Statistics and Application 统计学与应用, 2017, 6(4), 472-481
Published Online October 2017 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2017.64053

Hans汉斯

Economic Comprehensive Evaluation and Regional Characteristics Analysis of Cities in Shandong Province

—Based on Big Data Analysis

Minggang Sui, Xiangdong Liu

School of Economics, Jinan University, Guangzhou Guangdong
Email: tliuxd@jnu.edu.cn

Received: Oct. 9th, 2017; accepted: Oct. 23rd, 2017; published: Oct. 30th, 2017

Abstract

Using the big data analysis, economic indicators in different cities of Shandong province in 2015 are investigated. Some economic indicators are visualized. Then the k-means cluster is used to cluster the cities. Both city rank and category are obtained by data mining, which provides the basis for the planning of city economy.

Keywords

Big Data Analysis, Economic Indicators, Visualization Method, Data Mining, k-Means Cluster

山东省各市经济指标和地区特征的综合评价

—基于大数据方法

眭铭刚，柳向东

暨南大学经济学院，广东 广州
Email: tliuxd@jnu.edu.cn

收稿日期：2017年10月9日；录用日期：2017年10月23日；发布日期：2017年10月30日

摘　要

对2015年山东省各市的多个经济指标进行大数据方法分析，利用可视化方法显示出各市的一些指标，然

文章引用: 眭铭刚, 柳向东. 山东省各市经济指标和地区特征的综合评价[J]. 统计学与应用, 2017, 6(4): 472-481.
DOI: 10.12677/sa.2017.64053

后利用k-均值聚类算法对各市进行聚类，并结合数据挖掘，得出各地区经济排名及分类，为山东省经济规划提供了依据。

关键词

大数据方法，经济指标，可视化分析，数据挖掘，k-均值聚类

1. 引言

对一个区域的经济分析，有利于该区域的经济规划，并为政府决策提供建议。本文主要研究了山东省 17 个市的经济状况。通过对 8 项经济指标(x_1 人均地区生产总产值(元/人)、x_2 人均工业总产值(元/人)、x_3 人均建筑业总产值(元/人)、x_4 人均实际使用外资(元/人)、x_5 人均入境旅游外汇收入(元/人)、x_6 人均批发零售贸易业营业利润(元/人)、x_7 人均公共财政收入(元/人)、x_8 人均农林牧渔业总产值(元/人))的研究和分析，利用大数据的方法探索出几个具有代表性的因子，并对各市进行分类，得出各市的经济排名以及区域特征，并根据研究得出一些指导性的建议。

陈伟[1]用多元统计分析中的主成分分析和因子分析法对武汉城市圈的经济状况进行了较科学的分析；左瑞琼[2]介绍了多元统计分析方法的主要内容以及在经济研究工作的应用；柳向东和陈锦岚[3]在大数据与数据可视化方法方面进行的研究。本文就在此基础上探索了可视化方法包括脸谱图法和星象图法，并利用数据挖掘中的 k-means 均值聚类以及因子探索分析等方法得出了一些经济指标的综合因子分类和相对应的政策建议。

2. 经济指标的建立

在选取指标时，主要考虑这些指标能从国民经济、对外经济、旅游、财政、工业、农业等方面反映地区经济特性，统计数据应可靠且相关性较小。由于每个地方的人口总数不一致，所以每项指标的总值并不能很好的代表每个地方的经济发展水平，所以本文将选取的 2015 年山东省统计年鉴中 17 个市 8 项经济指标进行人均化，建立如下的经济指标体系：x_1 人均地区生产总产值(元/人)、x_2 人均工业总产值(元/人)、x_3 人均建筑业总产值(元/人)、x_4 人均实际使用外资(元/人)、x_5 人均入境旅游外汇收入(元/人)、x_6 人均批发零售贸易业营业利润(元/人)、x_7 人均公共财政收入(元/人)、x_8 人均农林牧渔业总产值(元/人)。

3. 数据描述和可视化分析

1) 各地区经济的星象(图 1)。
2) 各地区经济的脸谱(图 2)。

4. 基于数据挖掘的统计分析

4.1. 主成分分析

根据所搜集整理的数据[4]，R3.3.2 [5]统计软件进行相关分析：

从表 1 中可看出 x_7 与 x_1、x_2、x_3、x_4、x_5 相关性较强，尤其是与 x_1，即人均公共财政收入与人均

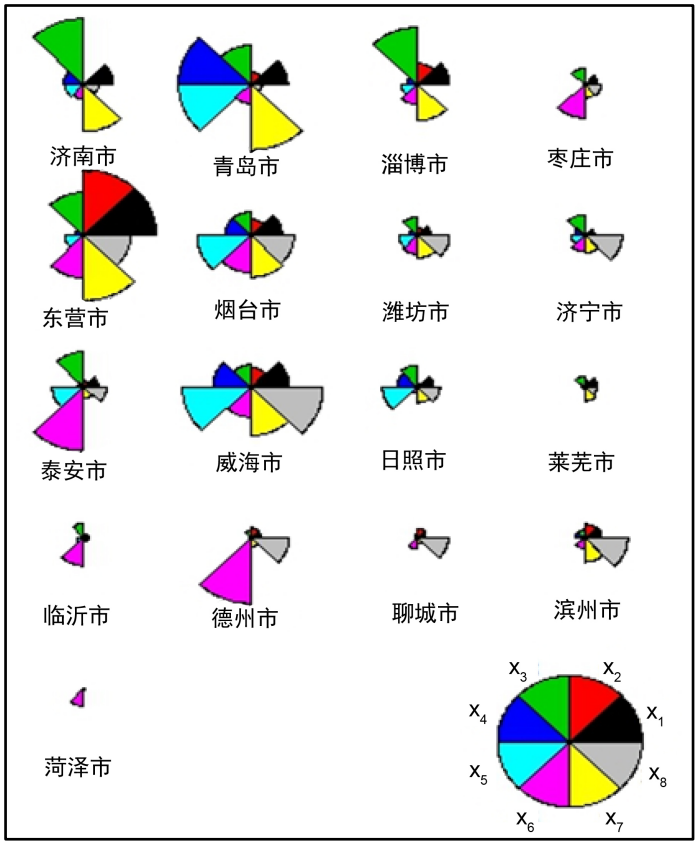

Figure 1. Star chart of economic analysis of cities
图 1. 各市经济分析的星象图

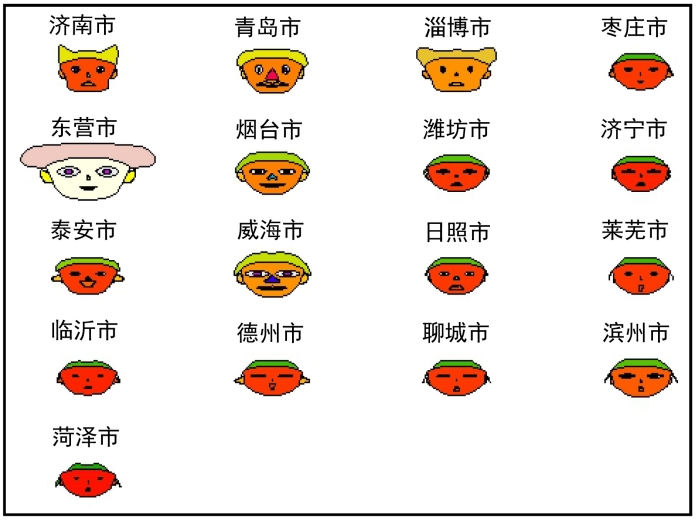

Figure 2. Face spectrum city economic analysis
图 2. 各市经济分析的脸谱图

地区生产总产值、人均工业总产值(元/人)、人均建筑业总产值(元/人)、人均实际使用外资(元/人)、人均入境旅游外汇收入(元/人)相关性较强，而尤其是与人均地区生产总值，这符合人均公共财政收入来源于这些项目的情况，故可把 x_7 删掉。

Table 1. Coefficient matrix
表 1. 系数矩阵

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1							
X_2	0.869241	1						
X_3	0.630158	0.313056	1					
X_4	0.46119	0.095492	0.391481	1				
X_5	0.493689	0.163757	0.366582	0.887571	1			
X_6	0.172857	0.270783	0.078619	−0.10866	0.102032	1		
X_7	0.909432	0.627116	0.690721	0.736312	0.674141	0.022955	1	
X_8	0.388064	0.397192	−0.15383	0.148329	0.346287	0.266407	0.318146	1

主成分分析的原理是设法将原来变量重新组合成一组新的相互无关的几个综合变量，同时根据实际需要从中可以取出几个较少的总和变量尽可能多地反映原来变量的信息，也是统计上处理降维的一种方法，是一种无指导学习方法。

下面做主成分分析，结果如下：

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.9999045	1.2467232	1.0808535	0.9118963
Proportion of Variance	0.4999523	0.1942899	0.1460305	0.1039444
Cumulative Proportion	0.4999523	0.6942421	0.8402727	0.9442170
	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	0.56468885	0.31901161	0.13468005	0.0865060867
Proportion of Variance	0.03985919	0.01272105	0.00226734	0.0009354129
Cumulative Proportion	0.98407620	0.99679725	0.99906459	1.0000000000

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
x_1	−0.464	0.194	0.209	0.155			0.367	0.732
x_2	−0.338	0.472	0.238	0.269	−0.436	−0.226	−0.450	−0.299
x_3	−0.330	−0.178	0.521	−0.313	0.596	−0.147	−0.335	
x_4	−0.363	−0.456	−0.285		−0.285	0.431	−0.498	0.243
x_5	−0.379	−0.296	−0.419	−0.197	−0.104	−0.692	0.194	−0.165
x_6		0.492	−0.141	−0.821	−0.136	0.189		
x_7	−0.484			0.110		0.467	0.480	−0.536
x_8	−0.209	0.405	−0.584	0.276	0.581		−0.177	

结果表明前 4 个主成分已达 94%的累积贡献率，这说明前 4 个主成分已经反映了信息的 94%，于是前 4 个因子可以作为评价山东省 17 个市的经济发展指标的综合变量。从而达到降维的目的，而损失的信息却不多。

上面 Loadings 反映了载荷的大小，它反映了原变量指标与主成分的相关关系，即反映了原变量对于主成分的重要程度。在解释主成分时，我们需要考察载荷，同时也需要考察一下原变量与主成分的相关

系数。前者是从多变量的角度，后者是从单变量的角度，因而前者应更值得重视[6]。而我们知道相关系数与载荷同符号，且成正比(图 3)。

下面得出前四个主成分：

F_1 代表反映地区的综合经济实力

$$F_1 = -0.464x_1 - 0.338x_2 - 0.330x_3 - 0.363x_4 - 0.379x_5 - 0.484x_7 - 0.209x_8$$

第 1 主成分对应载荷的符号相同，且其值都在 0.3 左右，差别不大，它反映了地区的综合经济实力。综合经济实力较强的地区，它的 8 项指标的值都较大，所以第 1 主成分的值较小(因为载荷均为负数)；而综合实力较弱的地区，它的 8 项指标的值都较小，因此第 1 主成分的值就较大。所以称第 1 主成分为综合经济因子。

F_2 代表反映批发零售贸易业、工业实力

$$F_2 = 0.194x_1 + 0.472x_2 - 0.178x_3 - 0.456x_4 - 0.296x_5 + 0.492x6 + 0.405x_8$$

第 2 主成分中 x_1、x_2、x_6、x_8 对应的载荷为正，载荷总和为 1.563，而 x_1、x_2、x_6、x_8 分别代表人均地区生产总值、人均工业生产总值、人均批发零售贸易业营业利润、人均农林牧渔业总产值，其中 x_2、x_6、x_8 对应的载荷较大；x_3、x_4、x_5 对应载荷为负，载荷绝对值总和为 0.93，x_4 有绝对值较大的负载荷，x_4 代表人均使用外资；结合变量的含义，第 2 主成分反映了批发零售贸易业、工业相对于外商投资的经济状况，称为批发零售、工业因子。

F_3 代表反映了农林牧渔业的实力

$$F_3 = 0.209x_1 + 0.238x_2 + 0.521x_3 - 0.285x_4 - 0.419x_5 - 0.141x_6 - 0.584x_8$$

第 3 主成分中 x_8 有绝对值较大的负载荷，x_3 有较大大的正载荷，其余变量的载荷较小，大(小)的 F 值意味着 x_8 有较小(大)的值，而 x_3 倾向于有较大的值，这个主成分基本上是 x_8 (人均农林牧渔业总产值)

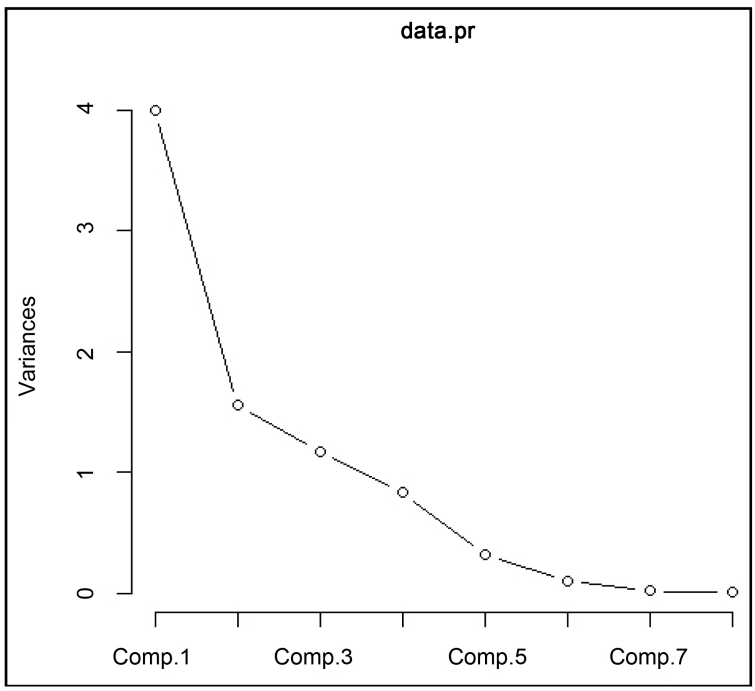

Figure 3. Scree plot
图 3. 碎石图

和 x_3 (人均建筑业总产值)的对比，反映了农林牧渔业相对于建筑业的经济状况，故称第 3 主成分为农林牧渔业因子。

$$F_4 = 0.155x_1 + 0.269x_2 - 0.313x_3 - 0.197x_5 - 0.821x_6 + 0.110x_7 + 0.276x_8$$

第 4 主成分中 x_6 的系数最大，其余都较小，故反映了批发零售业经济和其余产业的经济状况对比，从而称第 4 主成分为批发零售业和其他行业的对比因子。

进一步利用 R3.3.2 [5] [7]计算各市的主成分得分状况如表 2。

另外选择方差贡献率作为综合经济因子的权重，计算出综合经济因子，计算结果如下：

$$F = 0.4999523F_1 + 0.1942899F_2 + 0.1460305F_3 + 0.1039444F_4$$

从主成分结果可以看到，每个地区经济都有自己的特点，但不同的地区又存在着许多共性，可以将具有一定共性的地区划为一类，以利于更好地进行区域经济规划搞好经济建设，下面就对上述 17 个市进行因子分析。

4.2. Varimax 法旋转因子分析

主成分分析通过线性组合将原变量综合成几个主成分，用较少的综合指标来代替原来较多的指标。在多变量分析中，某些变量间可能会有关联，存在不能直接观测到的、但影响可观测变量变化的公共因子。因子分析法就是寻找这些公共因子的模型分析方法，它是在主成分的基础上，构造若干意义较为明确的公共因子，以它们为基础分解原变量，以此考察原变量间的联系与区别。

Table 2. Principal component score
表 2. 主成分得分

	F1	F2	F3	F4
济南市	−0.27947	−0.01078	2.63211	−1.25783
青岛市	0.249606	2.482547	0.231735	−1.43161
淄博市	0.109332	−0.50386	1.981636	0.504477
枣庄市	−0.1677	−0.75625	−0.1683	−0.39052
东营市	3.381101	−0.7833	0.328792	0.906512
烟台市	0.288237	1.288632	−0.48963	0.510039
潍坊市	−0.29527	−0.00267	−0.20038	−0.44693
济宁市	−0.57516	−0.1944	0.062933	−0.53233
泰安市	−1.13296	0.124425	0.744713	2.201258
威海市	0.576875	1.886549	−0.85554	0.790528
日照市	−0.76406	0.510012	−0.15058	1.103241
莱芜市	−0.01763	−0.66244	−0.54459	−0.63501
临沂市	−0.85128	−0.56436	−0.18538	0.382922
德州市	−0.11349	−0.82135	−0.7145	−0.05538
聊城市	−0.23723	−0.72037	−0.90494	0.088907
滨州市	0.531249	−0.5645	−0.79813	−1.41855
菏泽市	−0.70213	−0.70789	−0.96995	−0.31973

下面进行 varimax 法旋转因子[7]分析，然后得出结论(表 3)。

旋转后公共因子代表的意义较为明显，因子 F_1 在 x_1 人均地区生产总产值(元/人)、x_2 人均工业总产值(元/人)、x_7 人均公共财政收入(元/人)上载荷值较高，因此因子 F_1 代表地区的生产收入能力；因子 F_2 在 x_4 人均实际使用外资(元/人)、x_5 人均入境旅游外汇收入(元/人)上载荷值较高，因此因子 F_2 代表吸引外资能力；F_3 在 x_8 人均农林牧渔业总产值(元/人)上的载荷值较高，因此因子 F_3 代表农业生产能力；因子 F_4 在 x_6 人均批发零售贸易业营业利润(元/人)上载荷值较高，因此因子 F_4 代表批发零售贸易能力。可以看出，因子分析得到的公共因子的解释比对主成分的解释更为明确。

对比表 4 和表 5，旋转因子的排名和综合因子的排名并无太大差别，东营市、威海市、青岛市分别位居前三。

Table 3. Rotational factor loading
表 3. 旋转因子载荷

	Factor 1	Factor 2	Factor 3	Factor 4
x_1	0.9296	0.34679	0.03346	0.07760
x_2	0.9293	−0.05479	0.20984	0.13822
x_3	0.5680	0.37290	−0.63781	0.11761
x_4	0.1491	0.95527	−0.03491	−0.11989
x_5	0.1571	0.94652	0.12170	0.10731
x_6	0.1071	−0.02455	0.09402	0.98466
x_7	0.7515	0.63336	−0.05634	−0.04519
x_8	0.3084	0.21141	0.84677	0.18561

Table 4. Rotation factor score
表 4. 旋转因子得分

	F	rank
济南市	−0.01114	6
青岛市	0.70316	3
淄博市	−0.07838	8
枣庄市	−0.37727	13
东营市	1.07785	1
烟台市	0.58371	4
潍坊市	−0.12965	10
济宁市	−0.20942	12
泰安市	0.04102	5
威海市	1.01217	2
日照市	−0.16933	11
莱芜市	−0.54488	15
临沂市	−0.56809	16
德州市	−0.05270	7
聊城市	−0.41524	14
滨州市	−0.1081	9
菏泽市	−0.6634	17

Table 5. Ranking of cities by factor
表 5. 各城市综合因子排名

	F1	F2	F3	F4	F
东营市	3.381101	−0.7833	0.328792	0.906512	1.680441
威海市	0.576875	1.886549	−0.85554	0.790528	0.612184
青岛市	0.249606	2.482547	0.231735	−1.43161	0.492158
烟台市	0.288237	1.288632	−0.48963	0.510039	0.375987
淄博市	0.109332	−0.50386	1.981636	0.504477	0.298583
济南市	−0.27947	−0.01078	2.63211	−1.25783	0.111806
滨州市	0.531249	−0.5645	−0.79813	−1.41855	−0.10808
日照市	−0.76406	0.510012	−0.15058	1.103241	−0.19022
泰安市	−1.13296	0.124425	0.744713	2.201258	−0.20469
潍坊市	−0.29527	−0.00267	−0.20038	−0.44693	−0.22386
莱芜市	−0.01763	−0.66244	−0.54459	−0.63501	−0.28305
枣庄市	−0.1677	−0.75625	−0.1683	−0.39052	−0.29595
德州市	−0.11349	−0.82135	−0.7145	−0.05538	−0.32642
济宁市	−0.57516	−0.1944	0.062933	−0.53233	−0.37147
聊城市	−0.23723	−0.72037	−0.90494	0.088907	−0.38147
临沂市	−0.85128	−0.56436	−0.18538	0.382922	−0.52252
菏泽市	−0.70213	−0.70789	−0.96995	−0.31973	−0.66345

从图 4 和图 5 可以看出淄博在生产收入能力方面较强，而青岛市、威海市、烟台市在吸引外资能力方面较强，这与它们临海的便利交通有很大关系。

4.3. k-均值聚类分析法

利用 R3.3.2 统计软件对数据进行 k-均值聚类分析法，并分为 3 类。得出结果见图 6。

第 1 类：泰安市、德州市、枣庄市、临沂市、菏泽市、莱芜市、聊城市、日照市、滨州市、潍坊市、济宁市。

第 2 类：济南市、淄博市。

第 3 类：青岛市、东营市、烟台市、威海市。

5. 综合评价

本文在前人研究的基础上选用 2015 年山东省统计年鉴[4]中的 17 个市的 8 项经济指标，借用 R3.3.2 [5] [7]对山东省地区的经济进行主成分分析，得到了与事实较吻合的结果，在此基础上用 Ward 法进行分析，把各市的经济状况分成 3 类，再结合综合因子得出相应结论：

1) 东营市、威海市、青岛市、烟台市综合因子排名靠前，综合经济水平最高，济南市、淄博市综合因子排名次之，综合经济水平位于第二，其他城市的综合因子最低，综合经济水平最低。出现这样状况的原因是：东营市、威海市、青岛市、烟台市濒临海域，交通便利，对外开放程度高，利于经济发展；济南是省会城市，经济基础好，人口整体素质较高，竞争力强，是重要的政府机关、商业机构的集中地，就业机会多，城市人口比重大、居民有着比较稳定的收入；而淄博地处黄河三角洲高效生态经济区、山东半岛蓝色经济区两大国家战略经济区与省会城市群经济圈的重要交汇处，是中国城市 GDP40 强，位列

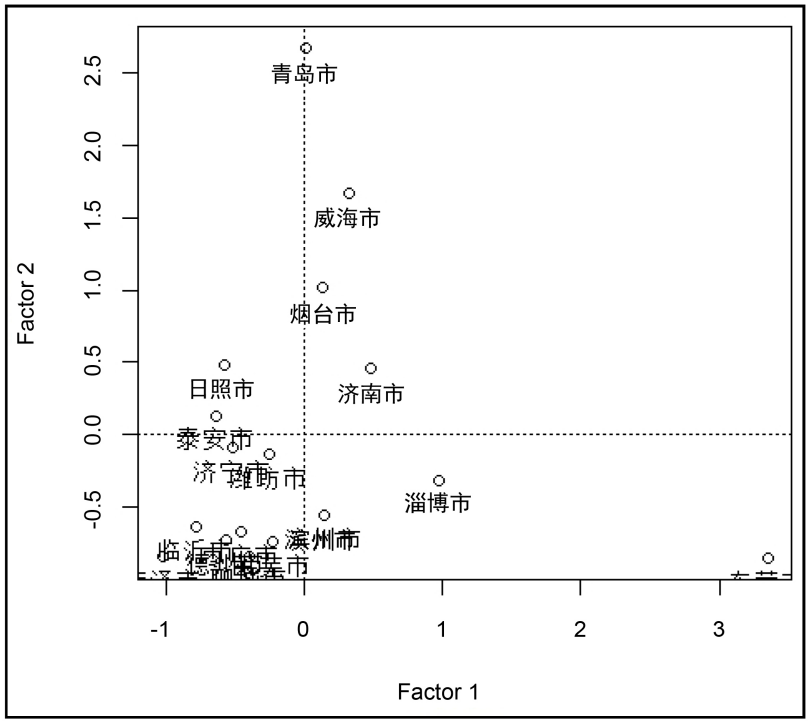

Figure 4. Rotating factor score chart
图 4. 旋转因子得分图

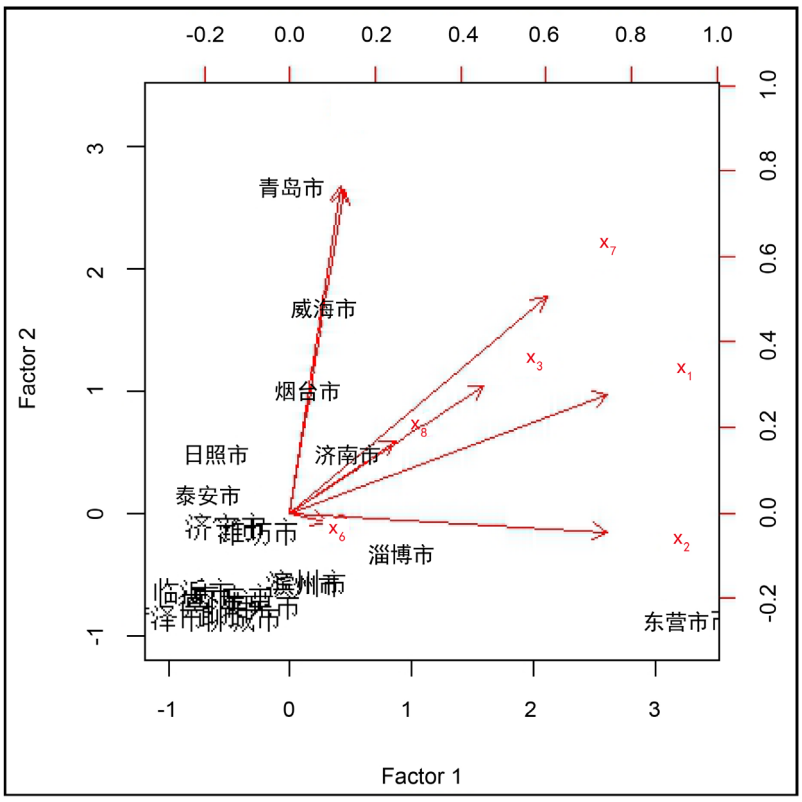

Figure 5. Spin factor information overlay
图 5. 旋转因子信息重叠图

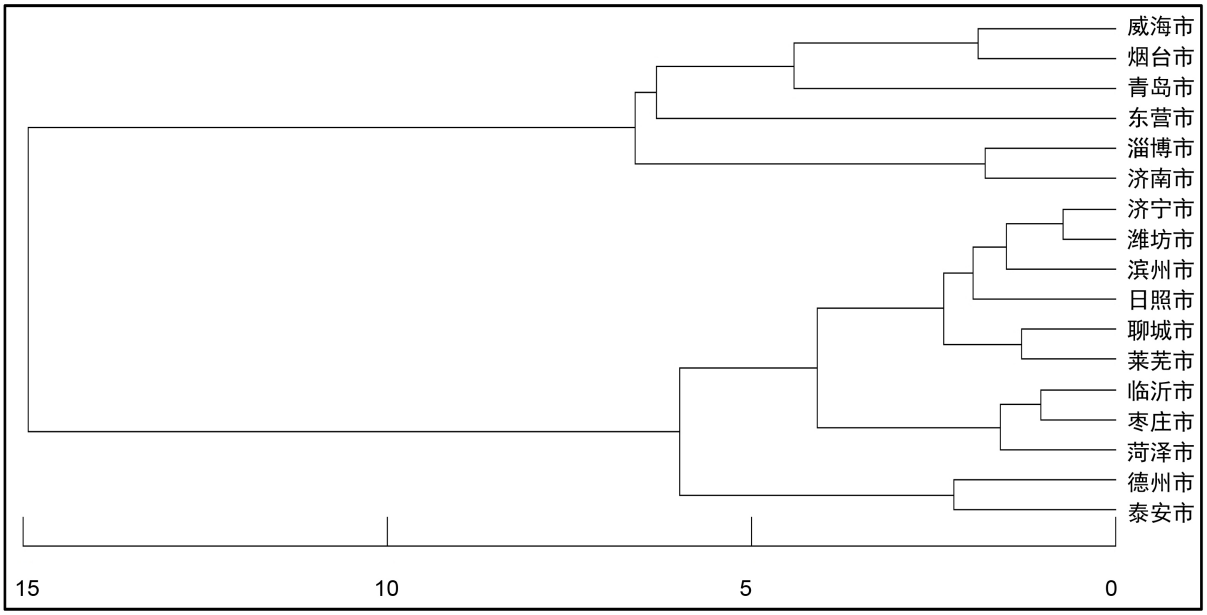

Figure 6. Shandong province city economic classification pedigree diagram
图 6. 山东省各市经济分类谱系图

社科院 2014 年中国城市综合经济竞争力排行榜第 34 名，是全国首批科技兴市试点市和国家级星火技术密集区，为全国重要的石油化工基地。

2) 威海市、青岛市、烟台市在批发零售贸易业和工业方面的水平明显高于其它地区，济南市、淄博市在建筑业以及工业方面水平高于其他地区，而菏泽市、滨州市、聊城市、莱芜市、威海市等城市相对农林牧渔业较发达，像威海、青岛等沿海城市，由于靠海，故其渔林业较为发达，而菏泽、滨州、聊城、莱芜等城市农业较为发达，工业相对落后，故积极推进这些城市的农村发展，加快科技教育发展，调整产业结构，注意发展地区特色经济，在生产发展的基础上增加城乡居民收入。再比较每个地区的特点，几乎各个地区的发展都不是均衡发展，在工业等方面青岛、烟台、济南等城市处于较高水平，源于国家的大幅投入和自身处于交通发达地区的地优势，值得注意的是菏泽、临沂、聊城等这样的城市在六项指标上水平都较低，全省在经济规划中要注意调整这些城市的经济结构，有关部门应出台相应扶持政策，促进其经济快速发展。

参考文献 (References)

[1] 陈伟. 多元统计分析在区域经济评价中的应用[D]: [硕士学位论文]. 武汉: 武汉科技大学, 2010.

[2] 左瑞琼. 多元统计分析方法介绍及在经济中的应用[J]. 时代经贸, 2007, 78(5): 23-14.

[3] 柳向东, 陈锦岚. 旅游电商对产品区域异质性的提升策略研究—基于大数据与数据可视化方法[J]. 统计与信息论坛, 2017, 32(8): 31-38.

[4] 山东省统计局. 山东统计年鉴-2015[M]. 北京: 中国统计出版社, 2015: 100-560.

[5] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007: 397-461.

[6] 王学民. 应用多元分析[M]. 第 4 版. 上海: 上海财经大学出版社, 2014: 192-193.

[7] Tan, P.N., Steinbach, M., Kumar. V. 数据挖掘导论(完整版)[M]. 范明, 范宏建, 等, 译. 北京: 人民邮电出版社, 2011.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org