

doi: 10.17586/2226-1494-2025-25-2-261-272

УДК 004.932

Кластеризация пространственных данных с неявно выраженной полигональной структурой на основе топологических подходов

Сергей Владимирович Еремеев✉

Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, 602264, Российская Федерация
sv-eremeev@yandex.ru✉, <https://orcid.org/0000-0001-8482-1479>

Аннотация

Введение. Кластеризация является одним из фундаментальных подходов для интеллектуального анализа данных, который в сфере геоинформатики и обработки изображений используется для поиска знаний и скрытых закономерностей пространственной информации. При автоматической векторизации объектов на спутниковых снимках из-за несовершенства этих технологий на линейных и полигональных объектах появляются пропущенные элементы, что препятствует полноценному анализу и визуализации данных. Представлен новый метод кластеризации геометрических примитивов с неявно выраженной полигональной структурой с возможностью устранения неполных данных в векторных моделях. **Метод.** Предлагаемый метод основан на итерационном формировании пространственных структур при растяжении исходных объектов линейного типа. В отличие от многих подходов кластеризации элементы группируются в кластеры не по принципу ближайшего евклидова расстояния, а за счет определения ближайшего пересечения между отрезками. Такой подход позволяет корректно разделить рядом стоящие линейные объекты в разные кластеры. Для формирующихся пространственных структур на каждой итерации в зависимости от коэффициента растяжения вычисляются их топологические особенности, что обеспечивает возможность обнаруживать и проводить фильтрацию неявно выраженных полигональных структур. **Основные результаты.** Разработанный метод апробирован для кластеризации линейных геометрических примитивов на векторных моделях городской инфраструктуры. Проведено сравнение работы метода с аналогами: k -средних, Density-Based Spatial Clustering of Application with Noise (DBSCAN) и агломеративной кластеризацией. Исследования показали, что с использованием метрики оценки качества кластеризации в виде индексов инерции и Жаккара предложенный метод имеет преимущество в случаях корректного разделения близко расположенных кластеров. **Обсуждение.** Полученные результаты могут быть использованы в геоинформационных системах для актуализации векторных данных после применения технологий автоматической векторизации объектов на спутниковых снимках.

Ключевые слова

кластеризация пространственной информации, топологический анализ данных, гомотетия, геоинформационные системы, оценка качества кластеризации

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда в рамках научного проекта № 23-21-10064.

Автор выражает благодарность магистру в области информационных систем Егай Марии Владимировне за помощь в подготовке исходных данных и проведении экспериментов.

Ссылка для цитирования: Еремеев С.В. Кластеризация пространственных данных с неявно выраженной полигональной структурой на основе топологических подходов // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 2. С. 261–272. doi: 10.17586/2226-1494-2025-25-2-261-272

Clustering of spatial data with implicit polygonal structure based on topological approaches

Sergey V. Ereemeev✉

Murom Institute (Branch) of Vladimir State University, Murom, 602264, Russian Federation
sv-eremeev@yandex.ru✉, <https://orcid.org/0000-0001-8482-1479>

Abstract

Clustering is one of the fundamental approaches for data mining, which in the field of geoinformatics and image processing is used to search for knowledge and hidden patterns of spatial information. During automatic vectorization of objects on satellite images due to imperfections of these technologies, missing elements appear on linear and polygonal objects, which prevent full-fledged data analysis and visualization. The paper considers the problem of clustering geometric primitives with implicit polygonal structure with the possibility of eliminating incomplete data in vector models. The proposed method is based on the iterative formation of spatial structures by stretching the original linear objects. Unlike many clustering approaches, elements are grouped into clusters not by the principle of nearest Euclidean distance, but by determining the nearest intersection between segments. This approach allows us correctly dividing adjacent objects into different clusters. For the spatial structures formed at each iteration, their topological features are calculated depending on the stretching coefficient, which makes it possible to detect and filter implicit polygonal structures. The developed method is tested for clustering of linear geometric primitives on vector models of urban infrastructure. The performance of the method is compared with its competitors: *k*-means, DBSCAN, and agglomerative clustering. The research has shown that using the clustering quality assessment metric in the form of inertia and Jaccard indices, the proposed method has an advantage due to the correct separation of closely located clusters.

Keywords

spatial information clustering, topological data analysis, homothety, geographic information systems, clustering quality assessment

Acknowledgements

This study was supported by the Russian Science Foundation, project no. 23-21-10064.

The author expresses gratitude to the Master of Information Systems, Maria Vladimirovna Egay, for her assistance in preparing the initial data and conducting the experiments.

For citation: Ereemeev S.V. Clustering of spatial data with implicit polygonal structure based on topological approaches. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2025, vol. 25, no. 2, pp. 261–272 (in Russian). doi: 10.17586/2226-1494-2025-25-2-261-272

Введение

Кластеризация является одним из фундаментальных подходов для интеллектуального анализа данных, под которым понимается группировка объектов по схожим признакам. В сфере геоинформатики и обработки изображений кластеризация используется для поиска знаний и скрытых закономерностей пространственной информации.

Сложность обрабатываемых данных с каждым годом возрастает [1, 2]. Это обуславливает потребность в разработке новых идей и компьютерных методов для извлечения ценной информации о пространственных объектах из быстрорастущего объема цифровых данных [3].

В настоящее время наблюдается активный рост научных работ, в которых предлагаются решения для автоматической векторизации объектов на спутниковых снимках с последующим их распределением по разным слоям. Однако существующие технологии векторизации несовершенны. Это подтверждает выполненный систематический обзор в работе [4] для выделения на изображении линейных отрезков, состоящих из двух соединенных между собой точек. В работе [4] обозначены следующие основные проблемы, которые препятствуют обнаружению линейных и полигональных объектов из растрового изображения: нестабильность конечных точек отрезка из-за шума, теней и освещенности изображения; нехватка наборов данных для машинного обучения; перекрытие объектов другими объектами.

В результате векторизации на линейных и полигональных объектах образуются пропущенные элементы, что затрудняет применение полноценного анализа и визуализации пространственных данных. Таким образом, на некоторых участках получается набор векторных объектов с неполными данными. Актуальной является задача разработки метода кластеризации линейных отрезков, которые представляют неявно выраженную пространственную структуру, и дальнейшего устранения пропущенных данных в векторных моделях.

Обзор существующих подходов

В качестве исходных данных рассмотрим наборы из линейных отрезков, которые потенциально образуют полигональные объекты. Пример таких данных показан на рис. 1. Заметим, что визуально отрезки образуют контуры трех зданий и могут быть представлены как полигоны.

Сформированные наборы данных получаются в результате работы различных технологий автоматического преобразования информации в векторный формат из спутниковых снимков или облака точек после лидарной съемки. В [5] для выделения отрезков на крышах зданий из облака точек применена спектральная кластеризация прямолинейных сегментов. В работе [6] из облака точек построены отрезки, формирующие полигональный объект, на основе семантической маркировки и кластеризации с использованием плотности

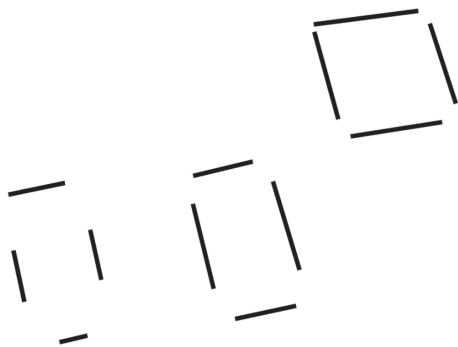


Рис. 1. Векторный набор данных из линейных отрезков с неявно выраженной полигональной структурой

Fig. 1. Vector dataset of linear segments with implicit polygonal structure

точек. Из точечных данных в [7] предложено выделять ключевые точки и относительно них формировать линейные отрезки. Также по опорным точкам и исходным спутниковым снимкам в работе [8] построены прямоугольные контуры зданий, но отметим, в данной работе опорные точки ставит пользователь вручную, что является дорогостоящим процессом при обработке больших объемов пространственной информации.

Результаты с достаточно высокой степенью точности по обнаружению отрезков на изображении представлены в работе [9], в которой предлагается включение сверточной нейронной сети в классический детектор Line Segment Detector (LSD).

Существует множество методов для кластеризации. Современный обзор методов кластеризации больших данных представлен в [10]. Наиболее широко используются следующие из них: k -средних, Density-Based Spatial Clustering of Application with Noise (DBSCAN), агломеративная кластеризация. Рассмотрим особенности этих методов и применим их для кластеризации векторных данных.

Метод k -средних разбивает исходное множество на k кластеров так, что внутри группы суммарное квадратичное отклонение объектов от центра кластера минимизируется. Полноценный обзор существующих разновидностей метода k -средних и существующие проблемы его применения показаны в [11]. Из-за использования евклидова расстояния в качестве метрики сходства одной из основных проблем является его ограничение при обнаружении других форм кластеров, что может повлиять на анализ пространственных объектов. Кроме того, для больших объемов пространственных данных определить оптимальное количество кластеров является достаточно сложной задачей.

Метод DBSCAN относится к подходам кластеризации на основе плотности, где объекты, расположенные в области с высокой плотностью, формируют отдельные кластеры, отделенные между собой областями с низкой плотностью [12].

Агломеративная иерархическая кластеризация начинает формировать отдельные кластеры для каждой точки исходных данных и объединяет кластеры с наименьшим расстоянием между ними, чтобы создать более крупные кластеры. Этот процесс повторяется до тех

пор, пока не будет сформирован один большой кластер, содержащий все данные.

Для группировки векторных объектов в потенциальные полигоны рассмотрим следующие подходы. При кластеризации линейных отрезков ведутся разработки для более выгодного их объединения [13]. Кроме того, в [14] рассматривается кластеризация отрезков в многомерном пространстве, что позволяет использовать результаты в статистическом анализе данных.

В работе [15] применено понятие связности для группировки объектов, что позволяет сделать разбиение на кластеры более оптимально. Похожая тенденция для кластеризации с использованием пространственных отношений наблюдается в [16], где для группировки объектов дополнительно применяется их атрибутивная информация. Кроме того, связи между объектами в виде топологических отношений в сочетании с агломеративной кластеризацией рассмотрены в работе [17].

Топологические отношения и топология в целом играют ключевую роль в анализе пространственных данных. Предлагаемый в настоящей работе метод также будет опираться на топологические подходы. Пространственные объекты, не связанные между собой, образуют сложную структуру, которая формируется по определенным и не известным заранее правилам. Это выражается в объединении пространственных данных в специальные формы кластеров, не зависящие от евклидова расстояния между исходными объектами. Существует множество ситуаций, когда рядом стоящие объекты относятся к разным кластерам. Большинство методов кластеризации не позволяют корректно разделить объекты на кластеры для таких исходных данных, потому что в одну группу объединяются ближайшие по расстоянию объекты. Это ограничивает обнаружение скрытых связей в пространственных структурах данных. Для исследования из множества пространственных данных выделим отдельный класс объектов линейного типа, которые неявным образом представляют собой полигональные объекты.

Целью настоящей работы является улучшение обнаружения неявно выраженных полигональных структур. Метод обеспечивает повышение эффективности кластеризации объектов линейного типа и выявление скрытых полигональных структур на основе доступной пространственной информации при соответствующей оценке качества полученных решений. В отличие от известных подходов, объекты объединяются в кластеры без вычисления ближайшего евклидова расстояния между ними. Вместо этого предлагается использовать принцип растяжения исходных линейных отрезков до их ближайшего пересечения между собой с последующим формированием структур полигонального типа.

Изложен принцип объединения линейных отрезков в кластеры, который основывается не на ближайшем расстоянии между отрезками, а на ближайшем пересечении между ними при их растяжении. Показан способ формирования пространственных структур, каждая из которых представляет собой кластер, образованный в результате пересечения отрезков при их растяжении.

Определяются топологические свойства образованных структур, которые дают возможность обнаруживать скрытые объекты полигонального типа.

Методы и материалы

Основные принципы кластеризации на основе топологических подходов. Топологические подходы при анализе данных используют инструменты топологии для изучения формы и связей между данными. При кластеризации данные абстрагируются в виде графа. Это позволяет находить шаблоны и структуры, такие как компоненты связности, циклы и другие, которые не могут быть выявлены с помощью статистических методов.

Графы могут быть использованы для представления пространственных объектов и их отношений на географических картах. Топология в этом контексте изучает свойства и структуру геометрических объектов в виде точек, линий и полигонов, а также определяет их связность и отношения.

Среди современных топологических подходов выделим топологический анализ данных. В последнее время он получил широкое распространение в разных областях, включая анализ пространственных данных. Этот подход рассматривает данные как множество точек в метрическом пространстве. Точки могут быть связаны между собой. Для этого вокруг каждой точки строится шар с постепенным увеличением радиуса. При пересечении шаров точки соединяются между собой отрезками, образуя топологические структуры. После этого анализируются диаграммы персистентности, которые

показывают, какая структура сохраняется в течение определенного промежутка пространства [18, 19].

Описание предлагаемого метода. Рассмотрим общую схему разрабатываемого метода кластеризации в сочетании с предварительной обработкой пространственной информации (рис. 2).

Исходными данными для предварительного этапа (этап 1) являются спутниковые снимки в видимом диапазоне в формате RGB (Red, Green, Blue). С использованием существующих подходов, наиболее популярные из которых преобразование Хафа и метод LSD, на снимках происходит выделение линейных отрезков. В силу различных причин, включая природные особенности и несовершенство подходов, в процессе формирования отрезков появляются ошибки, т. е. некоторые отрезки имеют пропуски, несмотря на то, что визуально их объединение представляет полигональную структуру. Заключительным этапом предварительной обработки является векторизация данных в отдельный слой геоинформационной системы. При этом отрезки хранятся в векторном виде в шейп-файлах (.shp).

Векторные данные поступают на вход для обработки (этап 2), который включает три основных шага: формирование пространственных структур (кластеров), вычисление признаков структур и их анализ для обнаружения полигональных объектов. Рассмотрим каждый из шагов более подробно.

На первом шаге (формирование пространственных структур) исходные данные представляют множество линейных отрезков $L = \{l_1, l_2, \dots, l_{|L|}\}$ на плоскости, каждый элемент которого $l \in L$ состоит из двух концевых точек (x_1, y_1) и (x_2, y_2) .

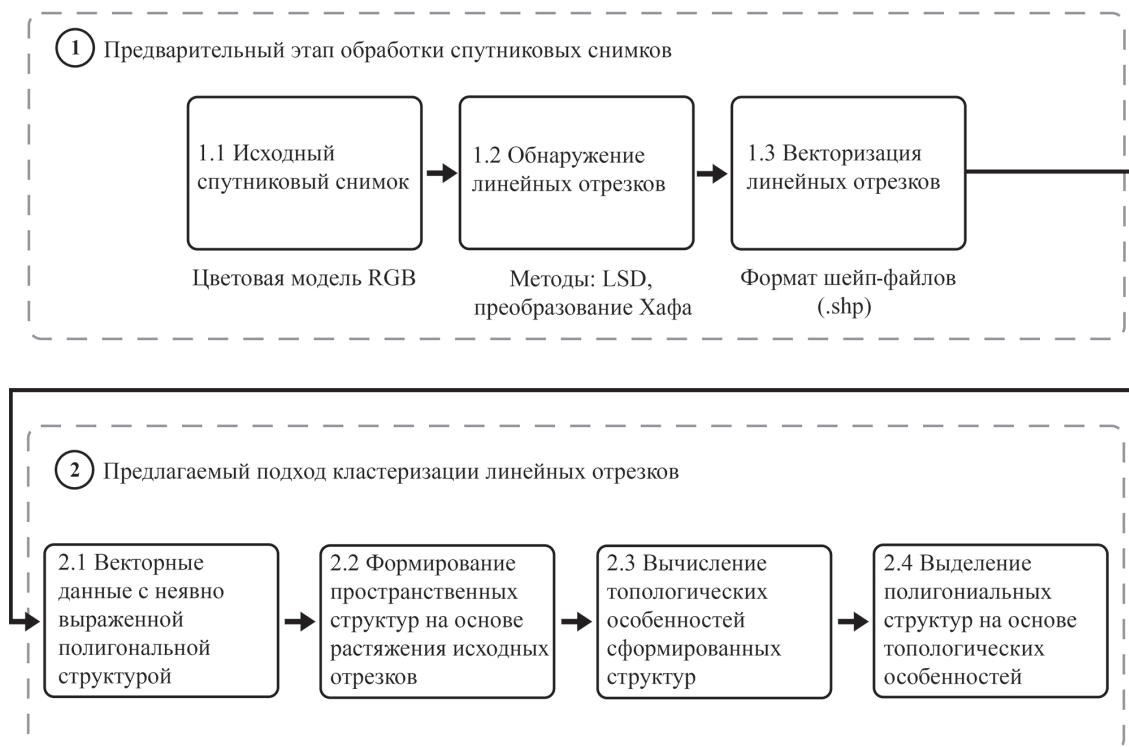


Рис. 2. Общая схема предлагаемого метода кластеризации объектов линейного типа с учетом предварительной обработки спутниковых снимков

Fig. 2. General scheme of the proposed linear object clustering approach taking into account pre-processing of satellite images

Для всех отрезков $l \in L$ введем растяжение в k раз. Для этого зададим гомотегию $H_{O,k}$ с центром в точке $O(x_0, y_0)$ и коэффициентом растяжения k . Координаты точки O для отрезка $l \in L$ вычисляются по формуле: $x_0 = \frac{x_1 + x_2}{2}$, $y_0 = \frac{y_1 + y_2}{2}$. Определим гомотегию в виде:

$$H_{O,k}: \begin{cases} x_i' = k(x_i - x_0) + x_0 \\ y_i' = k(y_i - y_0) + y_0 \end{cases}$$

где (x_i, y_i) и (x_i', y_i') — координаты концевых точек отрезка до и после растяжения ($i = 1, 2$); (x_0, y_0) — точка середины исходного отрезка; $k \in [1, +\infty]$ — коэффициент растяжения.

Применяя гомотегию $H_{O,k}$ к каждому отрезку из L с постепенным возрастанием коэффициента растяжения k , получим обновленные отрезки с увеличенной длиной. При этом отрезки начнут пересекаться, образовывать более сложные структуры: компоненты связности и циклы. При пересечении пространственных элементов выполним их группировку. При появлении циклов будут формироваться полигональные структуры, представляющие собой кластеры, образованные на основе исходных отрезков из L .

Для анализа данных при создании топологических структур построим соответствующие диаграммы их жизненного цикла, которые позволят узнать на каком шаге итерационного процесса были сформированы эти

структуры. Итерационный процесс применения гомотегии $H_{O,k}$ заканчивается тогда, когда будут построены все потенциальные полигональные объекты.

На рис. 3 показан пример итерационного процесса, где исходное множество отрезков состоит из 12 элементов (рис. 3, а). На начальном шаге при $k = 1$ имеем 12 отдельных компонент связности (рис. 3, б) в виде коротких штрихов. При постепенном растяжении каждого отрезка $l \in L$ в k раз происходит их пересечение, как, например, элементов 1 и 3 (рис. 3, с). В точке пересечения отрезки больше не растягиваются. Элементы 1 и 3 образуют полилинию, состоящую из двух отрезков. Будем при этом говорить, что одна компонента поглощает другую, т. е. в этой ситуации элемент 1 расширяется за счет присоединения элемента 3. У образованной полилинии сохраняется нумерация от поглощаемой компоненты связности. В данном случае приоритет при поглощении имеет элемент с наименьшим номером. В качестве приоритета можно также установить длину отрезка. Таким образом, элемент 3 прекращает свое существование при $k = 1,2$. Это показано на рис. 3, д, где все элементы имеют одинаковую длину штрихов, кроме элемента 3.

При $k = 1,4$ появляется первый замкнутый объект (рис. 3, е) и остается до конца итерационного процесса. Этому замкнутому объекту соответствует компонента связности 1 (полигон 1), так как при объединении с ней элементы 2, 3 и 4 прекратили свое существование, что

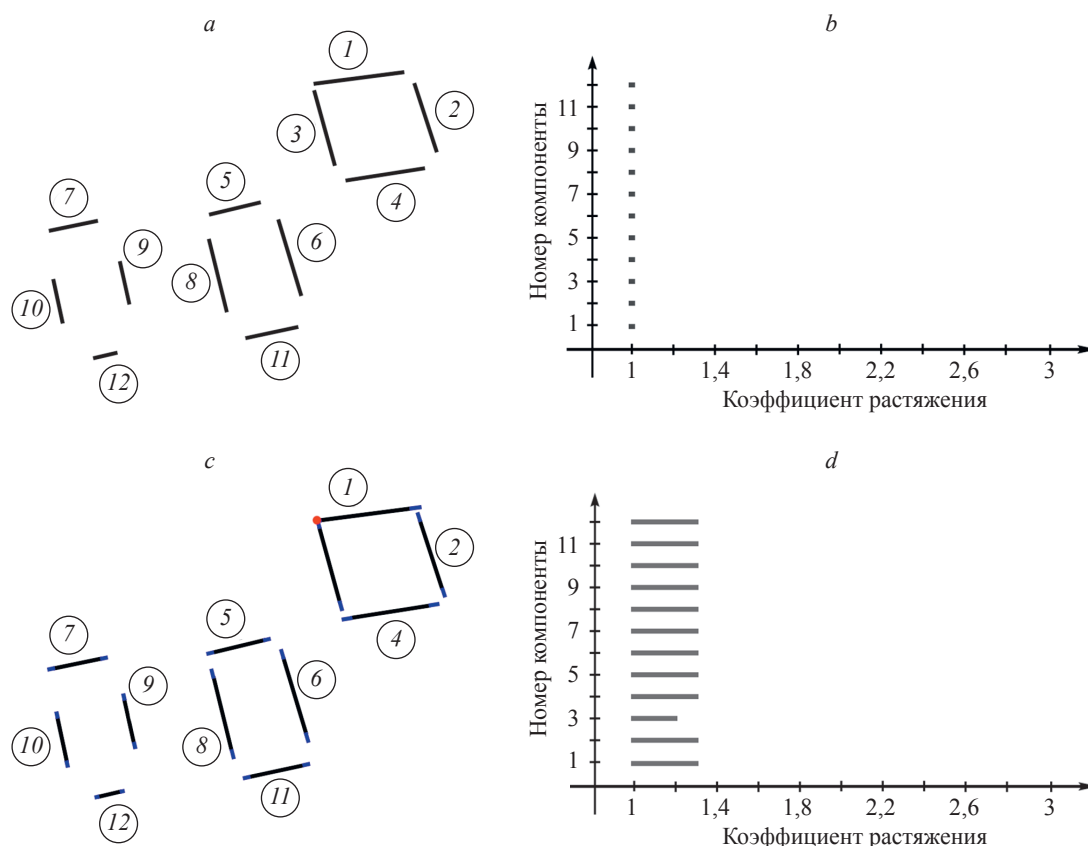


Рис. 3. Итерационный процесс формирования компонент связности и замкнутых объектов с отображением их топологических свойств на основе линейного растяжения отрезков

Fig. 3. Iterative process of formation of connectivity components and closed objects with representation of their topological properties on the basis of linear stretching of segments

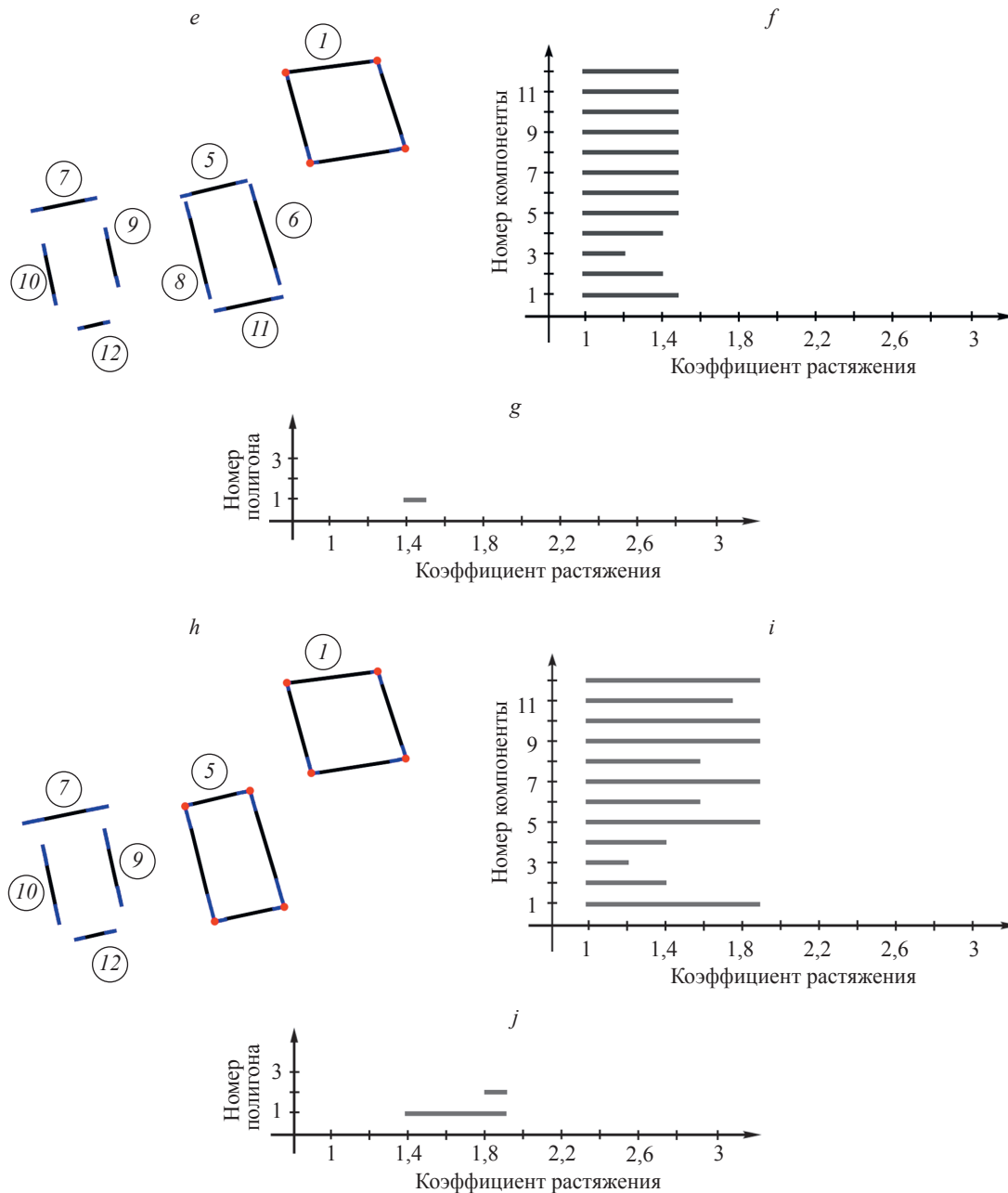


Рис. 3. Продолжение

Fig. 3. Continuation

наглядно показано на рис. 3, *f* в виде коротких штрихов. Эти штрихи у элементов 2, 3 и 4 увеличиваться дальше не будут. Информация о том, что появился замкнутый объект (рис. 3, *g*), показана на диаграмме, на которой продемонстрирован период существования созданных полигонов в зависимости от коэффициента растяжения.

Работа следующих шагов итерационного процесса, включая итоговый результат, показана на рис. 3, *h–m* с отображением текущих топологических характеристик элементов на диаграммах. Особое внимание обратим на рис. 3, *m*. Он показывает, при каком коэффициенте растяжения появились полигоны 1–3 при пересечении соответствующих отрезков. Самым последним появился полигон 3, ему соответствует компонента связности с номером 7. Заметим, что на рис. 3, *l* компоненты

связности с номерами 1, 5 и 7 имеют самые длинные штрихи. Это означает, что их не поглощали другие компоненты связности.

Диаграммы имеют практический потенциал. Они способствуют более детальному изучению эволюции пространственных объектов, а также помогают выявить паттерны и тенденции в их развитии, что облегчает интерпретацию данных.

На втором шаге (вычисление топологических особенностей) происходит расчет топологических свойств сформированных структур. Данный процесс является итерационным и при каждом увеличении коэффициента растяжения происходит изменение топологических особенностей структур. Объединение двух компонент связности осуществляется по принципу ближайшего

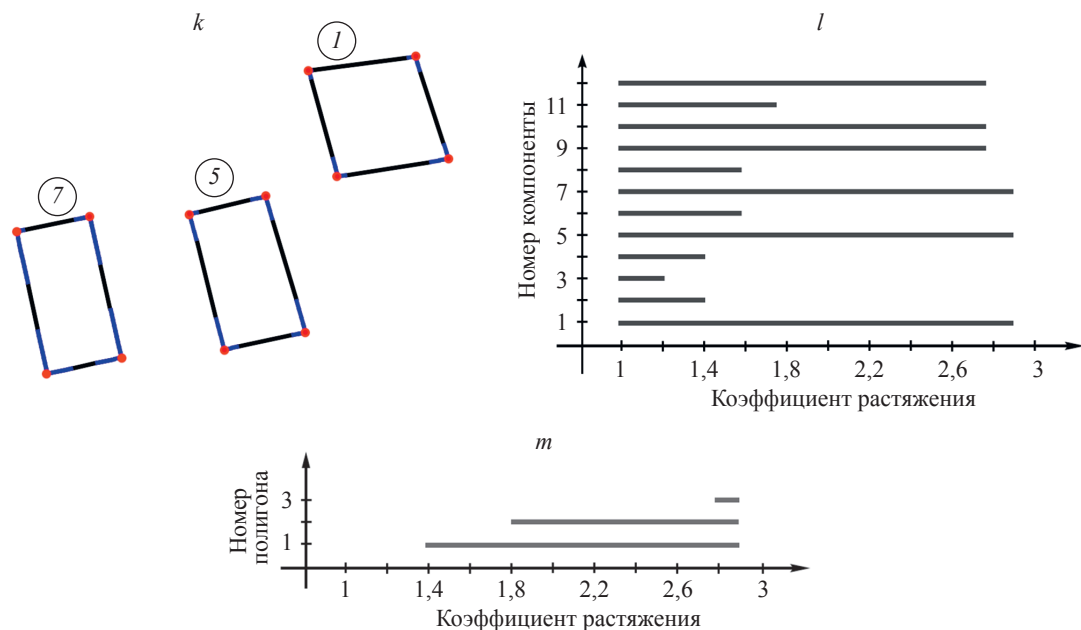


Рис. 3. Окончание

Fig. 3. End

пересечения их концевых отрезков. При таком подходе в кластер группируются не сами исходные отрезки, а их измененное представление после растяжения. Кластером в этом случае является полилиния, включающая несколько отрезков, в том числе и концевые. Именно концевые отрезки продолжают растяжение до тех пор, пока не пересекутся с другими полилиниями. Процесс останавливается, если концевые отрезки не могут достичь пересечения с другими отрезками, когда они параллельны или требуется слишком большое значение коэффициента растяжения. Всю топологическую информацию о пространственных структурах хранят диаграммы, отражающие период существования компонент связности и полигонов относительно коэффициента растяжения.

На третьем шаге (выделение полигональных структур) происходит анализ топологических свойств. Например, можно по этим свойствам отфильтровать только те полигоны, которые появились на начальных значениях коэффициента растяжения. Это говорит о том, что неявно выраженные полигональные структуры имеют небольшие пропуски и с помощью растяжения исходные отрезки объединяются в кластеры, дополняя геометрическую форму полигона недостающими данными.

Предложенный метод имеет ограничения в случае недостижимости замкнутых объектов. Потому его использование целесообразно для таких исходных векторных данных, которые имеют неявно выраженную полигональную структуру.

Результаты экспериментальных исследований

Критерии для оценки качества кластеризации. Для оценки качества кластеризации множества линейных отрезков с неявно выраженной структурой были

использованы следующие метрики: индексы Жаккара и инерции.

Индекс Жаккара — классическая мера сходства между двумя наборами данных, которая применяется в различных областях, включая геоинформатику [20]. Для двух сравниваемых полигональных объектов A и B индекс Жаккара вычислим как отношение площади их пересечения к площади объединения:

$$J(A, B) = \frac{S(A) \cap S(B)}{S(A) \cup S(B)}, \quad (1)$$

где $S(A)$ и $S(B)$ — площади объектов A и B .

Индекс инерции является одним из наиболее распространенных подходов для оценки качества кластеризации и используется во многих практических задачах, в том числе, и для оптимизации [21]. Он предназначен для измерения суммарного внутрикластерного разброса данных в кластерах. Индекс вычисляет сумму квадратов расстояний между каждым объектом и центроидом своего кластера:

$$InertiaIndex = \sum_{k=1}^K \sum_{y \in P_k} dist(y, c_k)^2, \quad (2)$$

где K — количество кластеров; $P_1, P_2, \dots, P_K$ — непересекающиеся кластеры, состоящие из объектов $y \in P_k$ ($k = 1, 2, \dots, K$); c_k — центроид кластера P_k ; $dist(y, c_k)^2$ — квадрат евклидова расстояния между объектом $y \in P_k$ и центроидом c_k из его класса.

При минимизации индекса инерции объекты внутри каждого кластера находятся близко друг к другу, а кластеры разделены между собой.

Проведение экспериментов и сравнение с аналогами. Проведение экспериментов базируется на схеме рис. 2. Первый этап является предварительным и предназначен для подготовки векторных данных из спут-

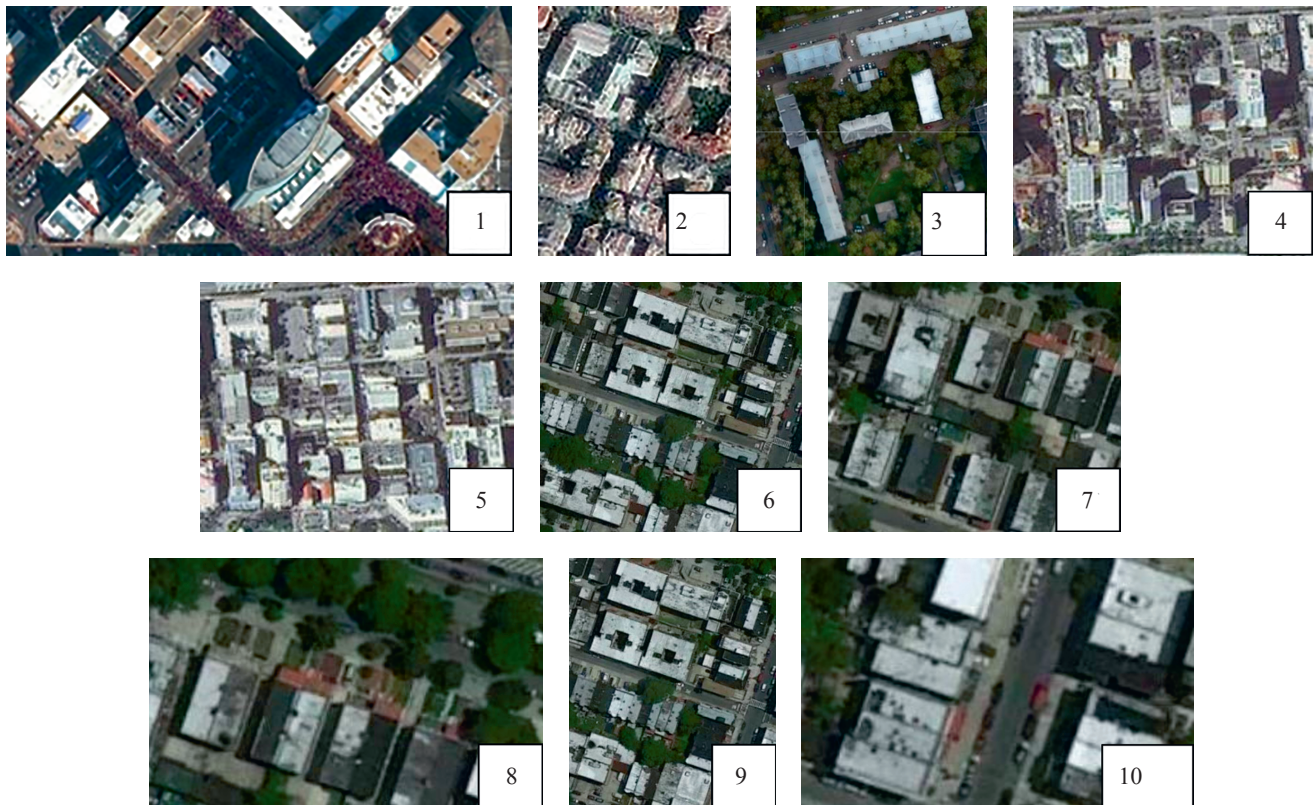


Рис. 4. Исходные изображения с разной плотностью расположения зданий

Fig. 4. Input images with different building densities

никовых снимков. Примеры исходных изображений показаны на рис. 4. Были отобраны снимки, которые содержат здания с разной плотностью застройки, включая очень близко расположенные.

На основе преобразования Хафа были выделены отрезки, которые описывают контур зданий, при этом часть отрезков были отредактированы вручную. Применена геоинформационная система QGIS¹, с помощью которой для каждого исходного снимка был создан векторный слой из отрезков. В результате выбраны для исследования работы метода только те векторные данные, которые представляют собой неявно выраженные контуры зданий. Исходные векторные данные (табл. 1) доступны по ссылке².

Для проведения экспериментов на основе предлагаемого метода были применены основные шаги. Постепенное растяжение исходных отрезков позволяет формировать пространственные структуры различной формы, соответствующие кластерам. Анализ топологических свойств дает возможность обнаружить циклы, чтобы выявить замкнутые контуры.

В качестве аналогов были взяты методы кластеризации: k -средних, DBSCAN и метод агломеративной кластеризации. Все эксперименты проводились на основе разработанного модуля в QGIS. Так как эти методы ра-

ботают с точечными данными, то в качестве исходных данных для них использованы концевые точки отрезков. Для улучшения визуализации для каждого кластера построены объемные контуры зданий (табл. 1). Анализ результатов показал, что методы k -средних, DBSCAN и метод агломеративной кластеризации имеют ограничения на выявление сложных кластеров с нестандартной формой или структурой.

С другой стороны, предлагаемый метод позволяет сгруппировать линейные отрезки в полигональные структуры с достаточно высокой точностью (табл. 1), но при определенных условиях. По сравнению с аналогами предлагаемый метод имеет преимущество в тех случаях, когда потенциальные полигональные объекты с неполными данными расположены рядом друг с другом. Если предполагаемые кластеры имеют значительное расстояние между собой, то, как показывает эксперимент 8, методы k -средних и DBSCAN имеют идентичный результат с предложенным методом.

Выполним сравнение ожидаемых кластеров с их предсказанными положениями, размерами и количеством. Кластер считается правильно установленным, если он покрывает площадь ожидаемой области свыше 75 %, но при этом не заходит на зону других кластеров более чем на 25 %. Для определения процента пересечения полигонов был использован индекс Жаккара (1). В случае DBSCAN кластером считается только тот объект, который содержит минимум три точки. Результаты приведены в табл. 2 в виде двух чисел.

Из табл. 2 видно, что методы k -средних и агломеративная кластеризация часто останавливаются слишком

¹ [Электронный ресурс] Режим доступа: <https://www.qgis.org/> (дата обращения: 05.12.2024).

² [Электронный ресурс] Режим доступа: <https://github.com/creems84/Howole-Geo/blob/main/VectorData.zip> (дата обращения: 05.12.2024).

Таблица 1. Исходные наборы векторных данных и визуализация результатов кластеризации различными методами

Table 1. Initial vector datasets and visualization of clustering results by different methods

Номер эксперимента	Векторные данные	k -means	DBSCAN	Агломеративная кластеризация	Предложенный метод
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

рано, либо неконтролируемо разрастаются, из-за чего в лучшем случае форма объекта становится некорректной, как в эксперименте 1, а в худшем — в один кластер объединяются два или более зданий. Особенно эта проблема проявляется при двух рядом стоящих объектах, как в эксперименте 6.

Метод DBSCAN сталкивается с такой же проблемой, но, из-за отсутствия количества кластеров, она усугубляется тем, что для длинных зданий метод останавливается слишком рано, из-за чего появляется слишком большое число кластеров малого объема.

Отметим, что предлагаемый метод, исправляет указанные недостатки. Для двух рядом стоящих зданий, если между ними есть хоть малейшее расстояние, то результат стремится к ожидаемому. Только в эксперименте 4, где начальные отрезки накладывались друг на друга, кластеры «слиплись». Исключением является эксперимент 1, где из-за слишком большого количества точек появились лишние кластеры, но при этом ожидаемый исход определен с высокой точностью.

Численные результаты экспериментов, полученные с помощью индекса инерции (2), приведены в табл. 3.

Таблица 2. Сравнение геометрических характеристик результатов кластеризации различными методами на основе индекса Жаккара

Table 2. Comparison of geometric characteristics of clustering results by different methods based on the Jaccard index

Номер эксперимента	Ожидаемое количество кластеров	Количество кластеров (правильно установленные/ошибочные)			
		<i>k</i> -means	DBSCAN	Агломеративная кластеризация	Предложенный метод
1	7	3/6	0/10	3/6	7/8
2	3	1/2	0/10	0/3	3/0
3	5	3/2	0/5	3/2	5/0
4	7	2/5	0/5	1/5	3/4
5	3	0/3	0/2	0/3	3/0
6	7	5/2	0/3	5/1	7/0
7	7	3/4	4/2	4/2	7/0
8	3	3/0	0/5	3/0	3/0
9	7	5/2	5/1	1/6	7/0
10	6	4/2	0/5	1/5	6/0
Итого	55	29/28	9/48	21/33	51/12

Примечание. Наилучшие значения показателей выделены жирным шрифтом.

Таблица 3. Индекс инерции для численной оценки качества кластеризации

Table 3. Inertia index for numerical estimation of clustering quality

Номер эксперимента	<i>k</i> -means	DBSCAN	Агломеративная кластеризация	Предложенный метод
1	12 892 364	10 934 472	12 872 668	17 602 038
2	3 180 503	3 190 121	3 186 581	3 084 090
3	2 107 060	1 768 042	1 939 976	1 626 280
4	590 958	591 609	522 601	476 504
5	734 823	785 081	759 291	711 394
6	38 730 687	46 023 242	36 596 174	46 316 400
7	1 103 283	1 166 625	1 181 062	1 152 520
8	579 360	613 335	518 255	640 936
9	10 381 677	10 221 419	10 566 786	10 447 165
10	10 522 955	10 636 685	10 120 163	10 233 920
Среднее	8 082 367	8 593 063	7 826 355	9 229 124

Примечание. Наилучшие значения показателей выделены жирным шрифтом.

По индексу инерции метод *k*-средних показал лучшие результаты в одном эксперименте, DBSCAN — в двух, метод агломеративной кластеризации — в трех и предложенный метод — в четырех. По среднему значению индекса инерции среди всех экспериментов лучший результат у метода агломеративной кластеризации. На среднее значение для предложенного метода сильное влияние оказал результат в эксперименте 1, который резко отличается от аналогов. Однако в остальных случаях наблюдается положительная тенденция.

Обсуждение

Рассмотрим ситуации, когда предлагаемый метод стоит использовать для кластеризации линейных отрезков и выявления скрытых полигональных структур,

а в каких случаях делать это менее целесообразно. На рис. 5 продемонстрированы два примера. Светлые линии на рис. 5, *a*, *c* показывают истинные объекты, а жирные — выделенные отрезки.

В первом случае (рис. 5, *b*) метод корректно создает два кластера из растянутых отрезков, что отражено в двух замкнутых объектах несмотря на то, что боковые стенки зданий находятся близко друг к другу. Во втором случае (рис. 5, *d*) показан некорректный результат работы метода. Вместо двух объектов метод все отрезки сгруппировал в один кластер с учетом их растяжения. Метод не обнаружил диагональную часть здания в левом верхнем углу и построил объект как пересечение вертикальной и горизонтальной линий. Кроме этого, метод не учел отсутствие боковых стенок у зданий и в итоге был сформирован один объект.

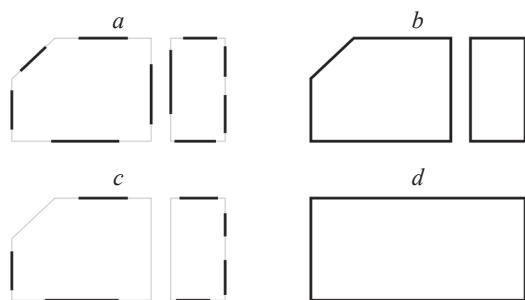


Рис. 5. Примеры корректного (a, b) и ошибочного (c, d) обнаружений полигональных структур

Fig. 5. Examples of correct (a, b) and erroneous (c, d) detection of polygonal structures

Таким образом, метод дает положительный результат, если неявно выраженные структуры содержат хотя бы один отрезок на каждой из своих сторон.

Для развития разработанного метода в дальнейшем планируется использовать предсказательные модели для определения более оптимального направления при растяжении исходных отрезков.

Заключение

В работе предложен метод для кластеризации линейных отрезков, которые образуют неявно выраженные полигональные структуры. Такие структуры появляются на векторных картах в геоинформационных системах при использовании технологий автоматической векторизации пространственных объектов на спутниковых снимках, так как эти технологии несовершенны.

Литература

1. Воробьев А.В., Воробьева Г.Р. Геоинформационная система динамической пространственной кластеризации распределенных источников данных // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 64. С. 61–73. <https://doi.org/10.17223/19988605/64/7>
2. Еремеев С.В., Абакумов А.В., Андрианов Д.Е., Богоявленский И.В., Никонов Р.А. Обнаружение многолетних бугров пучения с использованием декомпозиции цифровых моделей рельефа по топологическим признакам // Современные проблемы дистанционного зондирования Земли из космоса. 2023. Т. 20. № 6. С. 129–143. <https://doi.org/10.21046/2070-7401-2023-20-6-129-143>
3. Еремеев С.В., Абакумов А.В., Андрианов Д.Е., Шибакина Т.А. Метод векторизации спутниковых снимков на основе их разложения по топологическим особенностям // Информатика и автоматизация. 2023. Т. 22. № 1. С. 110–145. <https://doi.org/10.15622/ia.22.1.5>
4. Lin X., Zhou Y., Liu Y., Zhu C. A comprehensive review of image line segment detection and description: taxonomies, comparisons, and challenges // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. V. 46. N 12. P. 8074–8093. <https://doi.org/10.1109/tpami.2024.3400881>
5. Zhang C., He Y., Fraser C.S. Spectral clustering of straight-line segments for roof plane extraction from airborne LiDAR point clouds // IEEE Geoscience and Remote Sensing Letters. 2018. V. 15. N 2. P. 267–271. <https://doi.org/10.1109/LGRS.2017.2785380>
6. Zang D., Wang J., Zhang X., Yu J. Semantic extraction of roof contour lines from airborne lidar building point clouds based on multidirectional equal-width banding // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2024. vol. 17, pp. 16316–16328. <https://doi.org/10.1109/jstars.2024.3422973>

В основе метода лежит применение топологических подходов и идей для анализа данных. Показано, что гомотетия позволяет при постепенном растяжении линейных отрезков обнаружить и проанализировать топологические структуры из компонент связности и полигонов. Приведены результаты исследований, в которых на векторных моделях продемонстрирована работоспособность метода, что подтверждает численная оценка качества кластеризации с использованием индексов Жаккара и инерции. Предложенный метод позволяет устранить неполные данные и получить полигональные структуры, что очень необходимо для полноценного анализа и визуализации пространственной информации. Полученные результаты могут быть использованы для улучшения обнаружения неявно выраженных полигональных структур при автоматическом обновлении векторных карт по спутниковым снимкам. Эксперименты в сравнении с аналогами, такими как методы k -средних, DBSCAN и агломеративная кластеризация, показали, что предложенный метод позволяет обнаружить формы кластеров, отличающиеся от результатов существующих подходов в силу их ограничения на поиск наиболее близких точек в метрическом пространстве.

Преимуществом метода является корректное разделение близко расположенных кластеров. В перспективе метод может быть внедрен в технологии векторизации пространственных данных, а также использован при анализе сложной информации городской среды. Программная реализация метода предназначена для обработки данных в геоинформационной системе QGIS и вместе с тестовыми наборами данных доступна по ссылке¹.

¹ Электронный ресурс]. Режим доступа: <https://github.com/creems84/Howole-Geo.git> (дата обращения: 05.12.2024).

References

1. Vorobev A.V., Vorobeva G.R. Geoinformation system for dynamic spatial clustering of distributed data sources. *Tomsk State University Journal of Control and Computer Science*, 2023, no. 64, pp. 61–73. (in Russian). <https://doi.org/10.17223/19988605/64/7>
2. Eremeev S.V., Abakumov A.V., Andrianov D.E., Bogoyavlensky I.V., Nikonov R.A. Detection of perennial heaving mounds in digital elevation model images using decomposition by topological features. *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2023, vol. 20, no. 6, pp. 129–143. (in Russian). <https://doi.org/10.21046/2070-7401-2023-20-6-129-143>
3. Eremeev S.V., Abakumov A.V., Andrianov D.E., Shirabakina T.A. Vectorization method of satellite images based on their decomposition by topological features. *Informatics and Automation*, 2023, vol. 22, no. 1, pp. 110–145. (in Russian). <https://doi.org/10.15622/ia.22.1.5>
4. Lin X., Zhou Y., Liu Y., Zhu C. A comprehensive review of image line segment detection and description: taxonomies, comparisons, and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, vol. 46, no. 12, pp. 8074–8093. <https://doi.org/10.1109/tpami.2024.3400881>
5. Zhang C., He Y., Fraser C.S. Spectral clustering of straight-line segments for roof plane extraction from airborne LiDAR point clouds. *IEEE Geoscience and Remote Sensing Letters*, 2018, vol. 15, no. 2, pp. 267–271. <https://doi.org/10.1109/LGRS.2017.2785380>
6. Zang D., Wang J., Zhang X., Yu J. Semantic extraction of roof contour lines from airborne lidar building point clouds based on multidirectional equal-width banding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, vol. 17, pp. 16316–16328. <https://doi.org/10.1109/jstars.2024.3422973>

- Earth Observations and Remote Sensing. 2024. V. 17. P. 16316–16328. <https://doi.org/10.1109/jstars.2024.3422973>
7. Xin X., Huang W., Zhong S., Zhang M., Liu Z., Xie Z. Accurate and complete line segment extraction for large-scale point clouds // *International Journal of Applied Earth Observation and Geoinformation*. 2024. V. 128. P. 103728. <https://doi.org/10.1016/j.jag.2024.103728>
 8. Miroshnichenko S.Yu., Titov V.S., Dremov E.N., Mosin S.A. Hough transform application to digitize rectangular spatial objects on aerospace imagery // *SPIRAS Proceedings*. 2018. N 5 (60). P. 189–215. <https://doi.org/10.15622/sp.60.7>
 9. Teplyakov L., Erlygin L., Shvets E. LSDNet: Trainable modification of LSD algorithm for real-time line segment detection // *IEEE Access*. 2022. V. 10. P. 45256–45265. <https://doi.org/10.1109/ACCESS.2022.3169177>
 10. Shili H. Clustering in big data analytics: a systematic review and comparative analysis (review article). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2023. V. 23. N 5. P. 967–979. <https://doi.org/10.17586/2226-1494-2023-23-5-967-979>
 11. Ikotun A.M., Ezugwu A.E., Abualigah L., Abuhaija B., Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data // *Information Sciences*. 2023. V. 622. P. 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
 12. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. 1996. P. 226–231.
 13. Jang J.H., Hong K.S. Fast line segment grouping method for finding globally more favorable line segments // *Pattern Recognition*. 2002. V. 35. N 10. P. 2235–2247. [https://doi.org/10.1016/S0031-3203\(01\)00175-3](https://doi.org/10.1016/S0031-3203(01)00175-3)
 14. Gao J., Langberg M., Schulman L.J. Clustering lines in high-dimensional space: Classification of incomplete data // *ACM Transactions on Algorithms*. 2010. V. 7. N 1. P. 8. <https://doi.org/10.1145/1868237.1868246>
 15. Айдагулов Р.Р., Главацкий С.Т., Михалёв А.В. Модели кластеризации // *Фундаментальная и прикладная математика*. 2020. Т. 23. № 2. С. 17–36.
 16. Kang Y., Wu K., Gao S., Ng I., Rao J., Ye S., Zhang F., Fei T. STICC: a multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity // *International Journal of Geographical Information Science*. 2022. V. 36. N 8. P. 1518–1549. <https://doi.org/10.1080/13658816.2022.2053980>
 17. Alomari H.W., Al-Badarneh A.F., Al-Alaj A., Khamaiseh S.Y. Enhanced approach for agglomerative clustering using topological relations // *IEEE Access*. 2023. V. 11. P. 21945–21967. <https://doi.org/10.1109/access.2023.3252374>
 18. Edelsbrunner H., Letscher H., Zomorodian A. Topological persistence and simplification // *Discrete & Computational Geometry*. 2002. V. 28. N 4. P. 511–533. <https://doi.org/10.1007/s00454-002-2885-2>
 19. Абакумов А.В., Еремеев С.В. Оценка возможности применения метода декомпозиции изображений по топологическим признакам для уменьшения энтропии при их сжатии // *Научно-технический вестник информационных технологий, механики и оптики*. 2023. Т. 23. № 6. С. 1152–1161. <https://doi.org/10.17586/2226-1494-2023-23-6-1152-1161>
 20. Xu F., Beard K. A unifying framework for analysis of spatial-temporal event sequence similarity and its applications // *ISPRS International Journal of Geo-Information*. 2021. V. 10. N 9. P. 594. <https://doi.org/10.3390/ijgi10090594>
 21. Rykov A., De Amorim R.C., Makarenkov V., Mirkin B. Inertia-based indices to determine the number of clusters in K-Means: an experimental evaluation // *IEEE Access*. 2024. V. 12. P. 11761–11773. <https://doi.org/10.1109/ACCESS.2024.3350791>
 7. Xin X., Huang W., Zhong S., Zhang M., Liu Z., Xie Z. Accurate and complete line segment extraction for large-scale point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 2024, vol. 128, pp. 103728. <https://doi.org/10.1016/j.jag.2024.103728>
 8. Miroshnichenko S.Yu., Titov V.S., Dremov E.N., Mosin S.A. Hough transform application to digitize rectangular spatial objects on aerospace imagery. *SPIRAS Proceedings*, 2018, no. 5 (60), pp. 189–215. <https://doi.org/10.15622/sp.60.7>
 9. Teplyakov L., Erlygin L., Shvets E. LSDNet: Trainable modification of LSD algorithm for real-time line segment detection. *IEEE Access*, 2022, vol. 10, pp. 45256–45265. <https://doi.org/10.1109/ACCESS.2022.3169177>
 10. Shili H. Clustering in big data analytics: a systematic review and comparative analysis (review article). *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 5, pp. 967–979. <https://doi.org/10.17586/2226-1494-2023-23-5-967-979>
 11. Ikotun A.M., Ezugwu A.E., Abualigah L., Abuhaija B., Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 2023, vol. 622, pp. 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
 12. Ester M., Kriegel H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 226–231.
 13. Jang J.H., Hong K.S. Fast line segment grouping method for finding globally more favorable line segments. *Pattern Recognition*, 2002, vol. 35, no. 10, pp. 2235–2247. [https://doi.org/10.1016/S0031-3203\(01\)00175-3](https://doi.org/10.1016/S0031-3203(01)00175-3)
 14. Gao J., Langberg M., Schulman L.J. Clustering lines in high-dimensional space: Classification of incomplete data. *ACM Transactions on Algorithms*, 2010, vol. 7, no. 1, pp. 8. <https://doi.org/10.1145/1868237.1868246>
 15. Aidagulov R.R., Glavatsky S.T., Mikhalev A.V. Clustering models. *Journal of Mathematical Sciences*, 2022, vol. 262, no. 5, pp. 603–616. (in Russian). <https://doi.org/10.1007/s10958-022-05841-9>
 16. Kang Y., Wu K., Gao S., Ng I., Rao J., Ye S., Zhang F., Fei T. STICC: a multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical Information Science*, 2022, vol. 36, no. 8, pp. 1518–1549. <https://doi.org/10.1080/13658816.2022.2053980>
 17. Alomari H.W., Al-Badarneh A.F., Al-Alaj A., Khamaiseh S.Y. Enhanced approach for agglomerative clustering using topological relations. *IEEE Access*, 2023, vol. 11, pp. 21945–21967. <https://doi.org/10.1109/access.2023.3252374>
 18. Edelsbrunner H., Letscher H., Zomorodian A. Topological persistence and simplification. *Discrete & Computational Geometry*, 2002, vol. 28, no. 4, pp. 511–533. <https://doi.org/10.1007/s00454-002-2885-2>
 19. Abakumov A.V., Eremeev S.V. Assessing the possibility of using the method of image decomposition based on topological features to reduce entropy during their compression. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 6, pp. 1152–1161. (in Russian). <https://doi.org/10.17586/2226-1494-2023-23-6-1152-1161>
 20. Xu F., Beard K. A unifying framework for analysis of spatial-temporal event sequence similarity and its applications. *ISPRS International Journal of Geo-Information*, 2021, vol. 10, no. 9, pp. 594. <https://doi.org/10.3390/ijgi10090594>
 21. Rykov A., De Amorim R.C., Makarenkov V., Mirkin B. Inertia-based indices to determine the number of clusters in K-Means: an experimental evaluation. *IEEE Access*, 2024, vol. 12, pp. 11761–11773. <https://doi.org/10.1109/ACCESS.2024.3350791>

Автор

Еремеев Сергей Владимирович — кандидат технических наук, доцент, доцент, Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых», Муром, 602264, Российская Федерация, [sc 56673892800](https://orcid.org/0000-0001-8482-1479), <https://orcid.org/0000-0001-8482-1479>, sv-eremeev@yandex.ru

Author

Sergey V. Eremeev — PhD, Associate Professor, Associate Professor, Murom Institute (Branch) of Vladimir State University, Murom, 602264, Russian Federation, [sc 56673892800](https://orcid.org/0000-0001-8482-1479), <https://orcid.org/0000-0001-8482-1479>, sv-eremeev@yandex.ru