DOI 10.15826/vestnik.2025.24.2.020 Original Paper

УДК 336.67

Кластеризация российских производственных компаний по показателям их финансового состояния с использованием технологий машинного обучения

Л. А. Буланов[1], *А. В. Калина*[1, 2] ✉, *В. В. Криворотов*[1]

[1] *Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия*

[2] *Институт экономики Уральского отделения РАН, г. Екатеринбург, Россия*

✉ *alexkalina74@mail.ru*

Аннотация. Кластеризация объектов исследования и объединение их в сходные группы по совокупности признаков является важнейшим этапом при решении многих задач социально-экономического развития, в особенности задач, связанных с оценкой состояния социально-экономической системы, а также моделирования и прогнозирования показателей ее будущего развития. Целью настоящего исследования является относительная оценка финансового состояния крупных российских производственных компаний на основе данных форм бухгалтерской отчетности методами кластеризации, относящимися к категории машинного обучения без учителя. Результаты такой оценки впоследствии предполагается использовать для построения модели оценки финансового состояния компаний на основе одного из алгоритмов машинного обучения с учителем. В работе предложены ключевые показатели финансового состояния компаний, на основании которых предлагается выполнять их кластеризацию, которые были выделены как результат анализа современных методов и подходов к исследованию и оценке конкурентоспособности и конкурентной позиции компаний. При проведении кластеризации на основе предложенной совокупности показателей были использованы данные финансовой отчетности 2 249 российских производственных компаний по итогам 2023 г. В качестве крупных компаний рассматривались компании, которые имели оборот более 2 млрд руб. и штат сотрудников более 251 чел. В качестве алгоритмов кластеризации использовались K-Means++, иерархическая кластеризация и DBSCAN. С целью получения лучшего результата была проведена специальная предобработка данных и подбор необходимых гиперпараметров для алгоритмов кластеризации. Качество итоговой кластеризации оценивалось по индексам Дэвиса — Болдина (DBI) и Калински — Харабаса (CHI). Полученные результаты показали, что рассматриваемые производственные компании по показателям финансового состояния можно объединить в сравнительно небольшое число кластеров (обычно не более 3), что открывает широкие возможности для построения моделей финансового состояния компаний. По итогам использования трех методов кластеризации лучшим алгоритмом с небольшим отрывом оказался K-Means ++, сформированные центроиды которого можно назвать усредненной оценкой компаний с плохим, нормальным и хорошим финансовым состоянием. Качество итоговой кластеризации можно оценить как хорошее.

Ключевые слова: финансовый анализ; машинное обучение; показатели финансового состояния; крупная компания; кластеризация компаний; K-Means ++; иерархическая кластеризация; DBSCAN.

  ISSN 2712-7435

1. Введение

Финансовый анализ предприятий и компаний на сегодняшний день рассматривается как неотъемлемая часть формирования эффективной стратегии компании, поскольку он дает возможность объективно оценить текущее состояние ее финансовой деятельности, выявить ключевые факторы, влияющие на это состояние, а также провести оценку ключевых угроз экономической и финансовой безопасности компании. Однако классические методы финансового анализа, зачастую выполняемые вручную, становятся все более трудоемкими. Они затрудняют своевременное принятие управленческих решений и не всегда позволяют выявить скрытые паттерны, возникающие в объемных массивах данных, и нередко рассматривают компании изолированно, без учета их конкурентного окружения.

В современном мире стремительного развития технологий и повсеместной автоматизации особо актуальными становятся подходы, применяющие машинное обучение, в том числе кластеризацию, метод обучения без учителя. Подобные алгоритмы способны группировать компании по набору признаков, в частности формировать кластеры компаний со сходным финансовым состоянием, что может значительно ускорить процесс их оценки и улучшить качество проводимой экспертизы.

В последнее десятилетие огромное распространение при проведении экономических исследований в целом и финансового состояния компаний в частности получили технологии машинного обучения. Эти технологии показали существенные преимущества перед традиционными методами, позволяя получать качественные информативные оценки, характеризующие финансовое состояние компаний, в условиях больших объемов финансовой информации, большого числа и разнообразия финансовых показателей, неоднозначности и противоречивости многих данных и др.

В таких условиях использование традиционных классических методов и подходов к исследованиям и оценке финансового состояния компаний зачастую дают неопределенные противоречивые оценки, что не позволяет формировать качественные управленческие решения. Более того, во многих случаях оценки, получаемые с использованием таких методов и подходов, наоборот, могут искажать реальную картину и приводить к неверным управленческим решениям, не отвечающим фактически складывающийся ситуации.

С другой стороны, еще с 1990-х гг. в мире получили широкое распространение нейронные сети, лежащие в основе работы искусственного интеллекта и технологий машинного обучения. В первую очередь, полигоном для их применения стал финансовый сектор, что дало «новый прорыв» в исследовании финансового состояния компаний. Так, например, одним из первых случаев применения нейронных сетей в финансовом секторе было исследование Kryzanowski et al [1], посвященное прогнозированию показателей фондового рынка, которое показало преимущества новых методов и возможности их широкомасштабного использования при решении практических задач.

Однако, несмотря на широкомасштабное распространение технологий машинного обучения для решения задач, связанных с оценкой и анализом финансового состояния предприятий и компаний, наблюдаемое в последние годы, по-прежнему остается множество актуальных, а в некоторых случаях малоисследованных проблем. К одной из таких проблем, безусловно, относятся вопросы, связанные с моделированием финансового состояния компаний, которые к настоящему времени однозначного подхода к решению и практической реализации не получили. В этой связи исследования в этом направлении сохраняют высокую актуальность и значимость.

Целью статьи является относительная оценка финансового состояния крупных российских производственных компаний на основе данных форм бухгалтерской отчетности методами кластеризации, относящимися к категории машинного обучения без учителя. Полученные первоначальные метки будут служить основой для дальнейшего обучения моделей на базе современных алгоритмов машинного обучения с учителем, позволяя формировать более точные и оперативные модели принятия решений в области финансового управления. Этот подход представляется перспективным в условиях динамичных конкурентных рынков и стремительной цифровой трансформации бизнеса.

Гипотезы исследования:

H1. Использование кластерного анализа упростит и автоматизирует процедуру группировки компаний по их финансовому состоянию, что позволит повысить точность сравнительного анализа за счет уменьшения субъективных ошибок и детального учета метрик.

H2. Крупные отечественные производственные компании можно сгруппировать в несколько крупных кластеров по показателям их финансового состояния, в рамках которых действую общие закономерности и может быть применен единый подход к моделированию таких показателей.

Структура статьи построена в соответствии с поставленной целью и содержит шесть разделов. Во введении сформулированы актуальность, цель и гипотезы исследования. Во втором разделе представлен обзор основных исследований в области конкурентного анализа компаний и рассматриваемых методов кластеризации. В третьем разделе рассмотрена методология исследования и описание набора данных и особенностей используемых методов кластеризации. В четвертом и пятом разделах приведены результаты исследования и их обсуждение. В шестом разделе даны основные выводы по результатам проведенного исследования.

2. Обзор литературы

2.1. Формирование подходов к конкурентному анализу компаний

Конкурентный анализ компаний является одной из наиболее актуальных тем исследований с момента формирования экономики как науки. Однако общепринятого подхода к такому анализу до сих пор еще не существует.

С другой стороны, конкурентные позиции компании и связанный с ними конкурентный анализ являются ключевыми характеристики, определяющие ее финансовое состояние, а также экономическую и финансовую безопасность.

В этом направлении, в первую очередь, можно выделить модели Портера, в частности алмазную модель (Porter's Diamond) [2] и модель пяти сил (Five Forces Framework) [3]. Алмазная модель была предложена для объяснения причин, по которым одни страны обладают более выраженными конкурентными преимуществами в определенных отраслях по сравнению с другими, в то время как модель пяти сил позволяет выявить ключевые источники конкурентного давления и оценить перспективы прибыльности компании, учитывая совокупное воздействие различных экономических и организационных факторов.

Данные модели впоследствии использовались многими учеными и специалистами. Например, на основе модернизированной модели Портера был разработан и предложен к практической реализации механизм повышения конкурентоспособности аквакультурной индустрии Китая [4]. При этом авторами было проведено моделирование структурными уравнениями с использованием метода наименьших квадратов, а в качестве составляющих факторов конкурентоспособности рассматривались наличие ресурсов, научно-техническая среда, экономическая выгода и потенциал развития.

Другим примером современной реализации моделей Портера является работа Tao & Cai [5], посвященная оценке влияния цифровых финансов на конкурентоспособность экспорта китайских компаний. Проведенное исследование показало, что цифровое финансирование значительно повышает конкурентоспособность экспорта, в частности за счет расширения охвата и глубины использования цифровых финансов.

Можно привести примеры развития методов Портера для оценки конкурентоспособности современных высокотехнологичных секторов, в частности мобильной связи 5G. Liu et al. [6] анализируют конкурентоспособность китайских ТНК в сфере телекоммуникаций с трех точек зрения (то есть на уровне страны, отрасли и компании), что позволило сформировать стратегии развития 5G в Китае и определить политические последствия для развития 6G.

Кроме того, модели Портера успешно применялись в области возобновляемой энергетики. Fang et al. [7] оценивают конкурентоспособность возобновляемой энергетики в G20. Cibinskiene et al. [8] проводят оценку конкурентоспособности промышленных компаний, занимающихся производством компонентов для ветроэнергетики. Примерно аналогичные исследования провели Liu et al. [9], оценившие конкурентоспособность цепочки создания стоимости в ветроэнергетике Китая.

В целом можно констатировать, что модели и подходы, предложенные Портером, по-прежнему сохраняют высокую актуальность и используются многими учеными и специалистами для оценки конкурентоспособности и формирования конкурентных развития компаний, работающих в разных сферах деятельности.

Среди других подходов к оценке конкурентных позиций и конкурентоспособности компаний широкое распространение получили продуктовые методы, где главным критерием оценки выступают характеристики производимой компаниями продукции. При этом конкурентоспособность продукта может быть оценена с нескольких альтернативных позиций. Например, Lau et al. [10] используют рейтинговую оценку конкурентоспособности продукции на основании совокупности ее качественных параметров. Примерно аналогичных позиций придерживаются Lotfi & Karim [11] при оценке конкурентоспособности экспорта компаний.

Другой подход предлагает расчет комплексного индекса конкурентоспособности, учитывающий различные ее характеристики. Так, Li et al. [12] предлагают систему факторов, определяющих конкурентоспособность продукции, и разработан индекс конкурентных приоритетов, пригодный для проведения конкурентного анализа, который поможет усовершенствовать продукт и скорректировать стратегию. Такой подход авторы положили в основу динамического прогнозирования конкурентной позиции продукта.

Другим примером комплексной оценки конкурентоспособности продукции является исследование Kim et al. [13], где авторы разработали четырехквадрантную матрицу, позволяющую комплексно оценить сильные и слабые стороны продукта по сравнению с конкурентами, определить критические области, требующие немедленного улучшения, и предлагают стратегические рекомендации для повышения общей удовлетворенности клиентов применительно к гостиничному бизнесу.

Следующей вехой конкурентного анализа компаний можно назвать группу операционных методов. Операционные методы оценки конкурентных позиций и конкурентоспособности компаний базируются на принципах теории эффективной конкуренции, которая предусматривает всесторонний анализ деятельности всех подразделений компании. Эффективность работы каждой службы компании оценивается с точки зрения рационального использования ограниченных экономических ресурсов. В результате конкурентоспособность определяется на основе совокупности количественных показателей, охватывающих все ключевые аспекты деятельности компании. В частности, операционный подход использует метод американской консалтинговой фирмы Dun & Bradstreet. Метод основывается на анализе эффективности производственно-сбытовой деятельности, интенсивности использования основного и оборотного капитала, а также устойчивости финансовой деятельности. В предлагаемой модели конкурентоспособность оценивается через показатели, отражающие эти аспекты.

Среди примеров развития операционных методов для оценки конкурентоспособности компаний можно выделить работу Buckley et al. [14], где в основе лежит оценка эффективности производственной деятельности компании, которая эмпирически связывается с процессом принятия решений. Операционные результаты деятельности используются для оценки

конкурентоспособности компаний в работе Schefczyk [15]. Good et al. [16] в качестве основы для оценки конкурентоспособности применяют продуктивность и эффективность, а например, Parkan & Wu [17] предлагают использовать рейтинг операционной конкурентоспособности компании.

Комплексные методы предполагают оценку конкурентоспособности хозяйствующего субъекта как интегральной величины, включающей текущую (достигнутую) конкурентоспособность и конкурентный потенциал (перспективную конкурентоспособность). Для оценки текущей конкурентоспособности предполагается использовать продуктовые методы, определение конкурентного потенциала организации у большинства авторов сводится к оценке эффективности использования производственных, маркетинговых, финансовых, кадровых, управленческих и других ресурсов компании, аналогично операционным методам. Далее осуществляется агрегирование полученных оценок в единый интегральный показатель конкурентоспособности хозяйствующего субъекта. В частности, Fleisher & Bensoussan [18] представили 24 метода оценки конкурентоспособности компании, Prescott & Grant [19] формируют из 11 измерений 21 критерий конкурентного анализа и т. д.

Существуют и другие методы конкурентного анализа компаний, не вошедшие в рассмотренный краткий перечень. Но основные принципы и подходы, на которых они базируются, во многом схожи с перечисленными выше.

2.2. Обзор методов кластеризации

В самом общем представлении кластеризация представляет собой метод анализа данных и машинного обучения без учителя, который используется для группировки объектов в подмножества, кластеры, на основе их схожести по определенным признакам. Таким образом объекты — компании — в рамках одного кластера будут максимально похожи, а объекты из разных кластеров — максимально различны. На сегодняшний момент известно большое количество алгоритмов кластеризации, из которых можно выделить несколько наиболее популярных.

Алгоритм K-Means был независимо изобретен Steinhaus [20] и Lloyd [21] практически одновременно, но популярность приобрел после работы MacQueen [22]. Уже в первых двух в работах был представлен оригинальный принцип итеративной кластеризации, где центры кластеров пересчитываются как средние центроиды уже назначенных точек, однако были существенные различия в прикладном аспекте. Публикация Steinhaus [20] носила более абстрактный и теоретический характер, автор изучал существования и свойства оптимальных разбиений без привязки к конкретной прикладной задаче. В публикации Lloyd [21] акцент изначально был на практическом применении алгоритма в задачах векторного квантирования для оптимизации передачи сигналов. MacQueen [22] формализировал метод уже непосредственно в задачах многомерных наблюдений, именно в его работе появилось понятие

«кластер» в современном смысле, и после нее алгоритм получил признание в статистических и вычислительных задачах.

Первые версии K-Means имели большие проблемы в части выбора начальной позиции, к которой алгоритм принципиально чувствителен. Кроме того, он подвержен падению в локальные минимумы, что в комбинации приводило к тому, что различные стартовые положения давали совершенно разные результаты.

С момента создания было предложено множество модификаций изначального алгоритма, в частности Kanungo et al. [23] оптимизировали дорогостоящие вычисления расстояний между точками и центроидами через использования вспомогательных оценок. Ostrovsky et al. [24] исследуют сходимость алгоритма и определяют условия, при которых будет получено лучшее качество при наименьших временных затратах. Наконец, Ahmed et al. [25] охватывают и сравнивают многочисленные улучшения в нескольких направлениях: улучшение метода инициализации, оптимизации вычисления расстояний и улучшение сходимости, а также попытка адаптации к большим данным.

Одно из важнейших изменений было показано Arthur et al. [26], где появился K-Means++, в котором был радикально улучшен способ инициализации начальных центроидов, что значительно повлияло как на скорость схождения, так и на итоговый результат.

В настоящее K-Means остается популярным за простоту, скорость работы и приемлемый результат, все еще продолжаются эксперименты с модификациями под разные задачи. В частности, Abdulnassar et al. [27] аналогично пытаются модифицировать начальное положение центроидов, кроме того разрабатывают новую модель K-Means 9+, сочетающую в себе несколько глубоких модификаций. Ay et al. [28] исследуют возможность жестко зафиксировать центроиды при инициализации.

Существует и достаточно много недавних успешных сценариев применения данного алгоритма. Li et al. [29] используют одну из его модификаций в методике планирования низковольтной распределительной сети. He et al. [30] используют K-Means с байесовским анализом для моделирования системы поддержки ротора.

Иерархическая кластеризация является алгоритмом, основанным на построении иерархии кластеров [31], при этом сами наблюдения после формировании матрицы расстояний не нужны, используется только сама матрица. С другой стороны, нахождение оптимального случая гарантировано только при исчерпывающем поиске с вычислительной сложностью O(2^n), что неприемлемо. В целом основная проблема данного алгоритма именно в его высокой вычислительной сложности. Поэтому в специфических случаях существуют модификации SLINK [32] в случае одиночных связей и CLINK [33] в случае полных, где авторы предлагают решения для существенной оптимизации алгоритма через простой и эффективный способ построения дендрограммы на основе указателей. Существуют и другие методы оптимизации [34], но в целом алгоритм все еще трудно применим для больших наборов данных.

Из успешных примеров использования иерархической кластеризации можно выделить работу Riyahi & Martin [35], где авторы моделировали расширение генерации электроэнергии с учетом использования накопителей на примере энергетической системы Испании, в качестве методов использовалась иерархическая кластеризация в сочетании с методом локтя. Tang et al. [36] разрабатывали архитектуру для морских малых модульных реакторов с использованием нечеткой иерархической кластеризации и улучшенного генетического алгоритма. Zhang et al. [37] исследуют улучшение показателей энергосбережения и уменьшения задержек в сетях WSAN с использованием иерархической кластеризации.

DBSCAN — еще один популярный итеративный алгоритм кластеризации в основе которого лежит плотность распределения наблюдений [38], что сделало его более устойчивым к шуму и выбросам в данных. Кроме того, алгоритм способен обрабатывать кластеры произвольной формы. Идея и популярность алгоритма в задачах с кластерами произвольной формы [39] послужила основой для нескольких его модификаций. В частности, OPTICS, созданный Ankerst et al. [40] как расширение DBSCAN без гиперпараметра epsilon, и HDBSCAN, который был представлен Campello et al. [41] на примере анализа частых мест ДТП. Впоследствии метод расширен тем же коллективом авторов [42], что превратило его не только в инструмент кластеризации, но и обнаружения выбросов.

Обе модификации решают проблему DBSCAN на данных неоднородной плотности. Из примеров можно выделить работу Bíró et al. [43], в которой авторы разработали методику, позволяющую анализировать характеристики изображений с помощью алгоритма DBSCAN. Ozer et al. [44] определяли популярные места использования общественного транспорта с помощью пространственно-временной кластеризации алгоритмом DBSCAN. Последним примером отметим работу Mardani et al. [45], где авторы исследовали автоматическую сегментацию коронарных артерий с использованием аналогичного алгоритма.

Нейросетевые методы представляют собой важную группу подходов к решению задач кластеризации. Их можно разделить на два основных подхода. В рамках первого подхода нейронная сеть является самодостаточным инструментом для выполнения кластеризации. Примером такого метода является алгоритм самоорганизующихся карт (Self-Organizing Maps, SOM) [46], который совмещает функции снижения размерности и кластеризации данных. Другим примером является работа Urme et al. [47], где с помощью SOM исследовали содержание мышьяка, марганца и железа в подземных водах северо-западной части Бангладеш, а также их потенциальных риск для здоровья населения.

Во втором подходе нейронные сети используются исключительно для предварительного снижения размерности данных [48]. В этом случае применяется архитектура типа автоэнкодера, которая обучается так, чтобы исходные данные сначала сжимались в компактный скрытый вектор, а затем восстанавливались обратно с минимальными потерями информации. Такой процесс

позволяет исключить из данных избыточную информацию. Для выполнения кластеризации применяются скрытые векторы, сформированные энкодером, на которые затем накладывается любой из стандартных алгоритмов кластеризации. Например, в данной работе авторы использовали автоэнкодер совместно с K-Means [49].

Выбросы в данных — известная проблема в рамках использования многих алгоритмов работы с данными, в частности она подробно рассматривается в сборнике методов работы с зашумленными данными [50], в Chandola et al. [51] подробно разбираются методы обнаружения выбросов и работы с ними. В контексте анализа данных выбросы представляют собой наблюдения, которые значительно отклоняются от других значений в наборе данных. Эти аномальные значения могут существенно влиять на результаты статистического анализа и интерпретацию данных. Классическим способом борьбы с выбросами является их фильтрация либо установление верхнего порога.

3. Методология и данные

3.1. Характеристика набора данных для исследования

С использованием подходов к конкурентному анализу компаний был сформирован перечень ключевых показателей финансового состояния компаний, на основании которого предлагается проводить кластеризацию компаний. Все показатели были сгруппированы в четырех крупных блока, включающие:

1. Показатели финансовой устойчивости.
2. Показатели оборачиваемости.
3. Показатели финансовых результатов.
4. Показатели ликвидности.

Состав предлагаемых показателей по каждому из блоков представлен в табл. 1.

Таблица 1. **Ключевые показатели финансового состояния компаний, использованные при кластеризации**

Table 1. **Key indicators of the financial condition of companies used in clustering**

Показатели финансовой устойчивости	Показатели оборачиваемости	Показатели финансовых результатов	Показатели ликвидности
Коэффициент автономии	Оборачиваемость активов	Рентабельность активов	Коэффициент абсолютной ликвидности
Коэффициент капитализации	Оборачиваемость дебиторской задолженности	Рентабельность по чистой прибыли	Коэффициент быстрой ликвидности
Коэффициент обеспеченности запасов	Оборачиваемость запасов	Рентабельность продаж	Коэффициент текущей ликвидности

Окончание табл. 1

Показатели финансовой устойчивости	Показатели оборачиваемости	Показатели финансовых результатов	Показатели ликвидности
Коэффициент покрытия активов	Оборачиваемость кредиторской задолженности	Рентабельность задействованного капитала	Коэффициент обеспеченности собственными средствами
Коэффициент покрытия инвестиций	Оборачиваемость оборотных средств	Рентабельность собственного капитала	
Коэффициент покрытия процентов	Фондоемкость		
Коэффициент финансовой зависимости	Фондоотдача		
Коэффициент финансового левериджа			

Источник: составлено авторами.

Рабочий набор данных после очистки состоит из 2 249 наблюдений, которые представляют из себя отчетность крупных (имеющих доход более 2 млрд руб. в год и штатом от 251 сотрудника) производственных (относящихся к производственным группам ОКВЭД) компаний за 2023 г.

Распределение компаний по ОКВЭД представлено на рис. 1. Заметно, что порядка трети компаний по видам деятельности — это производство пищевых продуктов (около 30 %), остальные же группы значительно меньше, в частности следующая группа — это производство минеральной продукции с долей порядка 10 %. Интересно, что распределение в разрезе видов деятельности при анализе общих активов меняется не сильно, доля первой группы снизилась примерно на 5 %, доля остальных чуть возросла (рис. 2).

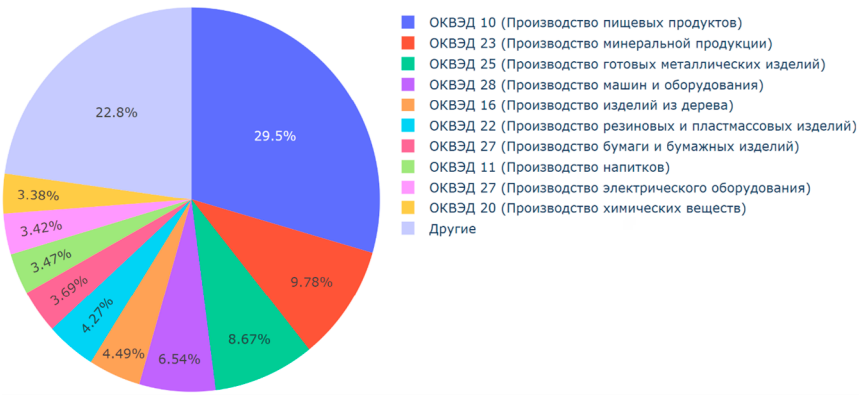

Рис. 1. Диаграмма распределения компаний по ОКВЭД

Figure 1. Diagram of the distribution of companies by OKVED

Источник: составлено авторами.

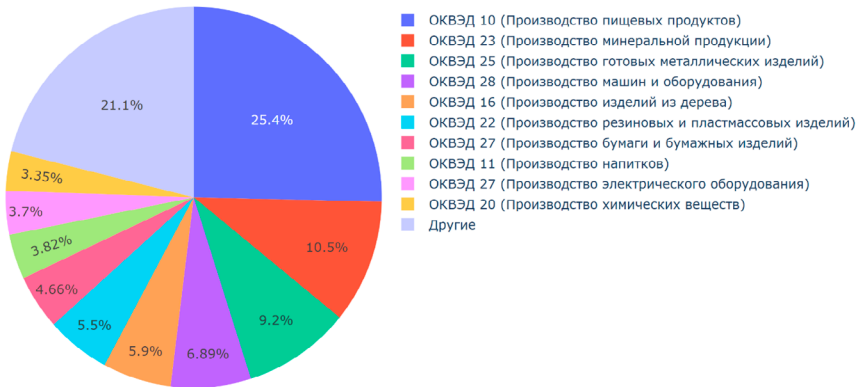

Рис. 2. Диаграмма распределения активов компаний по ОКВЭД

Figure 2. The diagram of the distribution of assets of companies by OKVED (Russian National Classifier of Types of Economic Activity)

Источник: составлено авторами.

Для оценки однородности данных рассмотрим «ящик с усами» по активам (рис. 3). Для того чтобы сделать его нагляднее, верхняя граница диапазона активов была несколько снижена, поэтому порядка 10 самых больших компаний график не показывает, тройка крупнейших компаний по активам в датасете представлена в табл. 2.

Таблица 2. **Тройка компаний-лидеров по сумме активов**

Table 2. **Top three companies in terms of total assets**

Компания	Активы, млрд руб.	Код ОКВЭД	Численность сотрудников
ООО «Амурский ГХК»	464,75	20	564
ООО «НМ-Тех»	216,77	26	398
ООО «РИ-Инвест»	183,47	19	697

Источник: составлено авторами.

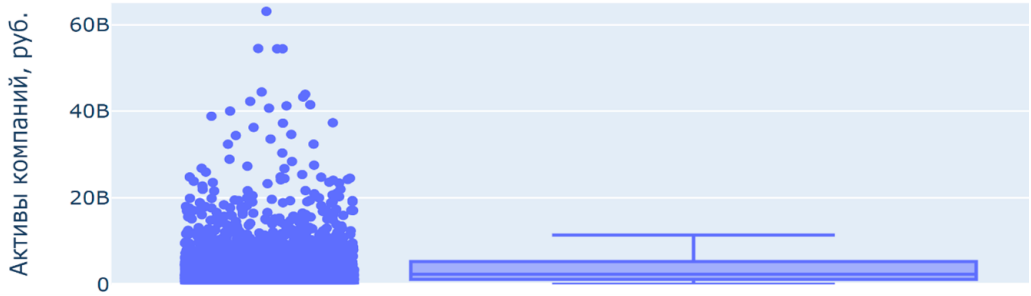

Рис. 3. «Ящик с усами» по активам компаний

Figure 3. The "mustache box" by company assets

Источник: расчеты авторов.

В целом в датасете оказалось порядка 8,5 % выбросов, что не слишком много. В тоже время дисперсия при этом относительно высокая, большая часть компаний расположены в промежутке от примерно одного до пяти миллиардов рублей активов, максимумы в лице гигантов, представленных в табл. 2, находятся очень далеко от медианных значений. Однако в данном исследовании это не проблема, поскольку для кластеризации использовались не балансовые данные, а посчитанные на их основе коэффициенты, которые лежали в основе фильтрации данных.

Для идентификации выбросов использовался метод, основанный на межквартильном размахе (IQR) [50]. Алгоритм этого метода следующий. Сначала вычисляются первый (Q1) и третий (Q3) квартили, после чего определяется IQR как разница между Q3 и Q1. Наблюдения, которые находятся ниже (Q1–1.5 • IQR) или выше (Q3 + 1.5 • IQR), классифицируются как выбросы. Этот метод позволяет эффективно выявлять аномалии и минимизировать их влияние на анализ данных.

3.2. Методология исследования

Ввиду особенностей работы алгоритмов кластеризации сложно предсказать, какой именно из них даст лучший результат на предлагаемом наборе данных. Поэтому предлагается использование трех алгоритмов с последующим сравнением результатов, полученным по каждому из них.

Первый и, вероятно, самый популярный алгоритм — K-Means. Алгоритм делит данные на k-кластеров, где k — заданный параметр. Примерную схему работы алгоритма можно представить следующим образом: сначала выбираются k начальных центроидов кластеров (случайно или на основе определенных правил, но могут и заданы вручную), затем каждая точка данных назначается ближайшему центроиду, формируя k-кластеров. После этого центроиды пересчитываются как среднее значение точек в каждом кластере.

Второй и третий шаг повторяется до сходимости, то есть момента, когда значения точек перестанут изменятся, но на практике часто задается максимальное количество итераций, чтобы алгоритм не занимал слишком много времени. K-Means является простым и вычислительно эффективным алгоритмом, который хорошо справляется с большими объемами данных, результаты которого представляют собой легкие для интерпретации сферические кластеры. Однако алгоритм чувствителен к инициализации. Неудачное количество кластеров или точка старта приведут к неоптимальным параметрам. Именно это было улучшено в K-Means++ [26]. Кроме того, алгоритм не способен работать с несферическими кластерами или кластерными структурами с разной плотностью, а также крайне чувствителен к шуму в данных [22].

Вторым алгоритмом, который будет использован, иерархическая кластеризация. Основная идея этого подхода заключается в построении иерархии кластеров одним из двух подходов: агломеративным (метод «снизу вверх») и дивизионным (метод «сверху вниз»). Основное различие состоит

в том, что дивизионный подход лучше оптимизирован для больших данных, но менее нагляден.

В проводимом исследовании будет использован агломеративный подход, при котором каждая точка данных начинается как отдельный кластер, которые затем последовательно объединяются на основе заданного критерия связи. Таким критерием выбран метод Уорда [52], суть которого заключается в минимизации суммы квадратов отклонений внутри кластеров. Метод позволяет создавать хорошо сбалансированные и однородные кластеры.

Основное преимущество данного вида кластеризации в том, что на начальном этапе нет необходимости как-то определять количество кластеров, дендрограмма позволяет выбрать число кластеров постфактум. Кроме того, метод в целом универсален, так как позволяет использовать различные критерии связи и метрики расстояния внутри кластеров. Из недостатков можно выделить большую вычислительную сложность, что делает использования данного метода затруднительным на больших объемах данных. Кроме того, большая вариативность позволяет испортить результат неправильно подобранным под задачу методом [31].

Последним используемым алгоритмом является DBSCAN — алгоритм кластеризации на основе плотности, который группирует точки, находящиеся в плотных областях, и помечает точки в разреженных областях как шум. Для старта алгоритм требует два параметра — радиус окрестности точки (Epsilon) и минимальное количество точек для создания кластера (MinPts). В процессе работы алгоритм выбирает произвольную начальную точку и находит все точки в пределах радиуса Epsilon. Если количество найденных точек больше или равно MinPts, то создается новый кластер, все точки которого становятся частью текущего кластера и рассматриваются для поиска новых соседей. Алгоритм продолжает работу, пока все точки не будут проверены.

Основным преимуществом данного метода является возможность работать с кластерами любой формы на зашумленных данных. Однако алгоритм требует выбора двух параметров, вычислительно сложен и хуже работает на наборах данных с переменной плотностью [38].

Характеристики алгоритмов кластеризации, предлагаемых в исследовании, приведены в табл. 3.

Таблица 3. **Сравнительная характеристика алгоритмов кластеризации**

Table 3. **Comparative characteristics of clustering algorithms**

Критерий	K-Means	Иерархическая кластеризация	DBSCAN
Механизм	Основан на центроидах	Основан на дендрограмме	Основан на плотности
Входные параметры	Количество кластеров (k)	Нет	Радиус (Epsilon), минимальное количество точек (MinPts)

Окончание табл. 3

Критерий	K-Means	Иерархическая кластеризация	DBSCAN
Форма кластера	Сферическая	Переменная	Произвольная
Обработка шума	Нет	Ограниченная	Да
Масштабируемость	Высокая	Низкая	Умеренная
Вычислительная стоимость	O (n)	O ($n3$) (алгомеративный)	O (nlogn)
Преимущества	Простота, эффективность	Нет стартовых параметров, гибкость	Гибкость, обработка шума
Недостатки	Требует параметр k, чувствителен к шуму	Вычислительно затратен, чувствителен к шуму	Чувствительность к параметрам, вариации плотности

Источник: составлено авторами на основании [22, 26, 31, 38].

Лучший результат будет определен через расчет индексов Дэвиса — Болдина (DBI) [53] и Калински — Харабаса (CHI) [54]. Индекс Дэвиса — Болдина измеряет компактность и разделение кластеров, чем меньше его значение, тем лучше кластеризация, индекс Калински — Харабаса измеряет отношение между разделением кластеров и их компактностью, чем больше его значение, тем лучше качество кластеризации. Подходы к расчету указанных индексов представлены ниже.

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d_{ij}}\right), \quad (1)$$

где k — количество кластеров; σ — среднее внутрикластерное расстояние; d_{ij} — расстояние между центроидами кластеров i и j.

Максимум берется по всем другим кластерам, кроме i-го.

$$CHI = \frac{\sum_{i=1}^{k} n_i \cdot d(c_i, c)}{\sum_{i=1}^{k}\sum_{x\in C_i} d(x, c_i)}, \quad (2)$$

где N_i — количество точек в кластере i; c_i — центроид кластера i; c — глобальный центроид всех данных; C_i — множество точек в кластере i; $d(x, y)$ — расстояние между точками x и y.

Начальное количество кластеров для K-Means определяется силуэтным анализом [55], этот метод основывается на вычислении коэффициента силуэта s для каждого объекта, который отражает, насколько хорошо объект вписывается в свой кластер по сравнению с другими кластерами.

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}, \tag{3}$$

где $a(i)$ — среднее расстояние от объекта i до всех остальных объектов в том же кластере (внутрикластерное расстояние); $b(i)$ — минимальное среднее расстояние от объекта i до объектов в других кластерах (межкластерное расстояние).

Общий коэффициент силуэта можно получить, усреднив результат каждого элемента.

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i), \tag{4}$$

где n — общее количество объектов в выборке; $s(i)$ — значение силуэта для объекта I.

Гиперпараметры для DBSCAN получены простым перебором.

4. Результаты

Оптимальным количеством кластеров оказалось три, что видно на графике силуэта (рис. 4).

4.1. K-Means++

Первый из используемых алгоритмов кластеризации стал K-Means++. Поскольку полноценно визуализировать большое количество измерений не представляется возможным, но авторам хотелось бы получить показатель визуальное различие между алгоритмами, размерность данных будет понижена методом главных компонент. Такой график позволит оценить различия, а также степень перекрытия кластерами друг друга. Получившийся график для K-Means++ представлен на рис. 5. Как можно заметить, присутствует три явных группы объектов, которые слегка перекрывают друг друга в месте соприкосновения.

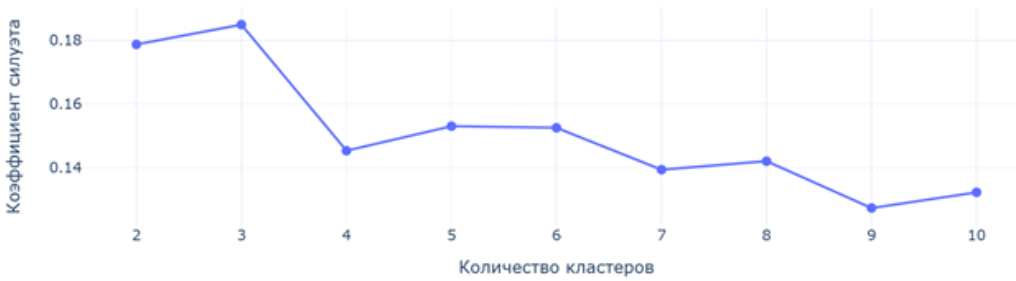

Рис. 4. Силуэтный анализ

Figure 4. Silhouette analysis

Источник: расчеты авторов.

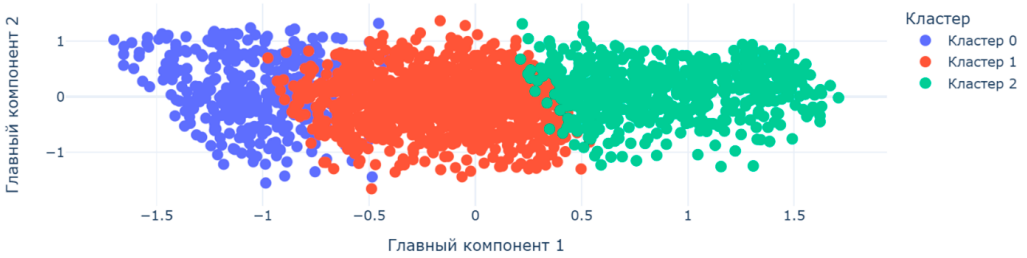

Рис. 5. Кластеризация алгоритмом K-Means++

Figure 5. Clustering by the K-Means++ algorithm

Источник: расчеты авторов.

Рассмотрим подробнее каждый из получившихся кластеров. Оценки компаний с точки зрения активов по кластерам представлены в табл. 4. Заметно, что самым большим оказался первый кластер, куда были отнесены порядка 1 200 компаний — порядка 53 % всего датасета. Нулевой и второй кластеры получили по 18 и 29 % соответственно.

Разброс значений в нулевом кластере больше чем в два раза выше, чем в остальных двух, что объясняется одним единственным выбросом в виде ООО «Амурский ГХК». Если убрать его из анализа, вариация между кластерами практически не отличается, хотя остается высокой. Это в целом ожидаемо — компании могут сильно отличатся по размеру даже внутри крупного бизнеса.

Рассмотрим значения центроидов, в K-Means++ они считаются в явном виде. Конкретный центроид представляет усредненное представление компаний внутри кластера (табл. 5).

Таблица 4. **Оценка размера компаний по кластерам K-Means++**

Table 4. **Estimation of the size of companies by K-Means++ clusters**

Кластер	Количество компаний	Среднее значение активов, млрд руб.	Медианное значение активов, млрд руб.	Среднеквадратическое отклонение
0	409	5,963	2,262	23,711
0 (без ООО «Амурский ГХК»)	408	4,838	2,259	6,721
1	1196	4,644	2,297	10,658
2	644	4,648	2,549	9,734

Источник: расчеты авторов.

Таблица 5. **Положение центроидов по алгоритму K-Means++**

Table 5. **The position of the centroids according to the K-Means++ algorithm**

Показатель	Центроид 0	Центроид 1	Центроид 2	Абсолютное отклонение между центроидами 0 и 2	Относит. отклонение между центроидами 0 и 2
Рентабельность активов	0,000	0,079	0,165	0,165	42663,846
Рентабельность по чистой прибыли	–0,005	0,053	0,133	0,139	2542,457
Рентабельность продаж	0,013	0,082	0,170	0,158	1223,855
Коэффициент быстрой ликвидности	0,081	0,210	0,913	0,832	1030,070
Коэффициент автономии	0,095	0,374	0,770	0,675	710,752
Коэффициент покрытия инвестиций	0,097	0,375	0,771	0,674	697,276
Коэффициент текущей ликвидности	0,802	1,968	4,674	3,872	483,004
Коэффициент покрытия активов	0,663	1,235	3,829	3,167	477,809
Коэффициент покрытия процентов	35,074	48,521	166,575	131,501	374,929
Собственные оборотные средства	–0,868	1,000	1,000	1,868	215,211
Коэффициент обеспеченности запасов	–2,078	0,095	2,100	4,178	201,077
Коэффициент обеспеченности собственными средствами	–0,821	0,013	0,626	1,446	176,242
Оборачиваемость кредиторской задолженности	4,685	7,096	12,266	7,581	161,825
Рентабельность задействованного капитала	0,136	0,221	0,266	0,130	95,656

ISSN 2712-7435

Окончание табл. 5

Показатель	Центроид 0	Центроид 1	Центроид 2	Абсолютное отклонение между центроидами 0 и 2	Относит. отклонение между центроидами 0 и 2
Коэффициент абсолютной ликвидности	0,755	0,844	1,088	0,332	44,024
Фондоотдача	5,868	9,268	7,322	1,455	24,791
Рентабельность собственного капитала	0,197	0,232	0,233	0,036	18,268
Оборачиваемость активов	1,472	1,613	1,375	–0,098	–6,629
Оборачиваемость запасов	8,645	7,344	7,554	–1,092	–12,626
Оборачиваемость дебиторской задолженности	9,840	7,398	8,120	–1,720	–17,478
Фондоемкость	0,359	0,208	0,247	–0,113	–31,337
Оборачиваемость оборотных средств	3,187	2,529	2,167	–1,020	–32,018
Коэффициент финансовой зависимости	3,515	3,257	1,342	–2,173	–61,833
Коэффициент капитализации	2,515	2,257	0,342	–2,173	–86,417
Коэффициент финансового левериджа	2,515	2,257	0,342	–2,173	–86,417

Источник: расчеты авторов.

В целом можно сказать, что компании были разделены на находящиеся в плохом, нормальном и хорошем финансовом состоянии, то есть отнесены к нулевому, первому и второму центроиду соответственно. Это очень удобно для интерпретации. Следует отметить, что алгоритмы кластеризации относят наблюдения к кластерам по принципу максимального подобия, то есть конкретная компания будет отнесена к кластерам по совокупности показателей, и некоторые из них могут кластеру не соответствовать.

Рассмотрим соотношение видов деятельности внутри кластеров (рис. 6–8). Во всех трех кластерах с большим отрывом лидирует группа производства пищевых продуктов, занимающая примерно треть кластера, но дальше появляются различия. В частности, в топ 10 ОКВЭД по количеству компаний в кластере 1 («нормальном») и 2 («хорошем») появляются виды деятельности, которые не представлены в большом количестве в кластере 0 («плохом»). Например, речь идет об ОКВЭД 22 «Производство резиновых и пластмассовых изделий», 27 «Производство электрического оборудования», 31 «Производство мебели» и 26 «Производство компьютеров, электронных и оптических изделий».

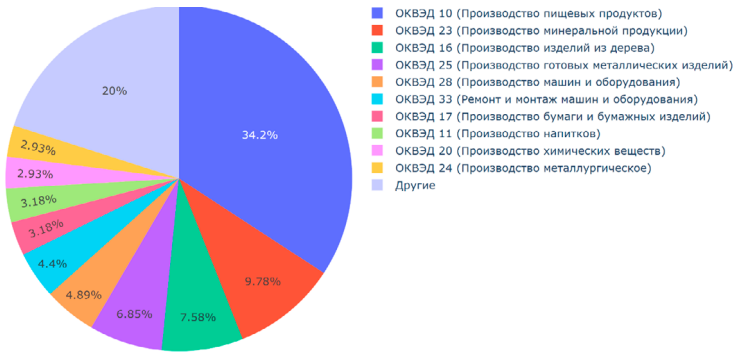

Рис. 6. Диаграмма распределения компаний по ОКВЭД, K-Means++, кластер 0

Figure 6. The diagram of the distribution of companies by OKVED, K-Means++, cluster 0

Источник: расчеты авторов.

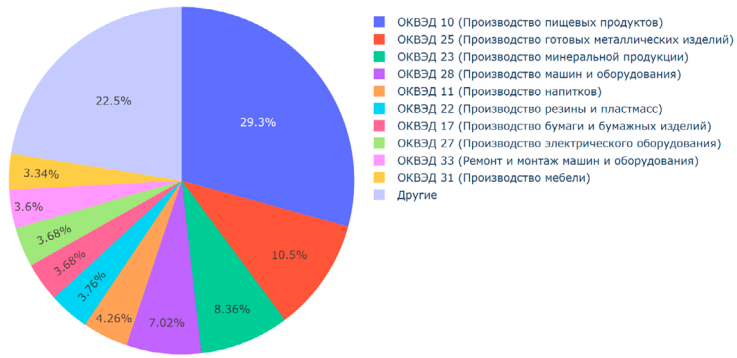

Рис. 7. Диаграмма распределения компаний по ОКВЭД, K-Means++, кластер 1

Figure 7. The diagram of the distribution of companies by OKVED, K-Means++, cluster 1

Источник: расчеты авторов.

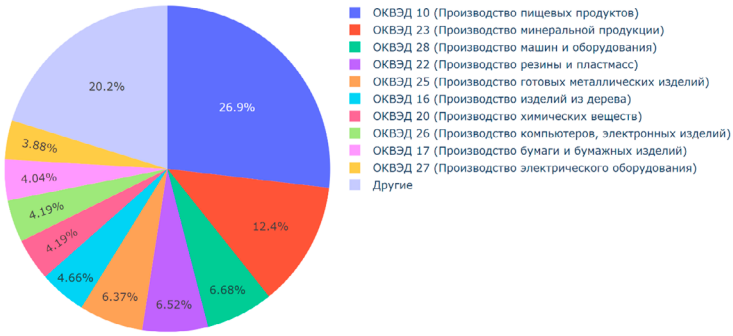

Рис. 8. Диаграмма распределения компаний по ОКВЭД, K-Means++, кластер 2

Figure 8. The diagram of the distribution of companies by OKVED, K-Means++, cluster 2

Источник: расчеты авторов.

В целом результаты работы рассматриваемого алгоритма выглядят хорошо, но все еще есть рад проблем, главной из которых является наложение кластеров 0 и 1, хорошо заметное на рис. 5. В остальном полученные результаты кажутся адекватными.

4.2. Иерархическая кластеризация

Дендрограмма, построенная на основании расчетов по алгоритму иерархической кластеризации, представлена на рис. 9. Евклидово расстояние, по которому будут разделены кластеры, было выбрано равным 20. Как следствие, формируются три кластера. Результаты иерархической кластеризации представлены на рис. 10.

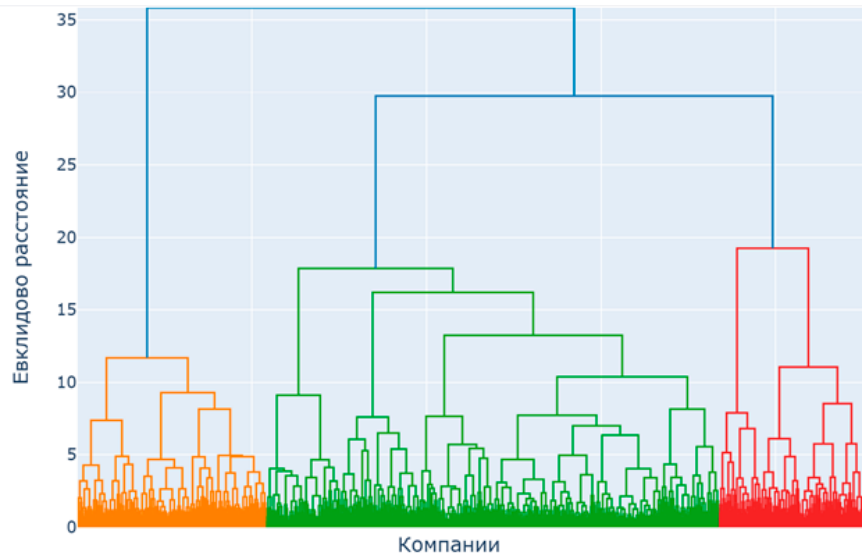

Рис. 9. Дендрограмма иерархической кластеризации

Figure 9. Hierarchical clustering dendrogram

Источник: расчеты авторов.

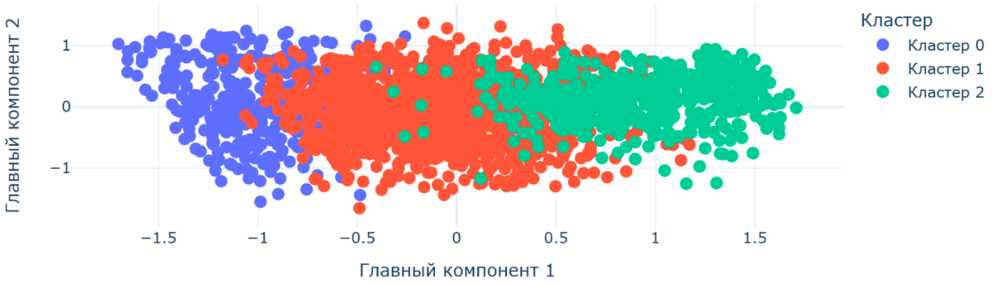

Рис. 10. Результаты иерархической кластеризации

Figure 10. Results of hierarchical clustering

Источник: расчеты авторов.

Следует отметить, что результат похож на K-Means++, но кластеры сильнее перекрывают друг друга. В меньшей степени это заметно для кластеров 0 и 1, в большей для кластеров 1 и 2, где пересечение практически до половины кластеров. Качество кластеризации этим алгоритмом на данном наборе данных заметно ниже.

Оценим размеры компаний внутри кластеров и их вариацию (табл. 6). В целом похоже на результаты предыдущего алгоритма. Большое среднеквадратическое отклонение в кластере 0 из-за выброса среди компаний — компании ООО «Амурский ГХК». Если его убрать вариация среди кластеров очень близка.

Иерархическая кластеризация не считает центроиды в явном виде, но, поскольку авторы использовали метод Уорда, который аналогично K-Means++ пытается составить компактные кластеры по сумме квадратов расстояний, а точка, в которой минимизируется квадрат расстояний — среднее арифметическое, которое можно посчитать для каждого кластера и которое

Таблица 6. **Оценка размера компаний по кластерам по алгоритму иерархической кластеризации**

Table 6. **Estimation of the size of companies by clusters according to the hierarchical clustering algorithm**

Кластер	Количество компаний	Среднее значение активов, млрд руб.	Медианное значение активов, млрд руб.	Среднеквадратическое отклонение
0	410	5,786	2,126	23,686
0 (без ООО «Амурский ГХК»)	409	4,664	2,107	6,696
1	1297	4,707	2,283	11,776
2	542	4,630	2,721	5,616

Источник: расчеты авторов.

допустимо сравнивать с центроидами K-Means++, чтоб оценить различия. Средние значения представлены в табл. 7. В целом прослеживается похожая динамика, финансовое состояние компаний улучшается от кластера 0 к кластеру 2, но грань между кластерами размыта, значительная часть показателей от кластера к кластеру могут отличаться несущественно, что видно при наложении кластеров друг на друга на рис. 10.

Рассмотрим распределение компаний по областям деятельности в разрезе кластеров (рис. 11–13). В целом распределение компаний схоже с предыдущим алгоритмом. Во всех трех кластерах больше всего представлены компании пищевой промышленности, доля которых несколько падает от кластера 0 к кластеру 2. Аналогично ряд видов деятельности, не представленные в кластере в топ-10 по частоте, появляются в кластерах 1 и 2. Различия тоже есть, в основном в конкретных долях.

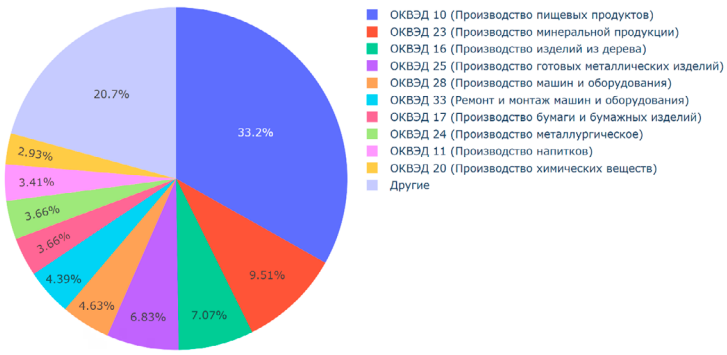

Рис. 11. Диаграмма распределения компаний по ОКВЭД, иерархическая кластеризация, кластер 0

Figure 11. The diagram of the distribution of companies by OKVED, hierarchical clusterization, cluster 0

Источник: расчеты авторов.

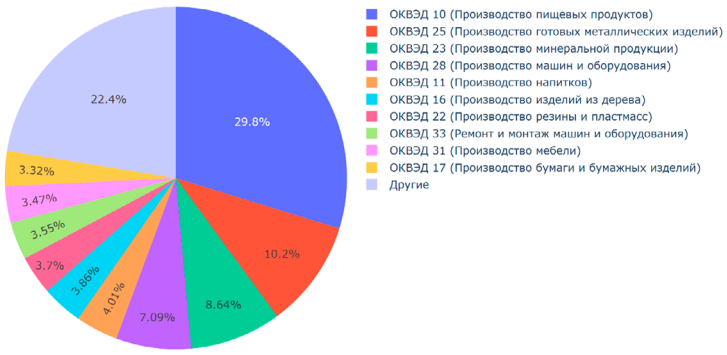

Рис. 12. Диаграмма распределения компаний по ОКВЭД, иерархическая кластеризация, кластер 1

Figure 12. The diagram of the distribution of companies by OKVED, hierarchical clusterization, cluster 1

Источник: расчеты авторов.

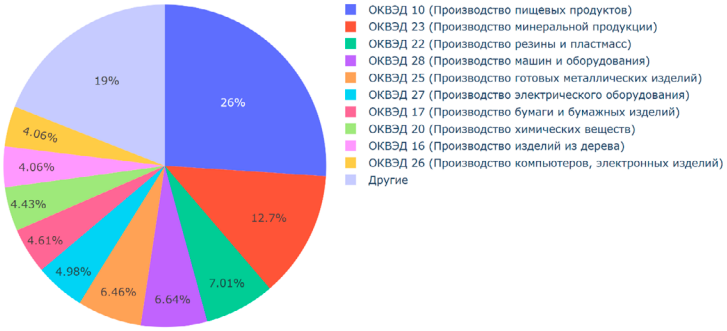

Рис. 13. Диаграмма распределения компаний по ОКВЭД, иерархическая кластеризация, кластер 2

Figure 13. The diagram of the distribution of companies by OKVED, hierarchical clusterization, cluster 2

Источник: расчеты авторов.

К сожалению, результаты работы иерархической кластеризации на данном наборе данных оказался значительно хуже предыдущего алгоритма ввиду попарного перекрытия кластеров 0 и 1, 1 и 2. Анализ средних показал эти же проблемы. Хотя тренд на улучшение финансового состояния заметен, некоторые значения оказались очень близкими.

4.3. DBSCAN

Последним рассмотрим результаты DBSCAN. Как указывалось ранее, данный алгоритм определяет количество кластеров самостоятельно при учете заданных параметров. Лучший результат у DBSCAN по индексам был получен при варианте кластеризации, указанном на рис. 14. Можно заметить, что алгоритм выделил один кластер, а четверть датасета отметил как выбросы (метка «–1»). К сожалению, такие результаты являются неудовлетворительными, более подробный анализ смысла не имеет. Возможно, что алгоритм показал такие результаты из-за высокой вариации признаков.

Рис. 14. Кластеризация алгоритмом DBSCAN

Figure 14. Clustering by the DBSCAN algorithm

Источник: расчеты авторов.

Таблица 7. **Средние значения показателей в кластерах иерархической кластеризации**

Table 7. **Average values of indicators in hierarchical clustering clusters**

Показатель	Кластер 0	Кластер 1	Кластер 2	Абсолютное отклонение между центроидами 0 и 2	Относит. отклонение между центроидами 0 и 2
Собственные оборотные средства	0	1	1	1	inf
Коэффициент быстрой ликвидности	0,021	0,079	0,874	0,853	4101,378
Коэффициент текущей ликвидности	0,117	0,260	0,755	0,638	544,724
Коэффициент покрытия активов	0,200	0,313	0,832	0,632	315,633
Оборачиваемость кредиторской задолженности	0,133	0,265	0,514	0,382	287,032
Коэффициент обеспеченности собственными средствами	0,248	0,646	0,873	0,625	252,108
Коэффициент обеспеченности запасов	0,238	0,498	0,700	0,461	193,673
Коэффициент автономии	0,396	0,614	0,873	0,477	120,331
Коэффициент покрытия инвестиций	0,399	0,614	0,875	0,475	119,034
Рентабельность продаж	0,358	0,458	0,673	0,314	87,668
Рентабельность по чистой прибыли	0,363	0,451	0,675	0,312	86,130
Рентабельность активов	0,360	0,460	0,626	0,266	73,817
Коэффициент покрытия процентов	0,375	0,386	0,618	0,242	64,585
Фондоотдача	0,155	0,282	0,221	0,067	43,138
Рентабельность задействованного капитала	0,398	0,475	0,528	0,129	32,501

Окончание табл. 7

Показатель	Кластер 0	Кластер 1	Кластер 2	Абсолютное отклонение между центроидами 0 и 2	Относит. отклонение между центроидами 0 и 2
Коэффициент абсолютной ликвидности	0,494	0,578	0,623	0,129	26,147
Рентабельность собственного капитала	0,460	0,473	0,495	0,035	7,614
Оборачиваемость запасов	0,303	0,269	0,293	–0,010	–3,399
Оборачиваемость активов	0,331	0,386	0,318	–0,013	–3,951
Оборачиваемость дебиторской задолженности	0,371	0,281	0,286	–0,084	–22,784
Фондоемкость	0,400	0,219	0,280	–0,121	–30,137
Оборачиваемость оборотных средств	0,448	0,367	0,301	–0,147	–32,769
Коэффициент финансового левериджа	0,602	0,490	0,366	–0,236	–39,223
Коэффициент капитализации	0,602	0,490	0,366	–0,236	–39,223
Коэффициент финансовой зависимости	0,602	0,490	0,366	–0,236	–39,223

Источник: расчеты авторов.

ISSN 2712-7435

4.4. Выбор лучшего алгоритма

Результаты работы алгоритмов по индексам Дэвиса — Болдина и Калински — Харабаса представлены в табл. 8.

Таблица 8. **Показатели качества кластеризации по различным алгоритмам**

Table 8. **Clustering quality indicators for various algorithms**

Алгоритм	Индекс Дэвиса — Болдина (меньше лучше)	Индекс Калински — Харабаса (больше лучше)
K-Means++	1,847	506,024
Иерархическая кластеризация	1,880	450,037
DBSCAN	2,023	378,825

Источник: расчеты авторов.

В целом значение индексов Дэвиса — Болдина ниже 2,5 и Калински — Харабаса выше 400 для трех кластеров и 11 измерений можно считать хорошим результатом. K-Means++ и иерархическая кластеризация показали близкий и результат по индексу Дэвиса — Болдина в 1,847 и 1,880, что говорит о хорошей компактности внутри кластеров и их слабом перекрытии.

По индексу Калински — Харабаса ситуация схожа с тем различием, что K-Means++ ощутимо опережает иерархическую кластеризацию, демонстрируя значение выше пяти сотен, что говорит о четко выделенных компактных кластерах. DBSCAN, к сожалению, на данном наборе данных показал себя плохо.

Таким образом, лучшим алгоритмом из трех рассмотренных становится K-Means++ с небольшим отрывом по индексу Дэвиса — Болдина и ощутимым по индексу Калински — Харабаса.

5. Обсуждение

Если представить конкретный центроид как усредненную характеристику компаний, которые к нему относятся, то можно сказать, что действительно получилось поделить компании на находящиеся в плохом, нормальном и хорошем финансовом состоянии, то есть отнесенные к нулевому, первому и второму центроиду соответственно. Следует отметить, что алгоритмы кластеризации относят наблюдения к кластерам по принципу максимального подобия. Поэтому конкретная компания будет отнесена к кластерам по совокупности показателей и некоторые из них могут кластеру не соответствовать. Все компании разные, и очень немногие из них находятся в абсолютно идеальном или в очень плохом финансовом состоянии.

Рассмотрим некоторые примеры компаний, отнесенные к тем или иным кластерам (табл. 9): АО «Краслесинвест» — представитель кластера 0, ОАО «Тамбовский хлебозавод» — представитель кластера 1 и АО «Тракья Гласс

Рус» — представитель кластера 2. Компании относятся к ОКВЭД 16, 10 и 23 соответственно и различаются по размерам, в частности самое крупное предприятие АО «Краслесинвест», за ним идет АО «Тракья Гласс Рус» и самым небольшим в данном сравнении является ОАО «Тамбовский хлебозавод». Уже здесь можно отметить, что распределение по кластерам не зависит от размера компаний.

Таблица 9. **Примеры компаний — представителей разных кластеров**

Table 9. **Examples of companies representing different clusters**

Показатель	АО «Краслесинвест»	ОАО «Тамбовский хлебозавод»	АО «Тракья Гласс Рус»
Кластер	0	1	2
Коэффициент автономии	0,855	0,567	0,942
Коэффициент капитализации	0,169	0,763	0,062
Коэффициент обеспеченности запасов	–2,002	0,765	9,377
Коэффициент покрытия активов	6,495	1,510	16,154
Коэффициент покрытия инвестиций	0,855	0,567	0,942
Коэффициент покрытия процентов	–7,914	5,930	inf
Коэффициент финансовой зависимости	1,169	1,763	1,062
Коэффициент финансового левериджа	0,169	0,763	0,062
Коэффициент абсолютной ликвидности	0,750	0,852	9,451
Коэффициент быстрой ликвидности	0,084	0,115	10,699
Коэффициент текущей ликвидности	1,369	1,479	12,259
Коэффициент обеспеченности собственными средствами	–1,307	0,280	0,915
Рентабельность по чистой прибыли	–0,154	0,004	0,213
Рентабельность активов	–0,027	0,018	0,093
Рентабельность задействованного капитала	–0,030	0,076	0,122

Окончание табл. 9

Показатель	АО «Краслесинвест»	ОАО «Тамбовский хлебозавод»	АО «Тракья Гласс Рус»
Рентабельность продаж	–0,159	0,010	0,263
Рентабельность собственного капитала	–0,032	0,031	0,099
Оборачиваемость активов	0,178	4,440	0,437
Оборачиваемость дебиторской задолженности	10,391	13,465	21,472
Оборачиваемость запасов	4,344	20,178	6,517
Оборачиваемость кредиторской задолженности	5,180	12,814	7,800
Оборачиваемость оборотных средств	2,835	7,385	0,636
Фондоемкость	3,927	0,071	0,528
Фондоотдача	0,255	14,011	1,893

Кратко охарактеризуем финансовое состояние этих трех компаний. Заметно, что АО «Краслесинвест» — представитель кластера 0 — демонстрирует признаки кризисного финансового состояния, компания убыточна — все показатели рентабельности ниже нуля, некоторые ощутимо ниже, в частности рентабельность по чистой прибыли (–15 %). Компания финансово неустойчива из-за отрицательного показателя обеспеченности собственными оборотными средствами, а два из трех показателя ликвидности, кроме абсолютной ликвидности, значительно ниже нормы.

Положение ОАО «Тамбовский хлебозавод» — компании из кластера 1 — ощутимо лучше, хотя компания все еще имеет ряд проблем, в частности ее рентабельность близка к нулю, а показатели быстрой и текущей ликвидности ниже нормы. Кроме того, долговая нагрузка компании находится на грани допустимого.

Что касается примера кластера 2, АО «Тракья Гласс Рус», проблем здесь как таковых нет, компания имеет высокую рентабельность, очень большой запас ликвидности и низкую долговую нагрузку.

Таким образом, обе гипотезы можно считать подтвержденными. У авторов действительно получилось группировать компании по финансовому состоянию средствами машинного обучения. В свою очередь это означает, что уже на текущем этапе другую компанию производственного сектора сравнимого размера можно определить к одному из кластеров, используя обученный на полном наборе данных классификатор.

Следует уточнить явные ограничения. Часть из них уже были озвучены ранее: компании относятся к конкретному кластеру по принципу

максимального подобия, возможны граничные случаи, возможно неполное соответствие всех показателей итоговому кластеру. Кроме того, с ходом времени полученная модель будет устаревать, что неминуемо скажется на итоговом качестве. Поэтому модель следует обновлять на актуальном наборе данных хотя бы раз в несколько лет.

6. Заключение

В данной работе авторы ставили перед собой цель — получить первичные метки классов для крупных производственных компаний России. Можно сказать, что цель достигнута, а рассматриваемые в работе гипотезы нашли свое подтверждение.

Были использованы несколько алгоритмов кластеризации, лучшим из которых оказался K-Means. Результаты работы алгоритмов были проанализированы на предмет аномалий, общую адекватность и соответствие поставленной задаче.

Следует отметить, что результаты работы K-Means++ и иерархической кластеризации в соответствии с критерием Уорда были во многом похожи, хотя кластеры после иерархической кластеризации оказались хуже разделены и частично перекрывали друг друга, что отразилось на средних значениях показателей соседних кластеров, грань между некоторыми из них очень тонка. Таким образом, лучшим алгоритмом с некоторым отрывом по индексу Калински — Харабаса оказался K-Means++.

Данные были разделены на три кластера в соответствии с силуэтным анализом. После анализа характеристик центроидов выяснилось, что компании действительно поделились на условно находящиеся в плохом, хорошем и отличном финансовом состоянии. Однако следует отметить, что компании были отнесены к конкретному кластеру по принципу максимального подобия по совокупности показателей и некоторые частные показатели отдельных компаний могут не соответствовать их кластеру.

В дальнейшем полученные метки классов с некоторыми ручными корректировками будут использованы для моделирования на базе машинного обучения с учителем с целью получить инструмент, способный быстро и достаточно качественно оценить финансовое состояние компании относительно других компаний в отрасли по данным стандартной финансовой отчетности.

Список использованных источников

1. *Kryzanowski L., Galler M., Wright D. W.* Using artificial neural networks to pick stocks // Financial Analysts Journal. 1993. Vol. 49, Issue 4. Pp. 21–27. DOI: https://doi.org/10.2469/faj.v49.n4.21

2. *Porter M. E.* The Competitive Advantage of Nations. New York : Free Press, 1990. 855 p. URL: https://archive.org/details/competitiveadvan0000port

3. *Porter M. E.* The five competitive forces that shape strategy // Harvard Business Review. 2008. Vol. 86, No. 1. Pp. 78–93. URL: https://sistemasgerenciales.wordpress.com/wp-content/uploads/2016/04/the-five-competitive-forces-that-shape-strategy.pdf

4. *Cao X., Shen X., Liu Q.* Mechanism of aquaculture competitiveness in China // Aquaculture Reports. 2024. Vol. 37. 102195. https://doi.org/10.1016/j.aqrep.2024.102195

5. *Tao X., Cai W.* The impact of digital finance on export competitiveness: Evidence from Chinese manufacturing enterprises // Finance Research Letters. 2025. Vol. 73. 106629. https://doi.org/10.1016/j.frl.2024.106629

6. *Liu Y.-L., Tian L., Li C., Wu Ya.* Analyzing the competitiveness and strategies of Chinese mobile network operators in the 5G era // Telecommunications Policy. 2024. Vol. 48, Issue 2. 102652. https://doi.org/10.1016/j.telpol.2023.102652

7. *Fang K., Zhou Y., Wang S., Ye R., Guo S.* Assessing national renewable energy competitiveness of the G20: A revised Porter's Diamond Model // Renewable and Sustainable Energy Reviews. 2018. Vol. 93. Pp. 719–731. https://doi.org/10.1016/j.rser.2018.05.011

8. *Cibinskiene A., Dumciuviene D., Bobinaite V., Dragašius E.* Competitiveness of industrial companies forming the value chain of wind energy components: The case of Lithuania // Sustainability. 2021. Vol. 13, Issue 16. 9255. https://doi.org/10.3390/su13169255

9. *Liu J., Wei Q., Dai Q., Liang C.* Overview of wind power industry value chain using diamond model: A case study from China // Applied Sciences. 2018. Vol. 8, Issue 10. 1900. https://doi.org/10.3390/app8101900

10. *Lau A. K.W., Baark E., Lo W. L.W., Sharif N.* The effects of innovation sources and capabilities on product competitiveness in Hong Kong and the Pearl River Delta // Asian Journal of Technology Innovation. 2013. Vol. 21, Issue 2. Pp. 220–236. https://doi.org/10.1504/IJTM.2012.047244

11. *Lotfi B., Karim M.* Competitiveness determinants of Moroccan exports: quantity-based analysis // International Journal of Economics and Finance. 2016. Vol. 8, Issue 7. Pp. 140–148. https://doi.org/10.5539/ijef.v8n7p140

12. *Li Ya., Yu H., Shen Z.* Dynamic prediction of product competitive position: A multisource data-driven competitive analysis framework from a multi-competitor perspective // Journal of Retailing and Consumer Services. 2025. Vol. 25. 104289. https://doi.org/10.1016/j.jretconser.2025.104289

13. *Kim S.-A., Park S., Kwak M., Kang C.* Examining product quality and competitiveness via online reviews: An integrated approach of importance performance competitor analysis and Kano model // Journal of Retailing and Consumer Services. 2025. Vol. 82. 104135. https://doi.org/10.1016/j.jretconser.2024.104135

14. *Buckley P. J., Pass C. L., Prescott K.* Measures of international competitiveness: A critical survey // Journal of Marketing Management. 1988. Vol. 4, Issue 2. Pp. 175–200. https://doi.org/10.1080/0267257X.1988.9964068

15. *Schefczyk M.* Operational performance of airlines: an extension of traditional measurement paradigms // Strategic Management Journal. 1993. Vol. 14, No. 4. Pp. 301–317. https://doi.org/10.1002/smj.4250140406

16. *Good D. H., Nadiri M. I., Roller L. H., Sickles R. C.* Efficiency and productivity growth comparisons of European and U.S. airlines: A first look at the data // The Journal of Productivity Analysis. 1993. Vol. 4. Pp. 115–125. https://doi.org/10.1007/BF01073469

17. *Parkan C., Wu M.-L.* Measurement of the performance of an investment bank using the operational competitiveness rating procedure // Omega. 1999. Vol. 27, Issue 2. Pp. 201–217. https://doi.org/10.1016/S0305-0483(98)00041-3

18. *Fleisher C. S., Bensoussan B. E.* Business and Competitive Analysis: Effective Application of New and Classic Methods. Second Edition. New Jersey : Pearson, 2015. 448 p. https://doi.org/10.24883/IberoamericanIC.v2i2.37

19. *Prescott J. E., Grant J. H.* A manager's guide for evaluating competitive analysis techniques // Interfaces. 1988. Vol. 18, Issue 3. Pp. 10–22. https://doi.org/10.1287/inte.18.3.10

20. *Steinhaus H.* Sur la division des corps matériels en parties // Bulletin L'Académie Polonaise des Science. 1956. Vol. 4, No. 12. Pp. 801–804. URL: http://laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf

21. *Lloyd S.* Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28, Issue 2. Pp. 129–137. https://doi.org/10.1109/TIT.1982.1056489

22. *MacQueen J.* Some methods for classification and analysis of multivariate observations // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press, 1967. Pp. 281–297. URL: https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1967-MacQueen-MSP.pdf

23. *Kanungo T., Mount D. M., Netanyahu N. S., Piatko C. D., Silverman R., Wu A. Y.* An efficient k-means clustering algorithm: Analysis and implementation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24, Issue 7. Pp. 881–892. https://doi.org/10.1109/TPAMI.2002.1017616

24. *Ostrovsky R., Rabani Yu., Schulman L. J., Swamy C.* The effectiveness of Lloyd-type methods for the k-means problem // Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). IEEE, 2006. Pp. 165–176. https://doi.org/10.1109/FOCS.2006.75

25. *Ahmed M., Seraj R., Islam S. M.S.* The k-means algorithm: A comprehensive survey and performance evaluation // Electronics. 2020. Vol. 9, Issue 8. 1295. https://doi.org/10.3390/electronics9081295

26. *Arthur D., Vassilvitskii S.* K-means++: The advantages of careful seeding // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007). Society for Industrial and Applied Mathematics3600 University City Science Center Philadelphia, 2007. Pp. 1027–1035. https://doi.org/10.1145/1283383.1283494

27. *Abdulnassar A. A., Nair L. R.* Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model // Measurement: Sensors. 2023. Vol. 25. 100666. https://doi.org/10.1016/j.measen.2023.100666

28. *Ay M.,* Özbakır *L., Kulluk S., Gülmez B.,* Öztürk *G.,* Özer *S.* FC-Kmeans: Fixed-centered K-means algorithm // Expert Systems with Applications. 2023. Vol. 211. 118656. https://doi.org/10.1016/j.eswa.2022.118656

29. *Li J., Li J., Wang D., et al.* Hierarchical and partitioned planning strategy for closed-loop devices in low-voltage distribution network based on improved KMeans partition method // Energy Reports. 2023. Vol. 9. Pp. 477–485. https://doi.org/10.1016/j.egyr.2023.05.161

30. *He J., Jiang D., Zhang D., Li J., Fei Q.* Interval model validation for rotor support system using Kmeans Bayesian method // Probabilistic Engineering Mechanics. 2022. Vol. 70. 103364. https://doi.org/10.1016/j.probengmech.2022.103364

31. *Nielsen F.* Introduction to HPC with MPI for Data Science. Springer, 2016. 282 p. https://doi.org/10.1007/978-3-319-21903-5

32. *Sibson R.* SLINK: an optimally efficient algorithm for the single-link cluster method // The Computer Journal. 1973. Vol. 16, Issue 1. Pp. 30–34. https://doi.org/10.1093/comjnl/16.1.30

33. *Defays D.* An efficient algorithm for a complete link method // The Computer Journal. 1977. Vol. 20, Issue 4. Pp. 364–366. https://doi.org/10.1093/comjnl/20.4.364

34. *Eppstein D.* Fast hierarchical clustering and other applications of dynamic closest pairs // Journal of Experimental Algorithmics (JEA). 2000. Vol. 5. Pp. 1-es. https://doi.org/10.1145/351827.351829

35. *Riyahi M., Martín A. G.* Optimizing Capacity Expansion Modeling with a Novel Hierarchical Clustering and Systematic Elbow Method: A Case study on Power and Storage Units in Spain // Energy. 2025. Vol. 323. 135788. https://doi.org/10.1016/j.energy.2025.135788

36. *Tang Z., Wang L., Guo S., Liang G., Zhang W., Zhang L., Rui M., Guan G., Wang Yu.* Study on modular design methodology of marine SMR system based on fuzzy hierarchical clustering and improved genetic algorithm // Progress in Nuclear Energy. 2025. Vol. 185. 105739. https://doi.org/10.1016/j.pnucene.2025.105739

37. *Zhang X., Wang Y.-L., Byun H.* Enhancing energy saving and reducing latency by divisive hierarchical clustering with Multi-UAVs in WSANs // Applied Soft Computing. 2025. 112861. https://doi.org/10.1016/j.asoc.2025.112861

38. *Ester M., Kriegel H. P., Sander J., Xu X.* A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the 1996 Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 1996. Pp. 226–231. URL: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

39. *Schubert E., Sander J., Ester M., Kriegel H. P., Xu X.* DBSCAN revisited, revisited: why and how you should (still) use DBSCAN // ACM Transactions on Database Systems. 2017. Vol. 42, No. 3. Pp. 1–21. https://doi.org/10.1145/3068335

40. *Ankerst M., Breunig M. M., Kriegel H. P., Sander J.* OPTICS: Ordering points to identify the clustering structure // Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD '99). ACM New York, 1999. Pp. 49–60. https://doi.org/10.1145/304182.304187

41. *Campello R. J., Moulavi D., Sander J.* Density-based clustering based on hierarchical density estimates // Advances in Knowledge Discovery and Data Mining. Pacific-Asia Conference on Knowledge Discovery and Data Mining / edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu. Springer, 2013. Pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

42. *Campello R. J., Moulavi D., Zimek A., Sander J.* Hierarchical density estimates for data clustering, visualization, and outlier detection // ACM Transactions on Knowledge Discovery from Data (TKDD). 2015. Vol. 10, No. 1. Pp. 1–51. https://doi.org/10.1145/2733381

43. *Bíró P. Kovács B. B.H., Novák T., Erdélyi M.* Cluster parameter-based DBSCAN maps for image characterization // Computational and Structural Biotechnology Journal. 2025. Vol. 27. Pp. 920–927. https://doi.org/10.1016/j.csbj.2025.02.037

44. *Ozer F. C., Tuydes-Yaman H., Dalkic-Melek G.* Increasing the precision of public transit user activity location detection from smart card data analysis via spatial-temporal DBSCAN // Data & Knowledge Engineering. 2024. Vol. 153. 102343. https://doi.org/10.1016/j.datak.2024.102343

45. *Mardani K., Maghooli K., Farokhi F.* Segmentation of coronary arteries from X-ray angiographic images using density based spatial clustering of applications with noise (DBSCAN) // Biomedical Signal Processing and Control. 2025. Vol. 101. 107175. https://doi.org/10.1016/j.bspc.2024.107175

46. *Kohonen T.* Self-organized formation of topologically correct feature maps // Biological Cybernetics. 1982. Vol. 43. Pp. 59–69. https://doi.org/10.1007/BF00337288

47. *Urme O., Reza S., Adham Md I., Sattar G. S.* Arsenic, manganese, and iron concentration in groundwater of northwestern part of Bangladesh using self-organizing maps: Implication for health risk assessment // Heliyon. 2025. Vol. 11, Issue 2. e41805. https://doi.org/10.1016/j.heliyon.2025.e41805

48. *Boubekki A., Kampffmeyer M., Brefeld U., Jenssen R.* Joint optimization of an autoencoder for clustering and embedding // Machine Learning. 2021. Vol. 110, No. 7. Pp. 1901–1937. https://doi.org/10.1007/s10994-021-06015-5

49. *Pulgar F. J., Charte F., Rivera A. J., del Jesus M. J.* AEkNN: An AutoEncoder kNN-based classifier with built-in dimensionality reduction // ArXiv Preprint arXiv:1802.08465. 2018. 35 p. https://doi.org/10.48550/arXiv.1802.08465

50. Understanding Robust and Exploratory Data Analysis / edited by D. C. Hoaglin, F. Mosteller, J. W. Tukey. John Wiley & Sons, 1983. 447 p. URL: https://archive.org/details/understandingrob0000unse/page/n7/mode/1up

51. *Chandola V., Banerjee A., Kumar V.* Anomaly detection: A survey // ACM Computing Surveys. 2009. Vol. 41, No. 3. Pp. 1–58. https://doi.org/10.1145/1541880.1541882

52. *Ward Jr J. H.* Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. 1963. Vol. 58, Issue 301. Pp. 236–244. https://doi.org/10.1080/01621459.1963.10500845

53. *Davies D. L., Bouldin D. W.* A cluster separation measure // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. Vol. PAMI-1, Issue 2. Pp. 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

54. *Caliński T., Harabasz J.* A dendrite method for cluster analysis // Communications in Statistics. 1974. Vol. 3, Issue 1. Pp. 1–27. https://doi.org/10.1080/03610927408827101

55. *Rousseeuw P. J.* Silhouettes: a graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. Pp. 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

ИНФОРМАЦИЯ ОБ АВТОРАХ

Буланов Лев Алексеевич

Аспирант кафедры экономической безопасности производственных комплексов Уральского федерального университета имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID https://orcid.org/0009-0001-0242-0127 e-mail: levbulanov2013@yandex.ru

Калина Алексей Владимирович

Кандидат технических наук, доцент кафедры экономической безопасности производственных комплексов Уральского федерального университета имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19), старший научный сотрудник Центра экономической безопасности Института экономики Уральского отделения РАН, г. Екатеринбург, Россия (620014, г. Екатеринбург, ул. Московская, 29); ORCID https://orcid.org/0000-0003-0376-2505 e-mail: alexkalina74@mail.ru

Криворотов Вадим Васильевич

Доктор экономических наук, профессор, заведующий кафедрой экономической безопасности производственных комплексов Уральского федерального университета имени первого Президента России Б. Н. Ельцина, г. Екатеринбург, Россия (620002, г. Екатеринбург, ул. Мира, 19); ORCID https://orcid.org/0000-0002-7066-0325 e-mail: v_krivorotov@mail.ru

ДЛЯ ЦИТИРОВАНИЯ

Буланов Л. А., Калина А. В., Криворотов В. В. Кластеризация российских производственных компаний по показателям их финансового состояния с использованием технологий машинного обучения // Journal of Applied Economic Research. 2025. Т. 24, № 2. С. 584–621. https://doi.org/10.15826/vestnik.2025.24.2.020

ИНФОРМАЦИЯ О СТАТЬЕ

Дата поступления 21 марта 2025 г.; дата поступления после рецензирования 7 апреля 2025 г.; дата принятия к печати 12 апреля 2025 г.

Clustering of Russian Manufacturing Companies by Indicators of Their Financial Condition Using Machine Learning Technologies

Lev A. Bulanov[1], ***Alexei V. Kalina***[1, 2] ✉, ***Vadim V. Krivorotov***[1]

[1] *Ural Federal University named after the First President of Russia B. N. Yeltsin, Yekaterinburg, Russia*

[2] *Institute of Economics, The Ural Branch of Russian Academy of Sciences, Yekaterinburg, Russia*

✉ *alexkalina74@mail.ru*

Abstract. Clustering of research objects and combining them into similar groups based on a set of characteristics is an important stage in solving many tasks of socio-economic development, especially tasks related to assessing the state of the socio-economic system, as well as modeling and forecasting indicators of its future development. The purpose of this study is a relative assessment of the financial condition of large Russian manufacturing companies based on data from accounting reporting forms using clusterization methods classified as machine learning without a teacher. The results of such an assessment are subsequently supposed to be used to build a model for assessing the financial condition of companies based on one of the machine learning algorithms with a teacher. The paper offers key indicators of the financial condition of companies, on the basis of which it is proposed to perform their clusterization. They were identified as a result of the analysis of modern methods and approaches to research and assessment of competitiveness and competitive position of companies. When conducting clusterization based on the proposed set of indicators, financial reporting data from 2,249 Russian manufacturing companies based on the results of 2023 were used. Companies with a turnover of more than 2 billion rubles and a staff of more than 251 people were considered as large companies.. K-Means++, hierarchical clustering, and DBSCAN were used as clustering algorithms. In order to obtain the best result, special data preprocessing and selection of the necessary hyperparameters for clustering algorithms were carried out. The quality of the final clustering was assessed using the Davies-Bouldin (DBI) and the Calinski–Harabasz (CHI) scores. The results showed that the production companies under consideration can be combined into a relatively small number of clusters (usually no more than 3) in terms of financial condition, which opens up wide opportunities for building models of the financial condition of companies. Based on the results of using 3 clustering methods, K-Means++ turned out to be the best algorithm by a small margin, the formed centroids of which can be called the average assessment of companies with poor, normal and good financial condition. The quality of the final clustering can be assessed as good.

Key words: financial analysis; machine learning; financial condition indicators; large company; clusterization of companies; K-Means++; hierarchical clusterization; DBSCAN.

JEL D22, G30, C45

References

1. Kryzanowski, L., Galler, M., Wright, D.W. (1993). Using artificial neural networks to pick stocks. *Financial Analysts Journal,* Vol. 49, Issue 4, 21–27. https://doi.org/10.2469/faj.v49.n4.21

2. Porter, M.E. (1990). *The Competitive Advantage of Nations.* New York, Free Press, 855 p. Available at: https://archive.org/details/competitiveadvan0000port

3. Porter, M.E. (2008). The five competitive forces that shape strategy. *Harvard Business Review,* Vol. 86, No. 1, 78–93. Available at: https://sistemasgerenciales.wordpress.com/wp-content/uploads/2016/04/the-five-competitive-forces-that-shape-strategy.pdf

4. Cao, X., Shen, X., Liu, Q. (2024). Mechanism of aquaculture competitiveness in China. *Aquaculture Reports,* Vol. 37, 102195. https://doi.org/10.1016/j.aqrep.2024.102195

5. Tao, X., Cai, W. (2025). The impact of digital finance on export competitiveness: Evidence from Chinese manufacturing enterprises. *Finance Research Letters,* Vol. 73, 106629. https://doi.org/10.1016/j.frl.2024.106629

6. Liu, Y.-L., Tian, L., Li, C., Wu, Ya. (2024). Analyzing the competitiveness and strategies of Chinese mobile network operators in the 5G era. *Telecommunications Policy,* Vol. 48, Issue 2, 102652. https://doi.org/10.1016/j.telpol.2023.102652

7. Fang, K., Zhou, Y., Wang, S., Ye, R., Guo, S. (2018). Assessing national renewable energy competitiveness of the G20: A revised Porter's Diamond Model. *Renewable and Sustainable Energy Reviews,* Vol. 93, 719–731. https://doi.org/10.1016/j.rser.2018.05.011

8. Cibinskiene, A., Dumciuviene, D., Bobinaite, V., Dragašius, E. (2021). Competitiveness of industrial companies forming the value chain of wind energy components: The case of Lithuania. *Sustainability*, Vol. 13, Issue 16, 9255. https://doi.org/10.3390/su13169255

9. Liu, J., Wei, Q., Dai, Q., Liang, C. (2018). Overview of wind power industry value chain using diamond model: A case study from China. *Applied Sciences,* Vol. 8, Issue 10, 1900. https://doi.org/10.3390/app8101900

10. Lau, A.K.W., Baark, E., Lo, W.L.W., Sharif, N. (2013). The effects of innovation sources and capabilities on product competitiveness in Hong Kong and the Pearl River Delta. *Asian Journal of Technology Innovation,* Vol. 21, Issue 2, 220–236. https://doi.org/10.1504/IJTM.2012.047244

11. Lotfi, B., Karim, M. (2016). Competitiveness determinants of Moroccan exports: quantity-based analysis. *International Journal of Economics and Finance,* Vol. 8, Issue 7, 140–148. https://doi.org/10.5539/ijef.v8n7p140

12. Li, Ya., Yu, H., Shen, Z. (2025). Dynamic prediction of product competitive position: A multisource data-driven competitive analysis framework from a multi-competitor perspective. *Journal of Retailing and Consumer Services,* Vol. 25, 104289. https://doi.org/10.1016/j.jretconser.2025.104289

13. Kim, S.-A., Park, S., Kwak, M., Kang, C. (2025). Examining product quality and competitiveness via online reviews: An integrated approach of importance performance competitor analysis and Kano model. *Journal of Retailing and Consumer Services,* Vol. 82, 104135. https://doi.org/10.1016/j.jretconser.2024.104135

14. Buckley, P.J., Pass, C.L., Prescott, K. (1988). Measures of international competitiveness: A critical survey. *Journal of Marketing Management,* Vol. 4, Issue 2, 175–200. https://doi.org/10.1080/0267257X.1988.9964068

15. Schefczyk, M. (1993). Operational performance of airlines: an extension of traditional measurement paradigms. *Strategic Management Journal,* Vol. 14, No. 4, 301–317. https://doi.org/10.1002/smj.4250140406

16. Good, D.H., Nadiri, M.I., Roller, L.H., Sickles, R.C. (1993). Efficiency and productivity growth comparisons of European and U.S. airlines: A first look at the data. *The Journal of Productivity Analysis,* Vol. 4, 115–125. https://doi.org/10.1007/BF01073469

17. Parkan, C., Wu, M.-L. (1999). Measurement of the performance of an investment bank using the operational competitiveness rating procedure. *Omega,* Vol. 27, Issue 2, 201–217. https://doi.org/10.1016/S0305-0483(98)00041-3

18. Fleisher, C.S., Bensoussan, B.E. (2015). *Business and Competitive Analysis: Effective Application of New and Classic Methods.* Second Edition. New Jersey, Pearson, 448 p. https://doi.org/10.24883/IberoamericanIC.v2i2.37

19. Prescott, J.E., Grant, J.H. (1988). A manager's guide for evaluating competitive analysis techniques. *Interfaces,* Vol. 18, Issue 3, 10–22. https://doi.org/10.1287/inte.18.3.10

20. Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bulletin L'Académie Polonaise des Science,* Vol. 4, No. 12, 801–804. (In French). Available at: http://laurent-duval.eu/Documents/Steinhaus_H_1956_j-bull-acad-polon-sci_division_cmp-k-means.pdf

21. Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory,* Vol. 28, Issue 2, 129–137. https://doi.org/10.1109/TIT.1982.1056489

22. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1. University of California Press, 281–297. Available at: https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1967-MacQueen-MSP.pdf

23. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 24, Issue 7, 881–892. https://doi.org/10.1109/TPAMI.2002.1017616

24. Ostrovsky, R., Rabani, Yu., Schulman, L.J., Swamy, C. (2006). The effectiveness of Lloyd-type methods for the k-means problem. *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06).* IEEE, 165–176. https://doi.org/10.1109/FOCS.2006.75

25. Ahmed, M., Seraj, R., Islam, S.M.S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics,* Vol. 9, Issue 8, 1295. https://doi.org/10.3390/electronics9081295

26. Arthur, D., Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007).* Society for Industrial and Applied Mathematics3600 University City Science Center Philadelphia, 1027–1035. https://doi.org/10.1145/1283383.1283494

27. Abdulnassar, A.A., Nair, L.R. (2023). Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model. *Measurement: Sensors,* Vol. 25, 100666. https://doi.org/10.1016/j.measen.2023.100666

28. Ay, M., Özbakır, L., Kulluk, S., Gülmez, B., Öztürk, G., Özer, S. (2023). FC-Kmeans: Fixed-centered K-means algorithm. *Expert Systems with Applications,* Vol. 211, 118656. https://doi.org/10.1016/j.eswa.2022.118656

29. Li, J., Li, J., Wang, D., et al. (2023). Hierarchical and partitioned planning strategy for closed-loop devices in low-voltage distribution network based on improved KMeans partition method. *Energy Reports*, Vol. 9, 477–485. https://doi.org/10.1016/j.egyr.2023.05.161

30. He, J., Jiang, D., Zhang, D., Li, J., Fei, Q (2022). Interval model validation for rotor support system using Kmeans Bayesian method. *Probabilistic Engineering Mechanics*, Vol. 70, 103364. https://doi.org/10.1016/j.probengmech.2022.103364

31. Nielsen, F. (2016). *Introduction to HPC with MPI for Data Science.* Springer, 282 p. https://doi.org/10.1007/978-3-319-21903-5

32. Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal,* Vol. 16, Issue 1, 30–34. https://doi.org/10.1093/comjnl/16.1.30

33. Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal,* Vol. 20, Issue 4, 364–366. https://doi.org/10.1093/comjnl/20.4.364

34. Eppstein, D. (2000). Fast hierarchical clustering and other applications of dynamic closest pairs. *Journal of Experimental Algorithmics (JEA),* Vol. 5, 1-es. https://doi.org/10.1145/351827.351829

35. Riyahi, M., Martín, A.G. (2025). Optimizing Capacity Expansion Modeling with a Novel Hierarchical Clustering and Systematic Elbow Method: A Case study on Power and Storage Units in Spain. *Energy,* Vol. 323, 135788. https://doi.org/10.1016/j.energy.2025.135788

36. Tang, Z., Wang, L., Guo, S., Liang, G., Zhang, W., Zhang, L., Rui, M., Guan, G., Wang, Yu. (2025). Study on modular design methodology of marine SMR system based on fuzzy hierarchical clustering and improved genetic algorithm. *Progress in Nuclear Energy,* Vol. 185, 105739. https://doi.org/10.1016/j.pnucene.2025.105739

37. Zhang, X., Wang, Y.-L., Byun, H. (2025). Enhancing energy saving and reducing latency by divisive hierarchical clustering with Multi-UAVs in WSANs. *Applied Soft Computing*, 112861. https://doi.org/10.1016/j.asoc.2025.112861

38. Ester, M., Kriegel, H.P., Sander, J., Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 1996 Knowl-edge Discovery and Data Mining (KDD'96).* AAAI Press, 226–231. Available at: https://cdn.aaai.org/KDD/1996/KDD96-037.pdf

39. Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems,* Vol. 42, No. 3, 1–21. https://doi.org/10.1145/3068335

40. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD '99).* ACM New York, 49–60. https://doi.org/10.1145/304182.304187

41. Campello, R.J., Moulavi, D., Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *In: Advances in Knowledge Discovery and Data Mining.* Pacific-Asia Conference on Knowledge Discovery and Data Mining. Edited by J. Pei, V. S. Tseng, L. Cao, H. Motoda, G. Xu. Springer, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

42. Campello, R.J., Moulavi, D., Zimek, A., Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD),* Vol. 10, No. 1, 1–51. https://doi.org/10.1145/2733381

43. Bíró, P. Kovács, B.B.H., Novák, T., Erdélyi, M. (2025). Cluster parameter-based DBSCAN maps for image characterization. *Computational and Structural Biotechnology Journal,* Vol. 27, 920–927. https://doi.org/10.1016/j.csbj.2025.02.037

44. Ozer, F.C., Tuydes-Yaman, H., Dalkic-Melek, G. (2024). Increasing the precision of public transit user activity location detection from smart card data analysis via spatial-temporal DBSCAN. *Data & Knowledge Engineering,* Vol. 153, 102343. https://doi.org/10.1016/j.datak.2024.102343

45. Mardani, K., Maghooli, K., Farokhi, F. (2025). Segmentation of coronary arteries from X-ray angiographic images using density based spatial clustering of applications with noise (DBSCAN). *Biomedical Signal Processing and Control,* Vol. 101, 107175. https://doi.org/10.1016/j.bspc.2024.107175

46. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* Vol. 43, 59–69. https://doi.org/10.1007/BF00337288

47. Urme, O., Reza, S., Adham, Md I., Sattar, G.S. (2025). Arsenic, manganese, and iron concentration in groundwater of northwestern part of Bangladesh using self-organizing maps: Implication for health risk assessment. *Heliyon*, Vol. 11, Issue 2, e41805. https://doi.org/10.1016/j.heliyon.2025.e41805

48. Boubekki, A., Kampffmeyer, M., Brefeld, U., Jenssen, R. (2021). Joint optimization of an autoencoder for clustering and embedding. *Machine Learning,* Vol. 110, No. 7, 1901–1937. https://doi.org/10.1007/s10994-021-06015-5

49. Pulgar, F.J., Charte, F., Rivera, A.J., del Jesus, M.J. (2018). AEkNN: An AutoEncoder kNN-based classifier with built-in dimensionality reduction. *ArXiv Preprint arXiv:1802.08465,* 35 p. https://doi.org/10.48550/arXiv.1802.08465

50. John Wiley & Sons (1983). *Understanding Robust and Exploratory Data Analysis.* Edited by D. C. Hoaglin, F. Mosteller, J. W. Tukey. John Wiley & Sons, 447 p. Available at: https://archive.org/details/understandingrob0000unse/page/n7/mode/1up

51. Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys,* Vol. 41, No. 3, 1–58. https://doi.org/10.1145/1541880.1541882

52. Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* Vol. 58, Issue 301, 236–244. https://doi.org/10.1080/01621459.1963.10500845

53. Davies, D.L., Bouldin, D.W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-1, Issue 2, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

54. Caliński, T., Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics,* Vol. 3, Issue 1, 1–27. https://doi.org/10.1080/03610927408827101

55. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* Vol. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

INFORMATION ABOUT AUTHORS

Lev Alexeevich Bulanov

Post-Graduate Student, Department of Economic Safety of Industrial Complexes, Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia (620002, Ekaterinburg, Mira street, 19); ORCID https://orcid.org/0009-0001-0242-0127 e-mail: levbulanov2013@yandex.ru

Alexei Vladimirovich Kalina

Candidate of Technical Sciences, Associate Professor, Department of Economic Safety of Industrial Complexes, Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia (620002, Ekaterinburg, Mira street, 19), Senior Researcher, The Center of Economic Security, Institute of Economics, The Ural Branch of Russian Academy of Sciences, Yekaterinburg, Russia (620014, Yekaterinburg, Moskovskaya street, 29); ORCID https://orcid.org/0000-0003-0376-2505 e-mail: alexkalina74@mail.ru

Vadim Vasilyevich Krivorotov

Doctor of Economics, Professor, Head of Department of Economic Safety of Industrial Complexes, Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Russia (620002, Ekaterinburg, Mira street, 19); ORCID https://orcid.org/0000-0002-7066-0325 e-mail: v_krivorotov@mail.ru

FOR CITATION

Bulanov, L.A., Kalina, A.V., Krivorotov, V.V. (2025). Clustering of Russian Manufacturing Companies by Indicators of Their Financial Condition Using Machine Learning Technologies. *Journal of Applied Economic Research,* Vol. 24, No. 2, 584–621. https://doi.org/10.15826/vestnik.2025.24.2.020

ARTICLE INFO

Received March 21, 2025; Revised April 7, 2025; Accepted April 12, 2025.