

基于聚类分析的数字普惠金融区域差异性研究

胡 勇

上海工程技术大学管理学院, 上海

收稿日期: 2024年1月20日; 录用日期: 2024年3月19日; 发布日期: 2024年6月21日

摘 要

普惠金融是指立足机会平等要求和商业可持续原则, 以合理的成本为有金融服务需求的社会各阶层和群体提供适当、有效的金融服务。近年来, 我国数字普惠金融的发展呈现出快速增长的趋势, 为实体经济的发展提供了支持。在全球化浪潮的带动下, 数字普惠金融必将打破重重壁垒, 引领世界范围内各阶层和群体平等共享金融服务。目前, 数字普惠金融产品和服务已涉及个人征信、消费金融、数字化支付、投资理财等诸多领域, 对城乡收入差距、减缓贫困、经济增长和居民消费都会产生影响。为了加快区域平衡发展, 应减少不同地区数字普惠金融水平的差异性。本文为了了解近几年全国31个省市的数字普惠金融发展水平, 以便发展水平较差的地区可以加快发展。收集了2012~2018年数字金融覆盖广度、数字金融使用深度和普惠金融数字化程度三个指标的“北京大学数字普惠金融指数”数据, 先确定最佳聚类数目, 然后对截面数据和多指标面板数据分别进行聚类。得出我国数字普惠金融水平整体呈现上升趋势, 但我国31个省市的数字普惠金融发展水平差异较大, 其中华东地区发展最好, 华南和华中地区整体情况发展水平较好, 华北, 西北, 西南和东北地区整体发展水平较差, 但具有省域差异性。针对这一现象, 提出相关的政策建议以期可以使数字普惠金融更好的发展。

关键词

数字普惠金融, 系统聚类法, 可视化, 区域差异性

A Study of Regional Differences in Digital Inclusive Finance Based on Cluster Analysis

Yong Hu

School of Management, Shanghai University of Engineering Science, Shanghai

Received: Jan. 20th, 2024; accepted: Mar. 19th, 2024; published: Jun. 21st, 2024

Abstract

Inclusive finance refers to the provision of appropriate and effective financial services at reasona-

文章引用: 胡勇. 基于聚类分析的数字普惠金融区域差异性研究[J]. 运筹与模糊学, 2024, 14(3): 391-398.

DOI: 10.12677/orf.2024.143277

ble costs to all strata and groups in society that have a need for financial services, based on the requirements of equal opportunity and the principle of commercial sustainability. In recent years, the development of digital inclusive finance in China has shown a trend of rapid growth, providing support for the development of the real economy. Driven by the wave of globalization, digital inclusive finance is bound to break down barriers and lead the way for all classes and groups around the world to share financial services on an equal footing. Currently, digital inclusive financial products and services have covered many areas, including personal credit, consumer finance, digital payments and investment and financial management, and have had an impact on the urban-rural income gap, poverty alleviation, economic growth and consumption. In order to accelerate balanced regional development, it is important to reduce the disparity in the level of digital financial inclusion in different regions. This paper aims to understand the national (31 provinces or cities) development level of digital inclusive finance in recent years, so that regions with a lower level of development can accelerate their development. Peking University Digital Inclusive Finance Index (DIFI) data on three indicators, namely, the breadth of digital finance coverage, the depth of digital finance use, and the degree of digitalization of inclusive finance, were collected from 2012 to 2018, and the optimal number of clusters was determined first, and then clustering was carried out on the cross-sectional data and the multi-indicator panel data respectively. It is concluded that the overall level of digital inclusive finance in China shows an upward trend, but the development level of digital inclusive finance in China (31 provinces or cities) varies greatly, with the best development in East China, better development in South China and Central China, and worse development in North China, Northwest China, Southwest China and Northeast China, but with provincial differences. In view of this phenomenon, relevant policy recommendations are proposed to enable the better development of digital inclusive finance.

Keywords

Digital Inclusive Finance, Systematic Clustering Method, Visualization, Regional Differentiation

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

2020 年十四五规划提出加快数字规划建设数字中国，数字普惠金融是将普惠金融数字化，为人民提供更加便利的金融服务。数字普惠金融指数是反映数字普惠金融发展水平的综合指标，从 2012 年到 2018 年我国的数字普惠金融水平整体呈上升趋势，如下图 1 所示，但发展水平仍然有待提高[1]。为了加快数字普惠金融发展水平，本文将关注点放到我国发展水平的区域差异性，以期政府可以采取更多的措施补齐短板，缩小区域差异性，使数字普惠金融发展水平整体提高。

然而对数字普惠金融的研究，主要集中在数字普惠金融的发展历程，指标测算以及对经济，创新，企业等产生的影响，对数字普惠金融的区域差异性进行研究的学者较少，但了解区域差异性可以对数字普惠金融的发展起到积极的作用。已经存在的对我国各省的数字普惠金融差异性的研究多半都是使用某一年的数据进行聚类分析，虽然可以体现出我国当年各省的数字普惠金融水平差异，但缺少时间维度，没有体现出近些年我国数字普惠金融发展的综合水平差异。多指标面板数据包含时间，样本，指标三个维度，可以更好地，更全面地反映我国各省近些年数字普惠金融的发展。因此，本文将多指标面板数据聚类应用到数字普惠金融方面。

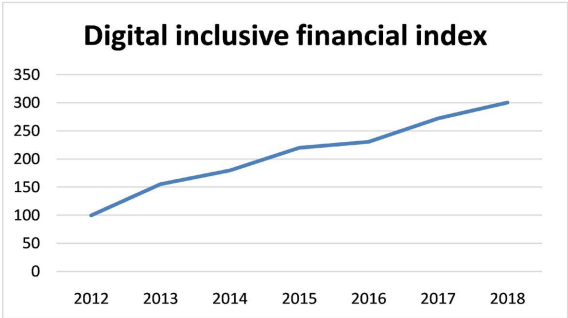


Figure 1. Development trends in the level of digital inclusive finance

图 1. 数字普惠金融水平发展趋势

2. 文献综述

聚类属于无监督学习，可将无标签的样本分为不同的类别，目前已被广泛应用到模式识别，客户分类，数据预处理等方面。聚类算法包括普通的系统聚类法，动态聚类法，以及与智能算法相结合的聚类算法等等。虽然聚类算法发展比较成熟，但对多指标面板数据聚类算法的研究及其应用仍然较少。吕晨[2] (2020 年)对于反映企业财务状况的多指标面板数据先用主成分分析提取特征降维，将多指标面板数据结构转换为时间序列数据，然后利用系统聚类方法进行聚类。党耀国等人[3] (2016 年)针对聚类常用的欧式距离存在的缺点，定义了可以反映出时间特性的特征，包括全时绝对量特征，全时波动特征，全时峰度特征，全时偏度特征和全时趋势特征，利用改进的特征提取后运用动态聚类方法对我国各地区的交通事故进行了聚类。林秀梅等人[4] (2016 年)利用吸引子传播聚类融合半监督学习，对多指标面板数据聚类，并与 AP 模型，SOM 神经模型和系统聚类结果进行了比较，证明了智能信息处理的多指标面板数据聚类结果更加精确。任娟[5] (2013 年)利用因子分析提取多指标面板数据中的重要信息，然后对提取出的公共因子进行有序聚类分析。郑宾云[6] (2008 年)对多指标面板数据聚类提出了“欧式时空距离”新概念以及将截面数据的离差平方和函数扩展到可以反映时间维度的离差平方和函数，然后对企业生产效率进行了实证分析，证明了方法的可行性。

通过对以上学者研究成果的学习与借鉴，本文将多指标面板数据聚类应用到数字普惠金融方面，以期可以对提高全国数字普惠金融发展水平提供思路。

3. 多指标面板数据的数据结构及特征

3.1. 多指标面板数据结构

面板数据是将横截面数据与时间序列数据整合起来的一种数据类型，有时间序列和截面两个维度，分为单指标面板数据和多指标面板数据。单指标面板数据与截面数据相似，可以用一个二维表来表示。多指标面板数据比单指标面板数据结构复杂，是一种三维结构数据，在平面上可将其转换为多个二维数据表示。多指标面板数据的表示方法如表 1 所示：

Table 1. Multi-indicator panel data

表 1. 多指标面板数据

时间	个体	1	2	3	...	k	...	m
1		$x_{1,1,n}$	$x_{1,2,n}$	$x_{1,3,n}$	...	$x_{1,k,n}$	...	$x_{1,m,n}$
2		$x_{2,1,n}$	$x_{2,2,n}$	$x_{2,3,n}$	...	$x_{2,k,n}$	...	$x_{2,m,n}$

续表

3	$x_{3,1,n}$	$x_{3,2,n}$	$x_{3,3,n}$	...	$x_{3,k,n}$	...	$x_{3,m,n}$
4	$x_{4,1,n}$	$x_{4,2,n}$	$x_{4,3,n}$	...	$x_{4,k,n}$	...	$x_{4,m,n}$
...	...	...	...	...	...	...	...
t	$x_{t,1,n}$	$x_{t,2,n}$	$x_{t,3,n}$	...	$x_{t,k,n}$	...	$x_{t,m,n}$

表中, $x_{i,j,n}$ 表示时间 i , 第 j 个样品的第 n 个指标所对应的样本数据

其中, $i=1,2,3,\dots,t$ $j=1,2,3,\dots,k,\dots,m$ $n=1,2,3,\dots,s$

$x_{i,k} = [x_{i,k,1}, x_{i,k,2}, x_{i,k,3}, \dots, x_{i,k,s}]^T$ 表示在时间 i , 样本 k 的所有指标的观测值

$x_{i,n} = [x_{i,1,n}, x_{i,2,n}, x_{i,3,n}, \dots, x_{i,m,n}]^T$ 表示在时间 i , 第 n 个指标所有样本的观测值

$x_{k,n} = [x_{1,k,n}, x_{2,k,n}, x_{3,k,n}, \dots, x_{t,k,n}]^T$ 表示第 k 个样本的第 n 个指标所有时间的观测值。

3.2. 多指标面板数据的数字特征

1) 样本 k 在时间 i 内的均值和方差分别为

$$\bar{x}_{i,k} = \frac{1}{s} \sum_{n=1}^s x_{i,k,n} \quad \text{Var}(x_{i,k}) = \frac{1}{s-1} \sum_{n=1}^s (x_{i,k,n} - \bar{x}_{i,k})^2$$

2) 时间 i , 第 n 个指标所有样本的均值和方差分别为

$$\bar{x}_{i,n} = \frac{1}{m} \sum_{j=1}^m x_{i,j,n} \quad \text{Var}(x_{i,n}) = \frac{1}{m-1} \sum_{j=1}^m (x_{i,j,n} - \bar{x}_{i,n})^2$$

3) 第 k 个样本的第 n 个指标的均值和方差分别为

$$\bar{x}_{k,n} = \frac{1}{t} \sum_{i=1}^t x_{i,k,n} \quad \text{Var}(x_{k,n}) = \frac{1}{t-1} \sum_{i=1}^t (x_{i,k,n} - \bar{x}_{k,n})^2$$

4) i 时刻总体样本的均值和方差分别为

$$\bar{x}_i = \frac{1}{ms} \sum_{j=1}^m \sum_{n=1}^s x_{i,j,n} \quad \text{Var}(x_i) = \frac{1}{ms-1} \sum_{j=1}^m \sum_{n=1}^s (x_{i,j,n} - \bar{x}_i)^2 \quad i=1,2,3,\dots,t$$

5) 样本 j 的均值向量和方差分别为

$$\bar{x}_j = \frac{1}{ts} \sum_{i=1}^t \sum_{n=1}^s x_{i,j,n} \quad \text{Var}(x_j) = \frac{1}{ts-1} \sum_{i=1}^t \sum_{n=1}^s (x_{i,j,n} - \bar{x}_j)^2 \quad j=1,2,3,\dots,m$$

6) 指标 n 的均值向量和方差分别为

$$\bar{x}_n = \frac{1}{tm} \sum_{i=1}^t \sum_{j=1}^m x_{i,j,n} \quad \text{Var}(x_n) = \frac{1}{tm-1} \sum_{i=1}^t \sum_{j=1}^m (x_{i,j,n} - \bar{x}_n)^2 \quad n=1,2,3,\dots,s$$

4. 聚类方法

4.1. 聚类数目 k 值的确定

聚类算法属于无监督学习, 聚类数目需要人为确定, 聚类数目过多过少都会对聚类结果产生负面影响。手肘法可以确定最优 k 值, 当聚类数目小于最佳聚类数时, 增加 k 值会使 SSE 大幅度减少; 当聚类数目大于最佳聚类数目时, 虽然聚类数目越多聚合程度会提高, 但 SSE 值下降的幅度会骤减, 逐渐趋于

平行。

误差平方和 SSE:

$$SSE = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \quad i=1,2,3,\dots,k, j=1,2,3,\dots,n \quad (4-1)$$

式子中 i 表示第 i 类, 总共有 k 类, x_{ij} 表示在 i 类中第 j 个样本, i 类中总共有 n 个样本, $\bar{x}_i$ 表示第 i 类 n 个样本的均值(也称为 i 类的质心)。

由图 2 可知各地区的数字普惠金融差异最佳可分为 4 类。

4.2. 系统聚类法

聚类方法有 K-means 聚类, 模糊聚类, 动态聚类以及与智能算法相结合的聚类, 其中系统聚类法是聚类分析方法中使用最多的, 主要内容包括点与点之间的距离和类与类之间的距离。点与点之间的距离

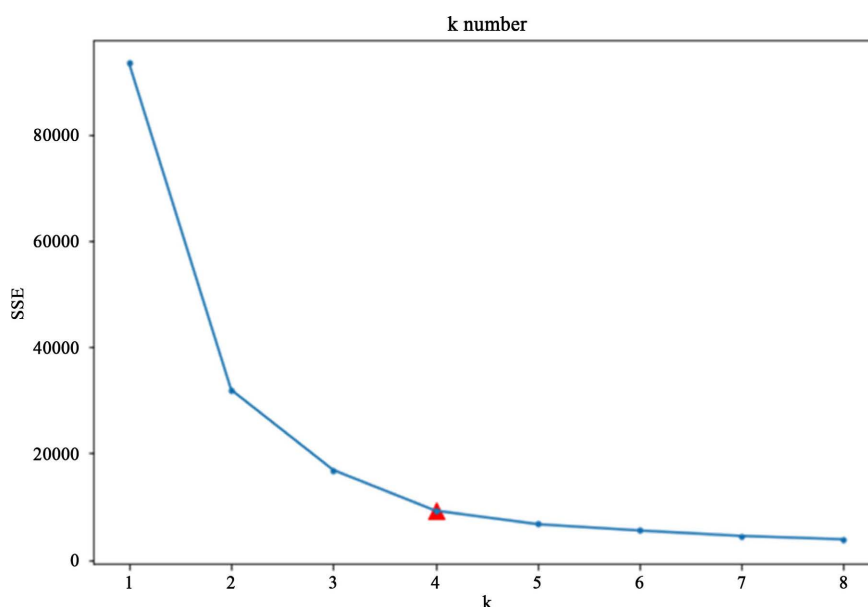


Figure 2. Number of clusters k value

图 2. 聚类数目 k 值

包括欧式距离, 切比雪夫距离, 明氏距离, 绝对值距离, 类与类之间的距离包括最短距离法, 最长距离法, 重心法, 离差平方和法。考虑数据的特点以及分布, 选择欧式距离来描述样本与样本之间的距离, Ward 方法来衡量类与类之间的距离。

1) x 点与 y 点的欧式距离:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4-2)$$

2) Ward 方法:

假设将 n 个样本分为 k 类 $L_1, L_2, L_3, \dots, L_k$, x_{it} 表示类 L_t 中第 i 个样本

n_t 表示类 L_t 中样本的个数, $\bar{x}_t$ 是类 L_t 的重心, 类 L_t 中的样本的离差平方和为:

$$S_t = \sum_{i=1}^{n_t} (x_{it} - \bar{x}_t)' (x_{it} - \bar{x}_t), \text{ 整个类内离差平方和为: } S = \sum_{t=1}^k S_t。$$

4.3. 多指标面板数据聚类

加入时间维度，点与点之间的距离变为横截面与横截面的距离，将欧式距离拓展，面与面之间的距离也即面中所有的点距离求和[7]。同时，需要考虑样本指标值随时间的变化速度和样本指标值随时间的波动程度。使用到的距离公式分别为全时“绝对量”距离，全时“增长速度”距离，全时“变异系数”距离，全时“综合”距离。

样本 j_1 与 j_2 之间的全时“绝对量”距离为：

$$d_1(j_1, j_2) = \left[\sum_{i=1}^t \sum_{n=1}^s (x_{i,j_1,n}, x_{i,j_2,n})^2 \right]^{1/2} \tag{4-3}$$

样本 j_1 与 j_2 之间的全时“增长速度”距离为：

$$d_2(j_1, j_2) = \left[\sum_{i=1}^t \sum_{n=1}^s \left(\frac{\Delta x_{i,j_1,n}}{x_{i-1,j_1,n}} - \frac{\Delta x_{i,j_2,n}}{x_{i-1,j_2,n}} \right)^2 \right]^{1/2} \tag{4-4}$$

$$\Delta x_{i,j_1,n} = x_{i,j_1,n} - x_{i-1,j_1,n} \quad \Delta x_{i,j_2,n} = x_{i,j_2,n} - x_{i-1,j_2,n}$$

样本 j_1 与 j_2 之间的全时“变异系数”距离为：

(4-5)

样本 j_1 与 j_2 之间的全时“综合”距离为：

$$d(j_1, j_2) = \alpha d_1(j_1, j_2) + \beta d_2(j_1, j_2) + \gamma d_3(j_1, j_2) \tag{4-6}$$

4.4. 聚类结果与分析

本文采用北京大学数字普惠金融指数数据，选取的指标为“数字金融覆盖广度”“数字金融使用深度”“普惠金融数字化程度”。指标是北京大学金融研究中心根据 24 个具体指标计算出的一级指标，因此三个指标可以充分反应每个地区的数字普惠金融水平。基于 2012~2018 年每个省的数字普惠金融指数数据，首先对我国 31 个省市每年的数字普惠金融水平运用系统聚类方法进行聚类，然后将每个省 7 年的数字普惠金融水平综合考虑，运用多指标面板数据聚类方法进行聚类[8]。本文截面数据的聚类使用的软件为 python，多指标面板数据聚类使用的软件为 R 语言。聚类结果如下表 2 所示：

Table 2. Clustering results
表 2. 聚类结果

		First	Second	Third	Fourth
截面数据	2012	3, 25, 27, 28	2, 4, 6, 7, 15, 19, 23, 24, 26, 29, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	0, 1, 8, 9, 10, 12, 18
	2013	2, 3, 4, 6, 7, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	1, 9, 12, 18	0, 8, 10
	2014	4, 23, 25, 27, 28	2, 3, 6, 7, 15, 19, 24, 26, 29, 30	1, 5, 9, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22	0, 8, 10
	2015	4, 23, 27, 28, 29	2, 3, 6, 7, 15, 19, 24, 25, 26, 30	5, 11, 13, 14, 16, 17, 20, 21, 22	0, 1, 8, 9, 10, 12, 18

续表

截面数据	2016	2, 6, 17, 25, 27, 28, 29, 30	3, 4, 7, 15, 19, 23, 24	1, 5, 11, 13, 14, 16, 20, 21, 22, 26	0, 8, 9, 10, 12, 18
	2017	3, 4, 23, 25, 27, 28, 29, 30	2, 5, 6, 7, 11, 13, 14, 15, 17, 19, 20, 21, 22, 24, 26	1, 9, 12, 16, 18	0, 8, 10
	2018	3, 4, 6, 7, 23, 25, 27, 28, 29, 30	2, 5, 11, 13, 14, 15, 17, 19, 20, 21, 22, 24, 26	1, 9, 12, 16, 18	0, 8, 10
多指标面板数据		2, 3, 4, 6, 7, 15, 19, 23, 24, 25, 26, 27, 28, 29, 30	5, 11, 13, 14, 17, 20, 21, 22	1, 9, 12, 16, 18	0, 8, 10

注：0 北京市；1 天津市；2 河北省；3 山西省；4 内蒙古自治区；5 辽宁省；6 吉林省；7 黑龙江省；8 上海市；9 江苏省；10 浙江省；11 安徽省；12 福建省；13 江西省；14 山东省；15 河南省；16 湖北省；17 湖南省；18 广东省；19 广西壮族自治区；20 海南省；21 重庆市；22 四川省；23 贵州省；24 云南省；25 西藏自治区；26 陕西省；27 甘肃省；28 青海省；29 宁夏回族自治区；30 新疆维吾尔自治区。

多指标面板数据聚类树状图如图 3 所示。

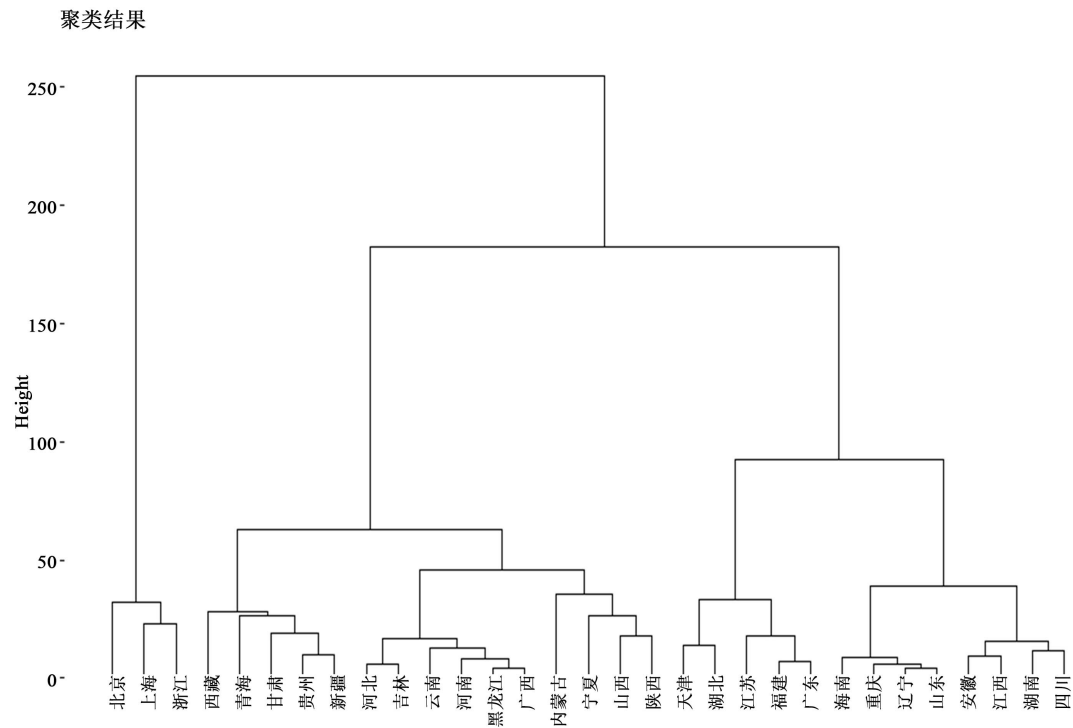


Figure 3. Multi-indicator panel data clustering tree diagram
图 3. 多指标面板数据聚类树状图

如图 3 所示，可以将我国 31 个省市的数字普惠金融分为四类，第一类为数字普惠金融发展水平不好的地区有西藏，青海，甘肃，贵州，新疆，河北，吉林，云南，河南，黑龙江，广西，内蒙古，宁夏，山西，陕西；第二类为数字普惠金融发展水平一般的地区有海南，重庆，辽宁，山东，安徽，江西，湖

南,四川;第三类为数字普惠金融发展水平较好的地区有天津,湖北,江苏,福建,广东;第四类为数字普惠金融发展水平最好的地区有北京,上海,浙江。由表 2 可知,截面数据的聚类结果显示,个别地区在不同的年份所属的类别也不一样,缺乏稳定性。多指标面板数据聚类将所有年份综合考虑,既排除了单个年份聚类结果的偶然性,也排除了个别城市聚类结果的波动性,使结果更具可靠性与稳定性,短期内可依据这个聚类结果制定相应的政策措施,减少我国数字普惠金融发展水平的区域差异性。

5. 结论与建议

本文根据数字金融覆盖广度、数字金融使用深度、普惠金融数字化程度三个综合性指标将 31 个省市的数字普惠金融发展水平基本分为 4 类。不难看出,我国数字普惠金融区域差异性较大,且发展好的省份偏少,大部分省份尤其西部地区发展水平较低。

如何根据这 4 类地区的经济发展水平,人口密度以及网民覆盖率等等影响因素提高我国整体数字普惠金融发展水平,减少区域差异性都是我们的政府、企业和学界应关注的重要内容。本文希望通过对全国各个地区的数字普惠金融发展水平进行聚类分析,能使得每个地区对自己的定位有一个清晰的认识,使同类地区的联系和各自的特点更加清楚,让各界对这些地区的关注有的放矢。同时,政府也应该大力补齐短板,使数字普惠金融发展水平较差的地区加快发展。同时,也应不断提高数字普惠金融的监管。数字技术与金融业务的加速融合下,金融机构间的互联互通更频繁,系统内的耦合度和关联性更加紧密,金融风险可以进一步地突破时间和空间上的局限,表现为更强的复杂性,传染性和跨地域性特征,使得传统监管难以适应这些新变化。因此,监管机构也应该与时俱进,创建新的监管模式,使数字普惠金融可以健康发展。

参考文献

- [1] 何文秀. 中国数字普惠金融的区域差异分析[J]. 统计理论与实践, 2020(3): 17-22.
- [2] 吕晨. 多指标面板数据下的企业财务状况聚类分析[J]. 商场现代化, 2020(6): 131-132.
- [3] 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.
- [4] 林秀梅, 孙海波, 王丽敏. 智能信息处理的多指标面板数据聚类方法及其应用[J]. 数理统计与管理, 2016, 35(4): 641-648.
- [5] 任娟. 多指标面板数据融合聚类分析[J]. 数理统计与管理, 2013, 32(1): 57-67.
- [6] 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.
- [7] 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.
- [8] 赵伟, 张红兵. 基于面板数据分析我国不同地区人均消费支出[J]. 统计与管理, 2016(6): 46-48.