приятия. Аналитик (ЛФР) совместно с экспертами для каждой функциональной зоны осуществляет SWOT-анализ и определяет семейство КФН. Используя процедуру метода анализа иерархий он совместно с экспертами определяет глобальные приоритеты КФН каждой функциональной зоны по критерию «Неудачи» и формирует семейство главных КФН.

Используя иерархическую модель (рис. 3), они определяют по критерию «Неудачи» глобальные приоритеты функциональных зон и ГКФН, строят множества K_0, Φ_0 и формируют изображение текущей ПС, позволяющее устанавливать прецеденты ПС в прошлом, формировать для ЛПР диагноз и варианты управленческих решений по преодолению текущей ПС.

5. Выводы

Создаваемая ИТ диагностического анализа текущего состояния развивающегося предприятия позволяет оценивать в текущий момент времени его стратегический потенциал и стратегические условия развития, выявлять и распознавать возникающие ПС, формировать эффективные управленческие решения по их преодолению.

Литература

1. Саати, Т. Принятие решений. [Текст] / Т. Саати // Метод анализа иерархий: пер. с англ. – М.: Радио и связь. 1993. – 320 с.
2. А. В. Шукалович. Концептуальные основы информационно-аналитической поддержки диагностики текущей деятельности предприятия. [Текст] / В. Л. Лисицкий // Восточно-европейский журнал передовых технологий. - 2007.-№ 3/5 (27).-С.31-34.

Розглянуто питання застосування алгоритмів неконтрольованої класифікації ISODATA та k-Means для обробки даних дистанційного зондування

Ключові слова: алгоритм, кластеризація, методи кластеризації, супутник

Рассмотрены вопросы применения алгоритмов неконтролируемой классификации ISODATA и k-Means для обработки данных дистанционного зондирования

Ключевые слова: алгоритм, кластеризация, методы кластеризации, спутник

The questions of application of algorithms of out-of-control classification of ISODATA and k-Means are considered for processing of data of the remote sensing

Keywords: algorithm, clusterization, methods of clusterization, space satellite

УДК 528:061.3

КЛАССИФИКАЦИЯ КОСМИЧЕСКИХ СНИМКОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

Ф.Т. Шумаков
Старший преподаватель*
E-mail: shumakov@ksame.kharkov.ua
В.А. Толстохатько
Кандидат технических наук, профессор*
Контактный тел. (057) 707-31-04
E-mail: tolstochatko@rambler.ru
А.Ю. Малец*
*Кафедра геоинформационных систем и геодезии
Харьковская национальная академия городского хозяйства
ул. Революции, 12, г. Харьков, Украина, 61002

Введение

Методы кластерного анализа широко используются в процессе цифровой автоматизированной обработки и классификации космических снимков, полученных со спутников в процессе дистанционного зондирования Земли. Классификация заключается в том, чтобы на основе спектральной информации из различных диапазонов проанализировать каждый пиксель изображения и отнести его к тому или иному классу объектов (пиксель – это наименьший и разрешимый элемент земной поверхности на космическом

снимке) [1]. Этот тип классификации называют также *распознаванием спектральных образов* [1]. В результате классификации на снимке выделяются контуры с неконтрастной структурой, например, растительность, водоемы и другие объекты. Результирующее изображение можно рассматривать как тематическую карту местности.

В процессе классификации различают *информационные* и *спектральные* классы. Информационные классы – это объекты, которые необходимо распознать на снимке: виды растительности, открытые почвы, геологические структуры, типы горных пород, водоемы и др. Спектральный класс – это группа пикселей обладающих приблизительно одинаковой яркостью в некотором спектральном диапазоне. Одна из основных целей классификации состоит в том, чтобы совместить спектральные классы с информационными классами.

Применяются два типа классификации [1]:

• контролируемая классификация (классификация с обучением);

• классификация без обучения (неконтролируемая классификация).

Контролируемая классификация основана на учете априорной информации о типах объектов и эталонных значениях спектральных характеристик этих объектов. В процессе классификации производится сравнение значения яркости текущего пикселя с эталонными признаками. По результатам сравнения пиксель относится к наиболее подходящему классу объектов. Обычно контролируемая классификация применяется, когда классы хорошо различаются на снимке и их число варьируется от 25 и выше [1].

Неконтролируемая классификация применяется при отсутствии априорной информации об объекте съемки. Классификация выполняются автоматически с использованием различных алгоритмов кластеризации. Кластеризация – это разбиение определенного множества объектов на не пересекающиеся подмножества (кластеры), чтобы каждый кластер содержал похожие объекты, а объекты разных кластеров различались между собой. К ним относятся алгоритмы кластеризации ISODATA и k-means [2]. Эти алгоритмы нашли широкое применение на практике благодаря включению их в специализированные программные продукты ERDAS IMAGINE и ENVI.

В статье рассматриваются особенности алгоритмов ISODATA и k-means, а также приводятся результаты классификации снимков с использованием программных продуктов ERDAS Imagine 2011 и ENVI 4.5.

Принцип классификации с использованием алгоритма ISODATA

Алгоритм ISODATA (Iterative Self-Organizing Data Analysis Techniques) реализует итеративный самоорганизующий метод анализа данных [2].

Данный метод кластеризации использует спектральные расстояния и производит классификацию пикселей в ходе нескольких итераций. На каждой итерации переопределяются критерии созданных классов, после чего классификация проводится повторно. При этом происходит постепенное слияние образцов, созданных на основе спектральных расстояний.

В алгоритме используются следующие исходные данные:

• $X = \{x_1, x_2, ..., x_N\}$ – набор данных, включающий спектральные характеристики N пикселей снимка;

• m – необходимое число классов (кластеров);

• Q_N – порог сходимости (относительное количество пикселей, которые не изменяют своей принадлежности к классу при переходе к следующей итерации);

• Q_s – параметр, характеризующий допустимое среднеквадратическое отклонение;

• Q_c –параметр компактности кластеров (определяет условие объединения кластеров);

• L – минимальное количество пикселей в кластере;

• M – допустимое число итераций.

Кластеризация состоит в разбиении набора X на K непересекающихся подмножеств $X_1, X_2, ..., X_m$ – кластеров. Все точки одного кластера должны состоять из «похожих» элементов, а точки разных кластеров существенно отличаться, т.е. $X_1 \cup X_2 ... \cup X_m = X$, $X_i \cap X_j = \varnothing$ для всех $i \neq j$.

Алгоритм ISODATA относится к классу эвристических алгоритмов. В него включены процедуры удаления, объединения и разделения кластеров. Каждая процедура выполняется при соблюдении некоторых условий. Допускается попарное объединение кластеров и разделение одного кластера на два кластера.

Процедура удаления кластеров выполняется, если число $|X_i|$ элементов в i-м кластере меньше заданного, т.е. при $|X_i| < L$. Элементы этого кластера распределяются по другим кластерам, а его центр ci удаляется из списка центров кластеров.

Процедура разделения кластеров выполняется, если разброс элементов от центра кластера достаточно большой, т.е. дисперсия i -го кластера D_i больше Q_s. В данном случае i -й кластер разделяется на два кластера. Для разделения кластера вычисляются покомпонентные дисперсии:

$$D_{ik} = \frac{1}{|X_i|} \sum_{x_j \in X_i} \| x_{jk} - c_{jk} \|^2, \; k = 1,...,n. \qquad (1)$$

где n – количество пикселей в i-м кластере.

Выбирается та l - я компонента, для которой $D_{il} > D_{is}$ для всех $l \neq s$ и осуществляется разделение i -го кластера по l -й компоненте. При этом пересчитываются новые центры кластеров **c**' и **c**''.

Слияние кластеров. Если расстояние между двумя центрами кластеров достаточно мало, то эти кластеры объединяются в один кластер. Для реализации этой процедуры вычисляется расстояние между двумя центрами кластеров:

$$l_{ij} = \| c_i - c_j \|, \text{ для всех } i \neq j. \qquad (2)$$

Если $l_{ij} < Q_c$, то кластеры X_i и X_j следует объединить. Новый центр кластера вычисляется по формуле

$$c = \frac{c_i |X_i| + c_j |X_j|}{|X_i| + |X_j|}. \qquad (3)$$

Классификация снимка выполняется в таком порядке.

В течение первой итерации кластеризации пространство признаков разбивается на области в виде вектора, центром каждой из которых являются средние значения яркости кластеров (рис. 1, а). Вектор

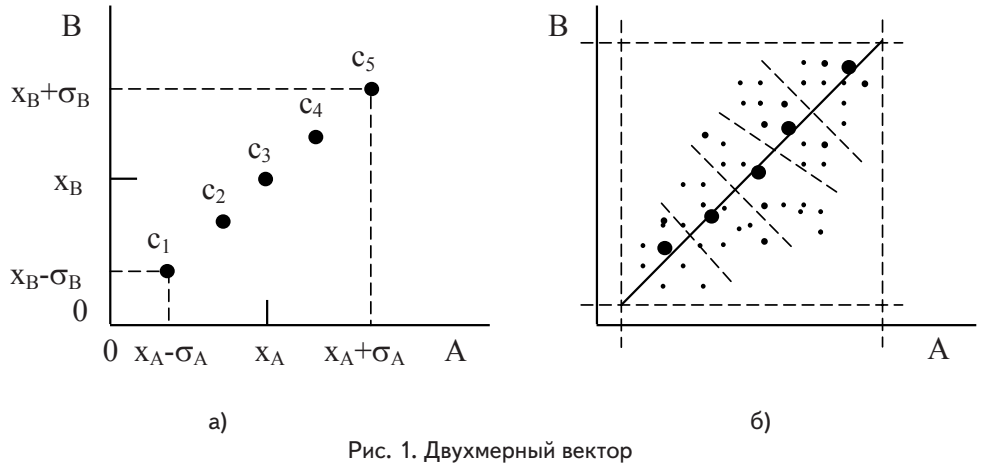
задается крайними точками со спектральными координатами по каналам A и B.

Первоначально, средние значения кластеров равномерно распределяются в пространстве признаков в виде вектора, задаваемого крайними точками со спектральными координатами по каналам A и B, и определяются центры кластеров $\mathbf{c}_1$, $\mathbf{c}_2$,...,c_m). На рис. 1, а показано распределения центров кластеров c_1, c_2, c_3, c_4, c_5 двухмерного вектора спектральных признаков.

На второй и следующей итерациях выполняется анализ пикселей последовательно от левого верхнего угла снимка к нижнему правому. Вычисляются спектральные расстояния между пикселями и средними значениями кластеров. Пикселя назначаются в те кластеры, где это расстояние минимально.

Статистики кластеров (например, средние значения яркости каждого кластера) рассчитываются с учетом нового спектрального положения пикселей. Эти средние значения используются для переопределения кластеров на очередном шаге.

В конце каждой итерации вычисляется процент пикселей, приписывание которых к определенному кластеру не изменилось по сравнению с предыдущим шагом. Когда этот показатель достигнет величины Q_N (порог сходимости), выполнение программы прекращается.



а) б)

Рис. 1. Двухмерный вектор

Возможна ситуация, когда процент пикселей, не сменивших кластер «приписки», никогда не достигнет порога Q_N. В таком случае выполняется максимальное количество итераций m и вычисления также прекращаются.

Принцип классификации с использованием алгоритма k-means

Алгоритм кластеризации k-means (алгоритм k-внутригрупповых средних) основан на минимизации функционала Q суммарной выборочной дисперсии, который характеризует разброс элементов относительно центров кластеров:

$$Q = \sum_i |X_i| \sum_{x \in X_i} d(x, c_i) \to \min,$$

где

$$c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$$ – центр кластера X_i.

Алгоритм выполняется методом итераций. На каждой итерации находятся центры кластеров, а также производится разбиение выборки на кластеры. Вычисления продолжаются до тех пор, пока функционал Q не перестанет уменьшаться.

Порядок выполнения алгоритма:

1. Выделяются начальные центры кластеров $c_1^{(0)}...c_m^{(0)}$ и полагается k=0.

2. Вся выборка разбивается на m кластеров по методу ближайшего соседа. Получаются некоторые кластеры $X_1^{(k)},...,X_m^{(k)}$.

3. Рассчитываются новые центры кластеров по формуле

$$c_i^{(k+1)} = \frac{1}{|X^{(k)}|} \sum_{x \in X_i^{(k)}} x .$$

4. Проверяется выполнение условия завершения вычислений: $c_i^{(k+1)} = c_i^{(k)}$ для всех k =1,...,m. Если условие не выполняется, то осуществляется переход к пункту 2.

Алгоритм k-means осуществляет локальную, но не глобальную минимизацию функционала Q [2]. Поэтому гарантии «хорошей» кластеризации этот алгоритм не дает.

Как показали результаты экспериментов, качество классификации алгоритма k-means гораздо ниже, чем у алгоритма ISODATA. Поэтому далее приведем результаты применения алгоритма ISODATA.

Классификация снимков с использованием алгоритма ISODATA

При проведении исследований использовался синтезированный снимок территории Харьковской области с высоким разрешением (рис. 2). Снимок получен со спутника SPOT-5, который предназначен для получения цифровых изображений земной поверхности с пространственным разрешением 2,5 м в панхроматическом режиме и 10 м в мультиспектральном режиме [1].

Снимок имеет координатную привязку и предоставлен по гранту компанией ESRI в рамках программы по изучению изменений климата. Этот и другие снимки использовались при изучении природы и антропогенных ландшафтов бассейна Северского Донца.

На снимке синтезированы «искусственные цвета». Растительность отображается в оттенках красного, городская застройка – зелено-голубая, а цвет почвы варьируется в диапазоне коричневого цвета. Хвойные леса выглядят более темно-красными или даже коричневыми по сравнению с лиственными. Эта комбинация очень популярна и используется, главным образом, для изучения состояния растительного покрова, мониторинга дренажа и почвенной мозаики,

а также для изучения агрокультур. В целом, насыщенные оттенки красного являются индикаторами здоровой растительности.

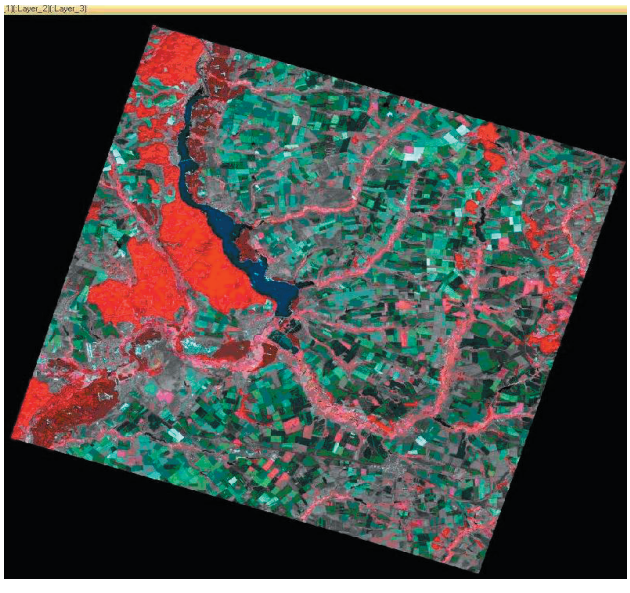


Рис. 2. Снимок территории Харьковской области

Растровые данные могут быть организованы в виде нескольких каналов цветовой информации. Каждый канал это подмножество файла данных представляющее отдельную часть электромагнитного спектра отраженного света или теплового излучения (красный, зеленый, синий инфракрасный, тепловой и прочие). Стандартная комбинация каналов 321 – «искусственные цвета». Растительность отображается в оттенках красного, городская застройка – зелено-голубая, а цвет почвы – темно или светло коричневый. Хвойные леса будут выглядеть более темно-красными или даже коричневыми по сравнению с лиственными. Эта комбинация очень популярна и используется, главным образом, для изучения состояния растительного покрова, мониторинга дренажа и почвенной мозаики, а также для изучения агрокультур.

После запуска программы Erdas Imagine 2011 в диалоговых окнах заданы имена файлов исходного изображения и выходного тематического растра, а также следующие значения входных параметров алгоритма:

1). Количество классов – m = 15 (это приблизительное число классов).

2). Максимальное число итераций – M = 25.

3). Множитель стандартного отклонения Q_s – = 2.

4). Порог сходимости Q_N = 0,950. Установка значения 0,950 означает, что если 95% пикселей изображения не изменили принадлежность к классу при переходе к следующей итерации, то процесс классификации завершается.

Кроме того, выбрана цветовая схема для раскраски классов в градациях серого тона, близких к исходному черно-белому изображению. Классификация снимка выполнена за 9 итераций.

В результате кластеризации методом ISODATA сформирован тематический растровый слой и набор

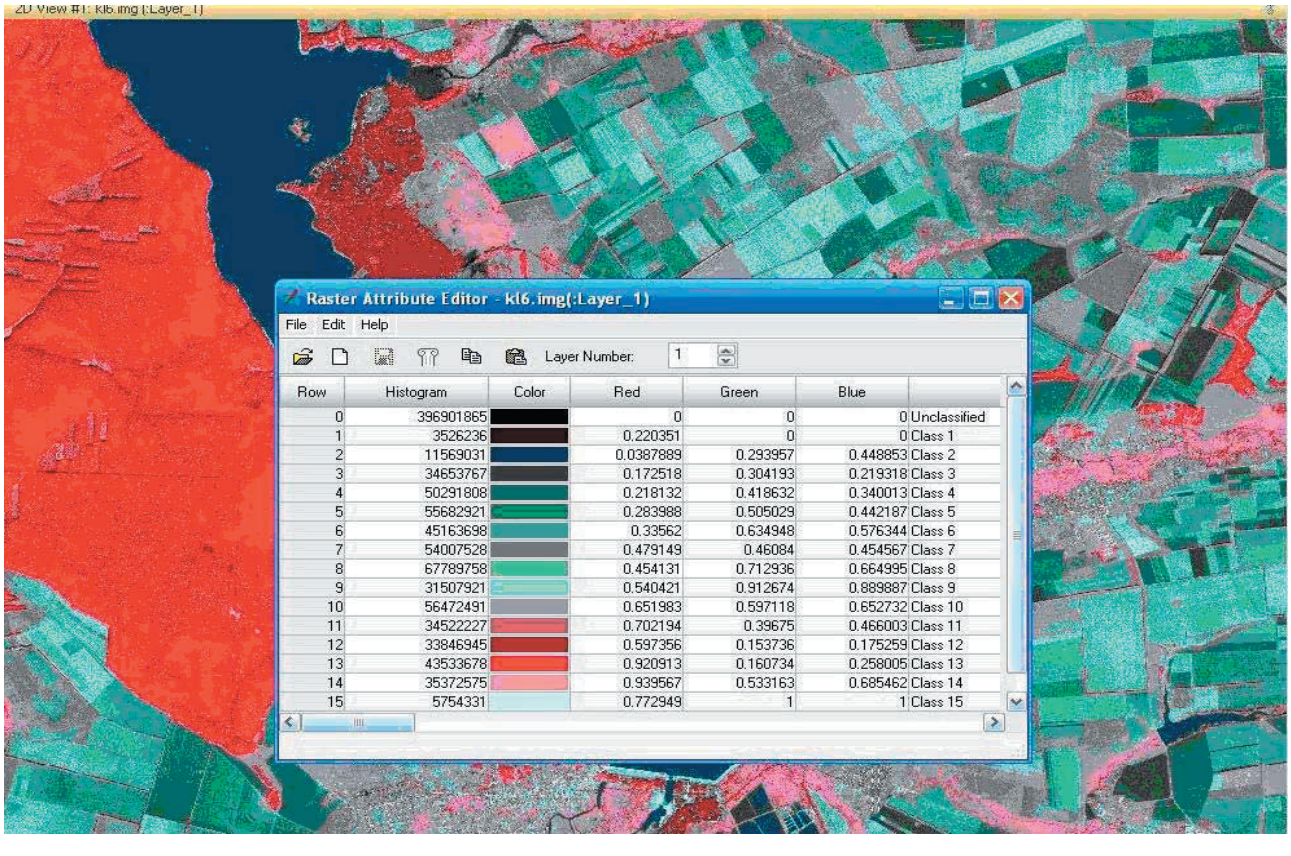


Рис. 3. Района Салтовского водохранилища

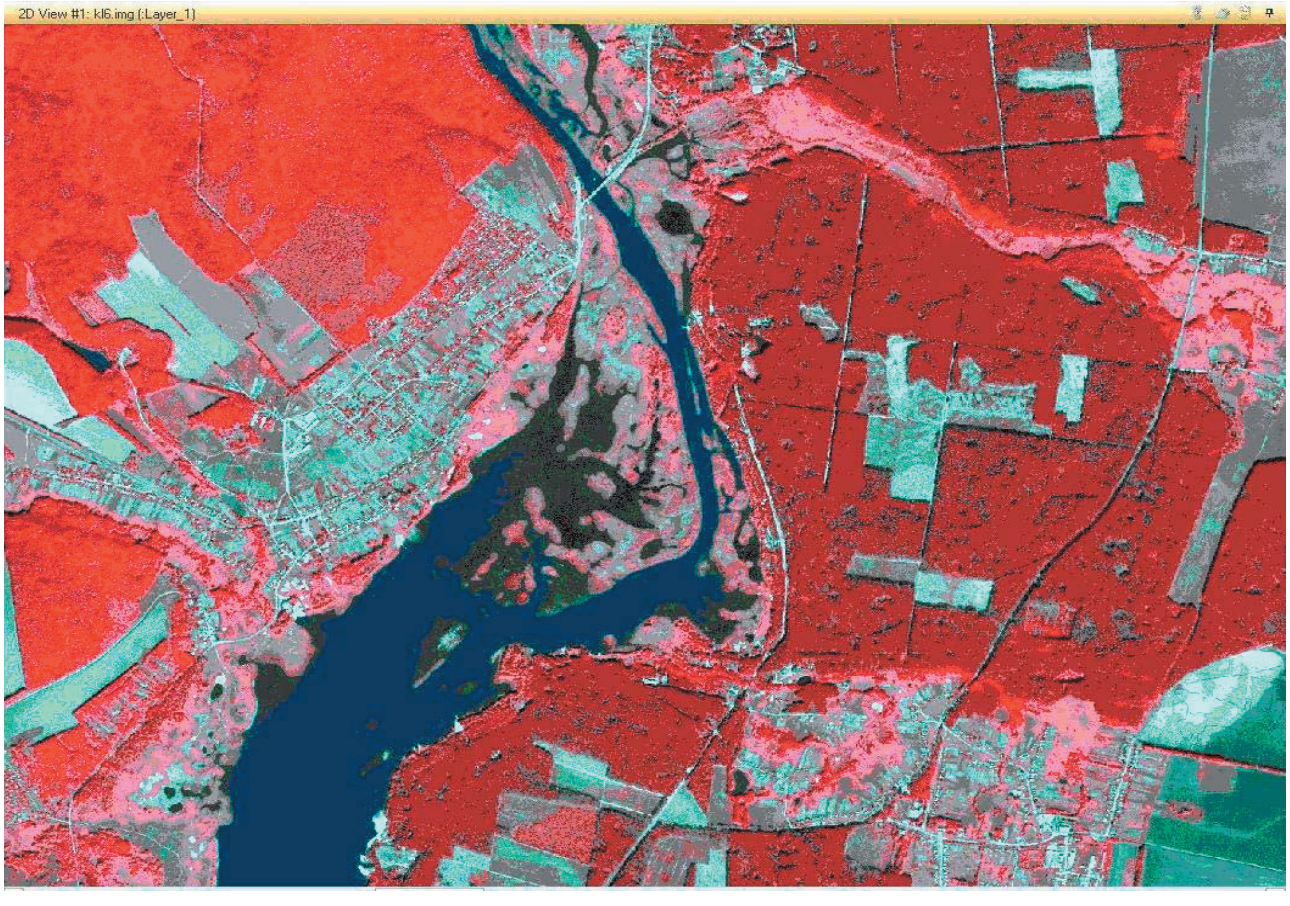


Рис. 4. Район Печенежского водохранилища

статистик для Харьковской области. Для иллюстрации качества классификации покажем фрагменты полученного тематического слоя.

На рис. 3 показан тематический растровый слой и набор статистик для района Салтовского водохранилища, а на рис. 4 – тематический растровый слой для района Печенежского водохранилища.

Анализ приведенных снимков позволяет сделать вывод о хорошем качестве классификации методом ISODATA. На приведенных рисунках хорошо видны водоемы, растительность, лес и многие другие классы объектов. Для облегчения анализа снимков целесообразно изменить цвета и названия классов, а также создать композиционную карту исследуемого района. Благодаря этому цифровые методы обработки многозональных космических изображений позволяют повысить оперативность мониторинга водохранилищ и лесных массивов, прогнозировать урожайность и контролировать состояние посевов, исследовать состояние растительного и почвенного покрова. Особую актуальность приобретет дистанционное зондирование для оперативной оценки площадей и динамики распространения пожаров, наводнений и других стихийных бедствий.

Таким образом, благодаря самоорганизации алгоритма ISODATA обеспечивается автоматическое распределение кластеров и приемлемое качество результирующего снимка.

Выводы

1. Алгоритмы неконтролируемой классификации менее зависимы от человеческого фактора, поскольку не требуют наличия априорной информации о свойствах исследуемой местности.

2. Классы, созданные рассмотренными алгоритмами кластеризации имеют более четкий спектральный состав, чем созданные алгоритмами контролируемой классификации.

Литература

1. Чандра, А.М. Дистанционное зондирование и географические информационные системы [Текст] / С.К. Гош – М: Техносфера, 2008. – 312 с.

2. Лепский, А.Е. Математические методы распознавания образов. Курс лекций [Текст] / А.Г. Броневич – Таганрог: ТТИ ЮФУ, 2009. – 155 с.