政府，应该如何发展自身的经济、教育等指标，从而达到更加合理健康的人口规模和结构。

本文数据主要来源于 16 年国家统计局发布出的数据，但是我们小组根据已知的数据得到的结论，对

比获得的部分 17 年、18 年的数据，发现很符合实际情况以及我们根据结论做出的预测，所以，我们相信 16 年的数据并不是一组特例，而是可以将结论应用到未来几年的相对普适性的例子。

本文的主要创新内容：1) 采用了因子分析的方法来构建综合得分模型，具体包括以下步骤，因子的提取、构建综合得分模型。2) 利用 K-means 聚类的方法对各个城市的综合得分进行一维聚类，得出最优分类结果。

参考文献

[1] 路锦非, 王桂新. 我国未来城镇人口规模及人口结构变动预测[J]. 西北人口, 2010, 31(4): 1-6+11.

[2] 国家数据[Z]. http://data.stats.gov.cn.