УДК: 339.9

ГИЧИЕВ НАБИЮЛА САПИЮЛАЕВИЧ
к.э.н., ведущий научный сотрудник
Института социально-экономических исследований ДФИЦ РАН,
e-mail: nabi-05@mail.ru

КЛАСТЕРНЫЙ АНАЛИЗ В ЭКОНОМИКЕ: ТЕОРЕТИЧЕСКИЙ АСПЕКТ

Аннотация. Цель работы. *Представить критический обзор научной литературы по кластерному анализу, выявить предпочтительность и недостатки различных методов оценкии алгоритмов кластеризации для повышения эффективности управления устойчивым социально-экономическим развитием региона.* ***Методы исследования.*** *В работе представлены общая процедура и этапы кластерного анализа, алгоритмы кластеризации, иерархическая кластеризация, кластеризация взвешенных K-средних, алгоритм Ллойда.* ***Результаты.*** *В статье рассматриваются актуальные проблемы генезиса методов кластерного анализа, позволяющего обосновать необходимость формирования и развития территориальных кластеров – востребованного сегодня метода обеспечения устойчивого регионального развития. Обзор представленной научной литературы выявил сильные и слабые стороны алгоритмов кластеризации, иерархического и других методов кластерного анализа, повышающих уровень качества и достоверности статистической информации при оценке и анализе эффективности государственного управления устойчивостью социально-экономических явлений и процессов, происходящих в российских регионах. Рассмотренные подходы к развитию методов кластеризации позволяют решить проблему возникновения кластера, зарождения отраслевых агломераций и объяснить процесс превращения в «критическую массу» предприятий, компаний и учреждений, необходимый для функционирования кластера.* ***Область применения результатов.*** *Разнообразие методов кластерного анализа свидетельствует об отсутствии единственно верного подхода к их практической операционализации. Полученные результаты компаративной оценки методов кластерного анализа следует учитывать при разработке программ и стратегий регионального развития. Отдельные положения кластерного анализа экономических процессов могут быть востребованы со стороны органов исполнительной власти региона при формировании разрабатываемой «Стратегии социально-экономического развития Республики Дагестан до 2035 г.».* ***Выводы.*** *Кластерный анализ является универсальным инструментом моделирования направлений социально-экономического развития. Результаты кластерного анализа представляются в наглядной форме, облегчающей принятие решений по определению оптимального числа факторов и взаимосвязи различных кластеров.*

Ключевые слова: *кластер, процедура кластерного анализа, алгоритм кластеризации, иерархическая кластеризация, кластеризация взвешенных K-средних, алгоритм Ллойда.*

GICHIEV NABIYULLA SAPIYULAEVICH
Ph. D. in Economics, leading researcher At the Institute
of socio-economic research of the Russian Academy of Sciences,
e-mail: nabi-05@mail.ru

CLUSTER ANALYSIS IN ECONOMICS: THEORETICAL ASPECT

Abstract. Purpose of work. *To present a critical review of the scientific literature on cluster analysis, to identify the preferences and disadvantages of various evaluation methods and clustering algorithms for improving the effectiveness of managing sustainable socio-economic development of the region.* ***Method of research.*** *The paper presents the General procedure and stages of cluster analysis, clustering algorithms, hierarchical clustering, weighted K-means clustering, and*

Lloyd's algorithm. ***Results.*** *The article deals with the current problems of the Genesis of cluster analysis methods, which allows us to justify the need for the formation and development of territorial clusters — a method of ensuring sustainable regional development that is in demand today. The review of the presented scientific literature revealed the strengths and weaknesses of clustering algorithms, hierarchical and other methods of cluster analysis that increase the level of quality and reliability of statistical information in assessing and analyzing the effectiveness of public management of sustainability of socio-economic phenomena and processes occurring in Russian regions. The considered approaches to the development of clustering methods allow us to solve the problem of cluster formation, the emergence of industry agglomerations and explain the process of transformation into a" critical mass " of enterprises, companies and institutions necessary for the functioning of the cluster.* ***The scope of the results.*** *The variety of cluster analysis methods indicates that there is no single correct approach to their practical operationalization. The results of comparative evaluation of cluster analysis methods should be taken into account when developing regional development programs and strategies. Certain provisions of the cluster analysis of economic processes may be in demand from the Executive authorities of the region when forming the "Strategy of socio-economic development of the Republic of Dagestan until 2035".* ***Conclusions.*** *Cluster analysis is a universal tool for modeling the directions of socio-economic development. The results of cluster analysis are presented in a visual form that facilitates decision-making to determine the optimal number of factors and the relationship of various clusters.*
Keywords: *cluster, cluster analysis procedure, clustering algorithm, hierarchical clustering, weighted K-means clustering, Lloyd's algorithm.*

Введение. В научной литературе кластеры рассматриваются как органические явления, возникающие из случайных событий, но многие политики и практики экономического развития склонны рассматривать кластеры как организованные структуры. Согласно этому технократическому пониманию, кластеры «создаются», а не воспринимаются как эмпирические явления. Следовательно, нет никакого различия между органическими кластерами, с одной стороны, и кластерными инициативами, кластерными организациями, кластерным управлением – все они в равной степени называются «кластерами». Такая практическая установка таит в себе опасность игнорирования основных сил агломерации и кластерной эволюции, попытки заменить их символической политикой и импульсивными действиями.

Разрыв между эволюционной перспективой кластеров, превалирующей в научных исследованиях, и технократическим пониманием, доминирующим в политике и практике, отражает особую рациональность в их функционировании, что серьезно затрудняет конструктивное взаимодействие не только между политико-административной и академической системой, но и адресатами кластерной политики в бизнесе и исследовательских организациях. Этим объясняется актуальность продолжающихся научных исследований данной проблематики, и трудности, с которыми сталкиваются многие политические инициативы, пытающиеся мобилизовать участие бизнеса.

Обзор литературы. Тематические исследования показывают, как в последнее время наблюдается ускоряющаяся тенденция развития кластерного анализа. Рассмотрим кластеризациюна основе взвешенных K-средних. Данный метод кластерного анализа (кластеризация K-средних), основанный на алгоритме секционной кластеризации, широко используется, поскольку его легко реализовать и интерпретировать. Взвешенная кластеризация K-средних является расширением традиционной кластеризации K-средних, при которой значительное улучшение производительности процесса кластерного анализа обеспечивается путем введения разнородных переменных весов при выполнении кластеризации K-средних [52, с. 23, 2247–2255].

Кластерный анализ или кластеризация – это общее обозначение множества процедур, разработанных для неконтролируемой классификации. Кластерный анализ определяет и классифицирует объекты по разным группам, так называемым «кластерам», на основании сходства или несходства набора характеристик, точнее, разбиения набора данных на подмножества, так чтобы данные в каждом подмножестве обладали некоторыми общими чертами. Результат кластерного анализа – это ряд групп, в которых есть существенные различия между группами, но

и сильное сходство внутри группы. К числу первых исследователей проблем кластерного анализа можно отнести R.C. Tryon [50], который использовал метод индивидуальных различий в исследовании психологии. Позднее, с середины 1950-х гг. R.C. Tryon использовал кластерный анализ применительно к социальной сфере. Иерархическая кластеризация [54, с. 58, 236–244] и секционная кластеризация [43, с. 4] – два основных типа кластерного анализа. В настоящее время кластерный анализ – очень важный и полезный метод анализ данных, широко используемый во многих областях, таких, как машинное обучение, интеллектуальный анализ, распознавание образов, анализ изображений, поиск документов и биоинформатика.

Универсальный кластерный анализ обычно включает следующие пять шагов [26]: 1) представление шаблона; 2) определение меры сходства / несходства, соответствующей области данных; 3) процесс кластеризации по заданному алгоритму; 4) абстракция данных; 5) проверка вывода. Следует отметить, что шаги 4 и 5, описанные выше, не являются обязательными для кластерного анализа.

Кластерный анализ – это исследовательский инструмент, за которым обычно следуют другие аналитические процедуры. Как правило, алгоритм кластеризации относится к первым трем шагам. На рис. 1 показана типичная последовательность первых трех шагов [27, с. 31, 264–323].

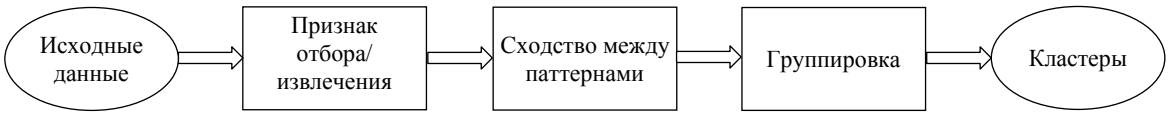

Рис. 1. *Алгоритм кластеризации*

После определения или измерения подобия следующий шаг процедуры кластеризации заключается в выборе правильного алгоритма кластеризации. В научной литературе предложены сотни алгоритмов кластеризации [27, с. 31, 264–323; 45, с. 59, 1–34]. Однако, как упоминалось ранее, большинство алгоритмов относятся к двум основным классам: иерархической кластеризации и секционной кластеризации.В иерархической кластеризации создается серия вложенных разделов, когда, выполняя иерархическую кластеризацию, пользователям не нужно с начала определять номер кластера. Вместо этого можно выбрать лучший раздел плюс соответствующий номер кластера. В иерархической кластеризации можно также выбрать либо агломеративный (снизу вверх) подход, либо разделительный (сверху вниз) подход в сочетании с различными способами измерения центров / расстояний кластеров, такими, как одиночный алгоритм связи, алгоритм полной связи или алгоритм средней связи [54, с. 58, 236–244; 33, с. 62, 86–101].

Алгоритмы частичной кластеризации обычно производят путем оптимизации определенной целевой функции. Алгоритм разбиения может быть классифицирован как жесткий алгоритм (каждый объект может быть размещен только в одном кластере), алгоритм K-средних [34, с. 281–297]; или мягкий алгоритм, когда каждый объект может быть отнесен к нескольким кластерам со степенями / вероятностями членства, например, нечеткая кластеризация C-средних (FCM) [41, с. 15, 22–32]. В то время как иерархическая кластеризация имеет только агломеративный алгоритм. Разделенная кластеризация имеет различные варианты алгоритма [27, с. 31, 264–323]. Без каких-либо предположений о распределении можно применить непараметрический подход, основанный на плотности кластеризации [26], например, кластеризацию ближайшего соседа [26]. При смешанном гауссовском предположении можно применить алгоритм максимизации ожидания (EM) [13, с. 39, 1–38] и алгоритм кросс-энтропии (CE) [7, с. 517–523]. Относительно недавно был разработан метод кластеризации подпространства корреляционной кластеризации специально для многомерных данных, чтобы справиться с проблемой размерности [3, с. 11, –-33].

Таким образом, К-кластеризация – это простейший и наиболее часто используемый алгоритм секционной кластеризации.Если цель исследования состоит лишь в простом определении количества кластеров или структуры набора данных, тогда часть набора данных является конечным продуктом, и нет необходимости в процедуре абстракции данных или проверке кластера. Однако в большинстве реальных приложений кластерный анализ обычно является пер-

вым шагом для изучения данных, а затем другие статистические методы или методы анализа данных применяются либо к каждому кластеру отдельно, либо к центрам кластеров. В кластерном анализе метод абстракции данных используется для общего представления каждого кластера. Самый популярный способ – метод «центр кластера» для представления каждого кластера [14, с. 47–94].

Валидация кластера используется для оценки результатов алгоритма кластеризации. Существует два типа проверки. При внешней оценке результат кластеризации оценивается с использованием внешних данных, которые не использовались для кластеризации (например, метки классов, если они есть; внешние тесты). Некоторые внешние критерии включают Rand-меру [40, с. 66, 846–850], F-меру [35; с. 25, 315–318] и матрицу путаницы [49, с. 9, 40–50]. С помощью методов внешней оценки рассчитывают результат кластеризации с дополнительными знаниями. При внутренней оценке, напротив, результат рассчитывается на основе данных, используемых для кластеризации, путем определения соответствия вывода данным. Однако, как указывает C. Manning и др. [35], высокий балл, по внутренней оценке, не обязательно приводит к эффективному восстановлению информации. Обычно внутренние критерии включают индекс Боулдина [11, с. 1, 224–227] и индекс Данна[17, с. 4, 95–104].

Алгоритм иерархической кластеризации также называется кластеризацией на основе подключения, который создает иерархию разделов. Результатом иерархической кластеризации является древовидная структура, называемая дендрограммой, представляющая вложенную группировку объектов и уровни сходства (рис. 2).

Первым шагом в иерархической кластеризации является определение подобия / несходства с использованием меры расстояния. Иерархическая кластеризация очень гибкая при выборе функций расстояния. Однако другое расстояние функции идентифицируют различные особенности данных и, следовательно, структуры дендрограммы, используя разные функции расстояния могут, естественно, различаться.

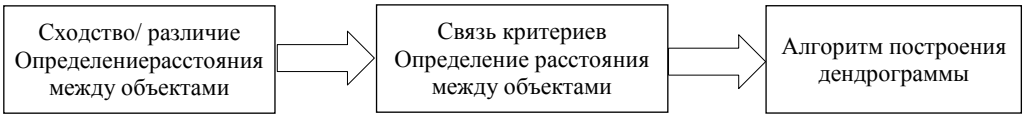

Рис. 2. *Общая процедура иерархической кластеризации*

Когда функция расстояния определена, правильно определяется расстояние между любыми двумя объектами. Затем иерархическая кластеризация предлагает несколько критериев связи для определения расстояния между двумя наборами объектов. Некоторые часто используемые критерии связи следующие:1) одиночная связь, также называется минимальной связью, определяемая как минимальное расстояние между любыми двумя объектами из двух групп; 2) полная связь, также называется максимальной связью, определяемая как максимальное расстояние между любыми двумя объектами из двух групп; 3) средняя связь, также называется средней связью или UPGMA, определяемая как среднее расстояние всех парных объектов из двух групп; 4) критерий Уорда, в критерии Уорда расстояние между двумя наборами объектов определяется как увеличение дисперсии при объединении двух наборов объектов.

Для одних и тех же данных разные критерии связи приведут к разной древовидной структуре даже с одинаковой мерой расстояния. Критерий Уорда работает хорошо, только если данные приблизительно нормальные. Отмечалось, что единственная связь всегда приводит к эффекту цепочки [38, с. 56: 836–62 (May 1968)], а при среднем сцеплении часто возникает эффект снежного кома [25, с. 13, 817–835]. Хотя такие ограничения можно частично исключить, остановив процесс кластеризации на другом уровне для разных данных, полная связь по-прежнему предпочтительнее и наиболее часто используется во многих приложениях. Полная связь иерархической кластеризации создает плотные и компактные кластеры, а также более значимые дендрограммы, чем метод одиночной связи [26]. Иерархическая кластеризация – это «пошаговый» алгоритм, позволяющий либо наращивать (агломеративная) или разрушать (разделяющая) иерархию кластеров.

Агломерационный алгоритм, также называемый подходом «снизу вверх», начинается в нижней части дерева как единый кластер. Затем на каждом шаге ближайшие два кластера, из-

меренные метриками расстояния и выбранной связи, объединяются в более крупный кластер, до тех пор пока все элементы не будут в одном кластере. Алгоритм разделения, также называемый подходом «сверху вниз», имеет обратный порядок и начинается с вершины дерева – единого кластера, включающего все объекты, который делится на каждом шаге рекурсивно, пока не прекратится процесс разделения. На практике более популярен агломеративный алгоритм.

Иерархическая кластеризация предоставляет широкий выбор по количеству кластеров, при котором нет необходимости заранее определять номер кластера. Каждый уровень в дендрограмме обеспечивает уникальное разделение данных, и окончательные кластеры могут быть определены, сравнивая все возможные результаты. Дендрограмма обеспечивает очень высокую интерпретируемость всей процедуры, которая делает иерархическую кластеризацию очень популярным методом. Однако древовидная структура очень чувствительна и нестабильна: различные методы связи, небольшое изменение данных может привести к значительной деформации в структуре дендрограммы.

Кластеризация K-средних – это самая ранняя и наиболее часто используемая секционная кластеризация. К 1960-м годам многие исследователи [47, с. 18, 267–276] предложили разделить данные, минимизируя внутри групповые вариации, чтобы групповые связи могли отражать определенный уровень однородности внутри кластера и неоднородность между кластерами. Термин «K-means» использовался Джеймсом Маккуином. Однако первоначальная идея кластеризации K-средних была предложена H. Steinhaus [43, с. 4.]. В отличие от иерархической кластеризации, кластеризация K-средних требует определение номера кластера K и производит только один раздел с K-кластерами. Когда номер кластера K зафиксирован, кластеризация K-средних – это фактически задача оптимизации поиска лучших K-подгрупп с K-центрами кластеров путем минимизации суммы внутригрупповой суммы квадратов (WGSS) следующим образом:

$$\sum_{G=1}^{k} \sum_{i \in I_g} \sum_{j=1}^{m} (x_{ij} - c_{gi})^2 \quad (1)$$

Стандартный алгоритм K-средних, также известный как алгоритм Ллойда (рис. 3), был первоначально предложен Стюартом Ллойдом в 1957 г. и впервые опубликован в 1982 г. В алгоритме Ллойда случайным образом инициализируются K кластерных центров, и определяется каждый субъект, ближайший к центру. Затем на основе назначения пересчитываются все субъекты для новых кластеров, до тех пор пока число K-центров кластера не останется без изменений.

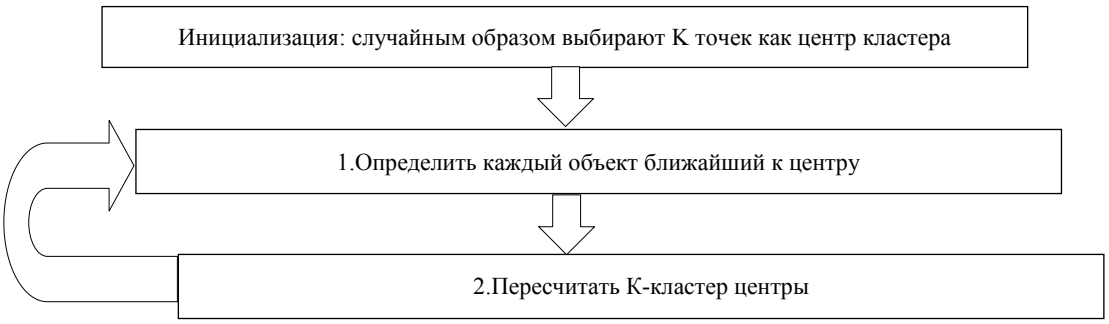

Рис. 3. *Алгоритм Ллойда*

С дополнительным предположением, что каждый кластер следует многомерному нормальному распределения, проблема K-средних может быть решена путем оценки конечной гауссовой модели. Тогда либо алгоритм ожидания-максимизации [13, с. 39, 1–38], или алгоритм кросс-энтропии [7, с. 517–523] используются для соответствия модели, и каждый объект относится к кластеру с наибольшей вероятностью. Эти алгоритмы на основе распределения превзошли стандартный алгоритм в определенных структурах данных. Однако это показало, что

алгоритм EM имеет проблему для определения ковариации внутри кластера матриц при решении задачи K-средних [12, с. 20, 1141–1147], и с данными высокой размерности (даже размерность > 2) алгоритмы EM и CE могут генерировать ложные кластеры (иногда называемые вырожденными кластерами) [36]. Кроме того, в литературе описано несколько альтернатив, использующих K-медиану [29], K-средние диапазоны [42] или K-режимы [24] вместо K-средних в качестве K-центров кластеров. Однако эти алгоритмы типа K-means работают должным образом только в особых случаях [8, с. 3–14], но у них было одно и тоже ограничение, представленное в исходной кластеризации K-средних [8, с. 3–14]. Хорошо известная проблема кластеризации K-средних заключается в том, что она может не обеспечить глобальный оптимум и очень чувствительна к случайно инициализированным центрам [44, с. 8, 294–304]. Даже несмотря на это, сообщалось, что алгоритм K-средних показывает очень хорошие свойства восстановления кластера [15, с. 67, 137–159]. K-метод кластеризации, безусловно, является удобным аналитическим инструментом, потому что его легко реализовать. Учитывая низкую временную сложность, можно запустить алгоритм K-средних с диапазоном номера кластера K и выбрать наиболее подходящий вариант позже, как и в случае иерархической кластеризации.

Кластерный анализ, как основной метод обучения, получил широкое распространение. Используется во многих дисциплинах, чтобы восстановить скрытую информацию.В биологии кластерный анализ применялся в транскриптомике, эволюционной биологии и биоинформатике. В транскриптомике кластерный анализ используется для построения групп генов с паттернами экспрессии генов [46, с.102, 15545–15550]. Данные группы часто содержат функционально связанные белки [53, с. 32, 151–155], выполняющие иерархическую кластеризацию транскриптома, протеома и эндометаболома. В эволюционной биологии и биоинформатике широко используется кластерный анализ в исследованиях, связанных с платформами высокопроизводительного генотипирования [23, с. 218–219], анализе данных микрочипов [25, с. 13, 817–835] и т.д. B. Andreopoulos и др. [4, с. 10, 297–314] рассмотрели около 40 алгоритмов кластеризации, применяемых в биоинформатике. В экологии кластерный анализ используется для выявления биогеографических или временных паттернов кластеризованными паттернами молекулярных последовательностей [55, с. 301, 976–978]. В медицинских исследованиях кластерный анализ используется в исследованиях медицинской визуализации для анализа данных, полученных с помощью функционального МРТ [20, с. 9, 298–310]. Кластерный анализ использовался для улучшения отношения сигнал / шум (SNR) динамических данных ПЭТ [31, с. 9, 554–561]. Группа S.L. Robinette применила алгоритм иерархической кластеризации и бикластеризации на ядерно-магнитных данных резонансной (ЯМР) визуализации для профилирования изменений в метаболическом составе биожидкостей. В бизнесе и маркетинге кластерный анализ применяется с 1960-х гг. [39, с. 20, 134–148]. Основное применение кластерного анализа в маркетинге состоит в обеспечении сегментации рынка [56, с. 15, 317–337]. Еще одно важное применение кластерного анализа в бизнесе и маркетинге состоит в том, чтобы понять поведение клиентов путем их группирования в однородные кластеры [30, с. 18, 233–239] и впоследствии принимать маркетинговые решения и стратегии. Относительно недавно была апробирована кластеризация Кохонена [51, с. 27, 757–764] для изучения поведения потребителей телекоммуникационных услуг. В информатике кластерный анализ является важным инструментом для обработки больших объемов данных. В атмосферных науках кластерный анализ применяется для определения режимов циркуляции и погодных условий [37, с. 93, 10927– 10952]. X. Gong и M.B. Richman [19, с. 8, 897–931] рассчитали эффективность кластерного анализа применительно к исследованию климата.

Сегодня предлагаются новые алгоритмы и прилагаются значительные усилия, направленные на повышение эффективности иерархической кластеризации и кластеризации K-средних. Так, T. Zhang, R. Ramakrishnan и M. Livny [57, с. 25, 103–114] предложили BIRCH, который сократил время выполнения и повысил эффективность по сравнению с другими иерархическими алгоритмами.Y. Cheng иG.M. Church [10, с. 93–103] создали «бикластеризацию», выполняющую иерархическую кластеризацию одновременно на уровне объектов и функций. Группа T. Kanungo [28, с. 24, 881–892] разработала алгоритм фильтрации для кластеризации K-средних с повышенной эффективностью по мере увеличения расстояния между кластерами. Совсем недавно нечеткая или перекрывающаяся кластеризация привлекла много внимания, позволив

каждому объекту принадлежать нескольким кластерам, в то время как кластеры взаимно исключают друг друга в классическом кластерном анализе [6, с. 532–537]. Кластеризация на основе знаний, включающая дополнительные базовые знания в области кластеризации становятся интересной темой в биоинформатике, в то время как классический кластерный анализ основан исключительно на числовых данных [22, с. 145–154.]. Помимо кластеризации числовых данных, некоторые разработки связаны с категориальными данными. Z. Huang [24, с. 2, 283–304] распространил алгоритм K-средних на категориальные данные. P. Andritsos совместно с другими исследователями [5, с. 531–532] создали новый алгоритм под названием «LIMBO» с новой мерой расстояний для категориальных данных и улучшенной масштабируемостью других иерархических алгоритмов кластеризации.

Кластерный анализ – это метод обучения, исследующий основную структуру данных без каких-либо предварительных знаний или информации. Подходы к многоцелевой кластеризации по-прежнему не решали эту проблему, сосредоточив внимание на только данных об экспрессии генов. Учитывая доступность различных источников биологических данных, кластерный анализ становится востребованным и для биологических исследований, чтобы получить биологически значимые кластеры [9, с. 14, 687–700]. Такие подходы называются кластеризацией на основе знаний. В иерархической кластеризации биологические знания обычно используются для определения биологического сходства между генами, а затем в сочетании со сходством экспрессии генов как общие метрики расстояния. Y. Cheng и другие исследователи [9, с. 14, 687–700] предложили биологическое сходство между двумя генами на основе общей границы и иерархическую кластеризацию по биологическому сходству, а также среднее биологическое сходство и евклидово расстояние.

Независимо от того, какой метод кластеризации используется, поиск наиболее разумного количества кластеров всегда имеет решающее значение. К сожалению, не существует стандартного подхода [18] к решению этой проблемы. Определение номера кластера до сих пор остается сложным и открытым вопросом. Было предложено множество подходов к поиску наиболее подходящего числа кластеров. В 2004 г. J. Cheng объединил существующие подходы к определению количества кластеров в пять групп: 1) перекрестная проверка; 2) оценка правдоподобия наложенного штрафа; 3) перестановочные тесты; 4) передискретизация; 5) нахождение изгиба кривой ошибок. S. Dudoit и J. Fridlyand [16, с. 3, 1–21] проанализировали большинство методов поиска кривой ошибок, включающих: 1) индекс Калински и Харабаша; 2) индекс Кшановского и Лая; 3) статистику Хартиганса; 4) Gapigap PC [48, с. 63, 411–423].

Вывод. В данной статье представлен обзор методологии статистического исследования процессов кластеризации экономики, рассмотрены значимые алгоритмы кластеризации экономики, рассматривается и обобщается зарубежный опыт проведения кластерной политики, масштабы распространения и разнообразие типов кластерных структур за рубежом. Применительно к экономической сфере использование кластерных методов анализа позволяет повысить эффективность использования новых моделей управления отечественными промышленными предприятиями в формате устойчиво развивающихся хозяйственных систем и координации социальной и экономической политики на региональном уровне.

Литература

1. Гичиев, Н. С. Новый этап экономического роста : региональные сценарии, прогнозы, модели : монография / Н.С. Гичиев. – М. : Изд-во «Перо», 2017.
2. Деневизюк, Д. А., Гимбатов, Ш. М., Тумсоев, А. Б., Кутаев, Ш. К., Ахмедова, Л. А., Султанов, Г. С., Патлис, В. В., Мартышенко, Н. С., Мартышенко, С. Н. Социально-экономическое развитие на современном этапе : проблемы и направления. – Москва, 2015. Книга 2.
3. Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. Automatic Subspace Clustering of High Dimensional Data // Data Mining and Knowledge Discovery, 2005. No.11. P. 5–33.
4. Andreopoulos, B., An, A., Wang, X., and Schroeder, M. A Roadmap of Clustering Algorithms : Finding a Match for a Biomedical Application // Brief Bioinform, 2009. No.10. P. 97–314.
5. Andritsos, P., Tsaparas, P., Miller, R., and Sevcik, K. Limbo : Scalable Clustering of Categorical Data Advances in Database Technology – Edbt 2004. Vol. 2992 ; eds. E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm and E. Ferrari. Springer Berlin. – Heidelberg, 2004. P. 531–532.
6. Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, RJ. Model-Based Over-lapping Clustering // Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. P. 532–537.
7. Botev, Z., and Kroese, DP. Global Likelihood Optimization Via the Cross-Entropy Method, with an Applica-

tion to Mixture Models. 2004. P. 517–523.
8. Carroll, J., and Chaturvedi, A. K-Midranges Clustering // Advances in Data Science and Classification ; eds. A. Rizzi, M. Vichi and H.H. Bock. – Berlin : Springer, 1998. P. 3–14.
9. Cheng, J., Cline, M., Martin, J., Finkelstein, D., et al. Knowledge-Based Clustering Algorithm Driven by Gene Ontology // J. Biopharm Stat, 2004. No.14. P. 687–700.
10. Cheng, Y/, and Church, G.M. Biclustering of Expression Data // Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. 2000. P. 93–103.
11. David, L.D., and Donald, W. B. A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979. No.1. P. 24–227.
12. De Backer, S., and Scheunders, P. A Competitive Elliptical Clustering Algorithm // Pattern Recognition Letters. 1999. No. 20. P. 1141–1147.
13. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data Via the Em Algorithm // Journal of the Royal Statistical Society. Series B (Methodological). 1977. No. 39. P. 1–38.
14. Diday, E., and Simon, J. C. Clustering Analysis // Digital Pattern Recognition; ed. KS Fu. – New York : Springer-Verlag, Inc., 1976. P. 47–94.
15. Dimitriadou, E., Dolničar, S., and Weingessel, A. An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets // Psychometrika, 2002. No. 67. P. 37–159.
16. Dudoit, S., and Fridlyand, J. A Prediction-Based Resampling Method for Estimating the Number of Clusters in a Dataset // Genome Biol, 2002. No. 3. P. 1–21.
17. Dunn, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions // Journal of Cybernetics. 1974. No. 4. P. 95–104.
18. Girman, C. J. Cluster Analysis and Classification Tree Methodology as an Aid to Improve Understanding of Benign Prostatic Hyperplasia // University of North Carolina at Chapel Hill, Dept. of Biostatistics. 1994.
19. Gong, X., and Richman, M. B. On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies // Journal of Climate. 1995. No. 8. P. 897–931.
20. Goutte, C., Toft, P., Rostrup, E., Nielsen, F.Å., and Hansen, L. K. On Clustering Fmri Time Series // NeuroImage. 1999. No. 9. P. 298–310.
21. Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., et al. Similarity Measures in Scientometric Research : The Jaccard Index Versus Salton's Cosine Formula // Information Processing & Management. 1989. No. 25. P. 315–318.
22. Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. Co-Clustering of Biological Networks and Gene Expression Data // Bioinformatics. 2002. No. 18. Suppl. 1. P. 145–154.
23. Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. Combinatorial Algorithms for Structural Variation Detection in High Throughput Sequenced Genomes // Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology. 2009. P. 218–219.
24. Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values // Data Mining and Knowledge Discovery. 1998. No. 2. P. 283–304.
25. Huth, R., Nemesova, I., and Klimperová, N. Weather Categorization Based on the Aver-age Linkage Clustering Technique : An Application to European Mid-Latitudes // International Journal of Climatology. 1993. No. 13. P. 817–835.
26. Jain, A. K., and Dubes, R. Algorithms for Clustering Data. – Prentice-Hall, Inc. 1988.
27. Jain, A. K., Murty, M. N., and Flynn, P. J. Data Clustering : A Review // ACM Comput. Surv. 1999. No. 31. P. 264–323.
28. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D, et al. An Efficient KMeans Clustering Algorithm : Analysis and Implementation / Pattern Analysis and Machine Intelligence // IEEE Transactions. 2002. No. 24. P. 881–892.
29. Kaufman, L., and Rousseeuw, P. Finding Groups in Data : An Introduction to Cluster Analysis // Wiley-Interscience. 1990.
30. Kiel, G. C., and Layton, R. A. Dimensions of Consumer Information Seeking Behavior // Journal of Marketing Research. 1981. No.18. P. 233–239.
31. Kimura, Y., Hsu, H., Toyama, H., Senda, M., and Alpert, N. M. Improved Signal-to-Noise Ratio in Parametric Images by Cluster Analysis // NeuroImage. 1999. No. 9. P. 554–561.
32. Kimura, Y., Hsu, H., Toyama, H., Senda, M., and Alpert, N. M. Improved Signal-to-Noise Ratio in Parametric Images by Cluster Analysis // NeuroImage. 1999. No. 9. P. 554–561.
33. King, B. Step-Wise Clustering Procedures // Journal of the American Statistical Association. 1967. No. 62. P. 86–101.
34. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations // Proc. 5th Berkeley Symp. Mathematical Statist. Probability. 1967. P. 281–297.
35. Manning, C., Raghavan, P., and Schütze, H. Introduction to Information Retrieval // Cambridge University Press. 2008
36. McLachlan, G. J., and Peel, D. Finite Mixture Models. – 1st ed. // Wiley-Interscience. 2000.
37. Mo, K., and Ghil, M. Cluster Analysis of Multiple Planetary Flow Regimes // J. Geophys. Res. 1988. No. 93. P. 10927–10952.
38. Nagy, G. State-of-the-Art in Pattern Recognition // Journal Name: Proc. IEEE (Inst. Elec. Electron. Eng.). 1968. No. 56. P. 836-62 (May 1968).
39. Punj, G., and Stewart, D. W. Cluster Analysis in Marketing Research : Review and Suggestions for Application // Journal of Marketing Research. 1983. No. 20. P. 134–148.
40. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association. 1971. No. 66. P. 846–850.

41. Ruspini, E. H. A New Approach to Clustering // Information and Control. 1969. No. 15. P. 22–32.
42. Späth, H. The Cluster Dissection and Analysis Theory Fortran Programs Examples // Prentice-Hall, Inc. 1985.
43. Steinhaus, H. Sur La Division Des Corps Matériels En Parties (in French) // Bull. Acad. Polon. Sci. 1957. No. 4. P. 4.
44. Steinley, D. Local Optima in K-Means Clustering : What You Don't Know May Hurt You // Psychological Methods. 2003. No. 8. P. 294–304.
45. Steinley, D. K-Means Clustering : A Half-Century Synthesis // British Journal of Mathematical and Statistical Psychology. 2006. No. 59. P. 1–34.
46. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., et al. Gene Set Enrichment Analysis : A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles // Proc Natl Acad Sci USA, 2005. No.102. P. 15545–15550.
47. Thorndike, R. Who Belongs in the Family? // Psychometrika. 1953. No. 18. P. 267–276.
48. Tibshirani, R., Walther, G., and Hastie, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001. No. 63. P. 411–423.
49. Townsend, J. Theoretical Analysis of an Alphabetic Confusion Matrix // Attention, Perception, & Psychophysics. 1971. No. 9. P. 40–50.
50. Tryon, R. C. Cluster Analysis; Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. – Ann Arbor, Mich.: Edwards brother, inc., lithoprinters and publishers. 1939.
51. Tsao, E. C.-K., Bezdek, J. C., and Pal, N. R. Fuzzy Kohonen Clustering Networks // Pattern Recognition. 1994. No. 27. P. 757–764.
52. Tseng, G. C. Penalized and Weighted K-Means for Clustering with Scattered Objects and Prior Information in High-Throughput Biological Data // Bioinformatics. 2007. No. 23. P. 224–2255.
53. Varela, C., Schmidt, S. A., Borneman, A. R., Krömer, J. O. et al. Systems Biology : A New Paradigm for Industrial Yeast Strain Development // Microbiology Australia. 2011. No. 32. P. 151–155.
54. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function // Journal of the American Statistical Association. 1963. No. 58. P. 236–244.
55. Whitaker, R. J., Grogan, D. W., and Taylor, J. W. Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea // Science. 2003. No. 301. P. 76–78.
56. Wind, Y. Issues and Advances in Segmentation Research // Journal of Marketing Research. 1978. No. 15. P. 317–337.
57. Zhang, T., Ramakrishnan, R., and Livny, M. Birch : An Efficient Data Clustering Method for Very Large Databases // SIGMOD Rec. 1996. No. 25. P. 103–114.

References:

1. Gichiev, N. S. Novyj etap ekonomicheskogo rosta : regional'nye scenarii, prognozy, modeli : monografiya / N.S. Gichiev. – M. : Izd-vo «Pero», 2017.
2. Denevizyuk, D. A., Gimbatov, SH. M., Tumsoev, A. B., Kutaev, SH. K., Ahmedova, L. A., Sultanov, G. S., Patlis, V. V., Martyshenko, N. S., Martyshenko, S. N. Social'no-ekonomicheskoe razvitie na sovremennom etape : problemy i napravleniya. – Moskva, 2015. Kniga 2.
3. Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. Automatic Subspace Clustering of High Dimensional Data // Data Mining and Knowledge Discovery, 2005. No.11. P. 5–33.
4. Andreopoulos, B., An, A., Wang, X., and Schroeder, M. A Roadmap of Clustering Algorithms : Finding a Match for a Biomedical Application // Brief Bioinform, 2009. No.10. P. 97–314.
5. Andritsos, P., Tsaparas, P., Miller, R., and Sevcik, K. Limbo : Scalable Clustering of Categorical Data Advances in Database Technology – Edbt 2004. Vol. 2992 ; eds. E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm and E. Ferrari. Springer Berlin. – Heidelberg, 2004. P. 531–532.
6. Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, RJ. Model-Based Over-lapping Clustering // Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005. P. 532–537.
7. Botev, Z., and Kroese, DP. Global Likelihood Optimization Via the Cross-Entropy Method, with an Application to Mixture Models. 2004. P. 517–523.
8. Carroll, J., and Chaturvedi, A. K-Midranges Clustering // Advances in Data Science and Classification ; eds. A. Rizzi, M. Vichi and H.H. Bock. – Berlin : Springer, 1998. P. 3–14.
9. Cheng, J., Cline, M., Martin, J., Finkelstein, D., et al. Knowledge-Based Clustering Algorithm Driven by Gene Ontology // J. Biopharm Stat, 2004. No.14. P. 687–700.
10. Cheng, Y/, and Church, G.M. Biclustering of Expression Data // Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. 2000. P. 93–103.
11. David, L.D., and Donald, W. B. A Cluster Separation Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979. No.1. P. 24–227.
12. De Backer, S., and Scheunders, P. A Competitive Elliptical Clustering Algorithm // Pattern Recognition Letters. 1999. No. 20. P. 1141–1147.
13. Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data Via the Em Algorithm // Journal of the Royal Statistical Society. Series B (Methodological). 1977. No. 39. P. 1–38.
14. Diday, E., and Simon, J. C. Clustering Analysis // Digital Pattern Recognition; ed. KS Fu. – New York : Springer-Verlag, Inc., 1976. P. 47–94.
15. Dimitriadou, E., Dolničar, S., and Weingessel, A. An Examination of Indexes for Determining the Number

of Clusters in Binary Data Sets // Psychometrika, 2002. No. 67. P. 37–159.
16. Dudoit, S., and Fridlyand, J. A Prediction-Based Resampling Method for Estimating the Number of Clusters in a Dataset // Genome Biol, 2002. No. 3. P. 1–21.
17. Dunn, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions // Journal of Cybernetics. 1974. No. 4. P. 95–104.
18. Girman, C. J. Cluster Analysis and Classification Tree Methodology as an Aid to Improve Understanding of Benign Prostatic Hyperplasia // University of North Carolina at Chapel Hill, Dept. of Biostatistics. 1994.
19. Gong, X., and Richman, M. B. On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies // Journal of Climate. 1995. No. 8. P. 897–931.
20. Goutte, C., Toft, P., Rostrup, E., Nielsen, F.Å., and Hansen, L. K. On Clustering Fmri Time Series // NeuroImage. 1999. No. 9. P. 298–310.
21. Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., et al. Similarity Measures in Scientometric Research : The Jaccard Index Versus Salton's Cosine Formula // Information Processing & Management. 1989. No. 25. P. 315–318.
22. Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. Co-Clustering of Biological Networks and Gene Expression Data // Bioinformatics. 2002. No. 18. Suppl. 1. P. 145–154.
23. Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. Combinatorial Algorithms for Structural Variation Detection in High Throughput Sequenced Genomes // Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology. 2009. P. 218–219.
24. Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values // Data Mining and Knowledge Discovery. 1998. No. 2. P. 283–304.
25. Huth, R., Nemesova, I., and Klimperová, N. Weather Categorization Based on the Aver-age Linkage Clustering Technique : An Application to European Mid-Latitudes // International Journal of Climatology. 1993. No. 13. P. 817–835.
26. Jain, A. K., and Dubes, R. Algorithms for Clustering Data. – Prentice-Hall, Inc. 1988.
27. Jain, A. K., Murty, M. N., and Flynn, P. J. Data Clustering : A Review // ACM Comput. Surv. 1999. No. 31. P. 264–323.
28. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D, et al. An Efficient KMeans Clustering Algorithm : Analysis and Implementation / Pattern Analysis and Machine Intelligence // IEEE Transactions. 2002. No. 24. P. 881–892.
29. Kaufman, L., and Rousseeuw, P. Finding Groups in Data : An Introduction to Cluster Analysis // Wiley-Interscience. 1990.
30. Kiel, G. C., and Layton, R. A. Dimensions of Consumer Information Seeking Behavior // Journal of Marketing Research. 1981. No.18. P. 233–239.
31. Kimura, Y., Hsu, H., Toyama, H., Senda, M., and Alpert, N. M. Improved Signal-to-Noise Ratio in Parametric Images by Cluster Analysis // NeuroImage. 1999. No. 9. P. 554–561.
32. Kimura, Y., Hsu, H., Toyama, H., Senda, M., and Alpert, N. M. Improved Signal-to-Noise Ratio in Parametric Images by Cluster Analysis // NeuroImage. 1999. No. 9. P. 554–561.
33. King, B. Step-Wise Clustering Procedures // Journal of the American Statistical Association. 1967. No. 62. P. 86–101.
34. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations // Proc. 5th Berkeley Symp. Mathematical Statist. Probability. 1967. P. 281–297.
35. Manning, C., Raghavan, P., and Schütze, H. Introduction to Information Retrieval // Cambridge University Press. 2008
36. McLachlan, G. J., and Peel, D. Finite Mixture Models. – 1st ed. // Wiley-Interscience. 2000.
37. Mo, K., and Ghil, M. Cluster Analysis of Multiple Planetary Flow Regimes // J. Geophys. Res. 1988. No. 93. P. 10927–10952.
38. Nagy, G. State-of-the-Art in Pattern Recognition // Journal Name: Proc. IEEE (Inst. Elec. Electron. Eng.). 1968. No. 56. P. 836-62 (May 1968).
39. Punj, G., and Stewart, D. W. Cluster Analysis in Marketing Research : Review and Suggestions for Application // Journal of Marketing Research. 1983. No. 20. P. 134–148.
40. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association. 1971. No. 66. P. 846–850.
41. Ruspini, E. H. A New Approach to Clustering // Information and Control. 1969. No. 15. P. 22–32.
42. Späth, H. The Cluster Dissection and Analysis Theory Fortran Programs Examples // Prentice-Hall, Inc. 1985.
43. Steinhaus, H. Sur La Division Des Corps Matériels En Parties (in French) // Bull. Acad. Polon. Sci. 1957. No. 4. P. 4.
44. Steinley, D. Local Optima in K-Means Clustering : What You Don't Know May Hurt You // Psychological Methods. 2003. No. 8. P. 294–304.
45. Steinley, D. K-Means Clustering : A Half-Century Synthesis // British Journal of Mathematical and Statistical Psychology. 2006. No. 59. P. 1–34.
46. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., et al. Gene Set Enrichment Analysis : A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles // Proc Natl Acad Sci USA, 2005. No.102. P. 15545–15550.
47. Thorndike, R. Who Belongs in the Family? // Psychometrika. 1953. No. 18. P. 267–276.
48. Tibshirani, R., Walther, G., and Hastie, T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001. No. 63. P. 411–423.

49. Townsend, J. Theoretical Analysis of an Alphabetic Confusion Matrix // Attention, Perception, & Psychophysics. 1971. No. 9. P. 40–50.
50. Tryon, R. C. Cluster Analysis; Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. – Ann Arbor, Mich.: Edwards brother, inc., lithoprinters and publishers. 1939.
51. Tsao, E. C.-K., Bezdek, J. C., and Pal, N. R. Fuzzy Kohonen Clustering Networks // Pattern Recognition. 1994. No. 27. P. 757–764.
52. Tseng, G. C. Penalized and Weighted K-Means for Clustering with Scattered Objects and Prior Information in High-Throughput Biological Data // Bioinformatics. 2007. No. 23. P. 224–2255.
53. Varela, C., Schmidt, S. A., Borneman, A. R., Krömer, J. O. et al. Systems Biology : A New Paradigm for Industrial Yeast Strain Development // Microbiology Australia. 2011. No. 32. P. 151–155.
54. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function // Journal of the American Statistical Association. 1963. No. 58. P. 236–244.
55. Whitaker, R. J., Grogan, D. W., and Taylor, J. W. Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea // Science. 2003. No. 301. P. 76–78.
56. Wind, Y. Issues and Advances in Segmentation Research // Journal of Marketing Research. 1978. No. 15. P. 317–337.
57. Zhang, T., Ramakrishnan, R., and Livny, M. Birch : An Efficient Data Clustering Method for Very Large Databases // SIGMOD Rec. 1996. No. 25. P. 103–114.