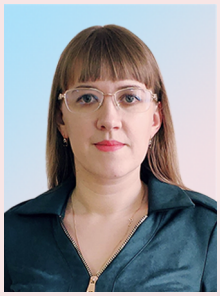
DOI: 10.15838/esc.2021.6.78.4
УДК 004.85, ББК 22.172

Кластеризация регионов Российской Федерации по уровню социально-экономического развития с использованием методов машинного обучения

Каролина Вячеславовна
КЕТОВА
Ижевский государственный технический университет
имени М.Т. Калашникова
Ижевск, Российская Федерация
e-mail: ketova_k@mail.ru
ORCID: 0000-0001-7143-1930; ResearcherID: AAB-9383-2020

Екатерина Васильевна
КАСАТКИНА
Ижевский государственный технический университет
имени М.Т. Калашникова
Ижевск, Российская Федерация
e-mail: e.v.trushkova@gmail.com
ORCID: 0000-0001-6596-0086; ResearcherID: M-6863-2016



Дайана Дамировна
ВАВИЛОВА
Ижевский государственный технический университет
имени М.Т. Калашникова
Ижевск, Российская Федерация
e-mail: vavilova_dd@mail.ru
ORCID: 0000-0002-2161-4402; ResearcherID: AAG-7809-2019

Для цитирования: Кетова К.В., Касаткина Е.В., Вавилова Д.Д. Кластеризация регионов Российской Федерации по уровню социально-экономического развития с использованием методов машинного обучения // Экономические и социальные перемены: факты, тенденции, прогноз. 2021. Т. 14. № 6. С. 70–85. DOI: 10.15838/esc.2021.6.78.4

For citation: Ketova K.V., Kasatkina E.V., Vavilova D.D. Clustering Russian Federation regions according to the level of socio-economic development with the use of machine learning methods. *Economic and Social Changes: Facts, Trends, Forecast*, 2021, vol. 14, no. 6, pp. 70–85. DOI: 10.15838/esc.2021.6.78.4

Аннотация. В работе решена задача кластеризации регионов Российской Федерации по социально-экономическому развитию с учетом отраслевой структуры валового регионального продукта. Инструментом решения задачи кластеризации являются классические методы машинного обучения. Исходная база данных включает реальные статистические данные по социально-экономическому развитию субъектов РФ и отраслевой структуре их валового регионального продукта за 2019 год. Для выявления кластеров регионов по социально-экономическому развитию применены современные методы машинного обучения, реализованные на высокоуровневом языке программирования Python с подключением библиотек для работы с данными: Pandas, Sklearn, SciPy и др. Выполнена предобработка исходной информации: оцифровка категорий данных, переход к удельным величинам, стандартизация показателей. Исходный набор данных за 2019 год содержит 5525 записей по 65 показателям социально-экономического развития 85 регионов РФ. На основе метода главных компонент выделено 15 базовых индикаторов социально-экономического развития региона, по ним методом k-средних определены пять региональных кластеров: первый кластер характеризируется высокой долей в структуре ВРП оптовой и розничной торговли, операций с недвижимым имуществом, профессиональной, научной и технической деятельности; второй кластер специализируется на обрабатывающем производстве, оптовой и розничной торговле, деятельности по операциям с недвижимым имуществом, сельском и лесном хозяйстве; третий можно описать как кластер со смешанной экономикой, для которого характерны средние значения по основным социально-экономическим показателям в РФ; в регионах, относящихся к четвертому кластеру, наблюдается высокий уровень безработицы, при этом выявлена высокая доля государственного управления и обеспечения военной безопасности, социального обеспечения; пятый кластер специализируется на добыче полезных ископаемых.

Ключевые слова: социально-экономические показатели, отраслевая структура, валовой региональный продукт, машинное обучение, кластерный анализ, метод главных компонент.

Введение

В настоящее время разработка эффективной стратегии развития регионов Российской Федерации требует оценки текущего состояния и перспектив изменения их социально-экономического развития. Данная задача является достаточно сложной, особенно при наличии значительных межрегиональных различий в социально-экономическом развитии, финансово-экономических возможностях, инновационном потенциале, качестве человеческого капитала и др. [1; 2; 3]. Одним из функциональных средств для формирования эффективной стратегии развития регионов выступает инструментарий кластерного анализа.

Кластерный анализ представляет собой один из методов многомерного статистического анализа данных, который позволяет выделять некоторые однородные группы объектов по различным параметрам [4; 5]. В целях нашего исследования использование кластерного анализа помогает определить группы российских регионов со схожим уровнем социально-экономического развития. Выявление подобных кластеров – это основа разработки дифференцированных и адресных мер поддержки от государства.

Следует отметить, что в отечественных научных исследованиях регионы выступают наиболее типичными объектами кластеризации и классификации по различным критериальным признакам: инновационное развитие [1], качество жизни [6], рождаемость [7], общественное здоровье [8], уровень человеческого капитала [9], эффективность сельского хозяйства [10], внешнеэкономическая деятельность [11], энергоэффективность [12], степень развития дорожно-транспортной системы [13] и т. д. Эти работы выполнены на основании кластеризации по отдельным показателям. Также существуют исследования, в которых кластеризация регионов осуществлена по совокупности показателей, содержащей 10–15 параметров (см., например, [14]).

Наша работа направлена на решение задачи кластеризации регионов по совокупности показателей, отражающих социально-экономическое развитие субъектов России, а также учитывающих отраслевую специфику развития экономики регионов. В представленном исследовании обрабатывается набор данных, содержащий 65 показателей.

Цель исследования – выделение однородных региональных кластеров методами анализа данных и машинного обучения для разработки платформы принятия правильных форм поддержки регионов, стимулирующих прорывной рост экономики РФ в целом. Для достижения указанной цели следует решить ряд задач, в частности:

– выявить структуру показателей, характеризующих социально-экономическое развитие регионов с учетом отраслевой специфики, за счет формирования укрупненных групп, на основе имеющейся на официальном сайте Федеральной службы государственной статистики информации;

– собрать и проверить качество большого набора исходных данных для проведения кластерного анализа регионов РФ;

– осуществить преданализ данных: заполнение пропусков, преобразование данных (переход к удельным величинам), стандартизация, выделение основных индикаторов в каждой укрупненной группе показателей методом главных компонент;

– выделить однородные региональные кластеры путем применения методов машинного обучения;

– проанализировать дифференциацию средних показателей развития региональных кластеров с целью верификации качества выполненной кластеризации.

Таким образом, научная новизна предложенного исследования заключается в решении задачи кластеризации на основе больших статистических данных, изучаемых в совокупности. Исследование также обладает практической значимостью, поскольку позволяет формулировать особенности социально-экономического развития групп регионов, на основе чего формируется стратегия их развития и политика инвестирования в актуальные на текущий момент сферы жизнедеятельности субъектов РФ.

Методы машинного обучения для решения задачи кластеризации

Машинное обучение (Machine Learning) представляет собой большой раздел из области изучения искусственного интеллекта; включает методы построения различных алгоритмов, способных самообучаться. Как правило, в научной литературе выделяют три группы классических методов машинного обучения, часто используемых для интеллектуального анализа данных [15–18]:

– обучение с учителем (регрессия, классификация);

– обучение без учителя (поиск правил, уменьшение размерности, кластеризация);

– обучение с подкреплением (генетический алгоритм, Q-learning и др.).

На практике для проведения кластеризации применяют следующие алгоритмы и методы машинного обучения [15; 19; 20; 21]:

1) эвристические графовые алгоритмы (алгоритм выделения связных компонент, алгоритм кратчайшего незамкнутого пути, FOREL алгоритм);

2) статистические алгоритмы, основанные на разбиении (метод k-средних (k-means), алгоритм DBSCAN, основанный на плотностях распределений изучаемых характеристик);

3) иерархические методы (агломеративные и дивизионные (алгоритмы CURE, ROCK, Chameleon, метод Варда (Ward clustering));

4) алгоритмы нечеткой кластеризации (FCM, FCS и MM алгоритмы).

Каждая группа методов кластеризации обладает своими преимуществами и недостатками. В частности, статистические алгоритмы, основанные на разбиении, эффективно работают с большими объемами данных, что не всегда можно отметить для графовых методов кластеризации. Алгоритмы нечеткой кластеризации имеют недостаток, заключающийся в невозможности корректного разбиения объектов на кластеры в случае наличия большой дисперсии по разным размерностям элементов [22].

Важным преимуществом в нахождении кластеров произвольной формы обладают иерархические методы, метод k-средних, алгоритм DBSCAN. Кластеризация элементов по указанным методам относится к итеративным методам эталонного типа [23]. Следует отметить,

что для метода *k*-средних и DBSCAN предварительно требуется принять решение о значениях гиперпараметров алгоритмов. Так, для метода *k*-средних необходимо знать число кластерных разбиений; для алгоритма DBSCAN нужно подбирать размер окрестности и минимальное число элементов в ней. Исследователь может принять решения, опираясь на собственную интуицию либо проведя предварительный поиск оптимальных значений необходимых гиперпараметров.

Вместе с этим чаще всего исследователи отдают предпочтение методу *k*-средних, поскольку он обладает такими сильными сторонами, как высокая эффективность при простоте его реализации, достаточный уровень качества выполненной кластеризации и возможность распараллеливания вычислительных процедур [24; 25]. Таким образом, применение данного алгоритма оправдано при работе с большими данными (Big Data) для извлечения новых знаний.

Предобработка исходного набора статистических данных для решения задачи кластеризации регионов

Статистическая информация по основным показателям развития регионов Российской Федерации предоставлена Федеральной службой государственной статистики[1]. Так как за последнее время содержание отчетности Федеральной службы государственной статистики по регионам менялось, как ввиду изменения методологии расчета показателей и общероссийского классификатора видов экономической деятельности, так и трансформаций в политико-территориальном устройстве, в качестве анализируемого периода выбран актуальный период 2015–2019 гг.

Исходный набор данных за 2019 год содержит 5525 записей по 65 показателям социально-экономического развития 85 регионов РФ. Выбранные для анализа и кластеризации регионов показатели приведены в *таблице 1*. Они объединены в укрупненные группы направлений социально-экономического развития. Аналогичный подход использовался в работе [26], в которой было выделено 8 групп показателей развития регионов. В настоящем исследовании определены укрупненные группы в соответствии с внедренными в статистическую практику общероссийскими классификаторами, применяемыми при составлении статистического сборника «Регионы России. Основные социально-экономические показатели».

Таблица 1. Показатели социально-экономического развития региона

Группа	Наименование показателя, единица изменения	Обозначение	Преобразование	Индикатор (главный компонент)
Федеральные округа	Центральный (ЦФО), Северо-Западный (СЗФО), Южный (ЮФО), Северо-Кавказский (СКФО), Приволжский (ПФО), Уральский (УФО), Сибирский (СФО), Дальневосточный (ДФО)	–	Фиктивные переменные	PCA_1
Основные социально-экономические показатели	Численность населения, тыс. чел.	X_1	–	PCA_2 PCA_3
	Стоимость основных фондов, млн руб.	X_2	$Y_1 = X_2/X_1$	
	Добыча полезных ископаемых, млн руб.	X_3	$Y_2 = X_3/X_1$	
	Сельское хозяйство, млн руб.	X_4	$Y_3 = X_4/X_1$	
	Обрабатывающие производства, млн руб.	X_5	$Y_4 = X_5/X_1$	
	Обеспечение электрической энергией, газом и паром; кондиционирование воздуха, млн руб.	X_6	$Y_5 = X_6/X_1$	
	Водоснабжение; водоотведение, организация сбора и утилизации отходов, деятельность по ликвидации загрязнений, млн руб.	X_7	$Y_6 = X_7/X_1$	
	Оборот розничной торговли, млн руб.	X_8	$Y_7 = X_8/X_1$	
	Сальдированный финансовый результат, млн руб.	X_9	$Y_8 = X_9/X_1$	

[1] Регионы России. Социально-экономические показатели. URL: https://rosstat.gov.ru/folder/210/document/13204

Продолжение таблицы 1

Группа	Наименование показателя, единица изменения	Обозначение	Преобразование	Индикатор (главный компонент)
Население	Соотношение мужчин и женщин, на 1000 мужчин приходится женщин	X_{10}	X_{10}	PCA_4 PCA_5
	Доля населения младше трудоспособного возраста, в процентах от общей численности населения	X_{11}	X_{11}	
	Доля населения в трудоспособном возрасте, в процентах от общей численности населения	X_{12}	X_{12}	
	Доля населения старше трудоспособного возраста, в процентах от общей численности населения	X_{13}	X_{13}	
	Общие коэффициенты рождаемости, число родившихся на 1000 человек населения	X_{14}	X_{14}	
	Общие коэффициенты смертности, число умерших на 1000 человек населения	X_{15}	X_{15}	
	Коэффициенты младенческой смертности, число детей, умерших в возрасте до 1 года, на 1000 родившихся живыми	X_{16}	X_{16}	
	Соотношение браков и разводов, на 1000 браков приходится разводов	X_{17}	X_{17}	
Занятость и безработица	Уровень безработицы, %	X_{18}	X_{18}	PCA_6
	Среднегодовая численность занятых, тыс. чел.	X_{19}	$Y_9 = X_{19}/X_1$	
	Потребность в работниках, заявленная работодателями, чел.	X_{20}	$Y_{10} = X_{20}/X_1$	
	Численность работников государственных органов и органов местного самоуправления, чел,	X_{21}	$Y_{11} = X_{21}/X_1$	
Уровень жизни населения	Средняя номинальная начисленная заработная плата работников организаций, руб./мес.	X_{22}	X_{22}	PCA_7
	Среднедушевые денежные доходы населения, руб./мес.	X_{23}	X_{23}	
	Потребительские расходы в среднем на душу населения, руб./мес.	X_{24}	X_{24}	
	Средний размер назначенных пенсий, руб./мес.	X_{25}	X_{25}	
	Жилищный фонд, млн кв. м	X_{26}	$Y_{12} = X_{26}/X_1$	
	Использование свежей воды, млн куб. м	X_{27}	$Y_{13} = X_{27}/X_1$	
Инвестиции	Поступление прямых иностранных инвестиций в РФ, млн руб.	X_{28}	$Y_{14} = X_{28}/X_1$	PCA_8
	Инвестиции в основной капитал, млн руб.	X_{29}	$Y_{15} = X_{29}/X_1$	
	Доля инвестиций в российскую собственность, %	X_{30}	X_{30}	
Образование	Численность воспитанников организаций дошкольного образования, чел.	X_{31}	$Y_{16} = X_{31}/X_1$	PCA_9
	Численность обучающихся общего образования, чел.	X_{32}	$Y_{17} = X_{32}/X_1$	
	Численность студентов, обучающихся по программам подготовки специалистов среднего звена, чел.	X_{33}	$Y_{18} = X_{33}/X_1$	
	Численность студентов бакалавриата, специалитета, магистратуры, чел.	X_{34}	$Y_{19} = X_{34}/X_1$	
	Численность аспирантов, чел.	X_{35}	$Y_{20} = X_{35}/X_1$	
	Численность учителей организаций, осуществляющих образовательную деятельность по программам начального, основного и среднего общего образования, тыс. чел.	X_{36}	$Y_{21} = X_{36}/X_1$	
	Численность профессорско-преподавательского состава организаций, осуществляющих образовательную деятельность по программам бакалавриата, специалитета, магистратуры, чел.	X_{37}	$Y_{22} = X_{37}/X_1$	
Здравоохранение	Численность врачей всех специальностей, тыс. чел.	X_{38}	$Y_{23} = X_{38}/X_1$	PCA_{10}
	Численность населения на одну больничную койку, чел.	X_{39}	X_{39}	
	Заболеваемость у пациентов с диагнозом, установленным впервые в жизни, на 1000 человек населения, чел.	X_{40}	X_{40}	

Окончание таблицы 1

Группа	Наименование показателя, единица изменения	Обозначение	Преобразование	Индикатор (главный компонент)
Культура, отдых и туризм	Численность зрителей театров и число посещений музеев на 1000 человек населения, чел.	X_{41}	$Y_{24} = X_{41}/X_1$	PCA_{11}
	Число спортивных сооружений, ед.	X_{42}	$Y_{25} = X_{42}/X_1$	
	Библиотечный фонд, экз.	X_{43}	$Y_{26} = X_{43}/X_1$	
	Численность российских туристов, обслуженных туристскими фирмами, чел.	X_{44}	$Y_{27} = X_{44}/X_1$	
	Количество зарегистрированных преступлений, ед.	X_{45}	$Y_{28} = X_{45}/X_1$	
Величина и структура валового регионального продукта	Валовой региональный продукт (ВРП), млн руб.	X_{46}	$Y_{29} = X_{46}/X_1$	PCA_{12}
	Отраслевая структура ВРП:			
	Добыча полезных ископаемых, доля	X_{47}	X_{47}	
	Торговля оптовая и розничная; ремонт автотранспортных средств и мотоциклов, доля	X_{48}	X_{48}	
	Деятельность в области информации и связи, доля	X_{49}	X_{49}	
	Деятельность по операциям с недвижимым имуществом, доля	X_{50}	X_{50}	
	Деятельность в области здравоохранения и социальных услуг, доля	X_{51}	X_{51}	
	Деятельность в области культуры, спорта, организации досуга и развлечений, доля	X_{52}	X_{52}	
	Деятельность домашних хозяйств как работодателей, доля	X_{53}	X_{53}	
	Сельское, лесное хозяйство, охота, рыболовство и рыбоводство, доля	X_{54}	X_{54}	PCA_{13}
	Обрабатывающие производства, доля	X_{55}	X_{55}	
	Строительство, доля	X_{56}	X_{56}	
	Деятельность финансовая и страховая, доля	X_{57}	X_{57}	
	Деятельность профессиональная, научная и техническая, доля	X_{58}	X_{58}	
	Государственное управление и обеспечение военной безопасности; социальное обеспечение, доля	X_{59}	X_{59}	
	Образование, доля	X_{60}	X_{60}	
	Водоснабжение; водоотведение, организация сбора и утилизации отходов, деятельность по ликвидации загрязнений, доля	X_{61}	X_{61}	PCA_{14}
	Транспортировка и хранение, доля	X_{62}	X_{62}	
	Деятельность административная и сопутствующие дополнительные услуги, доля	X_{63}	X_{63}	
	Обеспечение электрической энергией, газом и паром; кондиционирование воздуха, доля	X_{64}	X_{64}	PCA_{15}
	Деятельность гостиниц и предприятий общественного питания, доля	X_{65}	X_{65}	

Источник: разработано авторами.

В ходе исследования был выполнен переход к удельным величинам некоторых показателей социально-экономического развития региона. В частности, показатель стоимости основных средств заменяется на удельную величину основных средств на душу населения ($X_2 \rightarrow Y_1$), объем добычи полезных ископаемых в денежном выражении – на удельную величину добытых полезных ископаемых на душу населения ($X_3 \rightarrow Y_2$) и т. п. Однако одного лишь перехода к удельным величинам недостаточно, поскольку результаты кластерного анализа могут быть неадекватны в силу влияния различных единиц измерения величин. С целью приведения всех показателей к единому безразмерному формату и представлению, которое обеспечива-

ет корректное применение многомерной кластеризации, предлагается выполнить их стандартизацию [27]:

$$\tilde{x}_i^j = \frac{x_i^j - \bar{x}_i}{\sigma_{x_i}}, \quad (1)$$

где $\tilde{x}_i^j$ – стандартизированное значение x_i^j-показателя; x_i^j – исходное или удельное значение показателя для j-региона; σ_{x_i} – среднеквадратическое отклонение показателя x_i от его среднего значения по всем регионам; $i = \overline{1,\ 65};\ \ j = \overline{1,\ 85}$.

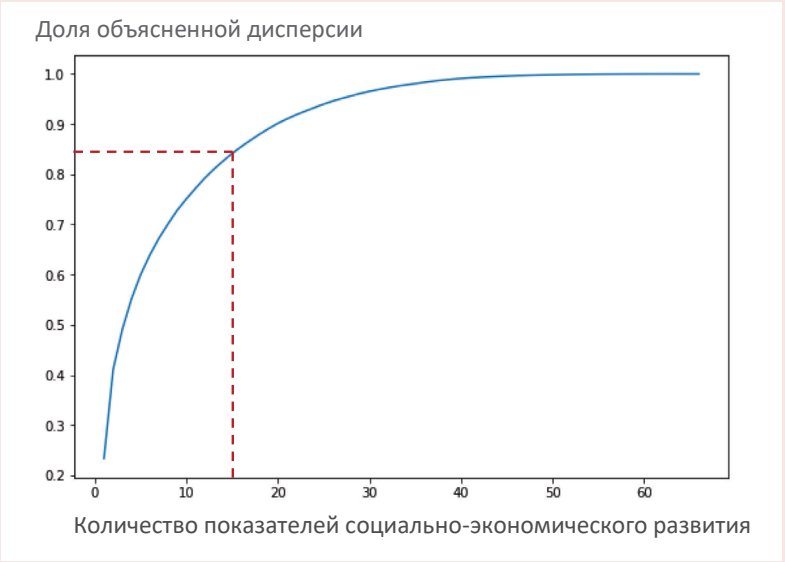
Далее в целях эффективного проведения кластеризации и выделения существенно влияющих на нее признаков предлагается понизить размерность исходного набора данных методом главных компонент (Principal Component Analysis, PCA) [28]. Алгоритм старается найти такие проекции в данных, которые сохраняют максимум дисперсии. Он позволяет снизить размерность, убрать неинформативные признаки и, тем не менее, сохранить способность разделять данные.

Для определения количества необходимых индикаторов (главных компонент), отражающих дифференциацию регионов по социально-экономическому развитию, следует построить график зависимости доли объясненной дисперсии от числа индикаторов. На *рисунке 1* представлена указанная зависимость для решаемой задачи. Она была построена с применением метода PCA, реализованного на языке Python с использованием библиотеки *Sklearn* и функции *decomposition.PCA*()[2].

Из полученного графика дисперсии в направлении собственного вектора, объясняемой каждым из компонентов, следует, что достаточно включить в анализ 15 индикаторов развития, которые будут описывать около 85% дисперсии.

Введем в исследование 15 главных компонент согласно выделенным в таблице 1 укрупненным группам. На *рисунке 2* представлено HeatMap-отображение коэффициентов корреляции Пирсона (библиотека *SciPy*) между удельными основными социально-экономическими показателями и полученными для них PCA-методом главными компонентами PCA_2

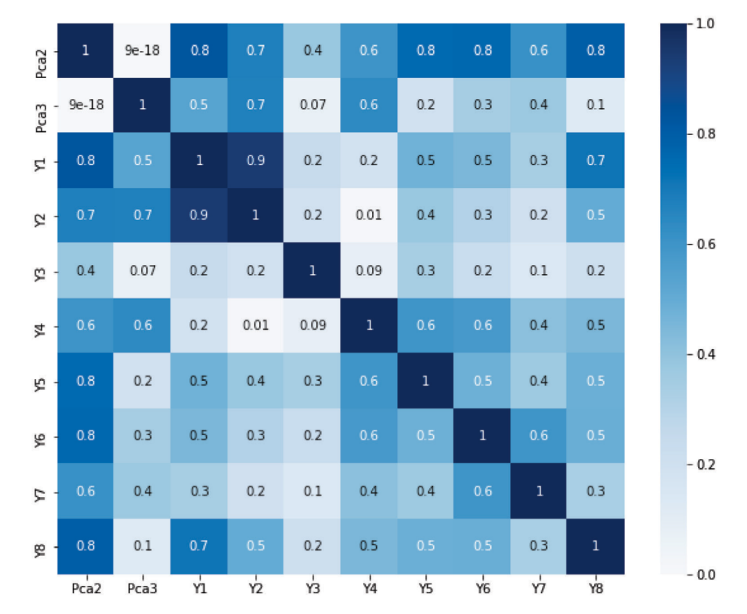
Рис. 1. Зависимость доли объясненной дисперсии от числа показателей социально-экономического развития регионов



Источник: разработано авторами.

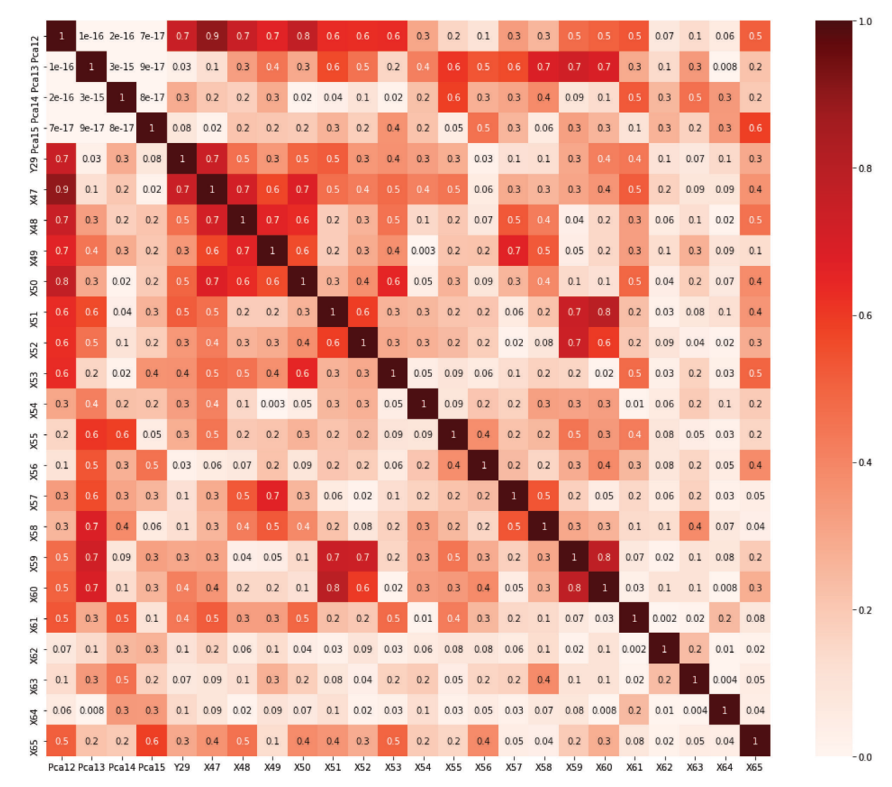
[2] Метод главных компонент. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Рис. 2. HeatMap-отображение коэффициентов корреляции Пирсона между социально-экономическими показателями и их главными компонентами



Источник: разработано авторами.

Рис. 3. HeatMap-отображение коэффициентов корреляции Пирсона между переменными (X_{47}–X_{65}) и главными компонентами (PCA_{12}–PCA_{15})



Источник: разработано авторами.

и PCA_3. По значениям коэффициентов корреляции между показателями видно, что главный компонент PCA_2 отвечает за переменные Y_1, Y_5, Y_6 и Y_8 (парные коэффициенты корреляции находятся в диапазоне 0,7–0,8 и указывают на сильную корреляционную связь), а главный компонент PCA_3 – за переменные Y_2 и Y_4.

На *рисунке 3* дано аналогичное HeatMap-отображение коэффициентов корреляции Пирсона между переменными, характеризующими отраслевую структуру ВРП, и полученными для них главными компонентами PCA_{12}, PCA_{13}, PCA_{14} и PCA_{15}.

Исследование показало, что после перехода к удельным переменным и снижения размерности исходных показателей социально-экономического развития регионов для кластеризации может быть использовано 15 индикаторов. В результате преобразованный набор данных (Dataset) содержит 1275 записей.

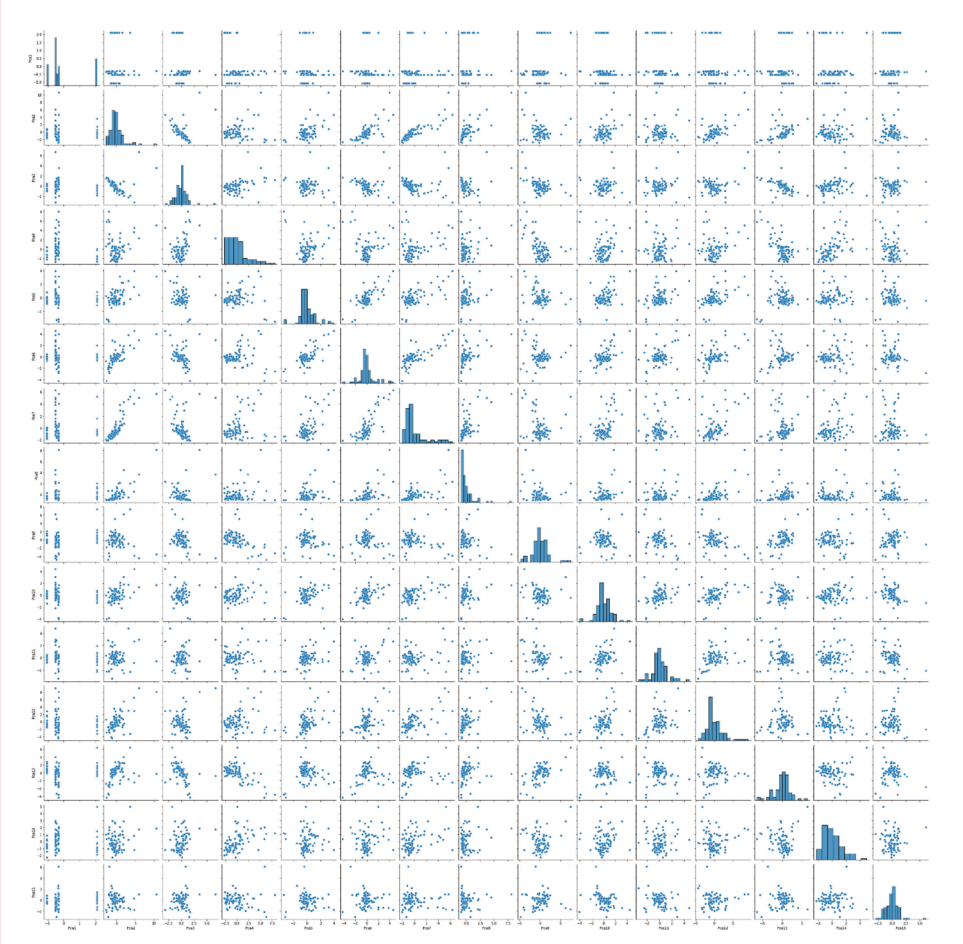
На *рисунке 4* представлены попарные графики корреляций для главных компонент преобразованного набора данных и их гистограммы.

Визуально не наблюдается тесных связей между индикаторами социально-экономического развития регионов, поэтому целесообразно проводить многомерную кластеризацию по всем индикаторам.

Результаты решения задачи кластеризации регионов по уровню социально-экономического развития

Как было сказано выше, к числу эффективных методов кластеризации относится метод *k*-средних, оптимальное число которых определяется исходя из анализа суммарного квадрата расстояний от предполагаемых центров

Рис. 4. Попарные графики индикаторов развития и их гистограммы

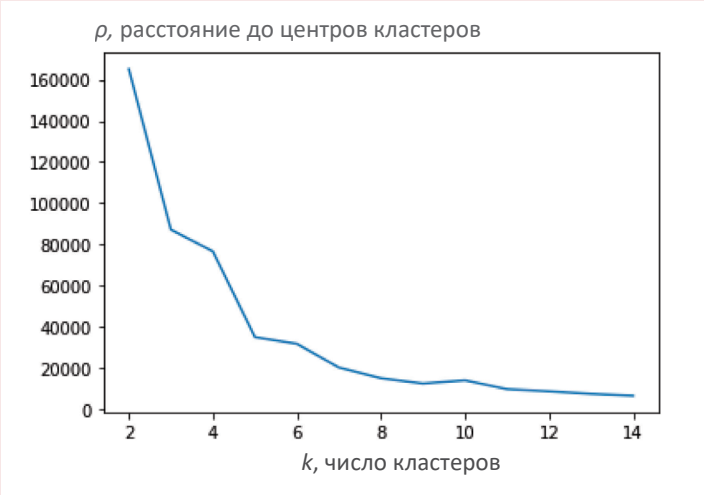


Источник: разработано авторами.

до регионов в кластере *(рис. 5)*. Реализация метода *k*-средних выполнялась на языке Python с использованием библиотеки *Sklearn* и встроенной в ней функции *cluster.KMeans*()[3].

Из графика, представленного на рисунке 5, видно, что при изменении числа кластеров с 4 до 5 резко сокращается суммарное расстояние от центров до объектов кластера, при этом для количества кластеров больше 5 данный показатель уменьшается незначительно; использовать $k > 5$ нецелесообразно (проверка выполнена по критерию Фишера с применением библиотеки *SciPy*). Таким образом, экономически целесообразно выделить 5 региональных кластеров *(табл. 2)*.

Рис. 5. Зависимость расстояния от предполагаемых центров до регионов в кластере и числа кластеров



Источник: разработано авторами.

Таблица 2. Кластеризация регионов по уровню социально-экономического развития с учетом отраслевой структуры

Кластер	Регионы
1	г. Москва, г. Санкт-Петербург
2	Белгородская область, Брянская область, Владимирская область, Воронежская область, Ивановская область, Калужская область, Костромская область, Курская область, Липецкая область, Московская область, Орловская область, Рязанская область, Смоленская область, Тамбовская область, Тверская область, Тульская область, Ярославская область
3	Алтайский край, Амурская область, Архангельская область, Астраханская область, Волгоградская область, Вологодская область, Еврейская автономная область, Забайкальский край, Иркутская область, Калининградская область, Камчатский край, Кемеровская область, Кировская область, Краснодарский край, Красноярский край, Курганская область, Ленинградская область, Мурманская область, Нижегородская область, Новгородская область, Новосибирская область, Омская область, Оренбургская область, Пензенская область, Пермский край, Приморский край, Псковская область, Республика Адыгея, Республика Башкортостан, Республика Карелия, Республика Коми, Республика Марий Эл, Республика Мордовия, Республика Татарстан, Республика Хакасия, Ростовская область, Самарская область, Саратовская область, Свердловская область, Ставропольский край, Томская область, Тюменская область, Удмуртская Республика, Ульяновская область, Хабаровский край, Челябинская область, Чувашская Республика
4	Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Алтай, Республика Бурятия, Республика Дагестан, Республика Ингушетия, Республика Калмыкия, Республика Крым, Республика Северная Осетия – Алания, Республика Тыва, Чеченская Республика, г. Севастополь
5	Магаданская область, Ненецкий автономный округ, Республика Саха (Якутия), Сахалинская область, Ханты-Мансийский автономный округ – Югра, Чукотский автономный округ, Ямало-Ненецкий автономный округ
Источник: расчеты авторов.	

[3] Машинное обучение. Кластеризация. KMeans. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

В первый кластер вошли города федерального значения Москва и Санкт-Петербург, во второй кластер – регионы только из ЦФО. Третий кластер на 30% состоит из субъектов ПФО, 20% – СЗФО, 17% – СФО, 13% – ДФО, 11% – ЮФО, 9% – УФО. В четвертом кластере 50% занимают субъекты СКФО, 25% – ЮФО, 17% – СФО, 8% – ДФО. Пятый кластер на 57% состоит из регионов ДФО, 29% – УФО, 14% – СЗФО.

Цветограмма кластерного распределения субъектов РФ по уровню социально-экономического развития представлена на *рисунке 6*.

В *таблице 3* приведены значения различных показателей, характеризующих социально-экономическое развитие регионов, в среднем по выделенному кластеру.

По анализу данных об уровне социально-экономического развития регионов с учетом отраслевой структуры можно сделать выводы, что:

– первый кластер характеризируется высокой долей в структуре ВРП оптовой и розничной торговли, высокой долей операций с недвижимым имуществом, профессиональной, научной и технической деятельности, отрасли информации и связи; для этого кластера характерна высокая доля занятых в экономике, низкий уровень безработицы, а также высокие среднедушевые денежные доходы и расходы;

– второй кластер специализируется на обрабатывающем производстве, оптовой и розничной торговле, деятельности по операциям с недвижимым имуществом, сельском и лесном хозяйстве;

– третий кластер, содержащий наибольшее количество регионов, можно описать как кластер со смешанной экономикой, для которого характерны средние значения по основным социально-экономическим показателям в РФ;

– четвертый кластер характеризуется низкими значениями социально-экономических показателей; в его регионах наблюдается высокий уровень безработицы, доля занятых в экономике составляет всего 38%; в отличие от дру-

Таблица 3. Средние значения ряда показателей по кластерам за 2019 год

Показатель	Кластер 1	Кластер 2	Кластер 3	Кластер 4	Кластер 5
Удельный ВРП (Y_{29}), тыс. руб./чел.	1250,5	430,7	496,8	238,4	3290,5
Удельная стоимость основных фондов (Y_1), тыс. руб./чел.	4670,4	1766,7	1827,3	904,2	11119,5
Доля занятых в общей численности населения, % (Y_9)	64,5	46,0	45,4	38,3	64,8
Уровень безработицы (X_{18}), %	1,4	4,0	5,2	11,6	4,7
Среднедушевые денежные доходы населения (X_{23}), руб./мес.	60611,0	29548,4	29074,0	22173,3	67221,7
Потребительские расходы в среднем на душу населения (X_{24}), руб./мес.	48040,0	24114,8	23720,0	16873,8	36079,6
Доля отрасли по добыче полезных ископаемых в ВРП (X_{47}), %	0,2	2,1	10,3	2,9	60,3
Доля оптовой и розничной торговли в ВРП (X_{48}), %	20,7	13,8	10,8	12,2	3,8
Доля отрасли в области информации и связи в ВРП(X_{49}), %	6,2	2,1	2,0	2,1	0,6
Доля деятельности по операциям с недвижимым имуществом в ВРП (X_{50}), %	14,8	11,1	9,4	11,3	2,1
Доля сельского и лесного хозяйства, охоты, рыболовства и рыбоводства в ВРП (X_{54}), %	0,1	10,1	7,3	11,5	2,0
Доля обрабатывающих производств в ВРП (X_{55}), %	14,8	22,5	19,0	5,1	1,6
Доля строительства в ВРП (X_{56}), %	3,6	5,7	5,7	9,1	6,6
Доля профессиональной, научной и технической деятельности в ВРП (X_{58}), %	8,2	2,6	2,7	1,2	0,9
Доля государственного управления и обеспечения военной безопасности, социального обеспечения (X_{59}), %	5,2	6,3	7,2	15,1	5,3
Доля образования в ВРП (X_{60}), %	2,9	3,7	3,6	7,2	2,3
Источник: расчеты авторов.					

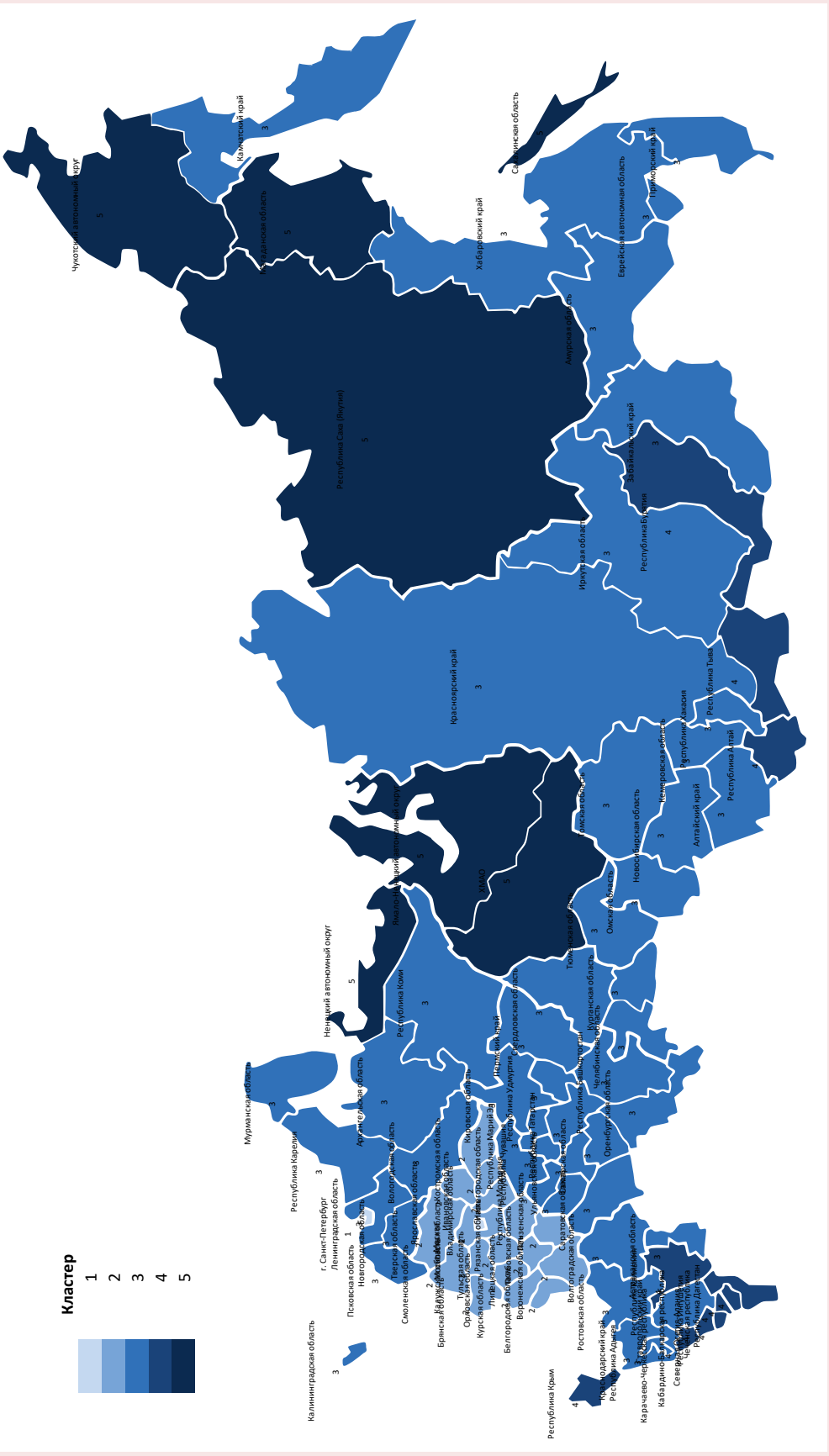


Рис. 6. Цветограмма кластерного распределения субъектов РФ по уровню социально-экономического развития

Источник: разработано авторами.

гих, в четвертом кластере выявлена высокая доля государственного управления, и обеспечения военной безопасности, социального обеспечения, образования, строительства;

– пятый кластер специализируется на добыче полезных ископаемых, для его регионов характерны максимальные среднедушевые денежные доходы населения в РФ.

Заключение

В ходе исследования выполнена кластеризация регионов России согласно уровню их социально-экономического развития и отраслевой структуре валового регионального продукта. Для осуществления кластерного анализа применялись такие методы машинного обучения без учителя, как методы главных компонент и *k*-средних.

В исходный набор данных вошли показатели развития регионов по укрупненным группам в соответствии с внедренными в статистическую практику классификаторами: основные социально-экономические показатели; население; занятость и безработица; уровень жизни населения; инвестиции; образование; здравоохранение; культура, отдых и туризм; величина и структура валового регионального продукта. Для показателей каждой укрупненной группы с применением метода главных компонент выявлены характерные индикаторы, за счет чего удалось снизить размерность исходного набора данных с 65 показателей до 15 индикаторов.

Было установлено, что целесообразно выделить пять региональных кластеров по уровню социально-экономического развития с учетом отраслевой структуры субъектов.

Методом *k*-средних получено, что первый кластер включает наиболее развитые города РФ: Москву и Санкт-Петербург. Для него характерна высокая доля занятых в экономике, низкий уровень безработицы, высокие среднедушевые денежные доходы и расходы. На территории кластера развита профессиональная научная и техническая деятельность, активно функционирует отрасль информации и связи, развита торговля.

Второй кластер содержит регионы Центрального федерального округа. Здесь присутствует развитое обрабатывающее производство, сельское и лесное хозяйство, торговля.

Третий кластер, наибольший по количеству регионов, состоит из субъектов Приволжского федерального округа (30% в структуре кластера), Северо-Западного федерального округа (20%), а также Сибирского (17%), Дальневосточного (13%), Южного (11%) и Уральского (9%) федеральных округов. Это кластер со смешанной экономикой, для которого характерны средние значения по основным социально-экономическим показателям в РФ.

Четвертый кластер содержит наименее развитые регионы РФ и, соответственно, характеризуется низкими значениями социально-экономических показателей. Наполовину состоит из субъектов Северо-Кавказского федерального округа, четверть – субъектов Южного федерального округа, 17% – Сибирского и 8% – Дальневосточного округов. В этом кластере присутствует высокая доля государственного управления и обеспечения военной безопасности, социального обеспечения, образования, строительства. Для его объектов характерен высокий уровень безработицы, доля занятых в экономике составляет всего 38%.

Пятый кластер специализируется на добыче полезных ископаемых. В него входят районы Дальневосточного (57 %), Уральского (29%), Северо-Западного (14%) округов. Для регионов пятого кластера характерны максимальные среднедушевые денежные доходы населения в РФ.

Таким образом, разработанная методика проведения кластерного анализа позволяет сформировать устойчивые региональные кластеры согласно социально-экономическому развитию субъектов РФ. Выполненная кластеризация, учитывающая отраслевую структуру экономики регионов, может использоваться при реализации кластерно-ориентированной государственной политики с целью поддержки ускоренного развития субъектов.

Литература

1. Golova I.M., Sukhovey A.F. Differentiation of innovative development strategies considering specific characteristics of the Russian regions. *Economy of Region,* 2019, vol. 15, pp. 1294–1308. DOI: 10.17059/2019-4-25

2. Mariev O., Pushkarev A. Clustering Russian regions by innovative outputs using a multi indicator approach. In: *Proceedings of the 7th International Conference Innovation Management, Entrepreneurship and Sustainability (IMES)*, 2019. Pp. 519–533.
3. Кетова К.В., Вавилова Д.Д. Оценка тенденций изменения человеческого капитала социально-экономической системы на основе применения алгоритма нейросетевого прогнозирования // Экономические и социальные перемены: факты, тенденции, прогноз. 2020. Т. 13. Вып. 6. С. 117–133. DOI: 10.15838/esc.2020.6.72.7
4. Shubat O.M., Bagirova A.P., Akishev A.A. Methodology for analyzing the demographic potential of Russian regions using fuzzy clustering. *Economy of Region*, vol. 15, pp. 178–190. DOI: 10.17059/2019-1-14
5. Кетова К.В., Трушкова Е.В. Решение логистической задачи топливоснабжения распределенной региональной системы теплоснабжения // Компьютерные исследования и моделирование. 2012. Т. 4. № 2. С. 451–470.
6. Локосов В.В., Рюмина Е.В., Ульянов В.В. Кластеризация регионов России по показателям качества жизни и качества населения // Народонаселение. 2019. Т. 22. № 4. С. 4–17.
7. Костина С.Н., Трынов А.В. Кластерный анализ динамики рождаемости четвертых и последующих детей в регионах Российской Федерации // Экономические и социальные перемены: факты, тенденции, прогноз. 2021. Т. 14. № 3. С. 232–245. DOI: 10.15838/esc.2021.3.75.14
8. Лавриненко П.А., Рыбакова Д.А. Сравнительный анализ региональных различий в сферах здоровья населения, экологии и здравоохранения // Экономические и социальные перемены: факты, тенденции, прогноз. 2015. № 5 (41). С. 198–210.
9. Петрыкина И.Н. Кластерный анализ регионов Центрального федерального округа по уровню развития человеческого капитала // Вестник Воронежского государственного университета. Экономика и управление. 2013. № 1. С. 72–80.
10. Демичев В.В., Маслакова В.В., Нестратова А.А. Кластеризация регионов России по уровню эффективности сельского хозяйства // Бухучет в сельском хозяйстве. 2020. № 12. С. 58–66. DOI: 10.33920/sel-11-2012-06
11. Аксенов И.А. Кластеризация внешнеэкономической деятельности регионов // Экономика и менеджмент систем управления. 2016. № 1-3. С. 309–315.
12. Марченко Е.М., Белова Т.Д. Кластеризация регионов с учетом показателей энергоэффективности // Региональная экономика: теория и практика. 2016. № 1 (424). С. 51–60.
13. Paul S., Alvi A.M., Nirjhor M.A., Rahman S., Orcho A.K., Rahman R.M. Analyzing accident prone regions by clustering. *Advanced Topics in Intelligent Information and Database Systems*, 2017, vol. 710, pp. 3–13.
14. Орлова И.В., Филонова Е.С. Кластерный анализ регионов Центрального федерального округа по социально-экономическим и демографическим показателям // Статистика и экономика. 2015. № 5. С. 111–115. DOI: 10.21686/2500-3925-2015-5-136-142
15. Ultsch A., Lotsch J. Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics*, 2017, vol. 66, pp. 95–104. DOI: 10.1016/j.jbi.2016.12.011
16. Khan I., Luo Z., Shaikh A.K., Hedjam R. Ensemble clustering using extended fuzzy k-means for cancer data analysis. *Expert Systems with Applications*, 2021, vol. 172, 114622. DOI: 10.1016/j.eswa.2021.114622
17. Ming F., Stephen T.A Machine learning based asset pricing factor model comparison on anomaly portfolios. *Economics Letters*, 2021, vol. 204, 109919. DOI: 10.1016/j.econlet.2021.109919
18. Blekanov I., Krylatov A., Ivanov D., Bubnova Y. Big data analysis in social networks for managing risks in clothing industry. *IFAC PapersOnLine*, 2019, vol. 52 (13), pp. 1710–1714. DOI: 10.1016/j.ifacol.2019.11.447
19. Arthur D., Vassilvitskii S. K-means++: The advantages of careful seeding. In: *Conference: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, *SODA 2007, New Orleans, Louisiana, USA*. DOI: 10.1145/1283383.1283494
20. Ozgur O., Akkoc U. Inflation forecasting in an emerging economy: Selecting variables with machine learning algorithms. *International Journal of Emerging Markets*, 2020. DOI: 10.1108/IJOEM-05-2020-0577
21. Faizullin R.V. Simulator of the navigation equipped with LIDAR of the mobile robot based on the neural network. *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 873, no. 1. DOI: 10.1088/1757-899X/873/1/012023

22. De Sousa J.M., Santos R.L.D., Lopes L.A., Machado V.P., Silva I.S. Automatic labelling of clusters with discrete and continuous data using supervised machine learning. In: *Proceedings of the 35th International Conference of the Chilean Computer Science Society (SCCC)*. 2016.
23. Lee C.H., Steigerwald D.G. Inference for clustered data. *Stata Journal*, 2018, vol. 18, no. 2, pp. 447–460. DOI: 10.1177/1536867X1801800210
24. Mitra D., Chu Y., Cetin K. Cluster analysis of occupancy schedules in residential buildings in the United States. *Energy and Buildings*, 2021, vol. 236, 110791. DOI: 10.1016/j.enbuild.2021.110791
25. Ofetotse E.L., Essah E.A., Yao R. Evaluating the determinants of household electricity consumption using cluster analysis. *Journal of Building Engineering*, 2021, vol. 43, 102487. DOI: 10.1016/j.jobe.2021.102487
26. Aivazian S., Afanasiev M., Kudrov A. Indicators of the main directions of socio-economic development in the space of characteristics of regional differentiation. *Applied Econometrics*, 2019, vol. 54, pp. 51–69. DOI: 10.24411/1993-7601-2019-10003
27. Касаткина Е.В., Вавилова Д.Д. Информационно-аналитическая система прогнозирования обобщающих показателей социально-экономического развития региона // Проблемы управления. 2015. № 4. С. 25–34.
28. Omuya E.O., Okeyo G.O., Kimwele M.W. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 2021, vol. 174, 114765. DOI: 10.1016/j.eswa.2021.114765

Сведения об авторах

Каролина Вячеславовна Кетова – доктор физико-математических наук, профессор, Ижевский государственный технический университет имени М.Т. Калашникова (426069, Российская Федерация, Удмуртская Республика, г. Ижевск, ул. Студенческая; д. 7; e-mail: ketova_k@mail.ru)

Екатерина Васильевна Касаткина – кандидат физико-математических наук, доцент, Ижевский государственный технический университет имени М.Т. Калашникова (426069, Российская Федерация, Удмуртская Республика, г. Ижевск, ул. Студенческая; д. 7; e-mail: e.v.trushkova@gmail.com)

Дайана Дамировна Вавилова – старший преподаватель, Ижевский государственный технический университет имени М.Т. Калашникова (426069, Российская Федерация, Удмуртская Республика, г. Ижевск, ул. Студенческая; д. 7; e-mail: vavilova_dd@mail.ru)

Ketova K.V., Kasatkina E.V., Vavilova D.D.

Clustering Russian Federation Regions According to the Level of Socio-Economic Development with the Use of Machine Learning Methods

Abstract. The paper solves the problem of clustering Russian Federation regions according to their socio-economic development, taking into account the sectoral structure of the gross regional product. Classical machine learning methods are a tool for solving the clustering problem. The object of the study is the differentiation of regions according to various socio-economic indicators. The subject of the study is the practice of using machine learning methods for clustering objects. The initial database for solving the problem of clustering regions includes actual statistical data on socio-economic development of RF constituent entities and the sectoral structure of their gross regional product as of 2019. We identify clusters of regions according to their socio-economic development with the use of modern machine learning methods implemented in Python, a high-level programming language, with the connection of libraries for working with data: Pandas, Sklearn, SciPy, etc. The preprocessing of the initial data was carried out: digitization of data categories, transition to specific values, standardization of indicators. The initial data set for 2019 contains 5,525 records on 65 indicators of socio-economic development for 85 regions of the Russian Federation. It identifies 15 basic indicators of socio-economic development of a

region, based on the principal component analysis. According to these indicators, five regional clusters were identified with the use of the k-means clustering: the first cluster is characterized by a high share of wholesale and retail trade, real estate transactions, professional, scientific and technological activities in the GRP structure; the second cluster specializes in manufacturing, wholesale and retail trade, real estate transactions, agriculture and forestry; the third cluster can be described as a cluster with a mixed economy, which is characterized by averages for the main socio-economic indicators in the Russian Federation; regions of the fourth cluster show a high level of unemployment and a high share of public administration, military and social security; the fifth cluster specializes in mining.

Key words: socio-economic indicators, industry structure, gross regional product, machine learning, cluster analysis, principal component analysis.

Information about the Authors

Karolina V. Ketova – Doctor of Sciences (Physics and Mathematics), Professor, professor of department, Kalashnikov Izhevsk State Technical University (7, Studen'cheskaya Street, Izhevsk, Udmurt Republic, 426069, Russian Federation; e-mail: ketova_k@mail.ru)

Ekaterina V. Kasatkina – Candidate of Sciences (Physics and Mathematics), Associate Professor, associate professor of department, Kalashnikov Izhevsk State Technical University (7, Studen'cheskaya Street, Izhevsk, Udmurt Republic, 426069, Russian Federation; e-mail: e.v.trushkova@gmail.com)

Diana D. Vavilova – Master of Applied Mathematics, Senior Lecturer, Kalashnikov Izhevsk State Technical University (7, Studen'cheskaya Street, Izhevsk, Udmurt Republic, 426069, Russian Federation; e-mail: vavilova_dd@mail.ru)

Статья поступила 31.08.2021.