УДК 544.623 DOI: 10.33075/2220-5861-2020-1-14-20

МЕТОДИЧЕСКИЕ АСПЕКТЫ РАЙОНИРОВАНИЯ ПРИРОДНЫХ ТЕРРИТОРИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Д.О. Кривогуз, Р.В. Боровская

Азово-Черноморский филиал ФГБНУ «ВНИРО» («АзНИИРХ»),
РФ, г. Керчь, ул. Свердлова 2
E-mail: krivoguz_d_o@azniirkh.ru

Рассмотрен современный подход к районированию природных территорий с помощью машинного обучения. Авторами детально проанализирован алгоритм проведения районирования с помощью применения кластеризации, а также выделены основные его этапы проведения. Были выделены как положительные стороны применения данного подхода (объективность, точность, простоту модифицируемости и настройки) так и отрицательные, к которым относится сильная зависимость от объема, точности и чистоты данных. В заключении авторами делается вывод о допустимости применения данного подхода, учитывая необходимость его совершенствования и оптимизации.
Ключевые слова: районирование территорий, рыбохозяйственное районирование, машинное обучение, экологические проблемы, кластеризация, нормализация данных.

Поступила в редакцию: 29.01.2020. После доработки: 12.02.2020.

Введение. Проблема районирования территорий всегда занимала важное место среди вопросов, решаемых географической наукой. В этом отношении под зоной или районом принято считать территориальную систему, являющуюся составной частью более крупного пространственного образования. Основываясь на вышеизложенном, под зонированием следует понимать процесс идентификации и изучения объективно существующей территориальной системы, которой свойственна иерархичность составляющих её комплексов [1].

Известно, что природные территории Российской Федерации подверглись существенным изменениям в результате постоянно увеличивающейся хозяйственной деятельности, что, несомненно, приводит к появлению природно-территориальных комплексов с рядом экологических проблем [2, 3]. Таким образом, изучение таких взаимодействий между обществом и природой, а также образуемые в результате экологические проблемы, прежде всего с помощью районирования, на данный момент представляет особый научный интерес [4].

Зонирование любой территории, как правило, включает в себя выполнение нескольких задач:

1. Определение существующих комплексов.
2. Картирование территории и факторов, обусловливающих её свойства.
3. Глубокое изучение комплексной организации исследуемой системы.
4. Изучение явлений, процессов и факторов, формирующих исследуемый комплекс.
5. Пространственную классификацию.
6. Определение и изучение любых взаимодействий между факторами или частями комплекса.
7. Разработка подходящих методов для зонирования территории.

Районирование территории базируется на интегральной оценке качества показателей окружающей среды при учете максимально возможного числа факторов, которые могут в полной мере описывать свойства и особенности анализируемой территории.

В целом машинное обучение в географических науках в последнее время получило значительное развитие. К примеру, Вагизов М.Р. [5] применял машинное обучение для разработки базы данных эталонов лесных насаждений, для определения их таксационных показателей. Первостепенной задачей в его

работе было определение границ лесотаксационного выдела.

Казаков Э. рассматривал применение машинного обучения для прогноза цветения фитопланктона [6], а также определения типов морского льда по спутниковым данным [7]. В контексте цветения фитопланктона машинное обучение применялось для исследования причин цветения, т.е. при каких условиях среды они происходят и какие комбинации параметров окружающей среды для них благоприятнее. Также важным аспектом их изучения была попытка спрогнозировать цветение фитопланктона в будущем с учетом данных климатических моделей.

Решением проблемы районирования была заинтересована Ружникова Н.Н., которая в своей работе [8] провела геоэкологическое районирование акватории Белого моря с целью осуществления нефтяных перевозок.

Особо важное значение в контексте развития Российской Федерацией новых территорий для добычи природных ресурсов имеет районирование арктического шельфа, проблему которого в своей работе [9] изложили Денисов В.В. и Ильин Г.В., акцентируя особое внимание на вылове рыбных ресурсов в Баренцево-Карском регионе.

В данном аспекте следует также отметить реализацию рыбохозяйственного районирования в Арктике Матишовым Г.Г., Балыкиным П.А. и Пономаревой Е.Н. с учетом интересов нефтегазодобычи и морского транспорта [10].

Таким образом, главная цель данной работы заключается в формировании современной математической методологии, основанной на методах машинного обучения, способной в значительной мере усовершенствовать современные подходы к зонированию любой территории.

Кластеризация, как метод зонирования территории. Кластеризация представляет собой задачу разделения совокупности данных на отдельные непохожие друг на друга группы схожих между собой объектов. Алгоритмы кластеризации можно разделить на 2 большие группы – иерархические и не иерархические [11, 12].

К иерархическим группам можно отнести агломеративные алгоритмы, суть которых заключается в том, что они начинают работу с отдельных элементов набора данных в последствии объединяя их, и сепарационные – начинающие работу с одного большого кластера, который впоследствии разделяют на более мелкие. К не иерархическим можно отнести алгоритмы, базирующиеся на теории градов, EM-алгоритмы, алгоритмы k-means и нечеткие алгоритмы.

Эффективность любого алгоритма кластеризации определяется достижением им гипотезы компактности, которая заключается в том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных [13].

Зонирование территории, используя кластеризацию, проходит в несколько этапов (рис. 1) и включает:

1. Выбор необходимых данных для анализа.

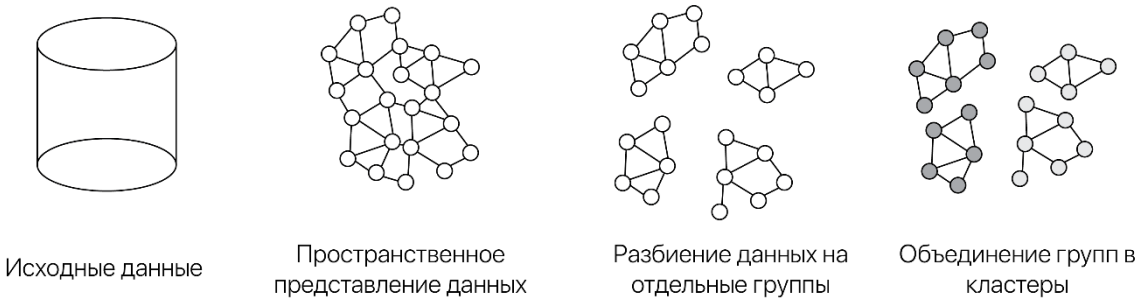

Рис. 1. Схематическое представление проведения зонирования территории с помощью кластеризации
Fig. 1. Schematic visualization of territory zoning using clustering

2. Нормализация данных.
3. Проверка предрасположенности данных к кластеризации.
4. Выбор оптимального количества кластеров.
5. Кластеризация и проверка итоговых результатов.

Нормализация данных. Нормализация данных представляет собой трансформацию данных, осуществляемую на этапе их первоначальной подготовки [14, 15]. В случае машинного обучения, нормализация – процедура первоначальной подготовки исходных данных, при которой значения переменных сводятся к узкому промежутку числовых значений, к примеру: [0...1] или [–1...1].

Важность нормализации данных исходит из особенностей функционирования алгоритмов и моделей в машинном обучении. Значения «сырых» данных могут находиться в довольно большом промежутке числовых значений и отличаться друг от друга на несколько порядков. Работа таких алгоритмов, используемых в машинном обучении, как искусственные нейронные сети или самоорганизующиеся карты Кохонена без проведенного нормирования данных будет абсолютно не правильной, что приведет к ложным результатам.

Существует достаточно много способов для проведения нормализации данных и приведению их к довольно узкому диапазону значений, чтобы их можно было использовать в машинном обучении. Основываясь на используемой в них функции, алгоритмы нормализации данных могут быть линейными и не линейными. Наиболее популярными способами нормализации являются:

Минимакс – линейная трансформация данных в промежутке [0...1], где минимальное значение соответствует 0, а максимальное – 1

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \ . \qquad (1)$$

Z-масштабирование основано на среднем и стандартном отклонении – разница между переменной и средним значением деленное на стандартное отклонение

$$Z = \frac{x - \mu}{\sigma} \ . \qquad (2)$$

Десятичное масштабирование осуществляется путем удаления десятичного разделителя из значений переменной.

На практике минимакс и Z-масштабирование имеют одинаковые области применения и, как следствие, взаимозаменяемы. С другой стороны, как показывает практика, для расчета расстояний между переменными применяется Z-масштабирование, тогда как минимакс в основном используется для визуализации данных.

Определение предрасположенности данных к кластеризации. Одной из наиболее актуальных проблем в данной сфере является то, что при кластеризации будет происходить образование групп, даже если исходные данные будут иметь совершенно случайную структуру. Поэтому первым шагом проверки кластеризации, который осуществляется еще до самой кластеризации, на этапе подготовки данных, является проверка предрасположенности исследуемых данных к кластеризации.

Существует два индикатора, которые показывают эту предрасположенность: статистика Хопкинса и визуальная оценка предрасположенности к кластеризации (VAT-диаграмма) [16, 17].

Для расчета статистики Хопкинса необходимо создать B псевдонаборов данных, которые создаются случайно на основании распределения с таким же стандартным отклонением, как и у исходного набора данных. От каждой переменной i из n, среднее расстояние к k ближайшим соседям рассчитывается следующим образом: w_i между настоящими переменными и q_i между сгенерированными и их ближайшими соседями. Тогда статистика Хопкинса рассчитывается как

$$H_{ind} = \frac{\sum_n w_i}{\sum_n q_i + \sum_n w_i} \ . \qquad (3)$$

Если $H_{ind} > 0{,}5$, то это соответствует

нулевой гипотезе о том, что q_i и w_i одинаковы и значения распределены случайно и равномерно. Если $H_{ind} < 0{,}25$, то это свидетельствует о том, что данные в исследуемом наборе имеют склонность к образованию групп.

Для визуальной оценки лучше подходит использование VAT-диаграммы. VAT алгоритм состоит из следующих этапов:

1. Рассчитать матрицу различий между объектами в наборе данных, используя Евклидово расстояние.
2. Перераспределить матрицу различий таким образом, чтобы похожие объекты были расположены близко друг к другу.
3. Перераспределенная матрица различий визуализируется как организованная диаграмма различий, которая и является визуальным представлением VAT.

Выбор оптимального количества кластеров. На данный момент существует два основных способа определить оптимальное количество кластеров у исследуемого набора данных – метод «локтя» (*elbow method*) и использование gap- статистики (*gap-statistics*).

Локтевой метод подразумевает присутствие шаблона отклонений дисперсии W_{total} с увеличением в количестве групп k. Объединив все найденные наблюдения n в одну группу, получится большая внутриклассовая дисперсия, которая будет уменьшаться к 0, когда $k \to n$. Точка, в которой это уменьшение дисперсии замедлится, называется «локтем».

Альтернативным решением локтевому методу служит применение gap-статистики, которая создается на основе передискретизации и метода Монте-Карло. К примеру, пусть $E_n^*\{log(W_k^*)\}$ означает оценку средней дисперсии W_k^*, полученой бутстрапинтом, когда k кластеров образовываются несколькими случайными объектами f из исходного набора данных размером n. Тогда gap-статистика будет рассчитываться как

$$Gap_n(k) = E_n^*\{log(W_k^*)\} - log(W_k) \ , (4)$$

где $Gap_n(k)$ означает отклонение наблюдаемой дисперсии W_n от ожидаемых значений, если исходные данные образуют только один кластер.

Проверка результатов кластеризации. На данный момент существует несколько оптимальных путей для проверки результатов кластеризации:

1. Внешняя проверка, которая заключается в сравнении результатов кластеризации с существующим проверочным набором данных.
2. Относительная проверка, которая заключается в определении структуры сформированных кластеров путем последовательного изменения параметров используемого алгоритма.
3. Внутренняя проверка, суть которой заключается в получении внутренней информации о проведенном процессе кластеризации.
4. Оценка устойчивости кластеризации путем передискретизации.

Часто для получения показателей точности кластеризации используются индексы, наиболее распространенными из которых являются индекс силуэта и индекс Калински-Харабаза [18].

Суть индекса Калински-Харабаза заключается в следующем.

Пусть $\overline{d}^2$ является среднеквадратичным расстоянием между элементами в кластеризируемом пространстве, а $\overline{d}_{c_i}^2$ – средний квадрат расстояния между элементами кластера c_i. Тогда расстояние внутри каждого кластера будет равным

$$WGSS = \frac{1}{2}\sum_{i=1}^{c}(nc_i - 1)^{\bar{d}^2} c_i \ , \quad (5)$$

а в свою очередь расстояние между кластерами будет равным

$$BGSS = \frac{1}{2}((c-1)\overline{d}^2 + (N-c)A_c) \ , \quad (6)$$

$$A_c = \frac{1}{N-c}\sum_{i=1}^{c}(n_{c_i} - 1)(\overline{d}^2 - \overline{d}_{c_i}^2) \ , \quad (7)$$

где A_c – взвешенная средняя разница расстояний между центрами кластеров и общим центром набора данных.

Отсюда индекс Калински-Харабаза будет равен

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{1+\frac{N-c}{c-1}a_c}{(1-ac)};\ a_c = \frac{A_c}{\bar{d}^2}\ . \quad (8)$$

Отсюда видно, что если все расстояния между точками похожи, тогда ac = 0, а *VRC* = 1. Если a_c = 1, то это характеризует идеальную кластеризацию. Максимальные значения *VRC* соответствуют наиболее оптимальной кластерной структуре.

Другой, широко применяемый способ оценить точность кластеризации, – использование индекса силуэта. Его значения отображают степень сходства между объектами и кластерами, к которым они принадлежат, в сравнении с другими кластерами [18].

Силуэт каждого кластера рассчитывается следующим образом.

Пусть объект x_j соответствует кластеру c_p. Обозначим среднее расстояние от этого объекта до другого объекта этого кластера c_p как a_{pj}, а среднее расстояние от этого объекта x_j до объекта из другого кластера $c_q, q \neq p$ как $d_{q,j}$. Пусть $b_{pj} = min_{q\neq p} d_{qj}$, что означает меру разнородности одиночного объекта от ближайшего кластера. Таким образом, силуэт каждого отдельного элемента кластера будет равен

$$S_{xj} = \frac{b_{pj}-apj}{\max(a_{pj},b_{pj})}\ . \quad (9)$$

Чем больше значение S_{xj}, тем более вероятна принадлежность элемента x_j к кластеру p. Расчет для всей кластерной структуры осуществляется путем усреднения значений элементов

$$SWC = \frac{1}{N}\sum_{j=1}^{N} S_{x_j}\ . \quad (10)$$

Лучшему результату кластеризации соответствуют большие значения *SWC*, что достигается при уменьшении расстояния внутри кластера a_{pj} и увеличении расстояния до соседнего кластера b_{pj}.

Наиболее оптимальным способом кластеризации пространственных данных является иерархическая кластеризация. Принцип её работы базируется на объединении наиболее похожих друг на друга с точки зрения атрибутивных характеристик объектов в отдельные группы, пока количество таких групп не будет равным искомому количеству кластеров. Расстояние между объектами в *N*-мерном пространстве, где *N* – количество атрибутивных характеристик каждого объекта, рассчитывается по формуле

$$\mathrm{d} = \sqrt{w_1(a_1 - b_1)^2 + \cdots + w_n(a_n - b_n)^2}, \quad (11)$$

где d - расстояние между объектами, a_n и b_n – координаты объектов a и b на осях *N*-мерного пространства, w_n – вес n-го признака.

В данном случае весовой параметр w_n не является обязательным и служит лишь для более точной настройки работы алгоритма в случаях, когда исследователю заранее известна значимость и вклад каждого исследуемого фактора.

В целом практическую реализацию зонирования территории с помощью кластеризации можно условно сформировать следующим образом.

Пусть имеется условный набор пространственных данных с двумя показателями, характеризующими некую территорию (рис. 2).

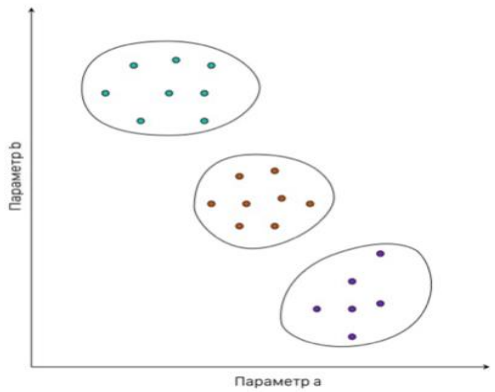

Рис. 2. Схематическая группировка кластеров в двумерном пространстве
Fig. 2. Schematic example of cluster grouping in two-dimensional space

Имея два анализируемых показателя *a* и *b*, получено двумерное пространство, где будут располагаться исследуемые объекты. Как видно из рис. 2 в пространстве четко выделяются 3 различных кластера со схожими между собой объектами.

Для объектов из «бирюзового кластера» характерны высокие показатели фактора *b* при низких значениях показателя *a*. Для «оранжевого» кластера характерны средние значения как фактора *a*, так и фактора *b*. Для «фиолетового» кластера характерны высокие значения показателя, а при низких значениях показателя *b*.

Заключение. Таким образом, методический подход, представленный в данной статье, позволяет в полной мере использовать его для проведения районирования территорий, выделяя среди них разнородные, не похожие друг на друга участки.

Как показал проведенный анализ, к неоспоримым преимуществам данного подхода можно отнести отсутствие субъективности, присущей человеку, точность проведения анализа, возможность постоянной модификации модели, как путем добавления новых данных полевых исследований, так и путем модификации самого алгоритма. Также следует отметить, что неоспоримым преимуществом данного подхода является универсальность его использования для любых территорий как по размеру, так и свойствам.

К недостаткам данного подхода следует отнести сильную зависимость от качества данных, проведенного нормирования, что при любых существенных отклонениях в данных аспектах может привести к значительному искажению полученных результатов. Это же касается и размера данных. При значительных объемах, используемых в модели данных, возможно возникновение затруднения проведения исследования и неспособности проводить работу алгоритма, что приведет или к смене используемого алгоритма, или к значительной редукции данных и как следствие – упрощению моделей и искажению реальных результатов.

В целом, следует отметить, что использование данного подхода оправдано в большинстве возникающих случаев, но необходимость в совершенствовании и дальнейшей оптимизации данного подхода является актуальным и на современном этапе.

СПИСОК ЛИТЕРАТУРЫ

1. *Кривогуз Д.О., Захарова Ю.Б.* Применение геопространственного анализа при прогнозировании эколого-экономического развития Керченского полуострова // Геоинформатика. 2018. № 1. С. 52–55.

2. *Bocharnikov V.* Ecological and geographical mapping of Russian economic regions based on GIS technologies // Vestnik Volgogradskogo gosudarstvennogo universiteta. Serija 3. Ekonomika. Ekologija. 2016. № 3 (3). С. 163–176.

3. *Matishov G.G., Bespalova L.A., Ivlieva O.V.* The Sea of Azov: Recent abrasion processes and problems of coastal protection // Doklady Earth Sciences. 2016. № 2 (471). С. 1269–1272.

4. *Войтов И.В.* Научно-инновационные принципы геоэкологического районирования административных территорий Беларуси / И.В. Войтов, М.А. Гатих, Л.С. Лис [и др.] // Вестник Белорусско-Российского университета. 2009. № 22 (1). С. 113–127.

5. *Михайлова А.А., Вагизов М.Р.* Методика обработки данных дистанционного зондирования Земли с применением информационных технологий и аллометрических зависимостей для определения лесотаксационных показателей древостоев // Успехи современного естествознания. 2018. № 4. С. 80–85.

6. *Kondrik D., Kazakov E., Pozdnyakov D.* A synthetic satellite dataset of the spatio-temporal distributions of Emiliania huxleyi blooms and their impacts on Arctic and sub-Arctic marine environments (1998–2016) // Earth System Science Data. 2019. № 1 (11). С. 119–128.

7. *Demchev D.* Sea ice drift tracking from sequential SAR images using accelerated-KAZE features / D. Demchev, V. Volkov, E. Kazakov [et al.] // IEEE Transactions on Geoscience and Remote Sensing. 2017. № 9 (55). С. 5174–5184.

8. *Ружникова Н.Н.* Геоэкологическое районирование акватории Белого моря при транспортировке нефтяных углеводородов // Известия высших учебных заведений. Северо-Кавказский регион. Естественные науки. 2012. № 6. С. 94–98.

9. *Денисов В.В., Ильин Г.В.* Районирование акваторий как инструмент оптимизации природопользования на Арктическом шельфе // Проблемы Арктики и Антарктики. 2008. № 79 (2). С. 134–144.

10. *Matishov G.G., Balykin P.A., Ponomareva E.N.* Fishery zoning is the first stage of spatial planning of marine activities in the Arctic // Science in the South of Russia. 2018. № 2 (14). С. 33–41.

11. *Семенова А.Ю.* Социально-экономические приоритеты сохранения и укрепления здоровья населения Республики Крым в контексте инновационного развития региона // Экономика и управление: теория и практика. 2018. № 3 (4). С. 49–54.

12. *Shi T., Horvath S.* Unsupervised learning with random forest predictors // Journal of Computational and Graphical Statistics. 2006. № 1 (15). С. 118–138.

13. *Качановский Ю.П., Коротков Е.А.* Предобработка данных для обучения нейронной сети // Фундаментальные исследования. 2011. № 1 (12). С. 117–120.

14. *Aydın A., Eker R.* Fuzzy rule-based landslide susceptibility mapping in Yığılca Forest District (Northwest of Turkey) // Journal Journal of the Faculty of Forestry Istanbul University. 2016. № 662 (66). С. 559–571.

15. *Mirzaei J.* Assessment of land cover changes using RS and GIS (case study: Zagros forests, Iran) / J. Mirzaei, A. Mohamadi, Z. Heidarizadi [et al.] // Journal of Materials and Environmental Science. 2015. № 6 (9). С. 2565–2572.

16. *Campello R.* Density-based clustering / R. Campello, P. Kröger, J. Sander [et al.] // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019. 15 с.

17. *Sivogolovko E., Thalheim B.* Semantic approach to cluster validity notion // Advances in Databases and Information Systems. 2013. Т. 186. С. 229–239.

18. *Сивоголовко Е.В.* Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании. 2011. № 4. С. 14–31.

METHODOLOGICAL ASPECTS OF NATURAL TERRITORIES ZONING USING MACHINE LEARNING

D.O. Krivoguz, R.V. Borovskaya

Azov-Black Sea Branch of the FSBSI «VNIRO» («AzNIIRKH»),
RF, Kerch, Sverdlova St., 2

This article focuses on modern approach to natural territories zoning using machine learning techniques. The authors analyze in detail the zoning algorithm by clustering, and also highlight its main stages. Both positive aspects of this approach (objectivity, accuracy, simplicity of modifiability and settings) and negative ones, which include a strong dependence on the volume, accuracy and purity of the data, are highlighted. In conclusion, the authors draw a conclusion that this approach is acceptable, but it the needs for improvement and optimization.
Keywords: territories zoning, fishery zoning, machine learning, environmental problems, clustering, data normalization.