Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 947-959
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141088

Hans汉斯

基于聚类和主成分分析的港口竞争力评价研究

陈涵怡[1]，金婷婷[1]，尚雨浩[1]，陈涵滢[2]

[1]南京信息工程大学数学与统计学院，江苏 南京
[2]桂林理工大学信息科学与工程学院，广西 桂林

收稿日期：2023年12月15日；录用日期：2024年1月5日；发布日期：2024年2月29日

摘　要

随着全球化进程的加快，各国之间的贸易联系愈加紧密，港口已不再仅仅是联结水路运输、铁路运输与公路运输的枢纽，更是资金、技术、信息的中转站，是城市面向世界的重要窗口。本文采用主成分分析和聚类分析相结合的方法对我国年吞吐量在1000万吨以上的13个沿海港口的港口综合竞争力进行了分析和评价。我们首先构建了港口综合竞争力评价指标体系，包含4个一级指标和13个二级指标。然后我们通过主成分分析法从13个影响指标中提炼出三个主成分作为影响港口竞争力的新指标，并得出各港口三个主成分的得分。在这基础上，我们对各港口三个主成分得分求加权平均数，最终得出各港口综合竞争力的得分和排名。我们把根据主成分分析方法得出的港口综合竞争力排名与真实排名作对比，发现两者虽有出入但相差不大。根据各港口三个主成分的得分，我们进行层次聚类和K均值聚类，通过聚类个数与类内平方和、类间平方和的关系曲线图，选定适合的聚类数目。层次聚类和K均值聚类的结果在聚成两类时相差较大，但在聚成三类时则比较相似。

关键词

主成分分析，层次聚类，港口综合竞争力，K均值聚类

Port Competitiveness Evaluation Based on Cluster and Principal Component Analysis

Hanyi Chen[1], Tingting Jin[1], Yuhao Shang[1], Hanying Chen[2]

[1]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu
[2]School of Information Science and Engineering, Guilin University of Technology, Guilin Guangxi

Received: Dec. 15th, 2023; accepted: Jan. 5th, 2024; published: Feb. 29th, 2024

文章引用: 陈涵怡, 金婷婷, 尚雨浩, 陈涵滢. 基于聚类和主成分分析的港口竞争力评价研究[J]. 运筹与模糊学, 2024, 14(1): 947-959. DOI: 10.12677/orf.2024.141088

Abstract

With the acceleration of the process of globalization, the trade links between countries have become more and more close. Ports are no longer just the hub connecting water transport, railway transport and road transport, but also a transit station for capital, technology and information, which is an important window into the world. In this paper, principal component analysis and cluster analysis are used to analyze and evaluate 13 coastal ports with an annual throughput of more than 10 million tons. First, we study the factors that influence the port competitiveness, which can be divided into four aspects: The scale of port development, the conditions of port infrastructure, the economic conditions in the hinterland of port and the development potential of port, from which we select 13 impact indicators, the evaluation index system of comprehensive competitiveness of ports is constructed. Then we use principal component analysis to extract three principal components from 13 impact indicators as a new indicator of port competitiveness, and get the scores of each port's three principal components. On this basis, we weighted the sum of the three principal component scores of each port to get the overall competitiveness of each port score and ranking. Compare the ranking of port comprehensive competitiveness based on principal component analysis with the real ranking, and find that there are differences but not much difference between them. According to the scores of the three principal components of each port, we conduct hierarchical clustering and k-means clustering, and select the appropriate number of clusters by the graph of the number of clusters and the relationship between the number of clusters and the sum of squares within and among clusters. The results of hierarchical clustering and K-means clustering are quite different when they are grouped into two groups, but similar when they are grouped into three groups.

Keywords

Principal Component Analysis, Hierarchical Clustering, Comprehensive Competitiveness of Port, K-Means Clustering

1. 引言

港口是位于江、海、河、湖沿岸，供来往船只停靠、装卸货物和上下旅客的地方，是水陆交通的结点，也是国际间贸易往来的门户。随着港口服务功能的逐渐完善，港口产业链的不断延伸，港口对区域经济的带动作用使得其重要性得到凸显，港口已不再仅仅是联结水路运输、铁路运输与公路运输的枢纽，更是资金、技术、信息的中转站，在国家发展全局中处于战略核心地位。

在港口逐渐成为国家战略性资源，国际与国内港口之间的竞争愈演愈烈的今天，通过研究港口竞争力，并对港口的港口竞争力做出评价，有助于港口找到自身发展中存在的缺陷与不足，突破发展瓶颈，从而在竞争中占据更多优势[1]。

2. 理论介绍

2.1. 主成分分析法

假设研究的数据集有P个指标，记为P维随机变量$X=\left(x_1,x_2,\cdots,x_p\right)^{\mathrm{T}}$，该变量的协方差矩阵记为$\Sigma$。

主成分分析就是通过线性变换，将原始的 P 维随机变量 X 转变为新的 P 维随机变量 Y，

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}=\begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \tag{1}$$

$$Y=\Gamma X,\ \Gamma=\begin{pmatrix} \gamma_1^{\mathrm{T}} \\ \gamma_2^{\mathrm{T}} \\ \vdots \\ \gamma_p^{\mathrm{T}} \end{pmatrix} \tag{2}$$

其中 Γ 为正交矩阵。

主成分分析的中心思想是让数据集的维度减少，但其中的信息要尽可能多地保留。信息可以通过方差来衡量，方差越大，信息就越多。

$$\mathrm{Var}(y_1)=\mathrm{cov}(y_1,y_1)=\left[\mathrm{cov}(Y)\right]_{11}=\left[\mathrm{cov}(\Gamma X)\right]_{11}=\left[\Gamma\Sigma\Gamma^{\mathrm{T}}\right]_{11}=\gamma_1^{\mathrm{T}}\Sigma\gamma_1 \tag{3}$$

第一主成分为 y_1，使其方差最大化就是求解下面这个优化问题：

$$\max_{\gamma_1}\ \gamma_1^{\mathrm{T}}\Sigma\gamma_1,\ \ \text{s.t.}\ \gamma_1^{\mathrm{T}}\gamma_1=1 \tag{4}$$

设 $L(\gamma_1,\lambda)=\gamma_1^{\mathrm{T}}\Sigma\gamma_1+\lambda\left(1-\gamma_1^{\mathrm{T}}\gamma_1\right)$，两边同时对 γ_1 求偏导：

$$\frac{\partial L(\gamma_1,\lambda)}{\partial\gamma_1}=2\Sigma\gamma_1-2\lambda\gamma_1=0 \tag{5}$$

即：

$$\Sigma\gamma_1=\lambda\gamma_1 \tag{6}$$

由此可知 λ 为 Σ 的一个特征值，γ_1 为这个特征值对应的特征向量。

上式两边左乘 γ_1^{T}，得：

$$\gamma^{\mathrm{T}}\Sigma\gamma_1=\lambda\gamma^{\mathrm{T}}\gamma_1=\lambda \tag{7}$$

因此，

$$\max\ \gamma^{\mathrm{T}}\Sigma\gamma_1=\max\ \lambda \tag{8}$$

所以 λ 为 Σ 的最大特征值，γ_1 为对应的特征向量。

使第二主成分 y_2 的方差最大化：

$$\max_{\gamma_2}\ \gamma_2^{\mathrm{T}}\Sigma\gamma_2,\ \ \text{s.t.}\ \gamma_2^{\mathrm{T}}\gamma_2=1,\ \gamma_2^{\mathrm{T}}\gamma_1=0,\ \gamma_1^{\mathrm{T}}\gamma_1=1 \tag{9}$$

设 $L(\gamma_2,\lambda_1,\lambda_2,\lambda_3)=\gamma_2^{\mathrm{T}}\Sigma\gamma_2+\lambda_1\left(1-\gamma_2^{\mathrm{T}}\gamma_2\right)+\lambda_2\gamma_2^{\mathrm{T}}\gamma_1+\lambda_3\left(1-\gamma_1^{\mathrm{T}}\gamma_1\right)$，两边同时对 γ_2 求偏导，得：

$$\frac{\partial L(\gamma_2,\lambda_1,\lambda_2,\lambda_3)}{\partial\gamma_2}=2\Sigma\gamma_2-2\lambda_1\gamma_2+\lambda_2\gamma_1=0 \tag{10}$$

(10)式两边左乘 γ_2^{T}，化简得：

$$\gamma_2^{\mathrm{T}}\Sigma\gamma_2=\lambda_1,\ \ \max\ \gamma_2^{\mathrm{T}}\Sigma\gamma_2=\max\ \lambda_1 \tag{11}$$

(10)式两边左乘 γ_1^{T}，得：

$$2\gamma_1^{\mathrm{T}}\Sigma\gamma_2+\lambda_2=0 \tag{12}$$

而 $\gamma_1^{\mathrm{T}}\Sigma=\left(\Sigma\gamma_1\right)^{\mathrm{T}}=\left(\lambda\gamma_1\right)^{\mathrm{T}}$，因此上式又可以写成：

$$2\lambda\gamma_1^{\mathrm{T}}\gamma_2+\lambda_2=0 \tag{13}$$

得出 $\lambda_2=0$。把 $\lambda_2=0$ 代回上式，得：

$$\Sigma\gamma_2=\lambda_1\gamma_2 \tag{14}$$

由以上可知 λ_1 为 Σ 的第二大特征值，γ_2 为对应的特征向量。

以此类推，可得第三个，第四个乃至第 P 个主成分，这些主成分之间线性无关，因为

$$\mathrm{cov}\left(y_i,y_j\right)=\gamma_i^{\mathrm{T}}\Sigma\gamma_j=\mu\gamma_i^{\mathrm{T}}\gamma_j=0 \tag{15}$$

在这 P 个主成分中选取前 r 个主成分代替原来 P 个指标，就实现了主成分方法的降维，其中 Σ 的特征值反映了 X 的信息量。

2.2. 聚类分析法

聚类一般可以分为对样本的聚类和对变量的聚类。在聚类之前，首先需要度量样本之间的距离和类间距离，常用的度量样本之间距离的方法有三种，假设有 n 个样本和 p 个变量：

欧式距离：

$$d_{ij}=\sqrt{\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2} \tag{16}$$

绝对距离：

$$d_{ij}=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right| \tag{17}$$

马氏距离：

$$d_{ij}=\sqrt{\left(x_i-x_j\right)^{\mathrm{T}}S^{-1}\left(x_i-x_j\right)} \tag{18}$$

其中，S 是由 $x_1,x_2,\cdots,x_n$ 得到的协方差矩阵 $S=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(x_i-\overline{x}\right)^{\mathrm{T}}$，$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$。

常用的度量类间距离的方法有四种：

最短距离法：

$$D\left(G_1,G_2\right)=\min_{\substack{x_i\in G_1\\ y_j\in G_2}}\left\{d\left(x_i-y_j\right)\right\} \tag{19}$$

它的直观意义为两个类中最近两点距离。

最长距离法：

$$D\left(G_1,G_2\right)=\max_{\substack{x_i\in G_1\\ y_j\in G_2}}\left\{d\left(x_i-y_j\right)\right\} \tag{20}$$

它的直观意义为两个类中最远两点间的距离。

重心法：

$$D\left(G_1,G_2\right)=d\left(\overline{x},\overline{y}\right) \tag{21}$$

式中：$\bar{x},\bar{y}$ 分别为 G_1,G_2 的重心。

类平均法：

$$D(G_1,G_2)=\frac{1}{n_1n_2}\sum_{x_i\in G_1}\sum_{x_j\in G_2}d(x_i,x_j) \tag{22}$$

它等于 G_1,G_2 中两样本点距离的平均，n_1,n_2 分别为 G_1,G_2 中的样本点个数。

主要的聚类方法有两种，分别是 K-means 和层次聚类。

K-means：假设要将 n 个样本分成 k 类，先在这 n 个样本中随机选取 k 个样本作为起始的聚类中心[2]，计算其余样本与这 k 个聚类中心的欧氏距离，将样本归于距离最小的那一类，然后重新计算聚类中心以及样本点与新的聚类中心的距离并再次进行归类……当聚类中心的位置不再变化时，迭代停止，此时的聚类结果就是最终的结果。

层次聚类：开始时每个样本自成一类，然后每次将类间距离最近的两类合并，合并后重新计算新类与其他类之间的距离，然后再合并……直到所有的样本都被覆盖，最终形成一棵有层次的聚类树。

3. 实证分析

本章主要利用主成分分析和聚类分析的相关理论，对 2019 年我国年吞吐量在 1000 万吨以上的 13 个沿海港口的港口综合竞争力进行分析和评价[3]，港口综合竞争力的指标体系如表 1 所示，数据主要来源于各地区统计年鉴和国家统计局的最新数据。

Table 1. Port comprehensive competitiveness evaluation index system
表 1. 港口综合竞争力评价指标体系

一级指标	二级指标	单位
港口发展规模	港口货物吞吐量	万吨
	港口集装箱吞吐量	万 TEU
港口基础设施条件	港口码头长度	米
	港口泊位数量	个
	港口万吨级泊位数量	个
港口腹地经济条件	港口所在城市 GDP	亿元
	港口所在城市第三产业总产值	亿元
	港口所在城市外贸进出口总额	亿元
	港口所在城市外商投资实际到位总额	万美元
港口发展潜力	货物吞吐量增长率	%
	港口所在城市 GDP 增长率	%
	港口所在城市外贸进出口总额增长率	%
	港口所在城市外商投资实际到位总额增长率	%

3.1. 港口综合竞争力主成分分析

首先对数据进行标准化[4]处理，公式如下：

$$y_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},\ i=1,2,\cdots,13;\ j=1,2,\cdots,13 \tag{23}$$

其中，x_{ij} 表示第 i 个港口第 j 个指标的原始取值，y_{ij} 表示标准化后第 i 个港口第 j 个指标的数值，$\overline{x}_j$ 表示第 j 个指标的平均值，计算公式如下：

$$\overline{x}_j = \frac{\sum_{i=1}^{13} x_{ij}}{13},\ j = 1,2,\cdots,13 \tag{24}$$

s_j 表示第 j 个指标的标准差：

$$s_j = \sqrt{\frac{\sum_{i=1}^{13}\left(x_{ij} - \overline{x}_j\right)^2}{12}},\ j = 1,2,\cdots,13 \tag{25}$$

计算标准化后数据的相关系数矩阵 R：

$$R = \frac{Y^{\mathrm{T}}Y}{12},\ Y = \left(y_{ij}\right),\ i = 1,2,\cdots,13;\ j = 1,2,\cdots,13 \tag{26}$$

将相关系数矩阵可视化，如图 1 所示。下三角中第 i 行第 j 列的数字表示第 i 个指标和第 j 个指标的相关系数，相关系数越大，两者的相关性越强，对应上三角中第 j 行第 i 列格子圆点的颜色就越深。由图 1 可知，这 13 个指标之间相关性较强。

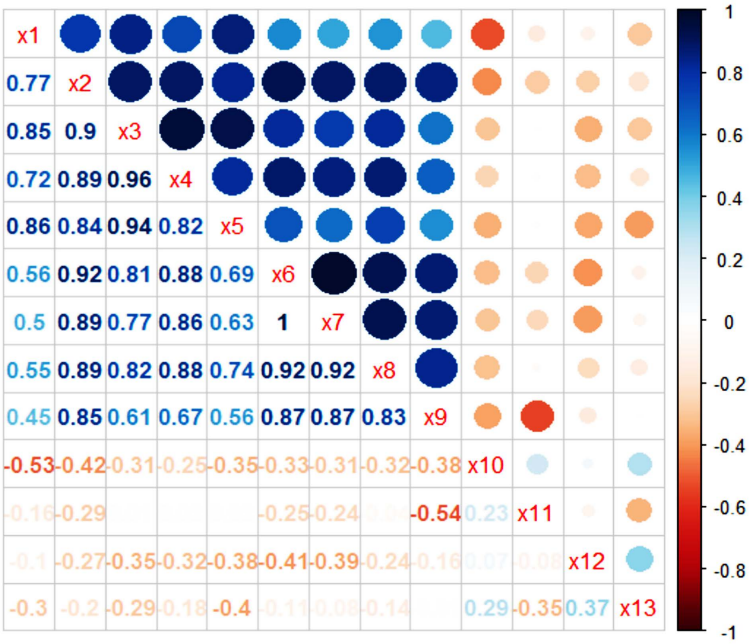

Figure 1. Diagram of the correlation
图 1. 相关图

计算相关系数矩阵的特征值和特征向量：

$$\left|R - \lambda E\right| = 0 \tag{27}$$

得出的特征值按从大到小的顺序排列，依次为 $\lambda_1 = 7.753$，$\lambda_2 = 1.818$，$\lambda_3 = 1.212$，……，$\lambda_{13} \approx 0$。根据对应的特征向量 $\alpha_1, \alpha_2, \cdots, \alpha_{13}$，构造主成分的表达式：

$$F_i = \alpha_i^{\mathrm{T}} X,\ i = 1,2,\cdots,13 \tag{28}$$

式中：$X=(X_1,X_2,\cdots,X_{13})^{\mathrm{T}}$，$F_1$为第 1 主成分，$F_2$为第 2 主成分，……，$F_{13}$为第 13 主成分。

此时还没有实现降维，我们需要从 13 个主成分中选取前$r(r\le 13)$个主成分作为新的综合性变量，前 r 个主成分的方差贡献率之和为累积方差贡献率[5]，要求大于等于 85%。第 i 个主成分的方差贡献率为$\lambda_i \Big/ \sum_{j=1}^{13}\lambda_j$，前 r 个主成分的累积方差贡献率为$\sum_{i=1}^{r}\lambda_i \Big/ \sum_{j=1}^{13}\lambda_j$。

Table 2. Variance contribution rate of principal component and cumulative variance contribution rate
表 2. 主成分的方差贡献率和累积方差贡献率

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
Proportion of Variance	0.5964	0.1398	0.0932	0.0737	0.047	0.032	0.0113
Cumulative Proportion	0.5964	0.7362	0.8294	0.9031	0.9501	0.9821	0.9934
	Comp8	Comp9	Comp10	Comp11	Comp12	Comp13	
Proportion of Variance	0.0035	0.0021	0.0011	0.0005	4.2e−05	3.4e−17	
Cumulative Proportion	0.9969	0.999	1	1	1	1	

由表 2 可知，当选取三个主成分时，方差的累积贡献率已经超过了 80%，当选取六个主成分时，方差的累积贡献率达到了 98%，再结合碎石图来决定主成分的个数。

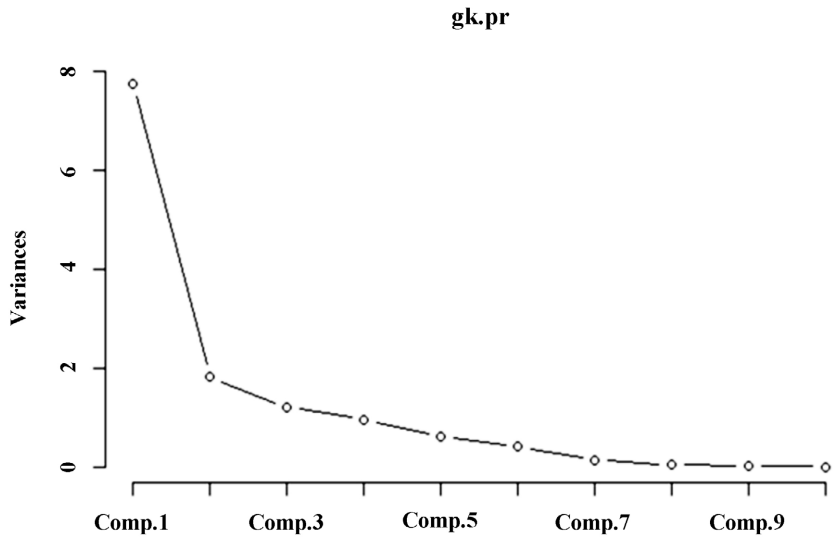

Figure 2. Diagram of the scree
图 2. 碎石图

从图 2 中我们可以很清晰地看到，转折点出现在第 2 个因子的位置，当主成分的个数大于 3 时，对应公共因子的特征值小于 1，其影响已经很微弱了，因此选取 3 个主成分较为合适，得到的 3 个主成分分别是

$$\begin{aligned}z_1 &= 0.278x_1+0.351x_2+0.336x_3+0.335x_4+0.313x_5+0.339x_6\\ &\quad +0.329x_7+0.33x_8+0.297x_9-0.159x_{10}-0.134x_{12}\\ z_2 &= 0.127x_1+0.162x_3+0.231x_5-0.126x_6-0.149x_7-0.342x_9\\ &\quad +0.588x_{11}-0.29x_{12}-0.552x_{13}\\ z_3 &= 0.419x_1-0.154x_4+0.106x_5-0.193x_6-0.231x_7-0.178x_8\\ &\quad -0.653x_{10}-0.268x_{11}+0.34x_{12}-0.225x_{13}\end{aligned}$$

将标准化后的港口数据带入主成分的表达式，可得各港口前 3 个主成分的得分，如表 3 所示，其中 C_1、C_2、C_3 分别表示第一、第二、第三主成分得分。

Table 3. Port principal component score
表 3. 港口主成分得分情况

港口	C_1	C_2	C_3
大连港	−0.4237415	1.5186033	−0.742662878
营口港	−2.4235852	0.3062559	−0.856675213
秦皇岛	−2.1778939	0.2752753	0.110616010
天津港	0.9385194	0.5017409	−0.248874813
烟台港	−0.7091304	0.9684601	−0.326861177
青岛港	0.8453752	−2.9530171	2.047027367
日照港	−1.5392386	0.5937900	1.437230347
上海港	7.1792928	−0.5915000	−1.601133921
连云港	−1.6640420	0.3602101	0.757527321
宁波港	3.1907179	1.9754907	1.590132291
汕头港	−2.7995006	0.2450488	−0.976107138
广州港	2.3519248	−0.7851723	−0.002451952
湛江港	−2.7686978	−2.4151856	−1.187766244

以三个主成分对应的特征值占比为权，最终得出的主成分得分加权就是各港口的综合得分。主成分得分的权重如表 4 所示：

Table 4. The weight of the principal component score
表 4. 主成分得分的权重

	第一主成分	第二主成分	第三主成分
特征值	7.753119	1.817703	1.211691
特征值占比	0.5963938	0.1398233	0.0932070

港口综合得分的计算公式如下：

$$C=\frac{0.5963938C_1+0.1398233C_2+0.0932070C_3}{0.5963938+0.1398233+0.0932070} \tag{29}$$

其中，C 为综合得分，C_1 为第一主成分的得分[6]，C_2 为第二主成分的得分，C_3 为第三主成分的得分。十三个港口的综合得分及排名如表 5 所示：

Table 5. Overall score and ranking of ports
表 5. 港口的综合得分及排名

港口	综合得分	排名	真实排名
大连港	−0.1321424	6	6
营口港	−1.7873096	11	8
秦皇岛	−1.5071691	10	12

续表

天津港	0.7314537	4	4
烟台港	−0.3833666	7	9
青岛港	0.3400833	5	3
日照港	−0.8451729	8	10
上海港	4.8825967	1	2
连云港	−1.0506709	9	7
宁波港	2.8059897	2	1
汕头港	−2.0813494	12	13
广州港	1.5585024	3	5
湛江港	−2.5314450	13	11

在根据主成分分析得出的 13 个沿海港口综合竞争力的排名中，上海港位列第一，宁波——舟山港则屈居第二，而真实的排名两者却正好相反。宁波——舟山港的货物吞吐量长年占据国内榜首，远远大于上海港，但上海港所在地上海市经济实力雄厚，交通发达，贸易往来频繁，这对于上海港的发展是极其有利的，也使得上海港即便在某些方面不如宁波——舟山港，但预测的排名仍然在宁波——舟山港之上。

天津港和大连港的预测排名和真实排名一致，但青岛港和广州港的排名两者正好调换了位置。无论从货物吞吐量、泊位数量、国民生产总值还是外贸进出口总额，广州港都力压青岛港，但青岛港的实际利用外资额、外贸进出口总额增长率、实际利用外资额都远远大于广州港，2019 年广州港的外贸进出口总额增长率为−2.4%，而青岛港的外贸进出口总额增长率为 11.4%，这些都进一步说明了青岛港这几年发展的势头要明显强于广州港，发展潜力巨大，因此青岛港的预测排名要优于实际排名。

总体来说，根据主成分分析得出的港口综合竞争力排名与真实排名相差不大，误差在前后一名到两名内浮动，结果的可信度较高。

3.2. 港口综合竞争力聚类分析

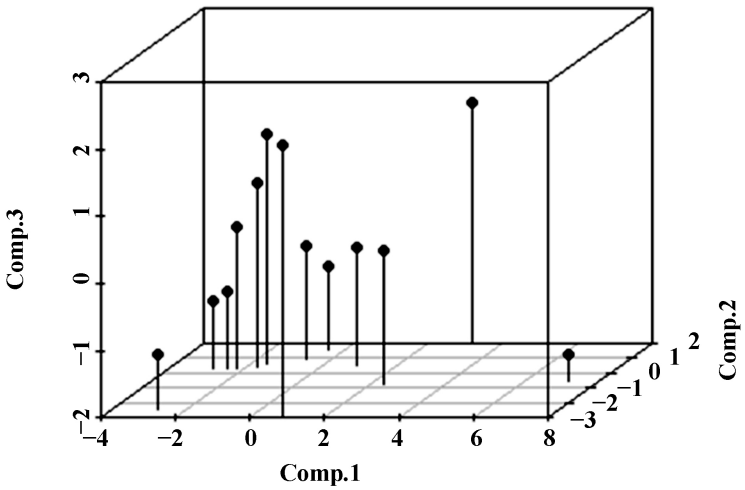

Figure 3. Three-dimensional diagram of port principal component score
图 3. 港口主成分得分的三维立体图

图 3 反映了这 13 个港口主成分的得分情况。在已经对变量进行主成分降维的前提下，我们对这 13 个港口进行聚类。

3.2.1. 层次聚类

层次聚类不需要预先确定聚类个数，但需要确定在一个聚类分析中类的最佳数目。一个好的聚类结果，其组内方差必须尽可能的小，而组间方差需要尽可能的大。R 语言的 NbClust 包能够帮助我们确定一个聚类分析中类的最佳数目，结果如表 6 所示：

Table 6. The selection of the number of hierarchical clustering clusters
表 6. 层次聚类聚类数目的选择

聚类个数	赞同数
2	6
3	6
4	1
6	2
10	2
11	7

聚类个数为 2、3 和 11 时，赞同数较多。由于我们的样本港口数只有 13，因此聚类个数为 2 或 3 比较适宜。

当聚类个数为 2 时，结果如图 4 所示：

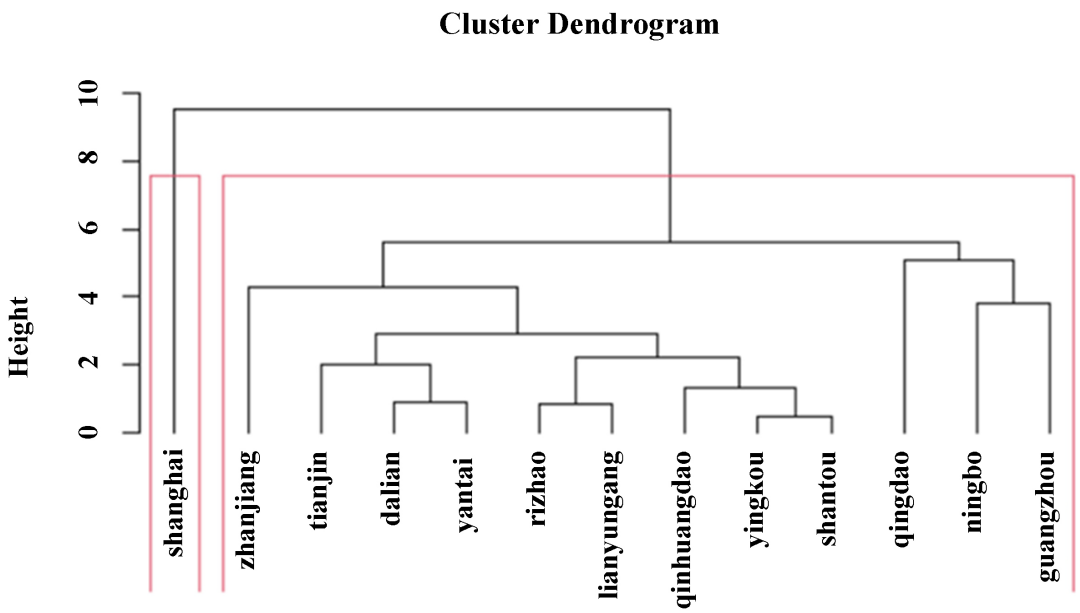

Figure 4. The cluster graph when the number of clusters is 2
图 4. 聚类个数为 2 时的聚类图

当聚类个数为 3 时，结果如图 5 所示：

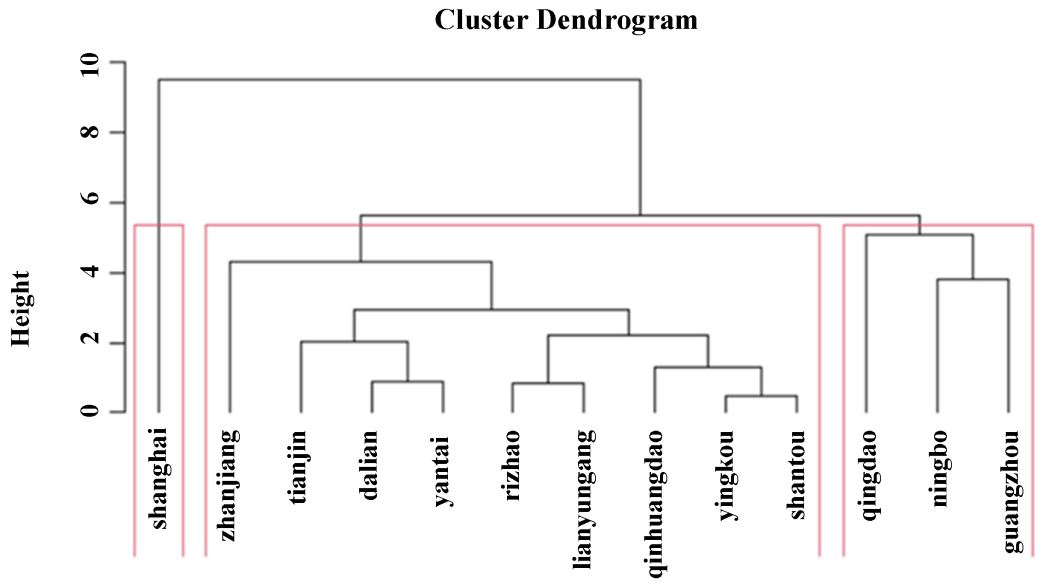

Figure 5. The cluster graph when the number of clusters is 3
图 5. 聚类个数为 3 时的聚类图

3.2.2. 均值聚类

均值聚类需要事先确定所要提取的聚类个数，我们同样可以调用 NbClust 包来帮助我们确定类的数目，结果如图 6 所示，其中聚类个数为 2 和 3 时，赞同数较多。

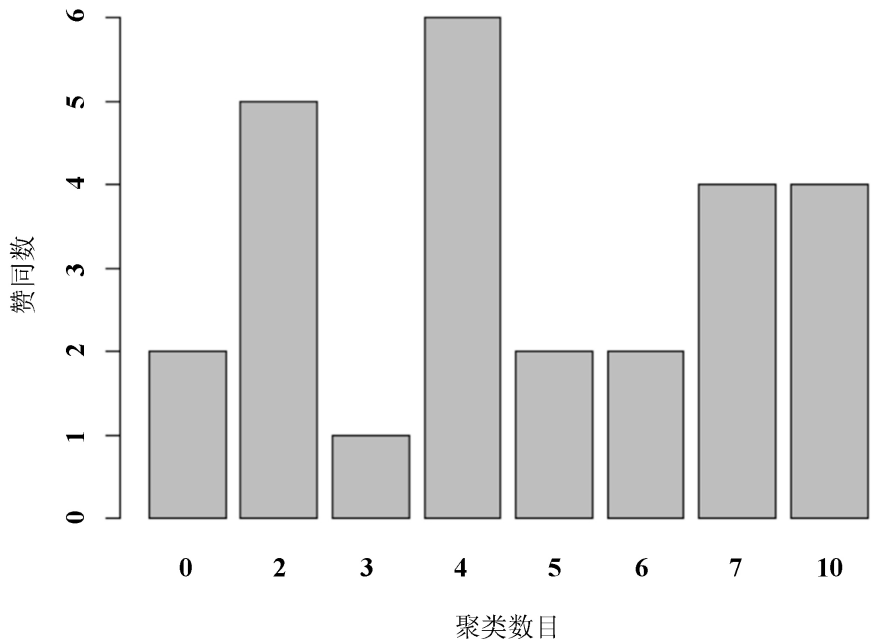

Figure 6. K-means the selection of the number of clusters
图 6. K-means 聚类数目的选择

也可以参考聚类个数与类内平方、类间平方和的关系曲线图，观察图中曲线的走势变化选择适当的聚类数量。

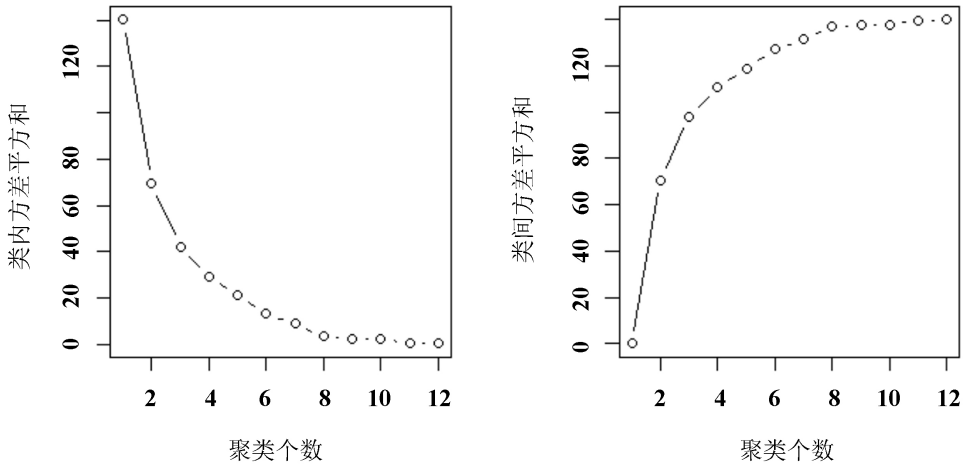

Figure 7. Graph of the relationship between the number of clusters and the squares within and between classes
图 7. 聚类个数与类内平方、类间平方的关系曲线图

图 7 中左表示随着聚类个数的增加，组内方差平方和有一个明显的下降趋势，但在三类之后，下降的趋势减弱。同样地随着聚类个数的增加，组间方差平方和也随之上升，但在三类之后上升的趋势减缓。结合调用 NbClust 包后所得结果，我们选择 2、3、4 为聚类的个数。

当聚类个数为 2 时，结果如表 7 所示：

Table 7. The result of port clustering when cluster is type 2
表 7. 聚为 2 类时的港口聚类结果

类别	港口
第一类	大连港、营口港、秦皇岛港、烟台港、日照港、连云港、汕头港、湛江港
第二类	天津港、青岛港、上海港、宁波－舟山港、广州港

当聚类个数为 3 时，结果如表 8 所示：

Table 8. The result of port clustering when cluster is type 3
表 8. 聚为 3 类时的港口聚类结果

类别	港口
第一类	天津港、青岛港、宁波－舟山港、广州港
第二类	上海港
第三类	大连港、营口港、秦皇岛港、烟台港、日照港、连云港、汕头港、湛江港

当聚类个数为 4 时，结果如表 9 所示：

Table 9. The result of port clustering when cluster is type 4
表 9. 聚为 4 类时的港口聚类结果

类别	港口
第一类	大连港、营口港、秦皇岛港、烟台港、日照港、连云港、汕头港、湛江港
第二类	青岛港
第三类	上海港
第四类	天津港、宁波－舟山港、广州港

在聚类分析中，层次聚类的结果和均值聚类的结果截然不同。当把 13 个港口分成两类时，层次聚类把上海港归为一类，而其他 12 个港口归为一类；均值聚类则把天津港、上海港、宁波舟山港、青岛港和广州港归为一类，观察可知，这五个港口就是主成分分析中综合得分排名前五的港口。虽然上海港在港口发展规模、港口基础设施建设方面一直处于比上不足，比下有余的状态，但上海港的 GDP、第三产业总产值、外贸进出口总额和其他港口相比简直是一骑绝尘，所以说上海港是“全能发展型选手”，把它另归为一类也无可厚非。当把 13 个港口分为三类时，层次聚类和均值聚类的结果非常相似，只是层次聚类将天津港从宁波舟山港、青岛港和广州港中踢了出去。天津港和青岛港在其他指标上差别不大，但青岛港的实际利用外资额、外贸进出口总额增长率远远大于天津港，所以投资对于港口发展来说是至关重要的。

4. 结论

本文对我国年吞吐量在 1000 万吨以上的 13 个沿海港口的港口竞争力进行了分析和评价，得到的排名从前往后依次为：上海港、宁波——舟山港、广州港、天津港、青岛港、大连港、烟台港、日照港、连云港港、秦皇岛港、营口港、汕头港和湛江港。这个预测的排名和真实的排名相差不大，误差在三名以内，说明这种运用主成分分析法对港口进行排名的方法结果可信度较高。

由各港口主成分的得分，我们进行层次聚类和 K 均值聚类。层次聚类和 K 均值聚类的结果在聚成两类时相差较大，但在聚成三类时则比较相似。

参考文献

[1] 施桦. 舟山港域竞争力评价[D]: [硕士学位论文]. 舟山: 浙江海洋大学, 2017.

[2] 刘翠玲, 王少敏, 吴静珠, 等. 基于太赫兹时域透射成像技术的葵花籽内部品质无损检测研究[J]. 光谱学与光谱分析, 2020, 40(11): 3384-3389.

[3] 陈辛. 基于主成分分析法的我国沿海港口竞争力评价研究[D]: [硕士学位论文]. 杭州: 浙江工业大学, 2011.

[4] 任东海. 主成分分析和聚类分析在高职学生成绩综合评价中的应用[J]. 计算机时代, 2023(11): 64-67+70.

[5] 韩智强, 左新黛, 周勇军, 等. 基于主成分-逐步回归的大跨弯连续刚构桥冲击系数计算[J]. 公路交通科技, 2022, 39(1): 72-80.

[6] 吴勇, 徐亚琼, 曾俞森, 等. 丘陵区耕地细碎化与种植多样性的空间相关性分析——以武胜县鸣钟乡为例[J]. 西华师范大学学报(自然科学版), 2023, 44(3): 305-310.