Journal of Siberian Federal University. Engineering & Technologies, 2017, 10(6), 714-722
~ ~ ~

УДК 528.852

Histogram Hierarchical Algorithm and the Reduction of the Dimensionality of the Spectral Features Space

Valerija S. Sidorova*
Institute of Computational Mathematics and Mathematical Geophysics SB RAS
6 Akademika Lavrentieva, Novosibirsk, 630090, Russia

Received 15.12.2016, received in revised form 16.05.2017, accepted 24.07.2017

This paper proposes the algorithm for dimension reduction of data in the process of hierarchical histogram clustering data of remote sensing of the Earth. Application of the algorithm is illustrated to multispectral data. Clustering large amount of data remote sensing is usually carried out in two ways: by K medium (in advance, you must know the number of clusters K and an approximation of the data distribution), and histogram. Here we propose a hierarchical histogram algorithm, which does not require to specify the number of clusters and is quick. This paper considers the issue of reducing the dimension of own space of features, obtained by hierarchical histogram algorithm. Getting clusters of multispectral image, pay attention to the fact that the different clusters corresponding to different objects on Earth can be characterized by different dimensionality of the data, i.e., the set of spectral channels coming from the satellite, it may be unnecessary for a number of objects. Also, the level of detail of clustering can be different in different clusters.

Keywords: remote sensing, clustering, multidimensional histogram, cluster rasilimali, own vectors space.

Citation: Sidorova V.S. Histogram hierarchical algorithm and the reduction of the dimensionality of the spectral features space, J. Sib. Fed. Univ. Eng. technol., 2017, 10(6), 714-722. DOI: 10.17516/1999-494X-2017-10-6-714-722.

* Corresponding author E-mail address: vsidorova@inbox.ru

Гистограммный иерархический алгоритм и понижение размерности пространства спектральных признаков

В.С. Сидорова
Институт вычислительной математики
и математической геофизики СО РАН
Россия, 630090, Новосибирск, пр. Академика Лаврентьева, 6

Предлагается алгоритм понижения размерности данных в процессе иерархической гистограммной кластеризации данных дистанционного зондирования Земли. Иллюстрировано применение алгоритма к многоспектральным данным. Кластеризация большого объема данных ДЗЗ обычно осуществляется двумя способами: по K средним (заранее должно быть известно число кластеров K и приближенное распределение данных) и гистограммными. Здесь предлагается иерархический гистограммный алгоритм, который не требует задания числа кластеров и является быстрым. В работе рассматривается вопрос о сокращении размерности собственного пространства признаков, полученного методом иерархического гистограммного алгоритма. Получая кластеры многоспектрального изображения, обратим внимание на то, что различные кластеры, соответствующие различным объектам на Земле, могут характеризоваться различной размерностью данных, т.е. множество спектральных каналов, поступающих со спутника, может оказаться лишним для ряда объектов. Детальность кластеризации также может оказаться различной в разных кластерах.

Ключевые слова: дистанционное зондирование, кластеризация, многомерная гистограмма, кластерная разделимость, собственное пространство векторов.

Введение

Учитывая иерархичность данных ДЗЗ, предлагаем иерархический гистограммный алгоритм [1]. Многомерная гистограмма рассматривается в нем как плотность вероятности векторов. Для алгоритма не задается число кластеров, но может задаваться детальность кластеризации.

На каждом этапе иерархии внутри каждого кластера применяется быстрый гистограммный алгоритм Нарендры [2], который делит пространство векторов на унимодальные кластеры. Идея построения унимодальных кластеров с помощью графов принадлежит Фукунаге [3]. Именно он предложил определение кластеров как областей, прилегающих к каждому модальному вектору, соответствующему каждому локальному максимуму гистограммы, до границ этого кластера, используя графы, тем самым решая одновременно три задачи: быстрый способ нахождения модальных векторов, границ между кластерами и собственно кластеризацию всех векторов, т.е. отнесение каждого вектора в определенный кластер. Нарендра [2] осуществил алгоритм технически. Он предложил построение многомерной гистограммы только для присутствующих на изображении векторов; для ускорения построения использовал ХЕШ функции векторов, а также предложил использовать Shell-сортировку, вариант её использования изложен также в [4]. Алгоритм является жестким: каждый вектор относится только к одному кластеру. Но кластеры могут по существу иметь множество общих векторов. Чтобы избежать неоднозначности, вектор, который лежит на границе кластеров, относится к кластеру, имеющему больший градиент гистограммы в данной точке пространства спектральных векторов.

Таким образом, границы кластеров находятся автоматически и лежат в долинах многомерной гистограммы, т.е. в областях низкой плотности векторов. Важно, чтобы гистограмма и различные вектора были записаны в виде списка, определенным образом упорядоченного так, чтобы соседние векторы оказывались в непосредственной близости друг от друга, и это обеспечивает автоматизм и быстроту их нахождения. В качестве соседства данного вектора отыскивается система векторов, отстоящих от него не далее чем на единицу по каждому спектральному направлению. В первоначальном виде алгоритм был реализован на машине БЭСМ-6 автором этой статьи [5], затем перенесен на ПК [6].

Иерархиченский кластерный алгоритм

Ценность алгоритма Нарендры в том, что он не требует задания a priory никаких параметров, формы и числа кластеров. Кроме того, он быстрый, т.е. линейно зависит от числа различных векторов. Остается лишь детальность кластеризации. Во времена Нарендры компьютеры обладали небольшой памятью и быстродействием. Поэтому даже гистограмма лишь присутствующих векторов могла не помещаться в оперативную память. Тогда в каждом байте, соответствующем определенному спектральному направлению (каналу) отсекались младшие (шумовые) биты. Тем самым осуществлялось квантование векторного пространства (каждое отсечение вдвое уменьшало число уровней квантования). Детальность кластеризации также, естественно, понижалась. Однако, отсекая разное число битов, мы получаем разные системы векторов и кластеров. С усовершенствованием компьютеров мы можем уже не отсекать биты, но тогда получается очень много векторов и очень много локальных максимумов. Можно ли рассматривать их не все? Помощь приходит, если привлечь разделимость кластеров. Большинство кластерных алгоритмов не гарантируют хорошую разделимость кластеров, а между тем качество кластеризации определяется именно ей [7].

Алгоритм [1] предлагает автоматизировать выбор детальности, основываясь на минимизации разделимости кластеров. В качестве мер разделимости выступают: мера отделимости для отдельного унимодального кластера $m^j(n)$ (1) и мера качества распределения в целом $m(n)$ по $K(n)$ кластерам (2):

$$m^j(n) = \frac{1}{B^j(n) \times H^j(n)} \sum_{i=1}^{B^j(n)} h_i^j(n), \quad (1)$$

$$m(n) = \frac{1}{K(n)} \sum_{j=1}^{K(n)} m^j(n), \quad (2)$$

где $h_i^j(n)$ – значение гистограммы в i-той точке границы кластера j; $B^j(n)$ – число точек границы кластера; $H^j(n)$ – максимальное значение гистограммы; n – число квантовых уровней [8].

Минимумы меры (2) соответствуют лучшим распределениям с наиболее разделенными кластерами в среднем. Всегда $m^j(n) \leq 1$ и $m(n) \leq 1$. Ценность этих мер в том, что они позволяют сравнивать распределения с тесно расположенными унимодальными кластерами, когда на их границах много общих векторов. Эти меры удовлетворяют условиям мер разделимости [7]: значения их убывает с увеличением расстояния между кластерами и ростом компактности кластеров (в смысле близости векторов кластера к модальному вектору). Кроме того, эти меры легко вычисляются, так как сравнивают скалярные значения гистограммы в центре и на

границах кластеров. Границы кластеров легко найти, используя списки соседей векторов, построенных как составная часть алгоритма Нарендры.

Алгоритм [1] представляет данные в их иерархической вложенности, последовательно увеличивая детальность рассмотрения. Сначала определяется минимум меры разделимости (2) и соответствующее число уровней квантования *n*. Затем для каждого полученного кластера, начиная с увеличенного *n*, снова находится минимум и подкластеры, ему соответствующие. Затем оценивается отделимость каждого подкластера. Данные подкластеров, для которых отделимость (1) больше *d,* объединяются для дальнейшего деления. Здесь не происходит остановки, потому что на более детальном уровне (большем этапе иерархии) эти данные могут удовлетворительно разделиться. Задаваемым параметром кластеризации является отделимость каждого полученного кластера *d* (одно для всех). Иерархический процесс деления кластеров заканчивается тогда, когда достигается максимальное значение признака по каждому спектральному каналу (обычно 255), либо по заданному числу иерархических этапов, либо из каких-то других физических соображений. Затем автоматически анализируются полученные результаты и производится возврат к тем детальностям, на которых отделимость кластера была не больше *d.* Кластеры, не удовлетворяющие этому условию ни на каком этапе, объединяются в один ложный кластер. Заметим, что для рассматриваемых данных тенденция такова, что с ростом детальности средняя разделимость кластеров уменьшается, соответственно (2) увеличивается. После кластеризации осуществляется глобальная сегментация изображения. В результате работы алгоритма [1] получаем существенно меньше кластеров, чем прямым алгоритмом Нарендры за счет того, что выбираем детальность рассмотрения области данных так, чтобы не нарушалось условие отделимости. Отделимость всех полученных унимодальных кластеров меньше заданного *d.* В то же время сохраняются достоинства алгоритма Нарендры: независимость от формы и задания числа кластеров, быстрота.

Выбор размерности данных

В данной статье предлагается учитывать не только то, что детальность в разных областях данных различна, но и то, что размерность данных в кластерах различна. Квантование пространства признаков может производиться по разным правилам. У Нарендры оно достигалось отсечением младших битов в каждом спектральном канале. Каждое отсечение уменьшает число уровней квантования вдвое. В работе [9] был предложен другой способ, более плавный, но по-прежнему в каждом спектральном направлении число уровней квантования сохранялось одинаковым. Однако в общем случае данные вытянуты вдоль какого-то направления, и правило квантования, обеспечивающее наименьшую потерю информации, требует различного подхода в различных направлениях, а именно: квантование должно сохранять ячейку квантования в форме гиперкуба (а не гиперпараллелепипеда). Это условие будет выполнено, если

$$\frac{N_{e1}}{S_{e1}} = \frac{N_{e2}}{S_{e2}} = \ldots = \frac{N_{ek}}{S_{ek}} \ , \tag{3}$$

где N_{e1}, N_{e2}, …, N_{ek} – числа уровней квантования вдоль для соответствующих собственных векторов по k – ортонормированным осям собственного пространства, а S^2_{e1}, S^2_{e2}, …, S^2_{ek} – собственные числа.

Зададим максимальное число уровней квантования в собственном пространстве равным N_{em}=255, таково обычное число уровней серого для данных дистанционного зондирования по каждому измерению. Тогда в соответствии с пропорциями (1) может быть найдено число уровней квантования и по другим осям собственного пространства. Для задач кластеризации это число должно быть больше или равно 2, иначе эта компонента одинакова для всех векторов и никакой роли в кластеризации не играет. Таким образом, если отношение $S_{em}/S_{ex} < 2$, то соответствующая ось x ортонормированного собственного пространства может не рассматриваться, и мы получаем сокращение размерности пространства признаков.

Ранее [9] было предложено сокращать размерность перед использованием алгоритма кластеризации. Пример: на рис. 1 представлено изображение (2772*1995) района Улан-Удэ с целью выделения области загрязнения территории. Это семиспектральное изображение Бурятии со спутника «Landsat-8», район Улан-Удэ. Исходный файл предоставлен Сибирским центром ФГПУ НИЦ «ПЛАНЕТА». Построение ковариационной матрицы спектральных данных для всего изображения и ее диагонализация методом Якоби [10] показало, что можно рассматривать три измерения без существенной потери информации (рис. 2). Сокращение размерности приводит к экономии компьютерного времени.

Также алгоритм применялся для картирования загрязнения отходами производств Омской области (рис. 3) по восьми спектральным каналам ИСЗ «Landsat-8” (3161*2590) от 08.02.2014 (разрешение 15 м).

Сокращение размерности производилось до кластеризации, анализ ковариационной матрицы данных позволил уменьшить число спектральных каналов с 8 до 3. Для 10 этапов иерархии получено 27 унимодальных кластеров. Углы розового и фиолетового кластеров на северо-востоке соответствуют загрязнению снега территории Омской области в соответствии

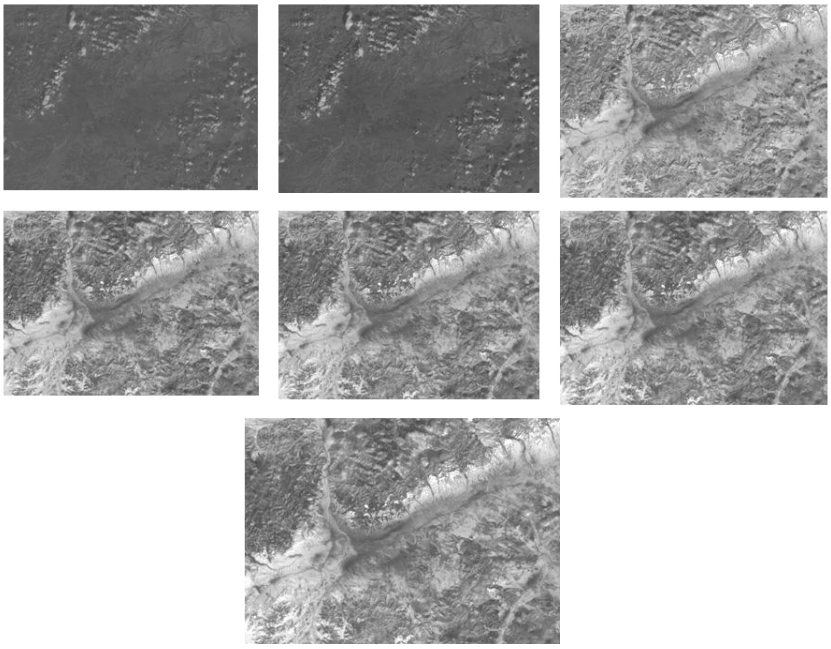

Рис. 1. Семиспектральное изображение со спутника «Landsat-8», район Улан-Удэ

Fig. 1. The semispectral image from the satellite “Landsat-8”, the area of Ulan-Ude

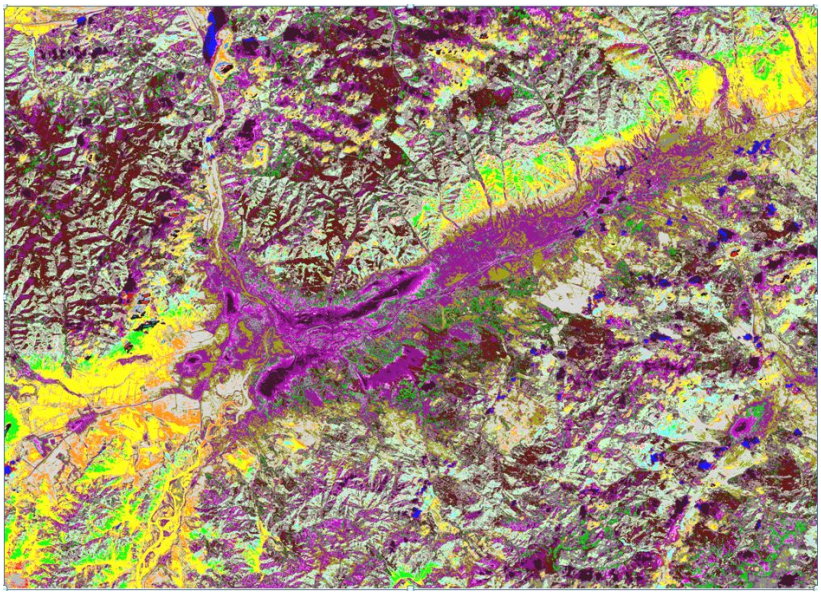

Рис. 2. Кластерная карта, полученная делимым иерархическим гистограммным алгоритмом. 15 этапов иерархии. *d*= 0,015. 54 кластера (включая маленькие вплоть до 1 пикселя). Загрязнение: лиловые и темно-зеленые оттенки

Fig. 2. The cluster map obtained divisible hierarchical histogram algorithm. 15 stages of the hierarchy. d= 0, 015. 54 cluster (including small up to 1 pixel). Pollution: purple and dark green shades

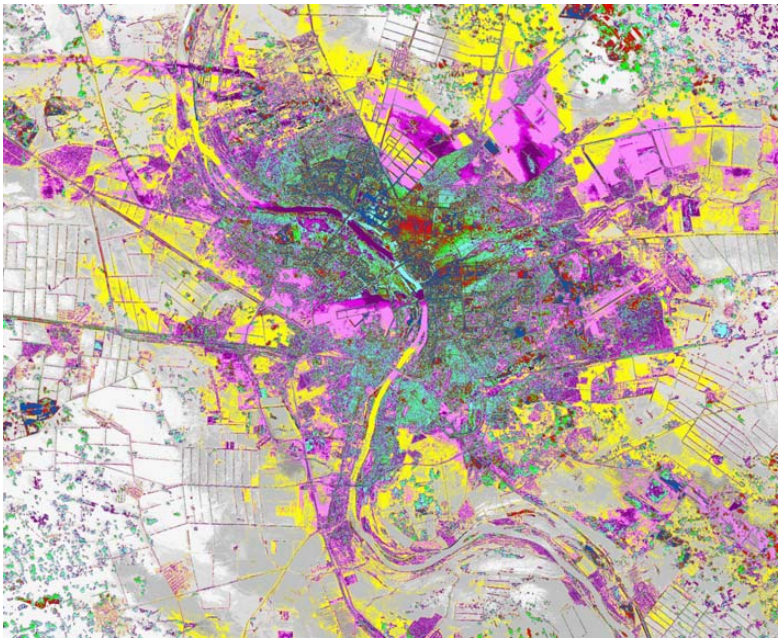

Рис. 3. Кластерная карта Омской области

Fig 3. Cluster map of Omsk region

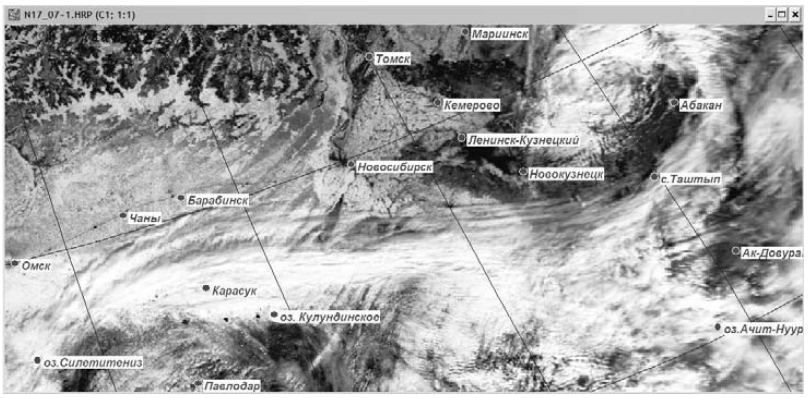

Рис. 4. Исходное изображение в видимой части спектра

Fig. 4. The original image in the visible spectrum

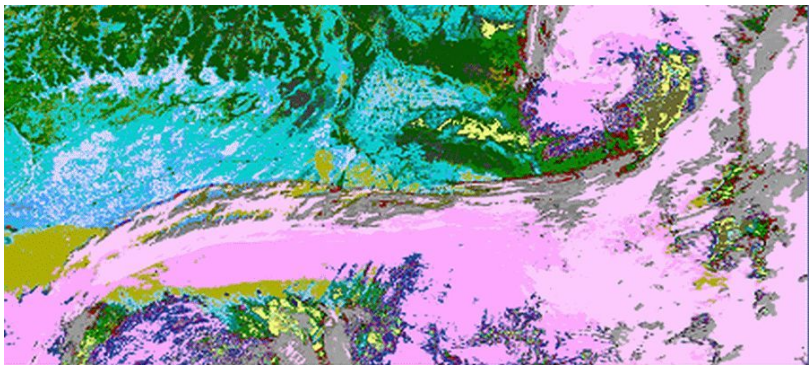

Рис. 5. Кластеризация иерархическим гистограммным алгоритмом с поиском размерности по кластерам; 7 этапов иерархии; задано *d*=0,12; получено 29 кластеров

Fig. 5. The hierarchical clustering of histogram algorithm finding the dimension of the clusters; 7 stages of the hierarchy; set d=0,12; received 29 clusters

с преобладанием юго-западных ветров в области. Предельная отделимость кластеров задана *d*=0,15.

Можно рассматривать сокращение размерности внутри каждого кластера. Ковариационная матрица строится для векторов данного кластера. Переходя от кластера к кластеру, каждый раз устанавливают максимальное значение *S*, а затем с использованием (3) определяют квантование других осей и размерность пространства.

Рассмотрим пример. Анализируется изображение поверхности Земли со спутника NOAA 17 от 7.04.2003, полный кадр (1328x624) пикселей представлен в пяти спектральных каналах (один в видимой части спектра, остальные в инфракрасной), объем около 4 мегабайт. На рис. 4 можно видеть изображение в одном из каналов.

Визуальный анализ изображений в каждом спектральном канале показывает, что снизу-вверх слева-направо на рис. 4 видно формирование вихря (по форме напоминает дракона); в верхней части рис. 4 в основном – тающие снега, тайга Сибири. Кластеры с площадью меньше 100 пикселей отнесены в фоновый. Для большинства подкластеров размерность собственного пространства оказалась равна трем, как и для всех данных в целом, но некоторые подкластеры

(соответствующие полупрозрачным облакам) потребовали пятиспектрального рассмотрения. Таким образом, сокращение размерности произошло более точно, в зависимости от характера области данных. Время вычислений оказалось в три раза меньше, чем для полного пятиспекрального варианта, и составило несколько минут на одноядерном компьютере РК 1.6 ГГц 512 МБ.

Резюме

Иерархический гистограммный алгоритм с заданием отделимости каждого полученного кластера позволяет не только автоматически выбрать детальность представления данных, различную для каждого полученного кластера, но и сократить размерность пространства признаков в зависимости от области рассмотрения. Это приводит к существенному сокращению времени вычислений. Единственный параметр *d* определяет отделимость каждого полученного унимодального кластера. Чем меньше этот параметр, тем лучше отделены кластеры. Но с уменьшением *d* увеличивается доля кластеров фона, т.е. не разделенных для данного *d*, и при некоторых *d* все кластеры сливаются в один фоновый. Доля этого кластера может служить индикатором остановки работы алгоритма.

Работа выполнена частично при финансовой поддержке Российского фонда фундаментальных исследований (проект № 16-07-00066) и Программы I.33П фундаментальных исследований Президиума РАН (проект № 0315-2015-0012).

Список литературы

[1] Sidorova V.S. Detecting Clusters of Specified Separability for Multispectral Data on Various Hierarchical Levels. *Pattern Recognition and Image Analysis,* 2014, 24(1), 151-155.

[2] Narendra P.M., Goldberg M. A non-parametric clustering scheme for LANDSAT. *Pattern Recognition,* 1977, 9, 207-215.

[3] Koontz W., Narendra P.M. and Fukunaga K. A graph theoretic approach to non-parametric cluster analysis, *IEEE Trans. Comput*. C-23, 936-944 (1967).

[4] Sidorova V. S. Separating of the Multivariate Histogram on the Unimodal Clusters. *Proceedings of the Second IASTED International Conference "Automation Control and Information Technology"*. 2005, 267-274.

[5] Сидорова В.С. Кластеризация многоспектральных изображений с помощью анализа многомерной гистограммы. *Математические и технические проблемы обработки изображений. СО АН СССР*, 1986, 52-57 [Sidorova V.S. Clustering of multispectral images using the analysis of multidimensional histograms. *"A collection of Mathematical and technical problems of image processing. OF SB AS USSR",* 1986, 52-57. (in Russian)]

[6] Сидорова В.С. Классификация многоспектральных космических изображений поверхности Земли с помощью разделения многомерной гистограммы по унимодальным кластерам. *Вестник КазНУ., сер. географическая,* 2004, 19(2), 206-210 [Sidorova V.S. Classification of multispectral space images of the Earth's surface by splitting multidimensional histograms for unimodal clusters. *The Bulletin Of KazNU., ser. geographic,* 2004, 19(2), 206-210. (in Russian)]

[7] Сидорова В.С. Оценка качества классификации многоспектральных изображений гистограммным методом. *Автометрия,* 2007, 43(1), 37-43. [Sidorova V.S. Quality assessment of classification of multispectral images histogram method. Quality assessment of classification of multispectral images histogram method. Autometry Quality assessment of classification of multispectral images histogram method. *Autometry,* 2007, 43(1), 37-43. (in Russian)]

[8] Halkidi M., Batistakis Y. and Vazirgiannis M. On clustering validation techniques. *Journal of Intelligent Information Systems,* 2001, 17(2-3), 107-131.

[9] Сидорова В.С. Гистограммная кластеризация данных дистанционного зондирования Земли. *Материалы II международной конференции "Региональные проблемы дистанционного зондирования Земли"*, 2015, 213-218 [Sidorova V.S. Histogram clustering data of remote sensing of the Earth. *Proceedings of the II international conference*, 2015, 213-218 (in Russian)]

[10] Калиткин Н.Н. *Численные методы*. Москва, Наука. 1978, 512 с. [Kalitkin N.N. *Numerical methods*. Moscow, Science, 1978, 512 p. (in Russian)]