Statistics and Application 统计学与应用, 2020, 9(5), 891-899
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.95093

Hans汉斯

基于多指标面板数据的中国经济发展省际差异聚类分析

王　超，赵明清

山东科技大学，山东　青岛
Email: 756163501@qq.com

收稿日期：2020年10月7日；录用日期：2020年10月22日；发布日期：2020年10月29日

摘　要

本文考虑面板数据时间维度上的“绝对量”、“波动”、“趋势”和“峰度”4个动态特征，用主成分分析法进行二次特征提取，采用熵值法对特征进行客观赋权，对我国经济发展的省际差异进行了更有效的聚类分析：东部沿海省份的经济发展状况最好，其次是中部内陆省份和东北三省，最后是西部和南部边疆省份，并对形成差异的原因进行了分析，针对以上分析还提出了相关政策建议。

关键词

面板数据，特征提取，熵值法，聚类分析

Cluster Analysis of Inter-Provincial Differences in China’s Economic Development Based on Multi-Index Panel Data

Chao Wang, Mingqing Zhao

Shandong University of Science and Technology, Qingdao Shandong
Email: 756163501@qq.com

Received: Oct. 7th, 2020; accepted: Oct. 22nd, 2020; published: Oct. 29th, 2020

Abstract

This paper considers the four dynamic characteristics of “absolute quantity”, “volatility”, “trend”

文章引用: 王超, 赵明清. 基于多指标面板数据的中国经济发展省际差异聚类分析[J]. 统计学与应用, 2020, 9(5): 891-899. DOI: 10.12677/sa.2020.95093

and "kurtosis" in the time dimension of panel data, uses principal component analysis to extract secondary features, and uses entropy method to objectively assign features. A more effective cluster analysis of the inter-provincial differences in my country's economic development was carried out. The result is that the eastern coastal provinces have the best economic development conditions, followed by the central inland provinces and the three northeastern provinces, and finally the western and southern border provinces. The reasons are analyzed, and relevant policy recommendations are put forward in response to the above analysis.

Keywords

Panel Data, Feature Extraction, Entropy Method, Cluster Analysis

1. 引言

近几十年来，中国经济得到了迅猛增长，综合国力也大幅提升。但随着经济的增长，区域之间的经济发展不平衡现象也开始出现，且越来越严重，同时贫富差距也在被不断拉大，这不利于资源的合理配置，也不利于经济长期、稳定、健康地发展。区域经济发展不平衡现象，既制约了我国经济水平进一步发展，也不利于社会的稳定与繁荣。因此，分析中国区域经济发展的差异及其形成原因对于解决区域经济发展不平衡现象具有重要积极意义。

关于区域经济差异的研究，乔慧[1]通过因子分析得到了总量因子、均量因子和价格因子，据此对各地区进行了聚类分析，并给出了一些合理化建议。杨英和李海萍等[2]提取了能够综合解释社会经济发展情况的 3 个主因子成分，并据此对各省市自治区的社会竞争力进行分类、比较和综合评价。朱莉莉[3]建立了区域经济的综合评价指标体系，用其提出的面板数据的有序聚类方法进行了聚类分析，并提出了相关建议。

面板数据具有截面数据和时间序列数据的特性，在以往研究中，主要从降维层面将面板数据的三维信息通过某种技术手段降为二维信息。朱建平和陈民恳[4]从面板数据描述层面出发，构造面板数据相似性指标，并提出面板数据聚类的单指标聚类方法。单指标面板数据自身具有简化面板数据的效果，且在现实中并不多见，因此该方法适用性较窄。郑兵云[5]根据聚类分析原理，重新构造了多指标面板数据的距离函数和离差平方和函数，并进行多指标面板数据的聚类分析。李因果和何晓群[6]从面板数据时序特征和截面特征出发，综合考虑面板数据"绝对指标"、"增量指标"和"时序波动"特征，提出了一套较为合理的面板数据聚类算法。刘汉丽和裴韬等[7]提取时间序列的趋势信息，并结合统计特征计算，将时间序列的趋势信息与几种统计特征组合在一起，构成 SOM 神经网络的输入向量，对时间序列进行聚类分析。党耀国和侯荻青[8]从特征提取的角度，将每个个体在时间维度上的不同指标的统计特征进行提取，以此来降低时间维度，并将所有不同指标的动态特征全部看作截面数据的指标维度，用传统的动态聚类方法来聚类。戴大洋和邓光明[9]对用面板数据提取的特征进行二次提取，消除了特征间的信息重叠，并通过对房地产数据的实证分析表明了二次提取可以优化聚类结果。

本文在建立省域经济发展评价指标体系的基础上，运用主成分分析对不同指标的"绝对量"、"波动"、"趋势"和"峰度" 4 个特征分别进行二次特征提取，对每个特征分别计算综合得分，再运用熵

值法计算各特征综合得分的权重，将赋权后的数据进行系统聚类和差异分析，并由此提出相关政策建议。本文在分析经济发展省际差异时所用方法更有效，所得结论更符合实际。

2. 面板数据及其特征提取

2.1. 多指标面板数据

多指标面板数据从横截面上看，是由若干个体在某一时刻构成的截面观测值，从纵剖面上看则是一组时间序列，其包含的每个数据点可用三下标变量表示：$X_i^k(t), i=1,2,\cdots,N, k=1,2,\cdots,P, t=1,2,\cdots,T$。其中，$N$ 表示面板数据中含有的个体数，P 表示指标的个数，T 表示时间序列的最大长度。

记个体 i 的第 k 个指标在 T 时期内的均值为

$$\mu_i^k = \frac{\sum_{t=1}^{T} X_i^k(t)}{T} \tag{1}$$

标准差为

$$\sigma_i^k = \left(\frac{\sum_{t=1}^{T} \left(X_i^k(t) - \bar{X}_i^k \right)^2}{N} \right)^{\frac{1}{2}} \tag{2}$$

由于面板数据各指标量纲或数量级不同会对聚类结果造成一定影响，故对 X_i^k 进行均值化的标准化处理，标准化公式为

$$Z_i^k(t) = \frac{X_i^k(t)}{\bar{X}^k} \tag{3}$$

其中，$\bar{X}^k = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_i^k(t)$。标准化后，每个指标的均值都为 1，方差为

$$Var\left(Z^k\right) = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{X_i^k(t)}{\bar{X}^k} - 1\right)^2 = \frac{Var\left(X^k\right)}{\left(\bar{X}^k\right)^2} = \left(\frac{\sigma^k}{\bar{X}^k}\right)^2 \tag{4}$$

这样标准化后各指标的方差是各指标变异系数的平方，不仅消除了量纲和数量级的影响，又保留了原指标的变异信息。为了方便表示，下文用 $X_i^k(t)$ 来表示标准化后得到的 $Z_i^k(t)$。

2.2. 特征提取

参照文献[3]，对于面板数据集 $\left\{X_i^k(t)\right\}$，提取“绝对量”、“波动”、“趋势”和“峰度”4 个特征，有关定义如下。

定义 1 样本 i 的第 k 个指标的全时“绝对量”特征，记为

$$AQF\left(F_i^k\right) = \frac{\sum_{t=1}^{T} X_i^k(t)}{T} \tag{5}$$

$AQF\left(F_i^k\right)$ 实际上是样本 i 第 k 个指标在总时期 T 内的均值，该特征反映了样本个体在整个分析时域内发展的绝对水平。

定义 2 样本 i 的第 k 个指标的全时“波动”特征，记为

$$VF\left(F_i^k\right)=\sqrt{\frac{\sum_{t=1}^{T}\left(X_i^k(t)-\bar{X}_i^k\right)^2}{T-1}} \tag{6}$$

其中，$\bar{X}_i^k=\frac{1}{T}\sum_{t=1}^{T}X_i^k(t)$，表示样本 i 第 k 个指标在总时期 T 内变量的均值；$VF\left(F_i^k\right)$ 表征了个体 i 在整个时期 T 内指标值随时间变化的波动程度。显然，如果两个样本相似性较大，则其波动程度应该相差不多，$VF\left(F_i^k\right)$ 的值也应比较接近。

定义 3 样本 i 的第 k 个指标的全时“趋势”特征，记为

$$TF\left(F_i^k\right)=\frac{\sum_{t=1}^{T}\left(X_i^k(t)-\bar{X}_i^k\right)\left(t-\frac{T}{2}\right)}{\sum_{t=1}^{T}\left(t-\frac{T}{2}\right)^2} \tag{7}$$

$TF\left(F_i^k\right)$ 用来描述指标值的长期变化情况(单调特征)，若两指标随着时间都呈同向变化，这种变化越协调，则两者越相似。本文对样本 i 的第 k 个指标所在序列 $X_i^k=\left[X_i^k(1),\cdots,X_i^k(t),\cdots,X_i^k(T)\right]$ 建立关于时间 t 的一元回归模型，即 $X_i^k=\alpha+\beta t$，采用最小二乘法估计线性方程的参数，用得到的斜率(系数 β)来衡量面板数据的趋势特征。

定义 4 样本 i 的第 k 个指标的全时“峰度”特征，记为

$$KCF\left(F_i^k\right)=\frac{\sum_{t=1}^{T}\left(X_i^k(t)-\bar{X}_i^k\right)^4}{T\left(\sigma_i^k\right)^4}-3 \tag{8}$$

该特征表征了个体 i 在整个时期 T 内指标值的集中程度或分布曲线的尖峭程度，若 $KCF\left(F_i^k\right)$ 大于 0，表示指标值的分布比正态分布更集中在平均值周围；若 $KCF\left(F_i^k\right)$ 小于 0，表示指标值的分布比正态分布更分散。

2.3. 二次特征提取

定义 5 [9] $F_1,F_2,\cdots,F_p$ 为 p 维指标向量 $\left(AQF\left(F^1\right),AQF\left(F^2\right),\cdots,AQF\left(F^p\right)\right)$ 提取的主成分，记 $\alpha_k\left(k=1,2,\cdots,p\right)$ 为主成分 F_k 的方差贡献率，则主成分降维后“绝对量”特征 $AQF\left(F^k\right)$ 的综合得分为

$$F\_AQF(F)=\sum_{k}^{p}\alpha_k F_k \tag{9}$$

同理，可分别定义“波动”、“趋势”和“峰度”3 个特征的综合得分为 $F\_VF(F),F\_TF(F),F\_KCF(F)$。

为避免前几个主成分计算综合得分时因信息损失而影响聚类效果，此处取所有主成分。为叙述方便，后面将 $F\_AQF(F),F\_VF(F),F\_TF(F),F\_KCF(F)$ 分别称为主成分“绝对量”特征、主成分“波动”特征、主成分“趋势”特征和主成分“峰度”特征。

3. 面板数据的聚类

上述从特征提取的角度减少了面板数据的时间维度，将面板数据转化为截面数据，因此可以直接用截面数据聚类方法对面板数据进行聚类。聚类前利用熵值法计算各特征权重，然后采用系统聚类法进行聚类，以保证聚类效果的稳定性。

3.1. 计算特征权重

本文中主成分“绝对量”特征、主成分“波动”特征、主成分“趋势”特征和主成分“峰度”特征对个体差异影响程度会有所不同，根据它们的影响程度必须赋予相应权重 $\omega_j\left(j=1,2,3,4\right)$，为了避免主观臆测，本文采取熵值法[10]客观赋权。

3.2. 聚类

先对 N 个个体的 4 项指标 $F\_AQF\left(F\right),F\_VF\left(F\right),F\_TF\left(F\right),F\_KCF\left(F\right)$ 在总体上进行 Z-Score 标准化，以消除数量级影响，标准化后 4 项指标分别记为 $F^*\_AQF\left(F\right),F^*\_VF\left(F\right),F^*\_TF\left(F\right),F^*\_KCF\left(F\right)$，然后再对加权后的 4 个指标 $\omega_1F^*\_AQF\left(F\right),\omega_2F^*\_VF\left(F\right),\omega_3F^*\_TF\left(F\right),\omega_4F^*\_KCF\left(F\right)$ 的值进行系统聚类。

系统聚类法的基本思想是[6]：首先定义样品间距离(或相似系数)和类与类之间的距离，N 个样品初时自成一类，此时，类间距离和样品间距离是等价的；然后将距离最近的两类合并为新类，并计算新类与其他类的类间距离，再按最小准则并类，这样，每次缩小一类，直到所有的样品都并成一类为止。

4. 实证分析

本文对 2013 年至 2018 年我国 31 个省份的经济发展状况进行聚类分析，说明我国省际间经济发展的差异，并对造成这种差异的可能原因进行分析，进一步给出相关政策建议。

4.1. 数据来源和指标选取

反映区域经济发展差异的指标很多，本文主要从宏观经济指标、居民收入和消费指标来构建省域经济发展评价指标体系，其中宏观经济指标包括人均 GDP、全社会固定资产投资、进出口总额、地方财政收入、社会消费品零售总额，居民收入是指居民人均可支配收入，消费指标是指居民人均消费支出。所构建指标体系如表 1 所示。

Table 1. Provincial economic development evaluation index system
表 1. 省域经济发展评价指标体系

一级评价指标	二级评价指标
宏观经济指标	人均 GDP (万元)
	全社会固定资产投资(亿元)
	进出口总额(千美元)
	地方财政收入(亿元)
	社会消费品零售总额(亿元)
居民收入指标	居民人均可支配收入(元)
居民消费指标	居民人均消费支出(元)

注：人均 GDP 等于各省 GDP (亿元)除以各省总人口(万人)。

本文使用的数据来源于国家统计局官网(《中国统计年鉴》2014~2019 年)。

4.2. 聚类分析

本文利用 R 语言计算所提取特征的权重：“绝对量”、“波动”、“趋势”和“峰度”4 个特征的权重分别为 0.382，0.474，0.085，0.059。可以看出，“绝对量”水平和“波动”水平对个体间差异的贡献程度都比较大，“趋势”水平和“峰度”水平的贡献程度相对而言比较小，可以理解为这些年的数据整体上都有一个较大的增长趋势，导致“趋势”和“峰度”对个体差异影响不大。

将得到权重后的数据集用 SPSS 进行系统聚类，其碎石图见图 1。从图 1 可以看出，到 4 类后，类间距离快速增大，形成极为“平坦的碎石路”，于是把分类数确定为 4，最终各省经济发展水平聚类结果见图 2 和表 2。

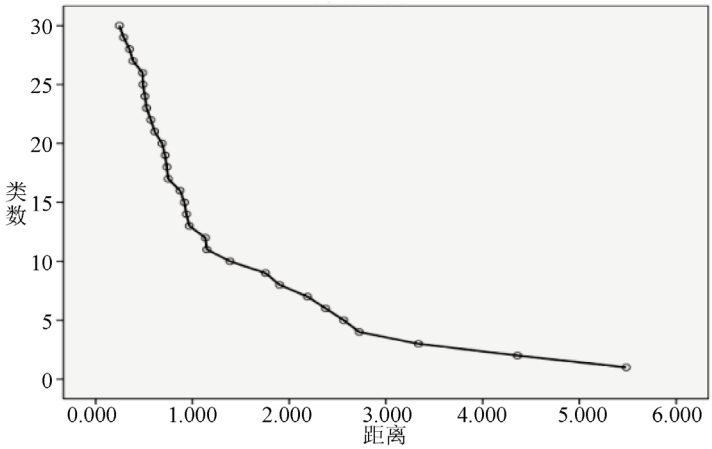

Figure 1. Cluster analysis gravel diagram of the economic development of 31 provinces, municipalities and autonomous regions
图 1. 31 个省市自治区经济发展的聚类分析碎石图

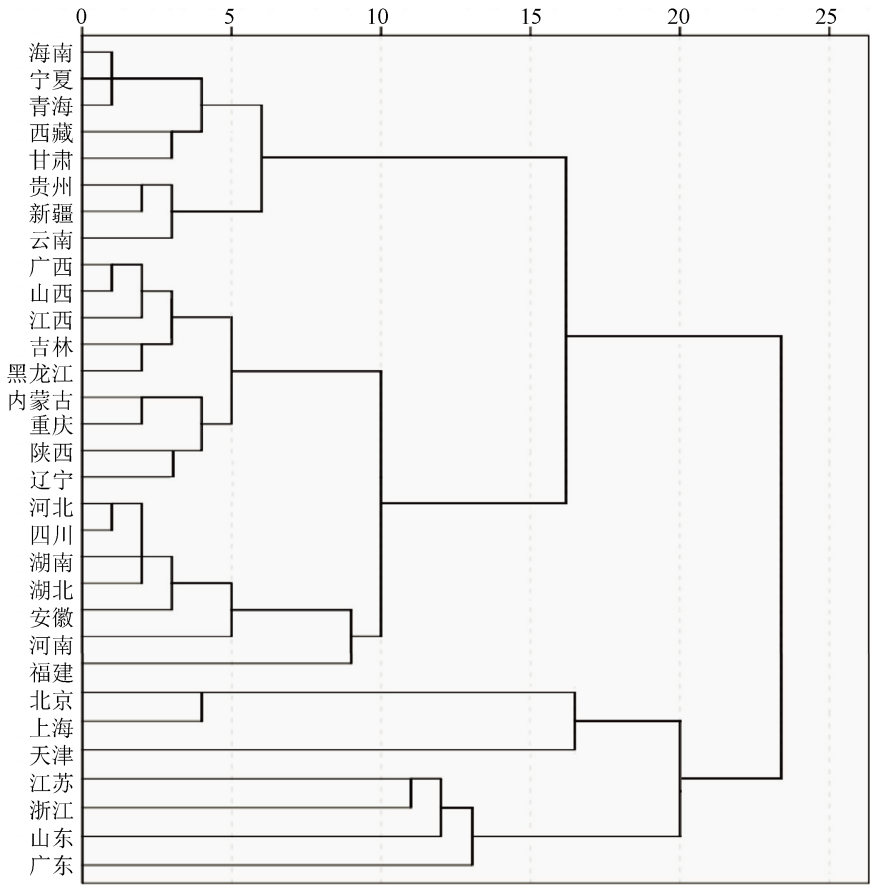

Figure 2. Cluster analysis tree diagram of the economic development level of each province
图 2. 各省经济发展水平聚类分析树状图

Table 2. Clustering results of panel data on the economic development level of each province
表 2. 各省经济发展水平面板数据聚类结果

类别	类
第 1 类	北京，上海，天津
第 2 类	江苏，浙江，山东，广东
第 3 类	福建，河南，安徽，湖南，湖北，辽宁，河北，四川，重庆，陕西，山西，黑龙江，江西，吉林，广西，内蒙古
第 4 类	海南，宁夏，青海，西藏，甘肃，贵州，新疆，云南

由表 2，第 1 类：北京、上海、天津，都是属于东部沿海的三个直辖市，人均 GDP 和人均可支配收入都领先全国其他省份，进出口贸易发达，经济和科技发展领先全国；第 2 类：江苏、浙江、山东、广东，都是属于东部沿海省份，地理位置优越，进出口贸易量大，人均 GDP 居于全国前列，地方财政收入高，经济和科技发展位于全国前列；第 3 类：福建、河南、安徽、湖南、湖北、辽宁、河北、四川、重庆、陕西、山西、黑龙江、江西、吉林、广西、内蒙古，大多属于我国的中部内陆地区和东北地区，资本流动和科技发展远不如东部沿海地区，经济发展处于中等水平；第 4 类：海南、宁夏、青海、西藏、甘肃、贵州、新疆、云南，多为西部和南部边疆地区，地广人稀，交通不便，地方财政收入不高，社会固定资产投资和进出口额度明显偏低，人均可支配收入和人均消费支出都相对较低，经济和科技相对落后，经济发展水平一般。

从聚类结果来看，本文所采用的面板数据聚类方法能较好地将各省的经济发展水平进行一个合理的划分。如果没有消除相同特征间的重叠信息就进行聚类的话，聚类结果(见表 3)会将绝大多数成员聚为一类，某些个体单独成类，聚类效果差，与实际情况不符。

Table 3. Clustering results of panel data without secondary feature extraction
表 3. 未二次特征提取的面板数据聚类结果

类别	类
第 1 类	浙江，江苏
第 2 类	广东
第 3 类	山东，四川，河南，湖南，湖北，河北，福建，上海，北京，安徽，辽宁，重庆，陕西，江西，广西，天津，内蒙古，云南，黑龙江，山西，吉林，宁夏，新疆，青海，贵州，海南
第 4 类	甘肃，西藏

5. 省际间经济发展差异的原因分析及政策建议

5.1. 原因分析

由聚类结果可以看出，我国不同区域之间经济发展是存在明显差异的，其中东部沿海省份的经济发展状况最好，其次是中部内部省份和东北三省，最后是西部和南部边疆省份。导致我国各省经济发展有明显差异的原因是多方面的，本文从以下四个方面分析其原因：

(一) 地理位置因素

我国东部地区地处平原，交通方便，基础设施较为完善，有利于工农业的发展，也有利于进行生产生活活动，沿海城市的众多天然港口为与其他国家进行贸易往来提供了便利条件，这些因素使得东部沿

海地区经济迅速发展；而我国西部地区大部分属高原、山脉、沙漠等地形，基础设施较为落后，交通不便，不利于工农业的发展，也不利于对外交流和吸引外资，地区发展受限。

(二) 国家政策因素

改革开放以来，国家对东部沿海地区从战略和政策上都有所倾斜，东部沿海省份获得了更优势的发展条件和资源环境，经济发展迅速，国家的经济重心也不断向东南沿海地区偏移；而中西部属于内陆地区，地理位置受限，政府引导支持不足，经济发展速度远不如东部地区迅速。

(三) 资本流动因素

优越的地理位置和政策支持吸引了大量的资本涌入，让东部沿海地区实现了经济高速发展；而中西部地区由于交通不便、市场不完善等因素影响，收到的投资较少，发展速度也较慢。

(四) 高素质人才区域分配不平衡

由于我国不同地区教育水平以及教育资源有差异，使得高考后众多高素质人才纷纷涌入东部沿海那些教育资源较多、教育水平较高的省份，这种情况随着人才受教育时间的增加而不断扩大着差距，而高校人才毕业后，也大多愿意留在一二线的大城市，使得东部省份的科技水平一直领先全国，金融和高新技术产业远多于中西部地区。这样的现象与选择造成了我国人才资源区域分配不平衡，进一步扩大了东西部之间的经济发展差距。

5.2. 政策建议

针对我国区域经济发展不平衡的问题，国家先后分别实施过西部大开发战略、振兴东北老工业基地战略、中部崛起战略，这些战略使区域经济发展不平衡得到过一定程度的缓解，但作用未能发挥持久。鉴于仍然存在区域经济发展不平衡的问题，本文根据聚类结果所得结论提出以下政策建议：

(一) 加快中西部地区基础设施建设

政府应加大对中西部地区的电力、交通、水利等基础设施建设的投入，基础设施越加完善，越有利于吸引资本流入和人才引进，越能促进地区的经济发展。营造良好的投资环境，为民营经济带来更多的资金投入，政府也要关注不发达区域的经济投资，加大对其人力物力及财力的支持，制定适当的区域经济扶持政策，有倾斜有针对性地带动区域经济的发展。

(二) 加快产业结构优化调整转型

中西部地区要加快对劳动密集型产业的转型，通过技术引进，提升此类产业的生产效率，加强与东部地区在技术、信息方面的合作，互通有无，发展能源产业，为其他地区输送产品，带动当地的经济发展；东北地区要重视国有企业改革，鼓励国企与外企与其它非国企之间的交互合作，激发市场活力的同时为民营经济带来更多的资金投入，振兴原有产业基础上重点发展农林产业，同时加大对新兴科技的关注与投入。

(三) 大力发展教育事业，重视人才引进

21 世纪各大城市都展开了人才争夺战，因为人才可以为当地的经济发展源源不断地注入活力。要想经济发展的好，经济发展落后地区一定要大力发展教育事业，完善人才机制，制定较为宽松的人才政策，重视引进高层次人才，充分发挥人才的知识溢出效应，为经济发展不断注入活力。

(四) 加强跨区合作，提升经济开放性

对于中西部地区来说，经济的腾飞必须加强与东部的互动，深化经济合作，借助东部的资源优势、技术优势和人才优势等，弥补自身发展的不足，带动自身的经济发展。充分利用好“一带一路”对周边的辐射作用，促进中西部地区和沿边地区的对外开放，在实现东西经济互动过程中促进中部心脏地带的崛起，进而形成东西互济、面向全球的开放新格局。

6. 总结

本文所采用的面板数据聚类方法综合考虑了面板数据时间维度上的“绝对量”、“波动”、“趋势”和“峰度”4 个动态特征，并用主成分分析法进行了二次特征提取，利用熵值法解决了这些特征的权重问题。从实证分析结果来看，该方法确实能较好解决指标间具有相似特征的多指标面板数据聚类问题。需要指出的是，本文仅仅构造了几个基础性的特征统计量来反映经济发展中面板数据的动态特征，其他更深层次更复杂的特征统计量还有待进一步研究；此外，经济发展评价指标体系还有待进一步地完善。

参考文献

[1] 乔慧. 关于我国 31 个省市自治区经济发展的多元统计分析[J]. 科技情报开发与经济, 2011, 21(1): 160-162+173.

[2] 杨英, 李海萍, 于向东, 屈玲玲. 基于因子和聚类分析的中国各省市竞争力分析与研究[J]. 河北工业科技, 2013, 30(5): 347-351.

[3] 朱莉莉. 区域经济发展差异的面板数据有序聚类分析[J]. 统计与决策, 2016(18): 148-150.

[4] 朱建平, 陈民恳. 面板数据的聚类分析及其应用[J]. 统计研究, 2007(4): 11-14.

[5] 郑兵云. 多指标面板数据的聚类分析及其应用[J]. 数理统计与管理, 2008(2): 265-270.

[6] 李因果, 何晓群. 面板数据聚类方法及应用[J]. 统计研究, 2010, 27(9): 73-79.

[7] 刘汉丽, 裴韬, 周成虎. 用于时间序列聚类分析的小波变换和特征量提取方法[J]. 测绘科学技术学报, 2014, 31(4): 372-376+382.

[8] 党耀国, 侯荻青. 基于特征提取的多指标面板数据聚类方法[J]. 统计与决策, 2016(19): 68-72.

[9] 戴大洋, 邓光明. 基于主成分特征提取的面板数据聚类方法[J]. 统计与决策, 2018, 34(21): 72-76.

[10] Zhang, H.P. (2015) Application on the Entropy Method for Determination of Weight of Evaluating Index in Fuzzy Mathematics for Wine Quality Assessment. *Advance Journal of Food Science and Technology*, **7**, 195-198. https://doi.org/10.19026/ajfst.7.1293