Телевизионные системы, передача и обработка изображений

DOI 10.24412/2221-2574-2021-444-49-56 УДК 519.172

МЕТОД ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ЦИФРОВЫХ МНОГОСПЕКТРАЛЬНЫХ ИЗОБРАЖЕНИЙ В ЗАДАЧАХ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Никитин Олег Рафаилович
доктор технических наук, профессор, заведующий кафедрой радиотехники и радиосистем ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»[1].
E-mail: olnikitin@mail.ru

Кисляков Алексей Николаевич
кандидат технических наук, доцент, доцент кафедры информационных технологий Владимирского филиала ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации»[2].
E-mail: ankislyakov@mail.ru

[1]*Адрес:* 600000, Российская Федерация, г. Владимир, ул. Горького, д. 87.
[2]*Адрес:* 600017, Российская Федерация, г. Владимир, ул. Горького, д. 59а.

Аннотация: Работа направлена на решение актуальной проблемы идентификации и интерпретации аномальных наблюдений при решении задач экологического мониторинга с помощью многоспектральных методов дистанционного зондирования земной поверхности. Предлагаемый в работе метод основан на использовании кластерного подхода к выявлению аномальных наблюдений на многоспектральных снимках. Кластеризация выполняется иерархическими методами, которые представляют собой совокупность алгоритмов упорядочивания изображений на основе их признаков и построения дендрограмм, состоящих из групп наблюдаемых изображений. В качестве метрики расстояний между числовыми и категориальными признаками изображений предлагается использовать расстояние Гауэра. Оценка качества кластеризации выполняется на основе показателя суммы квадратов метрических расстояний между признаками изображений внутри кластера и средней ширины силуэта, позволяющих выбрать оптимальное количество кластеров и оценить качество результатов разбиения. Выявление аномалий осуществляется путём анализа результатов иерархической кластеризации и выявления ветвей дендрограммы, располагающихся на начальных уровнях построения дендрограммы и не имеющих ветвлений. Методика позволяет более точно интерпретировать результаты кластеризации относительно выявления причин аномальных наблюдений в наборе данных, а также эффективно использовать наборы данных дистанционного зондирования, в том числе в целях комплексирования изображений.

Ключевые слова: информационная энтропия, фрактальная размерность, кластерный анализ, теория графов, многоспектральные снимки, дистанционное зондирование.

Актуальность задачи

В современном мире наиболее надёжными универсальными методами экологического мониторинга без сомнения являются методы дистанционного зондирования земной поверхности. Автоматизированные системы мониторинга позволяют получать наборы цифровых многоспектральных изображений земной поверхности в виде спутниковых снимков одновременно в различных диапазонах электромагнитного излучения, что с одной стороны, позволяет достаточно подробно изучить объект исследования путём оценки его характеристик по многоспектральным снимкам [1], а с другой стороны, порождает огромный поток данных, нуждающихся в детальном анализе в

целях идентификации аномальных объектов на изображениях и интерпретации причин их возникновения, а также кластеризации по признакам. При этом количество спектральных каналов обычно варьируется в зависимости от аппаратуры и назначения системы зондирования.

Многоспектральные снимки представляют собой набор цифровых изображений в различных диапазонах с привязкой к определённой местности и времени съёмки, позволяющие выявить характеристики объектов на основе соотношений яркости цифровых изображений в различных спектральных каналах, а также проследить изменения наблюдаемой сцены в динамике.

Однако такой огромный массив данных с трудом поддаётся детальному анализу каждого изображения, поэтому необходимо в первую очередь на основе обобщённых признаков выявить изображения с присутствующими на них аномальными объектами, а затем детально изучить небольшой набор изображений, представляющих особый интерес. Задача выделения из исходного набора данных изображений, содержащих аномальные объекты, является наиболее актуальной для совершенствования алгоритмов работы систем автоматизированного экологического мониторинга.

Целью работы является разработка методики признаковой кластеризации цифровых многоспектральных изображений земной поверхности на основе теории графов и изучение признаков групп объектов — кластеров.

Объекты и методы

Описанные системы экологического мониторинга позволяют получить массив снимков, который может содержать нехарактерные показатели или аномальные наблюдения [2], но не все из этих аномалий являются целью мониторинга. Существует достаточно способов выявления аномалий на снимках: от самых простых — визуальных, а также методов, основанных на статистических критериях, до более сложных, использующих алгоритмы морфологического анализа [3], и заканчивая рекуррентными нейронными сетями [4]. Однако все эти методы предполагают непосредственное изучение объектов на изображении, что в любом случае требует высокой производительности системы, а также необходимости детального анализа каждого изображения. Каждое изображение обладает косвенными характеристиками, которые достаточно легко вычисляются, но не позволяют с высокой долей надёжности идентифицировать и интерпретировать объекты на изображении. Эти признаки выражаются в виде яркостного разнообразия наблюдаемой сцены, сложности, однородности и многообразия форм объектов, спектрального диапазона, времени и места наблюдения.

Суть подхода состоит в возможности автоматически выделить набор изображений с аномальными признаками и изучить их более детально. При этом обычно выделяют следующие возможные источники аномальных наблюдений:

1. Ошибочное наблюдение, обусловленное шумами, помехами различного рода (например, облачный покров на изображениях видимого диапазона спектра) — ошибки первого рода.

2. Резкое изменение условий наблюдаемой сцены, связанных с формированием результата наблюдения – ошибки второго рода.

Основной проблемой анализа массива изображений является наличие в них огромного количества аномалий, не представляющих интереса для экологического исследования: это могут быть события рядового характера, многократно повторяющиеся на изображениях нескольких спектральных каналов во времени и пространстве, в то время как аномальные наблюдения, представляющие собой ошибки второго рода (например, лесной пожар) требуют от системы автоматизированного монито-

ринга незамедлительной реакции. В этой связи на первый план выходит задача не только поиска, идентификации, но и интерпретации аномальных наблюдений.

Для поиска изображений, содержащих аномалии, могут быть использованы методы кластерного анализа [5], суть которых заключается в оценке метрических расстояний между изображениями, характеризующимися векторами признаков. Если значение этого расстояния удалено от центров кластеров более чем на определённую величину, то может быть выдвинута гипотеза о том, что изображение содержит аномальные объекты. В этом понимании, кластерные методы схожи с метрическими и статистическими методами поиска аномальных наблюдений [6, 7], однако при более детальном рассмотрении кластерные методы позволяют решить более широкий круг задач, связанных ещё и с интерпретацией выявленных аномалий.

Для удобства локализации объектов интереса в наборе многоспектральных изображений необходимо выполнить кластеризацию этих изображений по признакам. Такая разбивка позволит не только выделить в отдельные кластеры изображения, содержащие аномалии и/или имеющие схожие признаки, но и выяснить, какие изображения, с точки зрения задач комплексирования снимков, наиболее отличны друг от друга [3]. Результат иерархической кластеризации представляется в виде направленного графа с бинарным разбиением — дендрограммы [8]. Для разбиения на кластеры необходимо определить метод объединения объектов и метрики разбиения. Существуют следующие варианты объединения объектов в группы [9]:

1. Агломеративная кластеризация начинается с n кластеров, где n – число изображений: предполагается, что каждое из них представляет собой отдельный кластер. Затем алгоритм пытается найти и сгруппировать наиболее схожие по совокупности признаков изображения в группы (кластеры).

2. Дивизионная кластеризация, наоборот, изначально предполагает, что все n изображений представляют собой одну большую группу, а далее наименее схожие из них разделяются на отдельные кластеры.

Таким образом, на основе метрических расстояний может быть сформирована иерархическая дендрограмма, что позволяет интерпретировать найденные аномалии: если изображение можно выделить в отдельный кластер и его набор признаков во всех спектрах наблюдения отделим от признаков изображений других кластеров, то данный снимок фиксирует аномальное наблюдение (объект) и является ошибкой второго рода, если же аномалия наблюдается только по одному или двум-трём признакам, то вероятнее всего снимок не будет отличаться от остальных по всему набору признаков.

При этом выделение аномальных наблюдений происходит на первых итерациях работы алгоритма на начальных уровнях построения дендрограммы, когда относительное расстояние между кластерами больше 0,4–0,5. Исследования [9] показали, что именно при этих значениях наблюдается выделение аномалий в отдельную ветвь. Вторым признаком аномальных объектов является отсутствие дальнейшего разветвления этой ветви дерева.

Преимуществом подхода является возможность автоматизированного определения количества групп интереса. В качестве метрики расстояний между изображениями в случае различных типов признаков, состоящих из числовых и категориальных переменных, используется расстояние Гауэра [9, 10]

$$S_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \qquad (1)$$

где W_{ijk} — весовой коэффициент. W_{ijk} принимает значение 1, если существует сопоставление объектов по признаку k, и 0 — если такое сопоставление отсутствует; S_{ijk} — весовой коэффициент, характеризующий вклад в сходство изображений, зависящий от того, учитывается ли признак k при сравнении объектов i и j. Чем меньше расстояние Гауэра, тем лучше классификация отражает структуру данных. На основе расстояния Гауэра вычисляется матрица отличий — таблица, заполненная значениями от 0 до 1. Значение «0» означает отсутствие связи между признаками изображений, значение «1» говорит об однозначной связи или полном сходстве признаков двух изображений. Для реализации высокого качества кластеризации необходимо, чтобы расстояние между точками внутри кластера (или компактность) было минимальным, а расстояние между группами (отделимость) — максимально возможным. Для этого необходимо выполнить расчёт основных характеристик качества кластеризации: суммы квадратов метрических расстояний между объектами внутри кластера и среднюю ширину силуэта [10,11]. Для каждого изображения ширина силуэта вычисляется следующим образом:

$$s_i = \frac{b_i - a_i}{\max\left(b_i, a_i\right)}, \qquad (2)$$

где a_i — среднее расстояние между i-м изображением и всеми изображениями кластера; b_i — среднее расстояние между i-м изображением и изображениями другого ближайшего кластера.

Ширина силуэта позволяет оценить степень согласованности признаков изображений внутри кластера и изменяется в пределах от –1 до 1. Чем ближе показатель ширины силуэта к нулю, тем менее однозначно можно отнести изображение к текущему кластеру.

Резкое изменение показателя суммы квадратов метрических расстояний между объектами внутри кластера (резкий изгиб на графике — рис. 2) при определённом количестве кластеров также позволяет сделать вывод о том, что дальнейшее разбиение на кластеры теряет смысл. Поэтому данные характеристики качества разбиения дополняют друг друга и позволяют автоматически определить количество кластеров.

Таким образом, предлагается выстроить следующую методику автоматической идентификации аномальных наблюдений на основе методов иерархической кластеризации с использованием показателей отделимости изображений друг от друга внутри кластера и показателя отделимости кластеров между собой:

1. Загрузка исходных данных с указанием упорядоченных значений признаков;
2. Вычисление матрицы отличий на основе расстояний Гауэра;
3. Выполнение иерархического разбиения данных агломеративным и дивизионным методом;
4. Построение дендрограммы для каждого вида кластеризации, первичная идентификация аномальных изображений;
5. Выбор количества кластеров на основе показателей разделимости признаков изображений внутри кластера и разделимости самих кластеров;
6. Построение дендрограммы с разбиением на кластеры, интерпретация идентифицированных изображений с аномалиями.

Результаты и обсуждение

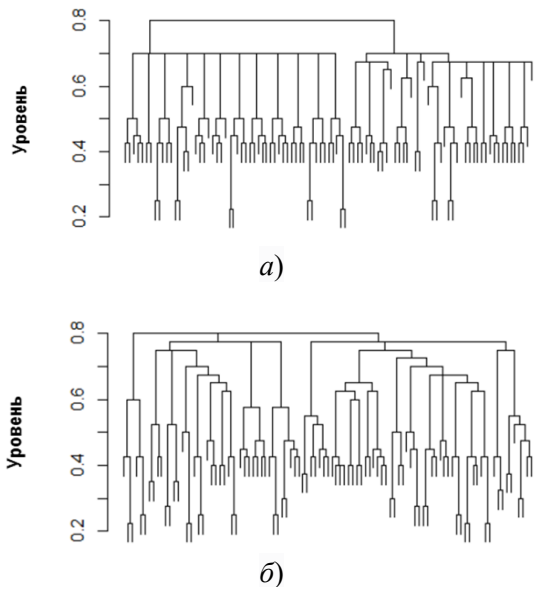
В качестве практического примера рассматривается возможность кластеризации временной серии изображений одной и той же местности, содержащей признаки цифровых многоспектральных изображений и состоящей из 99 изображений спутника Landsat 8. Аппаратура спутника имеет 9 спектральных каналов. Из открытых источников было взято 11 наборов

изображений для части территории Владимирской области в различное время года. В качестве признаков изображений используется: дата проведения съёмки; номер спектрального канала; информационная энтропия Шеннона [3], которая рассчитывается на основе гистограммы яркостей пикселей; фрактальная размерность Минковского, вычисленная на основе клеточного алгоритма [12]. Для программной реализации кластерных методов использовались пакетные функции на языке R в среде разработки RStudio: daisy(), diana(), clusplot() из библиотеки кластеризации. Данная выборка состоит из признаков, характеризующих изображения. Каждое из изображений является терминальным узлом дендрограммы («листья» дерева), а связи между этими вершинами характеризуются мерами схожести каждой пары изображений.

На рис. 1 наблюдаются терминальные узлы дендрограммы, которые перестают ветвиться на уровнях построения дерева больше или равных 0,4–0,5. Это и есть те самые изображения, на которых присутствуют аномалии.

Рис. 2 показывает, что для случая дивизионной кластеризации сложно определить оптимальное количество кластеров по показателю суммы квадратов расстояний внутри кластера. Несмотря на то, что в случае агломеративной кластеризации и наблюдается небольшой изгиб графика для 11 кластеров, необходимо дополнительно оценить разделимость между кластерами и вычислить ширину силуэта. Так, для дивизионной кластеризации ширина силуэта (рис. 3 *а*) показывает оптимальное количество кластеров равное либо 5, либо 11–13 (см. локальные максимумы на графике), а в случае с агломеративной кластеризацией целесообразно делить множество либо на 7, либо на 11 групп.

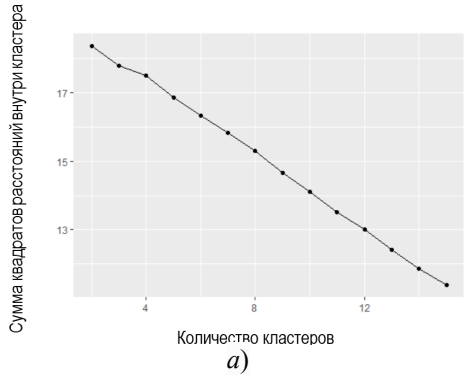
Ширина силуэта на рис. 3 *б* также имеет локальный максимум при 11 кластерах. Однако



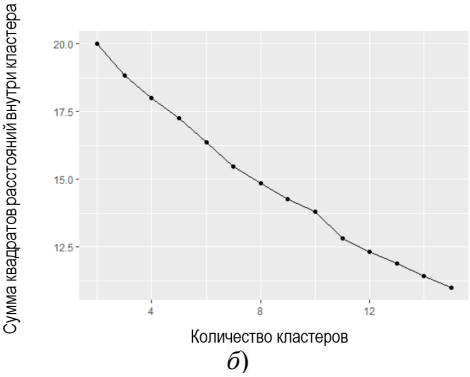
а)

б)

Рис. 1. Результат дивизионной (а) и агломеративной (б) кластеризации



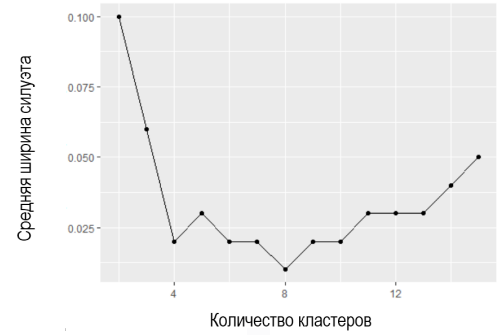
а)



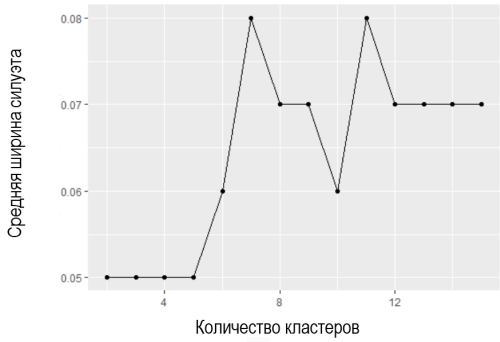
б)

Рис. 2. Сумма квадратов метрических расстояний между объектами внутри кластера для дивизионной (а) и агломеративной (б) кластеризации

меньшее количество кластеров позволит лучше интерпретировать результаты, поэтому следует остановиться на меньшем количестве в 7 кластеров ($k = 7$) (рис. 4).



а)



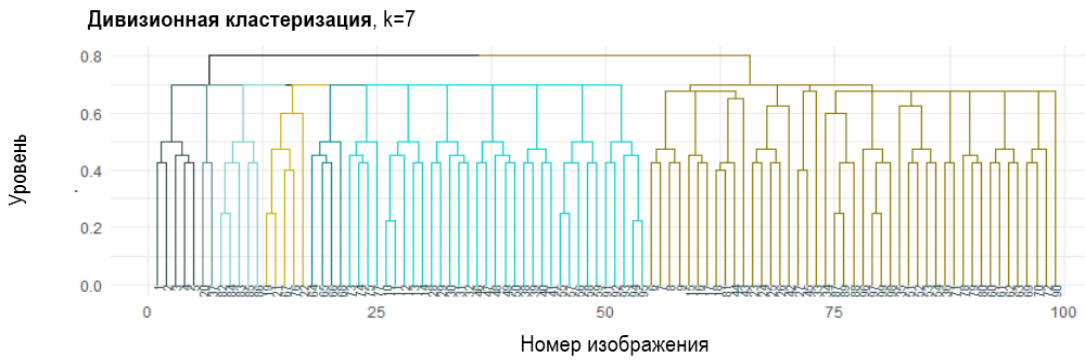
б)

Рис. 3. Средняя ширина силуэта для дивизионной (а) и агломеративной (б) кластеризации

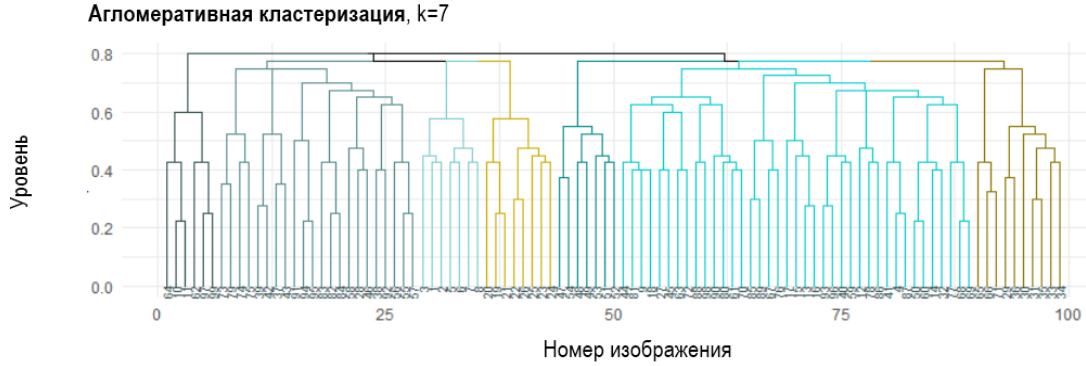
Дивизионная кластеризация выделяет в качестве аномалий изображения с номерами (22, 33, 34, 35, 43 и 90), кроме того есть один малочисленный кластер (изображения №20 и 37).В случае агломеративной кластеризации результат становится иным: аномальными наблюдениями считаются изображения №30 и 91, но при этом следует отметить, что и в случае дивизионной кластеризации изображения с номерами 30 и 91 не имеют ветвей на уровне иерархии меньшем 0,4, что позволяет сделать вывод о том, что данные изображения содержат наблюдения в целом не свойственные всему набору признаков. Детальное изучение содержимого этих изображений позволит более качественно оценить наличие в них аномальных объектов, а также установить их причины.

Заключение

Описанная методика дает возможность осуществлять поиск аномальных наблюдений на



а)



б)

Рис. 4. Результат дивизионной (а) и агломеративной (б) кластеризации

цифровых многоспектральных снимках, что позволяет повысить эффективность анализа данных в системах дистанционного зондирования, в том числе — для решения задач экологического мониторинга, таких как выявление очагов лесных пожаров, источников загрязнений подстилающей поверхности и т.п. Анализ дендрограммы позволяет определить наиболее крупные группы изображений и построить карту признаков наблюдаемых объектов. При этом методика имеет ряд особенностей:

1. Помехи, шумы, ошибочные наблюдения — это ветви дендрограммы, которые перестают ветвиться на уровне построения более 0,4–0,5 при дивизионной и агломеративной кластеризации.

2. Аномальными можно считать 3 изображения, которые идентифицируются при агломеративной и дивизионной кластеризации. На этих изображениях, в отличие от остальной выборки, присутствуют аномальные объекты, это могут быть лесные пожары, последствия экологической катастрофы и т.д.

3. При однородной тестовой выборке более важной характеристикой качества разбиения является ширина силуэта.

Литература

1. *Никитин О.Р., Кисляков А.Н.* Анализ информационного содержания цифровых многоспектральных изображений земной поверхности // Радиотехнические и телекоммуникационные системы. 2016. № 2 (22). С. 64–69.

2. *Бухтояров О.И., Несговорова Н.П., Савельев В.Г., Иванцова Г.В., Богданова Е.П.* Методы экологического мониторинга качества сред жизни и оценки их экологической безопасности: учебное пособие. Курган: Изд-во Курганского гос. ун-та. 2015. 239 с.

3. *Никитин О.Р. Кисляков А.Н.* Фрактальный анализ информационного содержания цифровых многоспектральных изображений в задачах экологического мониторинга // Теоретическая и прикладная экология. 2019. №2. С 32–38.

4. *Sotoca J.M., Pla F., Klaren A.C.* Unsupervised band selection for multispectral images using information theory // Proceedings of the 17th International Conference on Pattern Recognition (ICPR). 2004. No. 3. Pp. 510–513.

5. *Якимов В.Н., Шурганова Г.В., Черепенников В.В., Кудрин И.А., Ильин М.Ю.* Методы сравнительной оценки результатов кластерного анализа структуры гидробиоценозов (на примере зоопланктона реки Линда Нижегородской области) // Биология внутренних вод. 2016. № 2. С. 94–103.

6. *James R.G., Witten D., Hastie T., Tibshirani R.* An Introduction to Statistical Learning with Applications. Berlin: Springer, 2013. 440 p.

7. *Pan F., Chen D., Lu L.* Improved PSO based clustering fusion algorithm for multimedia image data projection. Multimed Tools Appl. 2019.

8. *Murtagh F., Contreras P.* Methods of Hierarchical Clustering // Computing Research Repository. CORR. 2011. 21 p.

9. *Kassambara A.* Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis) (Volume 1) 1st Edition / Publisher: CreateSpace Independent Publishing Platform. 2017. 188 p.

10. *Franklin J.* The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 2003, 27. Pp. 83–85.

11. *Tripathi S., Bhardwaj A., Eswaran P.* Approaches to Clustering in Customer Segmentation // International Journal of Engineering &Technology. 2018. № 7(3.12). Pp. 802–807.

12. *Шуплецов Ю.В., Ампилова Н.Б.* Алгоритм вычисления размерности Минковского для полутоновых изображений // Известия Российского государственного педагогического университета им. Герцена. 2014. № 165. С. 99–106.

Поступила 3 сентября 2021 г.

English

HIERARCHICAL CLUSTERING METHOD OF DIGITAL MULTI-SPECTRAL IMAGES FOR ENVIRONMENTAL MONITORING

Oleg Rafailovich Nikitin — Grand Dr. in Engineering, Professor, the Head of Department of Radio Engineering and Radio Systems, Federal State Budgetary Educational Institution of Higher Education “Vladimir State University named after A.G. and N.G. Stoletovs”[1].
E-mail: olnikitin@mail.ru

Aleksey Nikolaevich Kislyakov — PhD, associate Professor, Department of information technology, Russian Academy of National Economy and Public Administration under the President of the Russian Federation (Vladimir branch of RANEPA)[2].
E-mail: ankislyakov@mail.ru

[1]*Address:* 600000, Russian Federation, Vladimir, Gorky St., 87.

[2]*Address:* 600017, Russian Federation, Vladimir, Gorky St., 59a.

Abstract: The article is concerned with the study of possibility to identify anomalous surveillance in multispectral images through distinctive-feature clustering, as well as by using the developed method for environmental monitoring. This article's topic is significant since there are many problems associated with identifying anomalous surveillance in time series of images bound mainly to the need for detailed study of each picture. The paper proposes a method for estimating anomalous surveillance using summarized characteristics of satellite images, such as brightness entropy and Minkowski fractal dimension and others. Hierarchical clustering is executed based on the above characteristics and using Gower's distance breaking down as metrics. That enables to identify pictures with anomalous objects based on exponent for sum of squares of metric distances between image attributes inside the cluster and average silhouette width, which make it possible to select optimal number of clusters and evaluate the quality of breaking down results. Anomalies' detection is done by analyzing hierarchical clustering results and detecting dendrogram branches located at initial levels of dendrogram scheme and without branching. The authors did work on selecting distance metric when performing clustering and on studying characteristics of breaking down in case of partitional and agglomerative clustering. The examples of using methods as well as, distinctive features, advantages and disadvantages are presented. Computer mathematical modeling is also done using Daisy (), Diana(), clusplot() packages, R language and RStudio development environment.

Keywords: information entropy, fractal dimension, cluster analysis, graph theory, multi-spectral images, remote sensing.

References

1. *Nikitin O.R., Kislyakov A.N.* Analysis of the information content of the digital multispectral images of the earth surface. Radiotekhnicheskie i telekommunikacionnye sistemy. 2016. No. 2 (22). Pp. 64–69.

2. *Bukhtoyarov O.I., Nesgovorova N.P., Savelyev V.G., Ivantsova G.V., Bogdanova E.P.* Methods of ecological monitoring of the quality of living environments and assessment of their environmental safety: a textbook. Kurgan: Kurgan State University. 2015. 239 p.

3. *Nikitin O.R., Kislyakov A.N.* Fractal analysis of the information content of digital multispectral images in environmental monitoring tasks. Teoreticheskaya i prikladnaya ekologiya. 2019. No. 2. Pp. 32–38. doi:10.25750/1995-4301-2019-2-032-038

4. *Sotoca J.M., Pla F., Klaren A.C.* Unsupervised band selection for multispectral images using infor-mation theory // Proceedings of the 17th International Conference on Pattern Recognition (ICPR). 2004. No. 3. P. 510–513. doi: 10.1109/ICPR.2004.1334578

5. *Yakimov V.N., Shurganova G.V., Cherepennikov V.V., Kudrin I.A., Ilyin M.Yu.* Methods of compara-tive evaluation of the results of cluster analysis of the structure of hydrobiocenoses (on the example of zooplankton of the Linda river of Nizhny Novgorod region). Biologiya vnutrennih vod. Ed.: Russian Academy of Sciences. 2016. No. 2. Pp. 94–103.

6. *James R.G., Witten D., Hastie T., Tibshirani R.* An Introduction to Statistical Learning with Applications. Berlin: Springer. 2013. 440 p.

7. *Pan F., Chen D., Lu L.* Improved PSO based clus-tering fusion algorithm for multimedia image data projection. Multimed Tools Appl. 2019. doi: 10.1007/s11042-019-08015-z

8. *Murtagh F., Contreras P.* Methods of Hierarchical Clustering. Computing Research Repository. CORR. 2011. 21 p.

9. *Kassambara A.* Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis) (V. 1) 1st Edition. Publisher: CreateSpace Independent Publishing Platform. 2017. 188 p.

10. *Franklin J.* The elements of statistical learning: data mining, inference and prediction. The Mathe-matical Intelligencer. 2003. No. 27. Pp. 83–85. doi: 10.1007/BF02985802

11. *Tripathi S., Bhardwaj A., Eswaran P.* Approaches to clustering in customer segmentation. International Journal of Engineering &Technology. 2018. No.7 (3.12). Pp. 802–807. doi: 10.14419/ijet.v7i3.12.16505

12. *Shupletsov Yu.V., Ampilova N.* The algorithm of calculating Minkovsky Dimension for Grayscale images. Izvestiya Rossijskogo gosudarstvennogo pedagogicheskogo universiteta im. Gercena. 2014. No. 165. Pp. 99–106.