УДК 339.16.012.23

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА В РЕГИОНАЛЬНЫХ ИССЛЕДОВАНИЯХ

Прохоренков П.А., Регер Т.В., Гудкова Н.В.

ФГБОУ ВО «Финансовый университет при Правительстве Российской Федерации», филиал, Смоленск, e-mail: prohpavel@yandex.ru, tatjana-reger@bk.ru, finubi2020@yandex.ru

Статья посвящена исследованию социально-экономического развития регионов с использованием методов математической статистики и кластерного анализа. Регионы России, обладая своими историческими особенностями, природными условиями и целым рядом специфичных факторов, имеют свои особенности и в социально-экономическом развитии. Учет этих особенностей позволяет более эффективно управлять процессами модернизации экономики, обеспечивая рост валового регионального продукта и повышение качества жизни населения региона. Применение методов кластерного анализа позволяет на основе статистических данных по различным показателям социально-экономического развития регионов объединить отдельные регионы в целевые кластеры. Анализ полученных кластеров дает возможность выявлять и эффективно решать типичные проблемы регионального развития, придать ускорение экономическим процессам, сконцентрировать ресурсы на ключевых направлениях. В качестве объекта исследования в данной статье рассматриваются 16 регионов Российской Федерации. Вся процедура анализа состоит из этапа очистки данных, этапа нормализации данных и этапа непосредственного выделения кластеров. В данном исследовании на предварительном этапе использованы методы корреляционного анализа, позволившие убрать из рассмотрения факторы с сильной корреляционной зависимостью. В качестве непосредственного метода кластеризации использован алгоритм k-means, основанный на вычислении минимального евклидова расстояния между центром кластера и отдельными объектами, а также алгоритм иерархической агломеративной кластеризации. В качестве инструментальных средств проведения исследования использованы прикладные пакеты программ в среде Python, а также среда для анализа данных RStudio. Исследование проводилось по четырем группам факторных признаков, отражающих различные стороны развития регионов. Проведение ряда расчетов для разного числа кластеров позволило определить наиболее оптимальный уровень разбиения выборки. Все исследования проведены по данным статистических наблюдений за 2015, 2018, 2020 гг. По полученным кластерам каждой группы факторов дана содержательная интерпретация признаков включения объектов в соответствующий кластер. Полученные результаты дают многостороннюю характеристику регионов, прослеживают динамику отдельных изменений в направлениях развития, дают объективную картину регионального развития.

Ключевые слова: кластерный анализ, региональная экономика, корреляционный анализ

CLUSTER ANALYSIS METHODS IN REGIONAL STUDIES

Prokhorenkov P.A., Reger T.V., Gudkova N.V.

Financial University under the Government of the Russian Federation, branch, Smolensk, e-mail: prohpavel@yandex.ru, tatjana-reger@bk.ru, finubi2020@yandex.ru

The article is devoted to the study of the socio-economic development of regions using the methods of mathematical statistics and cluster analysis. The regions of Russia, having their own historical features, natural conditions and a number of specific factors, have their own characteristics in socio-economic development. Taking these features into account makes it possible to more effectively manage the processes of modernization of the economy, ensuring the growth of the gross regional product and improving the quality of life of the population of the region. The use of cluster analysis methods allows, on the basis of statistical data on various indicators of the socio-economic development of regions, to combine individual regions into target clusters. The analysis of the resulting clusters makes it possible to identify and effectively solve typical problems of regional development, speed up economic processes, and concentrate resources on key areas. This article considers 16 regions of the Russian Federation as an object of study. The entire analysis procedure consists of a data cleaning step, a data normalization step, and a direct clustering step. In this research, at the preliminary stage, methods of correlation analysis were used, which made it possible to remove factors with a strong correlation dependence from consideration. The k-means algorithm based on the calculation of the minimum Euclidean distance between the cluster center and individual objects, as well as the hierarchical agglomerative clustering algorithm, were used as direct clustering methods. Application packages in the python environment, as well as the environment for data analysis RStudio, were used as research tools. The research was carried out on four groups of factor characteristics, reflecting various aspects of the development of regions. Carrying out a series of calculations for a different number of clusters made it possible to determine the most optimal level of sample splitting. All studies were carried out according to statistical observations for 2015, 2018, 2020. Based on the obtained clusters of each group of factors, a meaningful interpretation of the signs of inclusion of objects in the corresponding cluster is given. The results obtained give a multilateral characterization of the regions, trace the dynamics of individual changes in the directions of development, and give an objective picture of regional development.

Keywords: cluster analysis, regional economy, correlation analysis

Успешное развитие экономики России определяется тем, насколько динамично развиваются ее отдельные регионы. Каждый регион России, обладая своими историческими традициями, природными условиями и целым рядом специфичных факторов,

имеет свои особенности и в социально-экономическом развитии. Учет таких особенностей позволяет более эффективно управлять процессами модернизации экономики, обеспечивая рост валового регионального продукта и повышение качества жизни населения региона.

Материалы и методы исследования

Любой регион России в самом общем случае можно рассматривать как сложную систему управления, подчиняющуюся действиям как внешних условий, так и внутренних факторов развития. Несмотря на особенности регионов, есть и много схожих процессов, свойственных ряду регионов. Одним из методов, позволяющих провести исследование процессов регионального развития, является кластерный анализ. Применение методов кластерного анализа позволяет на основе статистических данных по различным показателям социально-экономического развития регионов объединить отдельные регионы в целевые кластеры. Анализ таких кластеров позволяет выявлять и эффективно решать типичные проблемы регионального развития, придать ускорение экономическим процессам, сконцентрировать ресурсы на ключевых направлениях.

Методы кластерного анализа применяются уже достаточно давно, а сам термин предложен для данного вида анализа английским ученым Р. Трионом в 1939 г. Особую роль данный вид исследований получил с развитием цифровых технологий и баз данных. В современном цифровом пространстве все более заметную роль играют технологии «больших данных», где кластерный анализ занимает важную позицию. Методы кластерного анализа активно используются маркетологами, аналитиками банковской сферы, специалистами в области регионального планирования и ряде других областей науки и производства.

Методы кластерного анализа нашли широкое применение в различных областях науки и, в частности, в анализе и управлении социально-экономическими процессами. К достоинствам этих методов можно отнести универсальность, наличие большого числа алгоритмов, реализующих методы кластеризации, наличие универсальных и специализированных программных систем со встроенными сервисами кластеризации. В качестве примера использования методов кластеризации в медицинских исследованиях можно привести работы ряда авторов [1, 2]. Аналогичные методы применительно к области психологии рассматриваются в работе Т.Н. Савченко [3]. Кластерный анализ регионов России с позиций научного потенциала рассматривается в работе [4]. Целый ряд научных исследований посвящен выделению региональных кластеров по ряду критериев [5, 6]. В работах [7, 8] рассматриваются вопросы демографии и научных исследований.

Целью исследования в данной работе является выделение групп регионов со схожими социально-экономическими показателями и анализ изменений, происходящих в этих группах в течение последних лет.

Результаты исследования и их обсуждение

Успешность процедур кластеризации во многом определяется качеством данных, используемых в данном анализе. К основным требованиям, предъявляемым к выборке данных, можно отнести: отсутствие корреляции между используемыми показателями, отсутствие больших отклонений от средних значений, закон распределения факторов должен приближаться к нормальному. Кроме того, используемые в процедуре кластерного анализа показатели должны быть безразмерными и нормализованными. Таким образом, проведению кластерного анализа должны предшествовать процедуры очистки и нормализации данных.

Все методы кластеризации в качестве критерия объединения объектов кластеризации в кластер используют одну из мер близости объектов по рассматриваемому набору нормализованных признаков. Наиболее употребляемым методом оценки близости объектов можно считать квадрат евклидова расстояния. В частности, такой способ объединения объектов применяется в алгоритме k-means. Критерий объединения V вычисляется как сумма квадратов отклонений факторных признаков x_i от центра j-й группировки вектора признаков μ_j.

$$V = \sum_{i}^{k} \sum_{x_i \in K_j} \left(x_i - \mu_j \right)^2,$$

где x_i – факторный признак; u_j – центр группировки вектора признаков; k – число кластеров.

Число кластеров задается в начале процедуры кластеризации, центры кластеров вначале формируются произвольно, а затем на каждом шаге работы алгоритма уточняются. В ходе работы алгоритма исследуемые объекты закрепляются за теми кластерами, которые обеспечивают для них минимальное расстояние до центра кластера. Оптимальным считается разбиение, обеспечивающее минимальное значение критерия V.

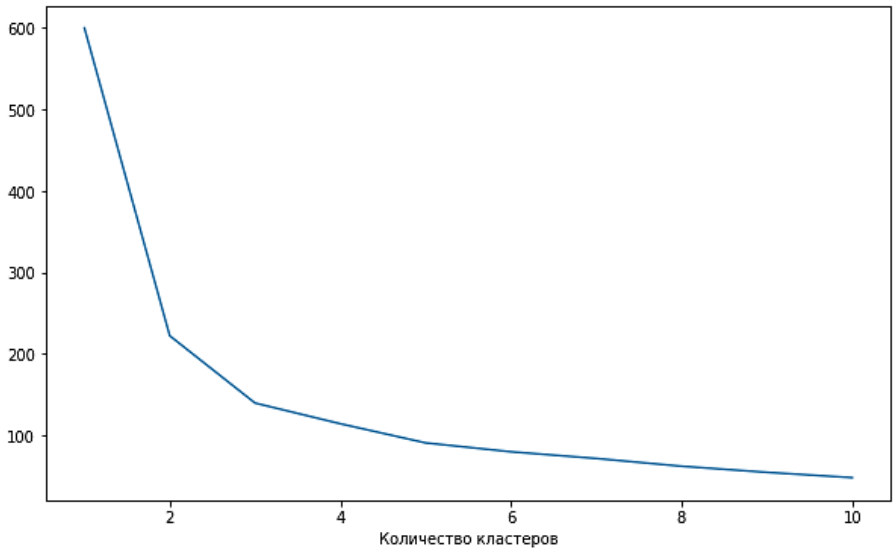

Рис. 1. Определение числа кластеров методом «локтя»

Схожими по построению алгоритма разбиения исходного множества объектов на кластеры K_j, но отличающимися выбором меры близости объектов являются такие методы, как k-medians, «расстояние городских кварталов», методы, основанные на расстоянии Чебышева.

Кроме нормализации исходных данных в начале исследования, приходится решать задачу выбора оптимального числа кластеров. Чаще всего для этих целей используется метод «локтя». Суть этого метода заключается в построении зависимости $V(k)$ и определении точки графика, после которой уменьшение критерия оказывается незначительным (рис. 1). В приведенном ниже примере такое число кластеров равно 3.

Для решения задач кластеризации с небольшим числом факторных признаков находит применение иерархическая кластеризация, которая, в свою очередь, делится на агломеративную и дивизимную. Алгоритм агломеративной кластеризации предусматривает пошаговое объединение объектов в классы, начиная с классов с минимальным расстоянием, и поэтапное объединение отдельных классов в более крупные. В качестве наглядного результата такого объединения строится дендрограмма (рис. 2). Алгоритм дивизимной кластеризации осуществляет формирование классов путем разбиения исходного множества объектов на отдельные подмножества, и далее деление промежуточных классов продолжается до полного деления всех объектов. Как и в предыдущем случае, результат может быть проиллюстрирован дендрограммой.

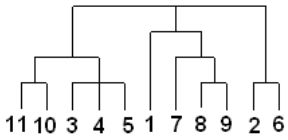

Рис. 2. Пример агломеративной кластеризации

Для выполнения алгоритмов кластеризации в данной работе рассматриваются регионы Центрального федерального округа, исключая Москву и Московскую область. Таким образом, в приведенных расчетах участвуют 16 объектов, характеризующихся набором факторных признаков по разным направлениям социально-экономического развития. Все факторы сгруппированы в четыре группы, каждая из которых имеет свой набор показателей:

1. Группа факторов общей характеристики объекта исследования:

– площадь территории p11;

– численность проживающего населения p12;

– валовой региональный продукт на душу населения p13;

– объем инвестиций p14.

2. Группа социальных факторов:

– уровень безработицы p21;

– фактическое потребление на душу населения p22;

– среднедушевой доход p23;

– число безработных p24;

– число заболевших на 10000 населения p25;

– количество коек на 10000 населения p26.

3. Характеристики инновационности:
– число организаций, занимающихся научными исследованиями, p31;
– затраты на научные исследования p32;
– использование инновационных производственных технологий p33;
– инновационная активность p34;
– наличие веб-представительств p35;
– подготовка бакалавров и специалистов p36.
4. Сферы деятельности:
– объем сельскохозяйственного производства p41;
– общий оборот сельскохозяйственной продукции p42;
– число предприятий обрабатывающей промышленности p43;
– объем обрабатывающей промышленности p44;
– транспортно-логистическая отрасль p45;
– объем транспортно-логистической отрасли p46;
– торговля p47;
– объем торговой отрасли p48.

В качестве данных для анализа использованы данные статистических наблюдений [7] за 2015, 2018 и 2020 гг. Кластерный анализ выполняется с использованием пакетов программ Python и пакете анализа RStudio. На первом этапе анализа выполняется нормализация данных с использованием формулы

$$x_{нор} = \frac{x - x_{мин}}{x_{мак} - x_{мин}},$$

где $x_{мин}$ и $x_{мах}$ определяются по каждому факторному признаку исходных таблиц данных. Как результат нормализации получаем четыре фрейма данных для каждой группы показателей в диапазоне [0–1].

Следующим шагом анализа является проведение корреляционного анализа с целью исключения факторных признаков, имеющих высокую корреляционную связь. Для реализации алгоритма воспользуемся пакетом Scipy из набора Python, а также пакет визуализации Seaborn.

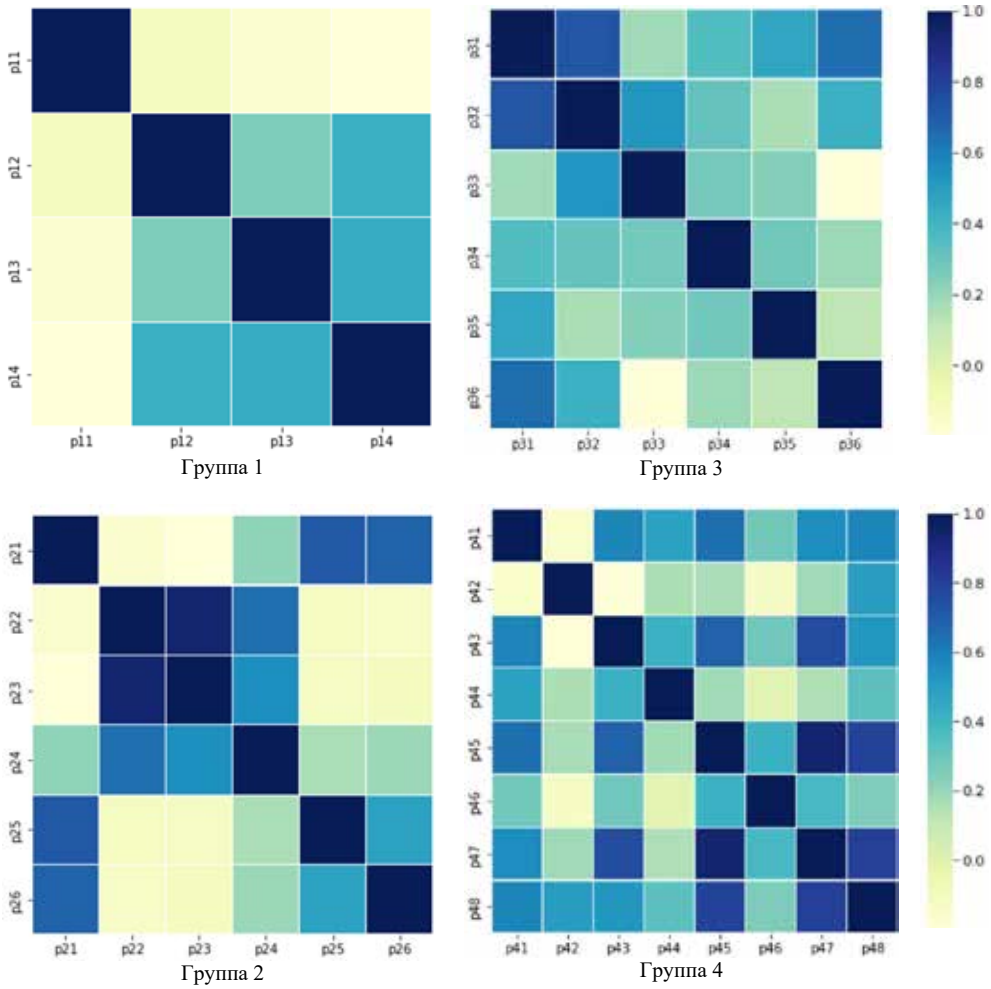

Рис. 3. Результаты корреляционного анализа

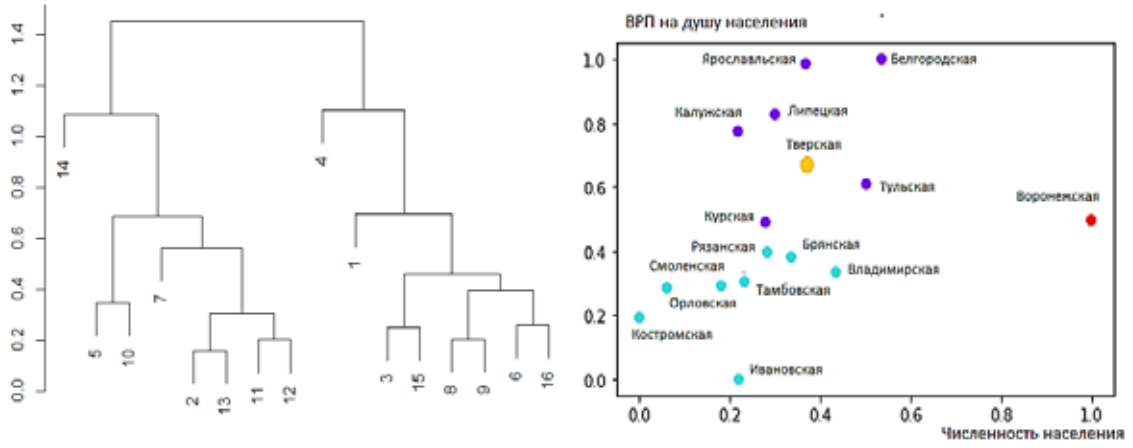

Рис. 4. Дендрограмма агломеративной кластеризации первой группы показателей

Анализ приведенных выше корреляционных диаграмм позволяет сделать следующие выводы. По первой группе факторных признаков высокой корреляционной зависимости не наблюдается. По второй группе наблюдается высокая корреляционная зависимость между фактическим потреблением и доходами на душу населения. Кроме того, достаточно высока связь между уровнем безработицы и числом заболевших на 10000 населения. Для дальнейшего анализа убираем факторные признаки p22 и p25. По третьей группе в силу высокой связи между факторами p31 и p32 оставляем для дальнейшего анализа факторный признак p32. По четвертой группе факторных признаков наблюдаем высокую корреляцию показателей p47 и p48 с остальными признаками, поэтому исключаем их из дальнейшего анализа.

Для проведения кластеризации по каждой группе факторных признаков в данной работе использовался алгоритм k-means в сочетании с агломеративной кластеризацией. На рис. 4 представлена дендрограмма для первой группы показателей.

Проведенные расчеты позволили выбрать наиболее рациональное число кластеров в разбиении. По всем четырем группам решается задача с четырьмя кластерами. Обозначим имена кластеров в каждой из четырех групп через Ai; Bi; Ci; Di, где *i* – номер группы факторных признаков. Проанализируем формирование кластеров в каждой группе факторов и признаки принадлежности к кластеру (табл. 1).

Результаты кластерного анализа по всем группам и периодам сведены в табл. 2.

Проведя комплексный анализ итоговой таблицы, можно проследить динамику социально-экономического развития регионов по рассматриваемым направлениям деятельности. Субъекты ЦФО условно можно распределить на 4 группы.

К первой группе можно отнести Воронежскую область – безусловного лидера по всем направлениям развития.

Вторую группу формируют Тверская, Белгородская, Калужская, Липецкая, Тульская, Ярославская области, имеющие в основном показатели выше среднего и максимальные по отдельным направлениям, обладающие потенциалом для дальнейшего роста и развития. Данные субъекты показывают стабильный рост ВРП. Анализ показывает, что следует усилить работу по реализации комплекса мер по увеличению инновационной составляющей экономики, привлечения инвестиций в экономику регионов.

В третьем кластере расположились Брянская, Владимирская, Курская, Рязанская, Тамбовская и Смоленская области, которые можно охарактеризовать как регионы с неравнозначным и нестабильным развитием. Их можно назвать умеренно стабильными. Основными направлениями деятельности по стабилизации социально-экономической ситуации в данной группе является деятельность по расширению и модернизации промышленного производства, повышению инновационной активности.

Четвертый кластер объединяет Ивановскую, Костромскую и Орловскую области, имеющие средние и ниже среднего показатели по большей части исследуемых факторов. Данный кластер характеризуется статичным или слабым социально-экономическим развитием.

Таблица 1

Характерные признаки объектов кластеризации в каждом кластере

A1	Максимальные значения показателей «численность населения» и «объем инвестиций». Среднее значение ВРП
B1	Наибольшая площадь территории и невысокая плотность населения
C1	Высокая плотность проживания населения. Высокие значения объема инвестиций (выше среднего). Наибольшее значение ВРП
D1	Низкие значения по таким показателям, как ВРП на душу населения и объем инвестиций
A2	Максимальные значения показателей «доход». Низкий уровень безработицы
B2	Уровень дохода, близкий к среднему. Низкий уровень безработицы
C2	Низкий уровень дохода. Высокий уровень безработицы
D2	Максимальные значения числа больничных коек на 10 000 населения.
A3	Максимальные показатели по затратам на научные исследования и количеству обучающихся в вузах. По всем остальным показателям значения выше среднего
B3	Высокие показатели по факторам «затраты на научные исследования» и «использование инновационных производственных технологий», а также значения инновационной активности близкие к среднему и выше
C3	Максимальные значения по показателю «наличие веб-представительств»
D3	Низкие показатели по факторам «затраты на научные исследования» и «использование инновационных производственных технологий»
A4	Лидеры по показателю «объем торговой отрасли», ведущие позиции по показателям «общий оборот сельскохозяйственной продукции» и «число предприятий обрабатывающей промышленности»
B4	Максимальные значения объема сельскохозяйственного производства. Высокие значения по числу предприятий обрабатывающей промышленности. Средние показатели объема торговой отрасли
C4	Высокие значения «объем сельскохозяйственного производства», при этом низкий показатель общего оборота сельскохозяйственной продукции. Высокий объем транспортно-логистической отрасли. Показатели объема торговой отрасли выше среднего значения.
D4	Низкие значения показателя «объем сельскохозяйственного производства». Показатели объема торговой отрасли ниже среднего значения

Таблица 2

Результаты кластеризации 16 объектов по четырем группам и трем периодам

№	Регион	1 группа			2 группа			3 группа			4 группа		
		2015	2018	2020	2015	2018	2020	2015	2018	2020	2015	2018	2020
4	Воронежская	A1	A1	A1	A2	A2	A2	A3	A3	A3	A4	A4	A4
6	Калужская	C1	C1	C1	A2	B2	A2	B3	B3	B3	A4	A4	A4
15	Тульская	C1	C1	C1	A2	A2	B2	C3	B3	B3	A4	A4	A4
1	Белгородская	C1	C1	C1	A2	B2	A2	C3	C3	B3	B4	A4	A4
14	Тверская	B1	B1	B1	C2	C2	C2	B3	B3	B3	A4	B4	B4
9	Липецкая	C1	C1	C1	A2	A2	A2	C3	C3	C3	B4	D4	C3
16	Ярославская	C1	C1	C1	C2	C2	C2	B3	B3	B3	C3	B4	B4
3	Владимирская	D1	D1	C1	D2	D2	D2	C3	B3	B3	A4	B4	B4
8	Курская	C1	C1	C1	A2	A2	B2	D3	D3	D3	D4	D4	C3
2	Брянская	D1	D1	D1	B2	D2	B2	C3	D3	C3	C3	C3	C3
13	Тамбовская	C1	C1	D1	B2	D2	B2	D3	C3	C3	D4	D4	C3
11	Рязанская	D1	D1	D1	B2	D2	D2	C3	C3	C3	A4	D4	D4
12	Смоленская	D1	D1	D1	C2	C2	C2	D3	D3	D3	D4	B4	B4
7	Костромская	B1	D1	B1	D2	C2	C2	D3	D3	D3	D4	D4	D4
10	Орловская	D1	D1	D1	C2	C2	D2	D3	D3	C3	D4	D4	C3
5	Ивановская	D1	D1	D1	D2	D2	D2	D3	D3	C3	D4	D4	D4

Таким образом, использование кластерного анализа позволяет анализировать влияние отдельных факторов на сбалансированность социально-экономического развития региона, что очень важно учитывать при разработке стратегий, программ и формировании социально-экономической политики региона. Кроме того, появляется возможность оценивать в перспективе последствия изменения показателей.

Выводы

По результатам исследования сделаны следующие выводы:

− Методы кластерного анализа являются эффективным инструментом при осуществлении комплексного анализа социально-экономического развития регионов.

− Проведенные исследования позволили выявить как общие особенности, так и различия в показателях исследуемых объектов, что позволяет оптимальным образом выстраивать стратегию и планы развития регионов.

− Анализ распределения регионов по отдельным кластерам за последние пять лет позволил оценить динамику развития регионов.

− Выделенные группы признаков при определении кластеров позволяют дать всестороннюю характеристику регионам.

Список литературы

1. Альбахели В.А. Сегментация магнитно-резонансных изображений на основе кластерного анализа // Тенденции науки и образования в современном мире. 2015. № 5 (5). С. 18–20.

2. Пономарев В.П., Белоглазова И.Ю. Применение факторного и кластерного статистического анализа в медицине // Перспективные информационные технологии: международная научно-техническая конференция. Самара, 26–28 апреля 2016 г. С. 589–592.

3. Савченко Т.Н. Применение методов кластерного анализа для обработки данных психологических исследований // Экспериментальная психология. 2010. Т. 3. № 2. С. 67–86.

4. Моденова А.А., Якимов И.М. Кластерный анализ регионов России по научной и инновационной активности // Научные исследования: от теории к практике. 2015. Т. 2. № 2. С. 69–72.

5. Дегтярева Т.Д., Чулкова Е.А., Торбина Е.С. Исследование дифференциации социального развития сельских территорий // Известия Оренбургского государственного аграрного университета. 2015. № 5. С. 212–216.

6. Фрумина И.Л., Цветкова Е.В. Исследование некоторых проблем аграрной экономики методом кластерного анализа // Известия Челябинского научного центра УРО РАН. 2007. № 4. С. 93–97.

7. Богорсукова Н.Я., Халафян А.А., Ракачев В.Н. Применение кластерного анализа при изучении динамики численности населения районов Краснодарского края // Вестник Северо-Кавказского федерального университета. 2014. № 2 (41). С. 142–146.

8. Овсянников А.О. Анализ внутренних затрат на научные исследования и разработки по субъектам Российской Федерации при помощи кластерного анализа RapidMiner // Научно-практический электронный журнал Аллея Науки. 2018. № 6 (22).

9. Регионы России. Социально-экономические показатели. 2021: Р32 Стат. сб. Росстат. М., 2021. 1112 с.