Statistics and Application 统计学与应用, 2020, 9(5), 754-761
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.95077

Hans汉斯

基于多维标度法的互联网基本资源发展状况评价

秦　丰

长安大学理学院，陕西　西安
Email: 974684882@qq.com

收稿日期：2020年9月8日；录用日期：2020年9月23日；发布日期：2020年9月30日

摘　要

简述了多维标度分析方法的思想及原理，采用多维标度法和聚类分析对全国31个省份互联网基本资源的发展状况进行综合的评价与分析，结果显示全国31个省份互联网基本资源的发展状况可分为四个层次，并以此分类为城市的交流与发展提供参考。

关键词

多维标度法，多维标度图，非度量MDS，聚类分析

Evaluation of the Development of Basic Internet Resources Based on Multidimensional Scaling

Feng Qin

School of Science, Chang'an University, Xi'an Shaanxi
Email: 974684882@qq.com

Received: Sep. 8th, 2020; accepted: Sep. 23rd, 2020; published: Sep. 30th, 2020

Abstract

This paper briefly introduces the idea and the principle of multidimensional scaling analysis method. The development status of Internet basic resources in 31 provinces of China is analyzed and evaluated by multi-dimensional scaling and clustering analysis. The result shows that the devel-

文章引用: 秦丰. 基于多维标度法的互联网基本资源发展状况评价[J]. 统计学与应用, 2020, 9(5): 754-761.
DOI: 10.12677/sa.2020.95077

opment status of Internet basic resources in 31 provinces of China can be divided into four levels, which can provide reference for the exchange and development of cities.

Keywords

Multidimensional Scaling Method, Non-Metric MDS, Clustering Analysis

1. 引言

实际问题中，得知城市之间的距离或得知了城市之间的距离大小次序，如何确定他们之间的相对位置；若得知了消费者对某些类品牌产品之间的差异程度的资料，如何确定他们在消费者认知中的相对位置，进而衡量消费者的偏好。在生活中我们通常会需要确定对象之间的相对距离来对其进行评价。本文主要针对互联网基础资源的发展状况进行评价，确定各个地区互联网发展的相对位置，并结合聚类分析结果为城市发展，企业投资提供参考。

2. 多维标度法概述

多维标度法是一种多元统计方法，其所要解决的问题是：当 n 个对象两两之间的相似性或距离确定之后，确定这些对象在一个合适的低维空间中的相对位置。低维空间中的任意一个点代表一个对象，通过该空间中点与点之间的距离来反映对象两两之间的相似性，通过多维标度法使得空间中点与点之间的距离与原对象两两之间的相似性尽可能一致，使得降维过程中发生的形变尽可能的小。

多维标度法的目的是通过客体间的距离或相似数据来表现他们之间的空间分布，进而通过空间相对位置来揭示实际客体间的亲疏和相似程度。多维标度法首先通过操作构建一个关键的维数，进而在该维数下表现样本的坐标，最后画出它的多维标度图。

根据基础数据的不同，多维标度法可以分为两类：一类为度量的多维标度法(metric MDS)，另一类为非度量的多维标度法(nonmetric MDS)。前者使用的数据是定量数据，即用间隔尺度或比率尺度测得的数据[1]。后者使用的数据是定性数据，即用次序尺度测得的相似数据[1]。

3. 多维标度法理论简述

3.1. 度量多维标度分析相关理论

一个距离矩阵 $D=\left(d_{ij}\right)_{n\times n}$ 称为欧式型的[2]，若存在某个正整数 p 及 p 维空间 R^p 中的 n 个点 $x_1,\cdots,x_n$，使得

$$d_{ij}^2=\left(x_i-x_j\right)'\left(x_i-x_j\right),\ i,j=1,2,\cdots,n \tag{1}$$

对于距离阵 $D=\left(d_{ij}\right)_{n\times n}$，设 p 和 R^p 中的 n 个点 $x_1,\cdots,x_n$，矩阵表示为 $X=\left(x_1,\cdots,x_n\right)'$，用 $\hat{d}_{ij}$ 表示 x_i 与 x_j 的欧式距离，$\hat{D}=\left(\hat{d}_{ij}\right)$，使得 $\hat{D}$ 与 D 在一定条件下相近，则称 X 为 D 的一个解，称 x_i 为 D 的一个拟合构造点，X 为拟合构图，D 的拟合距离阵为 $\hat{D}$，特别的，当 $\hat{D}=D$ 时，称 x_i 为 D 的构造点，X 为构

图。得出拟合构图就可以得出 n 个拟合构造点 x_i 的坐标，我们就可以画出多维标度图，进而对原始客体进行一个合理的统计解释。

令 $A=\left(a_{ij}\right)$，$a_{ij}=-\frac{1}{2}d_{ij}^2$，$B=H'AH$，$H=I_n-\frac{1}{n}1_n1_n'$，一个的距离阵 $D=\left(d_{ij}\right)_{n\times n}$ 是欧式型的充分必要条件是 $B\geq 0$。

对于必要性：

设 D 是欧氏型的，则由定义可知，存在 $x_1,\cdots,x_n\in R^p$，使得

$$d_{ij}^2=-2a_{ij}=\left(x_i-x_j\right)'\left(x_i-x_j\right) \tag{2}$$

可得

$$B=H'AH=A-\frac{1}{n}A1_n1_n'-\frac{1}{n}1_n1_n'A+\frac{1}{n^2}1_n1_n'A1_n1_n' \tag{3}$$

带入可得

$$b_{ij}=a_{ij}-\bar{a}_{i\cdot}-\bar{a}_{\cdot j}+\bar{a}_{\cdot\cdot} \tag{4}$$

其中 $\bar{a}_{i\cdot}=\frac{1}{n}\sum_{j=1}^n a_{ij}$，$\bar{a}_{\cdot j}=\frac{1}{n}\sum_{i=1}^n a_{ij}$，$\bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n a_{ij}$。

求得 a_{ij}，$\bar{a}_{i\cdot}$，$\bar{a}_{\cdot j}$，$\bar{a}_{\cdot\cdot}$ 带入可得

$$b_{ij}=\left(x_i-\bar{x}\right)'\left(x_i-\bar{x}\right),i,j=1,\cdots,n \tag{5}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^n x_i$。

将上式用矩阵表示并根据负定矩阵的性质有

$$B=\left(b_{ij}\right)_{n\times n}=\begin{pmatrix}\left(x_1-\bar{x}\right)'\\ \vdots\\ \left(x_n-\bar{x}\right)'\end{pmatrix}\left(\left(x_1-\bar{x}\right),\cdots,\left(x_n-\bar{x}\right)\right)'\geq 0 \tag{6}$$

其中，B 为拟合构图 X 的中心化内积矩阵。

对于充分性：

记 $p=rank\left(B\right)$，$\lambda_1,\cdots,\lambda_p$ 为 B 的正特征根，$x_{(1)},\cdots,x_{(p)}$ 为对应的特征向量。

由于 $B\geq 0$，则由谱分解定理

$$B=H'AH=\Gamma\Lambda\Gamma' \tag{7}$$

式中 $\Lambda=\mathrm{diag}\left(\lambda_1,\cdots,\lambda_p\right)$，$\lambda_1\geq\cdots\geq\lambda_p$ 为 B 的 p 个正特征值，Γ 的 p 个列为对应的 p 个标准正交化的特征向量。取 $X=\Gamma\Lambda^{1/2}$，其为 $n\times p$ 阶矩阵。把这个 X 写成 $X=\left(x_1,\cdots,x_n\right)'=\left(x_{(1)},\cdots,x_{(p)}\right)$，于是有：

$$X'X=\left(\Gamma\Lambda^{1/2}\right)'\left(\Gamma\Lambda^{1/2}\right)=\Lambda,B=XX' \tag{8}$$

即 $b_{ij}=x_i'x_j$。由此求得 x_i 与 x_j 两点的距离平方

$$\left(x_i-x_j\right)'\left(x_i-x_j\right)=b_{ii}-2b_{ij}+b_{jj}=a_{ii}-2a_{ij}+a_{jj}=-2a_{ij}=d_{ij}^2 \tag{9}$$

这表明存在正整数 p 和一个 $n\times p$ 阶矩阵 $X=\Gamma\Lambda^{1/2}$，使得 X 是 D 的构造点，即 D 是欧氏型的。

根据上述度量多维标度法的基本思想及方法，则其一般步骤：

(1) 计算对象两两之间的距离阵 $D=\left(d_{ij}\right)_{n\times n}$

(2) 由距离矩阵求得 $A=\left(a_{ij}\right)_{n\times n}, a_{ij}=-\frac{1}{2}d_{ij}^2$

(3) 令 $B=\left(b_{ij}\right)$，其中 $b_{ij}=a_{ij}-\overline{a}_{i\cdot}-\overline{a}_{\cdot j}+\overline{a}_{\cdot\cdot}$

(4) 求 B 的 r 个正特征值 $\lambda_1\geq\cdots\geq\lambda_r$ 和 r 个正特征值 $\lambda_1,\cdots,\lambda_r$ 对应的标准正交化的特征向量，维数空间 r 一般为 2、3 以达到可视化的效果。

(5) 根据 $X=\Gamma\Lambda^{1/2}$，得到 r 维拟合构图。

3.2. 非度量多维标度法的相关理论

在实际问题中，我们所能得到的 n 个客体的数据可能既不是相似系数也不是距离[3]，而只是他们之间某种差异程度的大小次序，其大小仅表明他们在排序队列中所处的位置，我们的目标是通过客体间的差异顺序找出一个拟合构图 X 拟合客体原本的差异关系。

非度量多维标度法，首先要构造一个可以反映样本信息的合适的 r 维空间，并用空间中的任意 n 个点来表示这 n 个客体，用 $X_i=\left(x_{i1},\cdots,x_{ir}\right)$ 表示第 i 个客体在 r 维空间的坐标，由这 n 个点组成的结构叫做初步图形结构，此时点间的距离数值大小次序不一定和原始客体之间的差异次序相同。接着，我们要一步步修正初步图形结构，使得这些代表客体的点之间距离的大小次序和原始客体之间的差异次序尽可能匹配。这其中的核心在于选择关键的维数和检验初步图形结构是否匹配进而进行修改。

对于坐标空间维数的确定[4]，理论上 n 个客体，$n-1$ 维空间可以完全反映出原本客体的次序，但是维数太高会使得计算复杂且结果不直观。实际中我们往往采用 2、3 维空间，然后去挑选出匹配程度最好的维数空间。

下面给出 Kruskal 非度量方法：

假定存在 n 个客体的不相似阵 $\left(\delta_{ij}\right)_{n\times n}$，首先用 r 维空间中任意不同点代表不同客体，$X_i=\left(x_{i1},\cdots,x_{ir}\right)$ 表示第 i 个客体在 r 维空间的坐标，用 d_{ij} 表示初步图形中客体 i 和客体 j 间的距离：

$$d_{ij}=\sqrt{\left(x_{i1}-x_{j1}\right)^2+\left(x_{i2}-x_{j2}\right)^2+\cdots+\left(x_{ir}-x_{jr}\right)^2} \tag{10}$$

若用 d_{ij} 确定的次序与原始客体的不相似次序不匹配，就逐步调整 n 个点的空间位置使得 d_{ij} 与不相似系数的 δ_{ij} 的完全一致[5]。即若：

$$\delta_{i_1j_1}\leq\delta_{i_2j_2}\leq\cdots\leq\delta_{i_mj_m}，\ 则\ d_{i_1j_1}\leq d_{i_2j_2}\leq\cdots\leq d_{i_mj_m}, m=\frac{1}{2}n\left(n-1\right) \tag{11}$$

问题的核心是 d_{ij} 与 δ_{ij} 的匹配性。Kruskal 采用最小二乘单调回归求解出 δ_{ij} 的单调正解 $\hat{d}_{ij}$，然后将 $\hat{d}_{ij}$ 与实际距离 d_{ij} 进行对比并作差，最后使用这个差值平方标准化之后作为匹配程度的度量，称之为应力(STRESS) [5]。

$$\text{STRESS}=\left[\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(d_{ij}-\hat{d}_{ij}\right)^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}d_{ij}^2}\right]^{\frac{1}{2}} \tag{12}$$

d_{ij} 与 $\hat{d}_{ij}$ 越接近，STRESS 指数就越小，表明拟合程度越好。实际中，我们往往当 STRESS 指数小于某个定值时，就认为该模型的拟合程度良好；若 STRESS 指数大于预先给定的临界值，就继续修改初始

图形，进而得到一个新的图形结构模型。一般采用迭代方法，找到使 STRESS 指数小于某个阈值的 r 维空间中的 n 个客体的坐标。

4. 多维标度法在互联网基本资源评价中的应用

互联网之于当今信息时代，就如同土地之于农业时代，机器之于工业时代，互联网作为一种重要的基础设施，正在对人类社会的变革发挥着巨大的作用，就像以蒸汽机为基础的机械制造时代，以电为原动力的电气化时代，以计算机为推动力的信息革命时代，如今互联网必将对今后的智能时代起到无法替代的作用，每个人已经和互联网深深结合在了一起，从 1994 年 4 月 20 日，我国正式接入国际互联网以来，在二十多年的奋斗历程中，我国互联网状况全面发展，“互联网+”的推进，都在告诉我们互联网已成为我国基础设施建设、生态建设、经济转型、城市发展、技术创新中不可或缺的内容。但是能够反映互联网发展的状况的指数众多，与此同时，各个地区的互联网发展的情况各异，各个指标此高彼低。因此必须对各地区互联网基本资源的发展状况进行综合的评价与分析。

本文分别利多维标度法和聚类分析[6]进行分析与评价并进行比对。其所依托的客体是 2017 年年底全国 31 个省份各省互联网主要指标发展情况。其所引用的资料来自于《中国统计年鉴 2018》，一共选取了 9 个指标：域名数 x_1 (万个)，网站数 x_2 (万个)，网页数 x_3 (万个)，IPv4 地址数 x_4 (万个)，互联网宽带接入端口 x_5 (万个)，互联网拨号用户 x_6 (万户)，移动互联网用户 x_7 (万户)，移动互联网接入流量 x_8 (万户)，互联网宽带接入用户 x_9 (万户)，具体数据见表 1。

Table 1. The development of major Internet indicators in 31 provinces
表 1. 全国 31 个省份各省互联网主要指标发展情况

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
北京	537.5	70.6	9,531,606.1	8633.6	1818.0	46.2	4639.7	78,117.9	541.9
天津	26.5	5.8	467,368.4	355.6	795.3	0.0	1309.6	25,341.8	339.3
河北	63.8	13.1	1,054,868.6	965.3	4126.9	0.0	6211.8	102,527.7	1910.1
山西	24.4	5.5	327,970.9	433.5	1840.1	3.4	2956.1	49,212.9	872.9
内蒙古	9.9	1.7	13,474.7	264.2	1294.1	0.0	2360.3	44,887.3	494.0
辽宁	47.5	12.3	209,977.1	1131.3	3118.8	16.7	3936.8	86,655.4	1058.6
吉林	24.6	3.3	152,994.0	409.8	1761.1	4.7	2349.6	65,369.0	501.5
黑龙江	20.1	4.3	258,253.2	409.8	1937.2	7.9	2858.7	56,155.5	664.6
上海	240.6	41.5	1,892,366.4	1527.6	1810.2	0.0	3393.1	48,162.7	681.3
江苏	161.6	28.9	1,291,048.1	1612.2	6531.7	3.7	9257.6	178,112.1	3106.1
浙江	207.6	40.0	3,316,217.1	2191.4	5455.1	29.6	7456.3	150,822.8	2464.6
安徽	72.2	8.1	206,187.6	558.9	2872.2	19.4	4639.7	85,024.9	1323.7
福建	882.5	30.3	827,546.3	657.1	2861.8	0.0	3508.8	71,383.3	1373.6
江西	33.2	4.4	212,189.3	586.0	1985.9	0.7	3079.3	57,687.9	997.1
山东	119.6	31.2	502,119.7	1656.3	5596.9	38.3	8508.0	118,818.1	2588.7
河南	123.2	23.5	1,187,272.6	890.8	4475.8	0.0	7542.4	134,407.2	2128.4

Continued

湖北	79.0	11.7	189,036.0	809.5	2605.5	12.7	4116.4	74,504.0	1242.9
湖南	112.4	8.8	152,334.9	799.3	2436.0	0.0	4922.5	76,087.2	1315.5
广东	397.9	77.7	3,274,458.2	3227.9	6482.3	68.8	14160.3	328,666.1	3246.8
广西	52.9	5.0	126,896.1	467.4	2216.4	7.3	3713.4	63,282.7	968.0
海南	38.6	2.5	84,745.1	159.2	571.6	0.0	911.7	22,937.8	228.7
重庆	43.8	5.5	77,438.7	569.0	1935.2	0.0	2831.6	54,842.8	866.9
四川	118.4	23.4	299,900.3	938.2	4702.8	7.7	6898.5	103,502.5	2167.5
贵州	25.5	2.0	17,166.7	149.0	1325.6	0.0	2938.5	84,612.3	568.6
云南	23.3	2.7	177,203.9	331.9	1661.8	5.6	3680.5	116,314.7	812.6
西藏	1.8	0.2	380.2	44.0	154.6	0.0	198.7	2420.2	61.2
陕西	39.6	6.9	1,660,042.5	552.1	1993.2	11.4	3668.3	78,748.9	903.2
甘肃	11.7	1.3	11,312.2	159.2	1099.9	9.8	2034.1	40,620.5	576.4
青海	2.0	0.4	1,529.2	61.0	310.5	0.9	533.3	17,331.4	120.1
宁夏	3.3	0.7	5714.8	94.8	415.0	1.4	682.5	20,823.4	159.2
新疆	8.2	1.1	10,284.0	203.0	1407.4	5.1	1855.8	21,999.4	569.9

下图和数据表示分别利用度量多维标度法和聚类分析进行处理的结果。

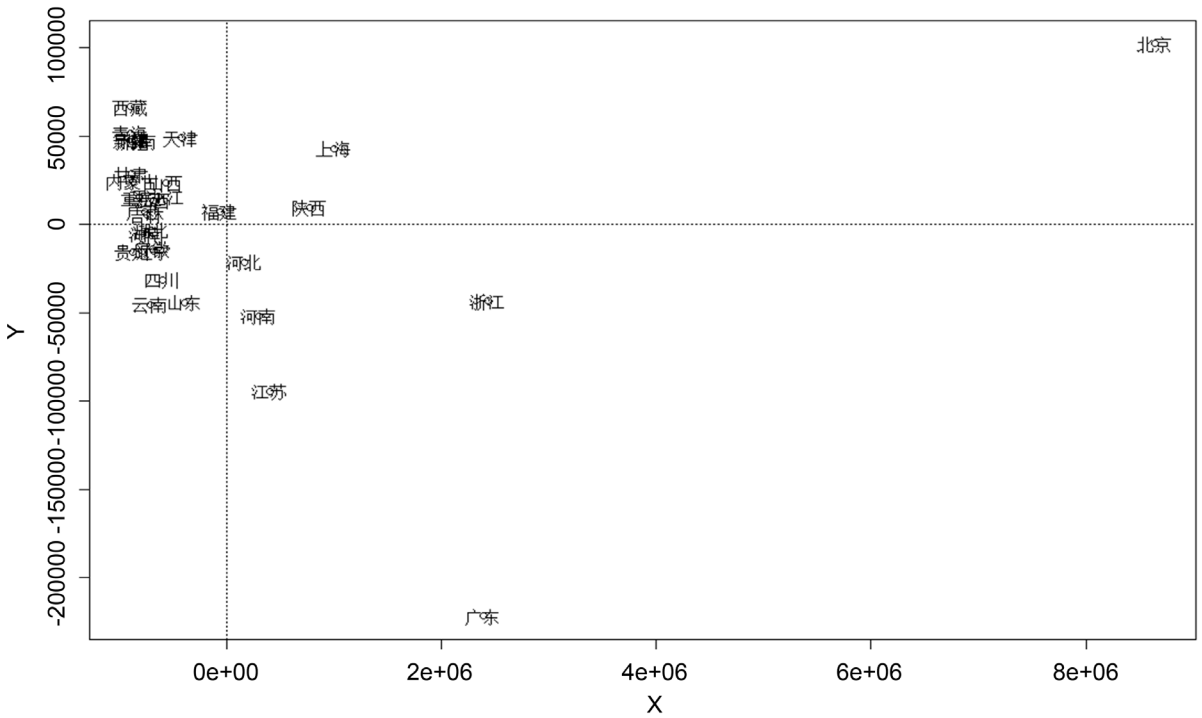

Figure 1. Multi-dimensional scale diagram
图 1. 多维标度图

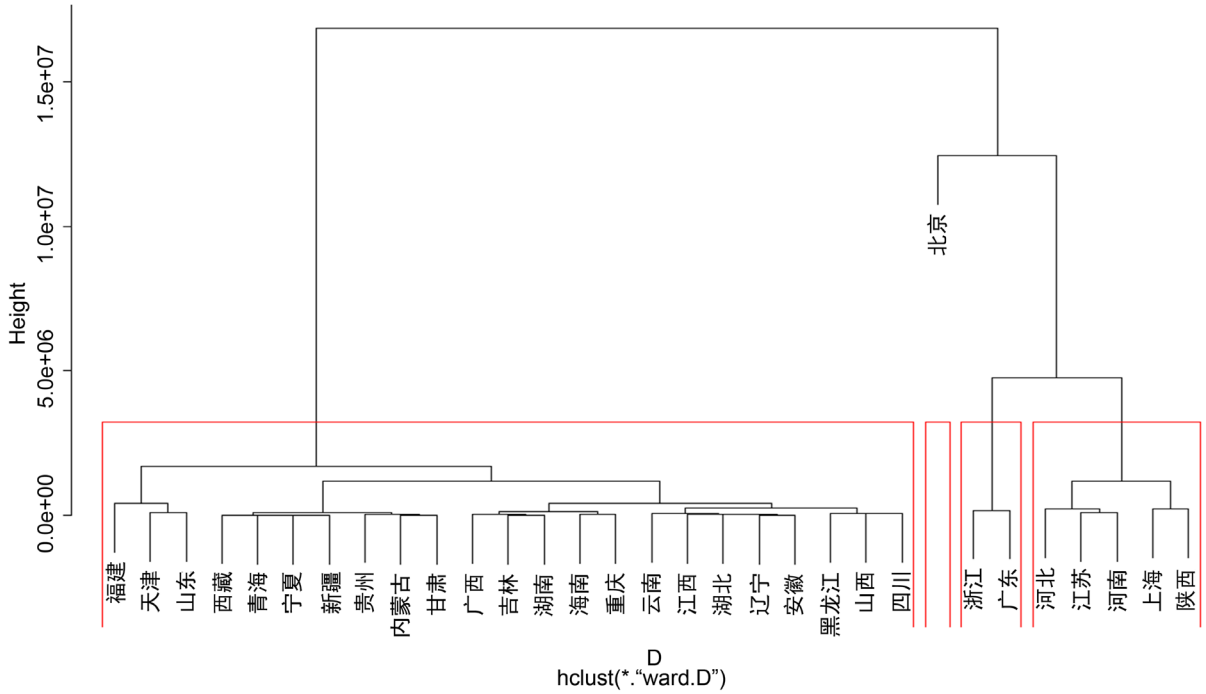

Figure 2. Cluster analysis results
图 2. 聚类分析结果

5. 总结

在多维标度图(图 1)中，由于我们在维数中选择了二维，即用二维平面可以比较直观地反映各地区的位置。在 PC 互联网中，北京表现比较出色，排名第一，紧接其后的是浙江和广东，评价结果与《中国互联网发展报告 2017》蓝皮书评估结果不谋而合。这是有其原因的，因为这三地分别立足本地实际，开展了一系列促进互联网发展的措施，由其北京作为首都，互联网应用广泛，在企业联网化，电子政务，以及网络管理队伍建设完善。广东地处粤港澳大湾区，互联网基础好，规模大，创新能力强，尤其以深圳和广州为代表。浙江作为网络建设强省，互联网基础资源雄厚，已由坐落于杭州的阿里巴巴公司为代表的互联网电子商务业务辐射至全省互联网发展的方方面面。在移动互联网方面，广东位列第一，其后是江苏、浙江和河南。广东在移动互联网方面表现优异，广东有 14 家企业入围了 2018 年中国互联网企业 100 强，腾讯和网易排名前五。江苏省移动宽带用户规模巨大，位居前列。在我国的“新四大发明中”，网购和支付宝都在浙江诞生、孕育与发展，移动互联网发展强势，蚂蚁金服、阿里巴巴在移动互联网发展中持续发力。河南在以云计算、物联网为代表的移动互联网中发展迅速。上海作为老牌强省，互联网发展均衡。陕西因为一带一路，高校科研单位云集，这在一定程度上促进了陕西互联网的发展。河北因为坐落在京津翼城市群，其互联网水平都得到长足性发展。

在聚类分析图(图 2)中，我们选择离差分析和法(Word)，类内离差平方和尽可能地小，类间离差平方和尽可能地大，具体步骤为：所有样本自成一类，每次减少一类，在类聚合的过程中，选择使方差增量最小的两类聚合，最后所有的样本聚合成一类。按照聚类图整理出聚类结果为表 2。

由聚类表(表 2)可得，互联网基础资源发挥发展状况最好的是首都北京。其次是互联网基因厚实的浙江和广东。接着是河北、江苏、河南、上海、陕西占据互联网发展新高地的省份，而剩余省份被分为一类。

Table 2. Cluster result tablet
表 2. 聚类结果表

	第一类	第二类	第三类	第四类
分四类	北京	浙江 广东	河北 江苏 河南 上海 陕西	福建，广西，吉林，江西，山西， 天津，甘肃，湖南，湖北，四川， 西藏，内蒙古，海南，辽宁，青海， 贵州，重庆，安徽，宁夏，新疆， 云南，山东，黑龙江

对比多维标度的结果和聚类分析的结果发现，两种分析方法的结果高度一致，北京处在互联网发展第一梯队，广东和浙江处在互联网基础资源发展第二梯队，江苏，河南，上海，河北，陕西处在第三梯队。分析结果对于我们了解各省市互联网发展状况有一定的帮助，为各省市之间的交流和发展，企业的投资和发展，城市经济的发展和转型提供了参考。

参考文献

[1] 王斌会. 多元分析及 R 语言建模[M]. 广州: 暨南大学出版社, 2011: 268-270.

[2] 张润楚. 多元统计分析[M]. 北京: 科学出版社, 2006: 291-294.

[3] 马慧, 魏立力. 基于多维标度和聚类的 CPI 数据结构分析[J]. 兰州文理学院学报(自然科学版), 2019, 33(3): 13-17.

[4] 赵静, 蒲越. 基于 MDS 对工业科技人才培养的研究分析[J]. 吉林化工学院学报, 2017, 34(3): 82-86.

[5] 曾薇, 赵守盈. 非计量多维尺度中的单调最小二乘回归技术[J]. 中国考试, 2011(10): 8-9.

[6] 揭水平. 多维标度法的聚类分析: 问题与解法[J]. 统计与决策, 2009, 24(11): 148-149.