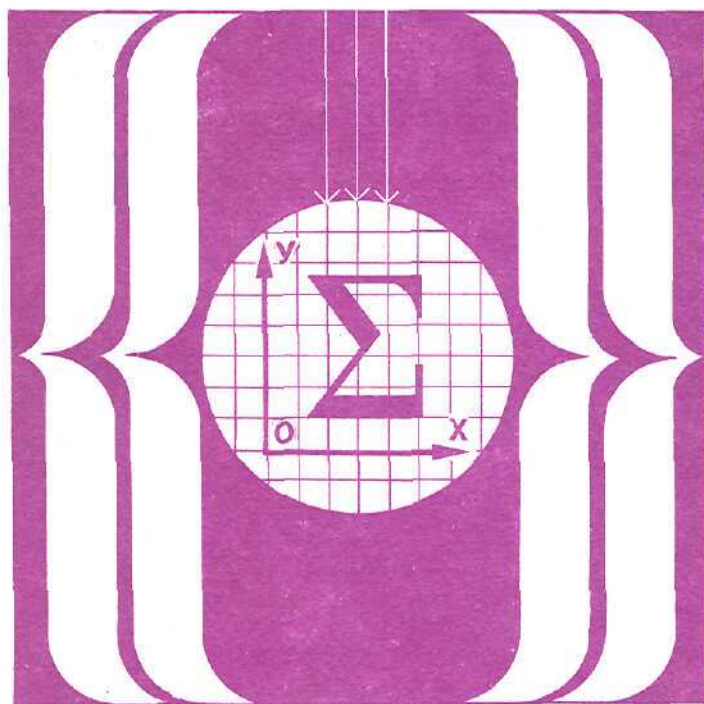


РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ МАТЕМАТИКЕ

Н.Н. АПРАУШЕВА, Л.А. БАКСТ, И.А. ГОРЛАЧ,
А.А. ЖЕЛНИН, С.В. СОРОКИН, Н.Н. ФЕДОРОВА

О РАСПОЗНАВАНИИ ТИПОВ
ФРОНТАЛЬНОЙ ОБЛАЧНОСТИ
ПО ДАННЫМ ИСКУССТВЕННЫХ
СПУТНИКОВ ЗЕМЛИ



ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА 1996

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

СООБЩЕНИЯ ПО ПРИКЛАДНОЙ МАТЕМАТИКЕ

Н.Н. АПРАУШЕВА, Л.А. БАКСТ, И.А. ГОРЛАЧ,
А.А. ЖЕЛНИН, С.В. СОРОКИН, Н.Н. ФЕДОРОВА

**О РАСПОЗНАВАНИИ ТИПОВ
ФРОНТАЛЬНОЙ ОБЛАЧНОСТИ
ПО ДАННЫМ ИСКУССТВЕННЫХ
СПУТНИКОВ ЗЕМЛИ**



ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РАН
МОСКВА 1996

Ответственный редактор
доктор техн. наук В.М. Солодов

Проведено распознавание основных типов облачности и подстилающей поверхности по многоспектральной информации искусственных спутников Земли (ИСЗ) с использованием методов кластер-анализа: предварительного оценивания числа кластеров, алгоритма Дзя-Шлезингера расщепления смеси на составляющие и алгоритма выбора оптимального числа кластеров из множества допустимых значений.

Рецензенты: С.К. Дулин,
И.А. Куликова

Научное издание

С

Вычислительный центр РАН, 1996, Св.план 1996, поз.19

"В небесах торжественно и чудно!

Спит земля в сияньи голубом..."

М.Ю.Лермонтов

При использовании спутниковой информации в анализе и прогнозе погоды возникает необходимость в физической (метеорологической) интерпретации количественных данных, получаемых по измерениям различных радиометров, установленных на ИСЗ. Задача интерпретации возникла при запуске первых ИСЗ, и она не решена окончательно до сих пор ввиду сложности, множественности процессов, происходящих в атмосфере и на подстилающей поверхности, и неоднозначности перехода от данных измерений ИСЗ к физическим параметрам атмосферы. Выполнено много исследований с целью создания "объективной" (независимой от опыта синоптика-дешифровщика), автоматической методики интерпретации данных ИСЗ.

В решении задачи объективной интерпретации данных ИСЗ обычно применяется один из трех подходов или различные их комбинации [1, 2]: 1) накопление априорной информации ИСЗ о

различных типах облачности в разных географических условиях и использование ее в качестве эталона при работе с новыми данными; 2) исследование процессов взаимодействия излучения различных участков спектра через облака с разными физическими свойствами и выработка пороговой классификации облачности; 3) автоматическая классификация с использованием статистического подхода.

Первые опыты по автоматическому дешифрированию данных ИСЗ с применением в основном первого подхода [3] оказались мало успешными. Как показали исследования последнего времени [4, 5], применение второго подхода в чистом виде также дает результаты с не очень высокой точностью.

Наиболее успешными оказались опыты по автоматическому дешифрированию облачности при использовании вероятностной нейронной сети (ВНС) с учетом первых двух подходов. Так, в работе [6] проведена классификация облачности по данным 5-канального радиометра AVHRR, установленного на ИСЗ NOAA. Выделялись 9 классов облачности и безоблачные участки над морской поверхностью. Исходные шаблоны для входа в ВНС были заранее вычислены на основании примерно 200 признаков, выделенных из спектральных, текстурных и физических измерений. Разработанная таким образом автоматическая классификация дала точность определения классов облачности 75-80%. Еще более высокая точность достигнута в работе [7] при использовании иерархии пороговых величин и ВНС. Здесь на первом этапе классификации производится автоматическое разделение участков изображения

по данным об их структурных особенностях различных масштабов - от синоптического до масштаба отдельных облаков. Точность классификации в [7] достигает 90%.

При дешифрировании относительно небольших по площади участков изображения, занятых фронтальной облачностью, возникают проблемы с выделением различных типов облачности на малых расстояниях.

Для решения такой задачи был использован статистический алгоритм автоматической классификации, который базируется на предположении, что для данной совокупности p -мерных наблюдений

$$X^n = \{X_1, X_2, \dots, X_n\}, \quad (1a)$$

$$X_j = (X_{j1}, X_{j2}, \dots, X_{jp}), \quad j=1, 2, \dots, n, \quad 1 \leq p \leq n, \quad (1b)$$

неизвестная функция плотности вероятности $f(Y)$ с допустимой точностью аппроксимируется смесью k нормальных распределений с различными векторами средних значений

$$\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip}), \quad i=1, 2, \dots, k,$$

и с равными ковариационными матрицами Σ

$$f(Y, \theta) = |\Sigma|^{-1/2} (2\pi)^{-p/2} \sum_{i=1}^k \pi_i \exp(-0,5(Y - \mu_i) \Sigma^{-1} (Y - \mu_i)') \quad (2)$$

π_i - априорная вероятность i -й компоненты смеси,

$$0 < \pi_i < 1, \quad \sum_{i=1}^k \pi_i = 1,$$

Y - вектор-строка, Y' - вектор-столбец. Множество неизвестных параметров смеси в (2) кратко обозначим r -мерным параметром θ ,

$$\theta = (\pi_1, \pi_2, \dots, \pi_{k-1}, \mu_1, \mu_2, \dots, \mu_k, \Sigma),$$

$$r = k-1 + kp + p(p+1)/2.$$

Тогда под классом или образом ω_i , $i=1, 2, \dots, k$, понимается генеральная совокупность, описываемая унимодальной функцией плотности вероятности нормального распределения

$$f(Y, \mu_i, \Sigma) = |\Sigma|^{-1/2} (2\pi)^{-p/2} \exp(-0,5(Y-\mu_i)\Sigma^{-1}(Y-\mu_i)').$$

Такая модель нашла широкое применение в решении различных научных и практических задач распознавания образов [8-10]. Для выделения классов в исследуемой совокупности наблюдений (1) необходимо разделить смесь (2) на ее составляющие, т.е. найти оптимальные оценки для неизвестных параметров θ , k , а затем провести классификацию элементов x_j в (1) по правилу Байеса [11, 12].

При известном значении k оптимальной оценкой θ_{opt} для параметра θ мы считаем оценку θ несмещенную, состоятельную и асимптотически эффективную [13], выбираемую по принципу максимума правдоподобия. Оптимальной оценкой для параметра k мы считаем наиболее вероятное значение в ряду $k=1, 2, \dots, l$, $l \ll n$.

Алгоритм автоматической классификации, используемый в решении поставленной задачи, состоит из двух частей.

1. Вычисление оптимальной оценки максимального правдоподобия для параметра смеси θ при фиксированном значении k , $k=1, 2, \dots, l$, $1 \leq l < n$. При $k=1$ такая оценка θ единственна [12], при $k \geq 2$ система уравнений правдоподобия имеет несколько решений $\{\theta\}$, которые можно получить по алгоритму Дэя-Шлезингера [11, 14]. Среди множества полученных решений оптимальным решением θ_{opt} считаем то, которое обращает в максимум функцию правдоподобия.

2. Определение оптимальной оценки k_{opt} путем последовательной проверки двух конкурирующих гипотез H_k и H_{k+1} , $k=1, 2, \dots, l$, $1 \leq l < n$, H_k - данная совокупность наблюдений (1) есть выборка из смеси k нормальных распределений. В ходе последовательной проверки принимается та гипотеза H_k , которая впервые не отвергается по критерию отношения правдоподобия Уилкса [13, 15]. По теореме Уилкса [13] в случае истинности гипотезы H_k при выполнении некоторых условий статистика

$$\lambda_{k,k+1} = -2 \ln(L(X^{(n)}, k, \theta_{opt}) / L(X^{(n)}, k+1, \theta_{opt})), \quad (3)$$

$$k \in \{1, 2, \dots, l-1\}, \quad 1 \leq l < n,$$

сходится по вероятности к χ^2 -распределению с числом степеней свободы s , $s=p+1$, $L(X^{(n)}, k, \theta_{opt})$ - значение функции правдоподобия исследуемой совокупности наблюдений при $\theta = \theta_{opt}$.

При получении оптимальных оценок k_{opt} , θ_{opt} для данной

выборки (1) классификация ее элементов проводится по статистическому решающему правилу [12]: X_j относится к такому классу ω_{i_0} , $i_0 \in \{1, 2, \dots, k_{opt}\}$, для которого значения его апостериорной вероятности

$$\bar{P}(X_j/\omega_1) = \frac{\pi_1 \exp(-0,5(X_j - \mu_1)' \Sigma^{-1} (X_j - \mu_1))}{\sum_{s=1}^k \pi_s \exp(-0,5(X_j - \mu_s)' \Sigma^{-1} (X_j - \mu_s))'}$$

максимально, т.е.

$$i_0 = \max_{1 \leq i \leq k_{opt}} \arg \bar{P}(X_j/\omega_i).$$

Для уменьшения объема вычислений при отыскании оптимального значения $k - k_{opt}$ целесообразно провести предварительный анализ структуры исследуемого множества [16]. Например, полу-

чить верхние оценки для числа k по гистограммам признаков и трех вариационных рядов расстояний между точками множества (1). Тогда, если k_0 - наименьшая из полученных верхних оценок, т.е. $k \geq k_0$, то при проверке гипотез H_k, H_{k+1} полагаем

$$k = k_0 - 1, k_0, \dots, 1, 1 < n.$$

Для проверки возможности распознавания типов облачности по данным ИСЗ по вышеизложенному алгоритму были выбраны два района. Первый район находился в зоне вторичного фронта средних широт в летнее время (30 июля 1991 г.). Размер выборки составил 140×80 пикселей (объектов) измерений, в качестве признаков использованы данные в пяти диапазонах спектра излучения: радиационная отражаемость в видимой (A1) и инфракрасной (A2) частях солнечного спектра (1 и 2 каналы AVHRR ИСЗ NOAA), радиационная температура в трех участках излучения в различных участках инфракрасного диапазона спектра (3, 4 и 5 каналы - T3, T4, T5). Радиационная отражаемость измеряется на ИСЗ NOAA в окне излучения 0.6 мкм и 0.9 мкм , радиационная температура измеряется при излучении 3.7 , 10.5 , 11.5 мкм .

Таким образом, каждый объект характеризовался пятью статистически зависимыми признаками,

$$X_j = (x_{j1}, x_{j2}, x_{j3}, x_{j4}, x_{j5}), \quad j=1, 2, \dots, 11200. \quad (4)$$

Каждый признак x_i , $i=1, 2, \dots, 5$, исходя из его физического смысла, можно рассматривать как реализацию некоторой непрерывной случайной величины ξ_i , принимающей любое значение из некоторого интервала $[a_i, b_i]$. Так что представление неизвестной функции плотности вероятности выборки (4) в виде смеси нормальных распределений (2) вполне корректно.

Второй район - фронт окклюзии в поле разрывающейся

облачности в средних широтах. Размер выборки составил 100×30 пикселей, в качестве признаков использованы данные в двух диапазонах спектра излучения: в инфракрасном диапазоне и в диапазоне излучения водяного пара. Данные получены со спутника METEOSAT 9 декабря 1991 г.

Таким образом, каждый объект характеризовался двумя непрерывными статистически зависимыми, слабо коррелированными признаками (их коэффициент корреляции равен 0.43),

$$x_j = (x_{j1}, x_{j2}), \quad j=1, 2, \dots, 3000. \quad (5)$$

В этом случае также допустимо использование смеси нормальных распределений для аппроксимации неизвестной функции плотности вероятности множества наблюдений (5).

Перейдем к детальному изложению распознавания типов облачности и подстилающей поверхности в первом районе по алгоритму автоматической классификации.

Анализ гистограмм пяти выбранных признаков (рис.1-5) показал, что все признаки являются информативными, на гистограммах существуют статистически значимые локальные минимумы [16]. Для числа классов получили оценку $k \geq 3$. По гистограмме основного вариационного ряда (ОВР) расстояний между точками подвыборки объема $n = 200$, полученной из выборки (4) методом случайного выбора без возвращения (СВБВ), имеем $k \geq 5$ [16]. Поэтому при определении оптимального числа классов k_{opt} проверку двух конкурирующих гипотез H_k и H_{k+1} можно про-

водить при $k = 3, 4, \dots, 1, 1 \ll 11200$.

Таблица 1

Признаки	1	2	3	4	5
1	1	0.97	-0.32	-0.68	-0.66
2	0.97	1	-0.37	-0.71	-0.69
3	-0.32	-0.37	1	0.86	0.87
4	-0.68	-0.71	0.86	1	0.98
5	-0.66	-0.69	0.87	0.98	1

В табл.1 приведена корреляционная матрица (r_{is}) , $i, s = 1, 2, \dots, 5$, исследуемой совокупности (4) до проведения классификации. По данным этой таблицы следует, что наибольшие коэффициенты корреляции имеют признаки: четвертый и пятый ($r = 0.98$), первый и второй ($r = 0.97$), а также третий и четвертый ($r = 0.86$). Наименьшие по модулю коэффициенты корреляции имеют признаки: первый и третий ($r = -0.32$), второй и третий ($r = -0.37$). После проведения классификации значения коэффициентов меняются.

Для распознавания видов изображения на множестве (4) алгоритм Дзэ-Шлезингера использовался в два этапа [11,14,15].

На I этапе две подвыборки, полученные методом СВБВ, разделялись последовательно на 3,4,5,6 классов при случайном за-

дании начальных условий. Из множества значений $k=3,4,5,6$ выделяли оптимальное значение $k_{opt} = 5$ методом последовательной проверки гипотез H_k и H_{k+1} , $k = 3,4,5,6$, с использованием статистики Уилкса и 5 % уровня значимости χ^2 -распределения, $\chi^2_{0.05}(6) = 12.59$ [13,15], число степеней свободы статистики $\lambda_{k,k+1}$ в этом случае равно 6.

На II этапе проводилась классификация множества X на 5 классов с использованием в качестве начального приближения θ_0 оценки θ , подсчитанной на первом этапе. Ввиду ограниченности оперативной памяти ЭВМ Эльбрус, II этап был разбит на 11 подэтапов, на каждом из которых при фиксированном значении θ_0 на 5 классов разделялась подвыборка $X(i)$, $i=1,2,\dots,11$ объема $n_i=1100$, формируемая по методу СВБВ.

Оценка вероятности ошибки классификации $P_{0ш}$ не превышала 5 %.

В табл. 2 даны значения коэффициентов корреляции, используемых признаков после проведения классификации. Сравнивая данные табл. 1, 2, мы видим, что статистически не изменились коэффициенты корреляции первого и второго, первого и третьего, четвертого и пятого признаков, но существенно изменились значения коэффициентов корреляции третьего и четвертого, третьего и пятого признаков.

Таблица 2

Признаки	1	2	3	4	5
1	1	0.98	-0.28	-0.71	-0.61
2	0.98	1	-0.27	-0.71	-0.62
3	-0.28	-0.27	1	0.56	0.51
4	-0.71	-0.71	0.56	1	0.91
5	-0.61	-0.62	0.51	0.91	1

Таблица 3

Номера классов	Средние значения по классам					Априорные вероятности
	μ_{11}	μ_{12}	μ_{13}	μ_{14}	μ_{15}	π_1
1	90	82	308	280	279	0.255
2	88	84	268	237	237	0.176
3	15	28	302	294	291	0.331
4	113	104	276	307	232	0.002
5	46	47	305	280	277	0.236

В результате проведенной классификации получены также оценки средних значений признаков по 5 классам $\bar{\mu}_{is}$ и априорных вероятностей классов π_i , $i = 1, 2, 3, 4, 5$, $s = i = 1, 2, 3, 4, 5$, помещенные в табл. 3.

Наиболее представительные классы - это класс 3, названный классом подстилающей поверхности - (радиационная отражаемость $A_1, A_2 \sim 10-30\%$, $T_3, T_4 \sim 280-310^\circ K$), класс 1 низкой слоистообразной облачности, ($A_1, A_2 \sim 80-90\%$, $T_3, T_4 \sim 275^\circ-310^\circ K$), класс 2 конвективных кучево-дождевых облаков ($A_1, A_2 \sim 85-100\%$, $T_3, T_4 \sim 230-240^\circ K$), класс 5 кучевой облачности ($A_1, A_2 \sim 46-47\%$, $T_3, T_4 \sim 270^\circ K$).

Кроме того, в табл. 4 представлена симметричная матрица расстояний Махаланобиса [12] ρ_{is} , $i, s = 1, 2, \dots, 5$.

$$\rho_{is} = (\mu_i - \mu_s) \Sigma^{-1} (\mu_i - \mu_s)'$$

Из этой таблицы следует, что классы достаточно далеко отстоят друг от друга, $\rho_{is} \gg 42$, наибольшее значение равно 143.

Таблица 4

Классы	1	2	3	4	5
1	0	72	95	72	56
2	72	0	126	79	89
3	95	126	0	143	42
4	72	79	143	0	110
5	56	89	42	110	0

Разделение многоспектральной спутниковой информации на классы по второму району проводилось по той же схеме, что и

для первого района.

Анализ гистограмм признаков выборки (5) показал, что оба признака являются информативными (рис. 6-7) и $k \geq 3$. По гистограмме ОВР подвыборки объема 150, полученной из выборки (5) методом СВБР, имели также $k \geq 3$.

На I этапе классификации подвыборка объема 300 последовательно разделялась по алгоритму Дзя-Шлезингера на 2, 3, 4, 5 классов при случайном задании начальных условий [11, 14]. Из множества значений $k = 2, 3, 4, 5$ нашли оптимальное значение $k_{opt} = 3$ методом последовательной проверки гипотез H_k и H_{k+1} , $k = 2, 3, 4$, с использованием статистики Уилкса $\lambda_{k, k+1}$ и 5% уровня значимости χ^2 -распределения с тремя степенями свободы, $\chi^2_{0.05}(3) = 7.81$ [13, 15].

На II этапе проводилась классификация множества (5) на 3 класса с использованием в качестве начального приближения θ_0 оценки θ , полученной на I этапе при $k = 3$.

Ввиду ограниченности оперативной памяти ЭВМ Эльбрус II этап был разбит на три подэтапа, на каждом из которых при фиксированном начальном значении θ_0 на 3 класса разделялась подвыборка $X(i)$, $i=1, 2, 3$, объема 1000, формируемая из (5) методом СВБВ. Оценка вероятности ошибки классификации не превышала 5%.

После проведения классификации коэффициент корреляции между признаками почти не изменился, он принял значение 0.37.

В табл. 5 приведены средние значения признаков по трем выделенным классам $\bar{\mu}_{is}$, $s = 1, 2$, а также априорные вероятности классов $\bar{\pi}_i$, $i = 1, 2, 3$.

Полученные значения расстояний Махаланобиса между классами: $\rho_{12}=116$, $\rho_{23}=75$, $\rho_{13}=54$, свидетельствуют о достаточной удаленности классов друг от друга.

Таблица 5

Номера классов	Средние значения (условные единицы)		Априорные вероятности π_1
	$\bar{\mu}_{11}$	$\bar{\mu}_{12}$	
1	149	68	0.155
2	120	180	0.052
3	165	119	0.793

Наиболее представительный класс - класс 3 подстилающей поверхности (радиационная температура $T_3 \sim 280-310$ K), класс 1 слоистообразной облачности ($T_3 \sim 285$ K) и класс 2 контуров береговой линии, выделенных программным способом.

В результате проведенных экспериментальных расчетов можно сделать вывод о реальной возможности применения описанного выше алгоритма автоматической классификации к распознаванию основных типов облачности и подстилающей поверхности по много-спектральной спутниковой информации.

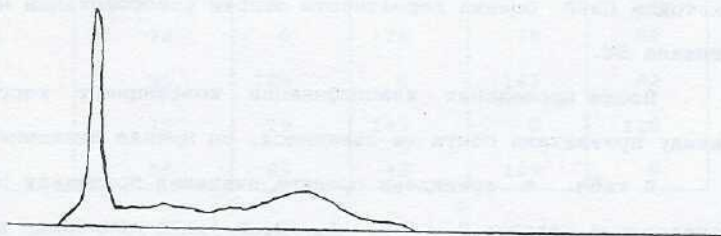


Рис. I

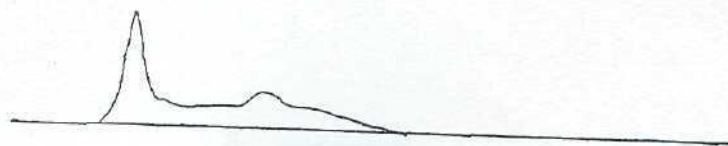


Рис. 2

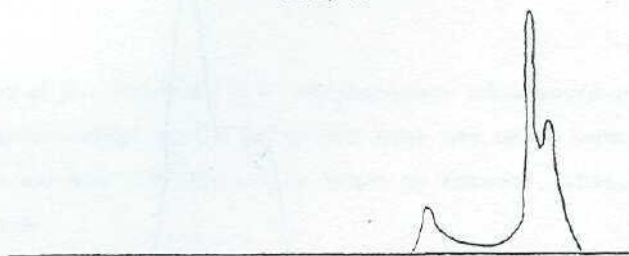


Рис. 3

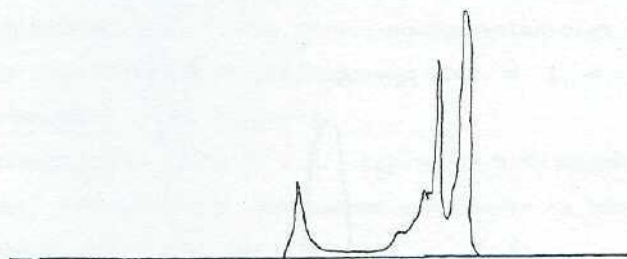


Рис. 4

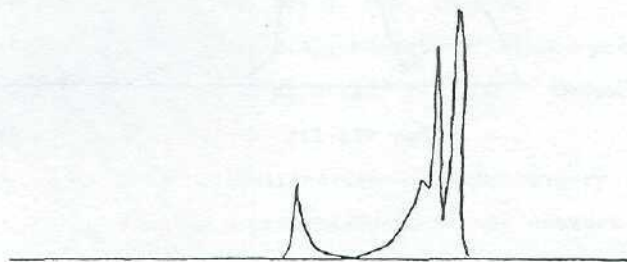


Рис. 5

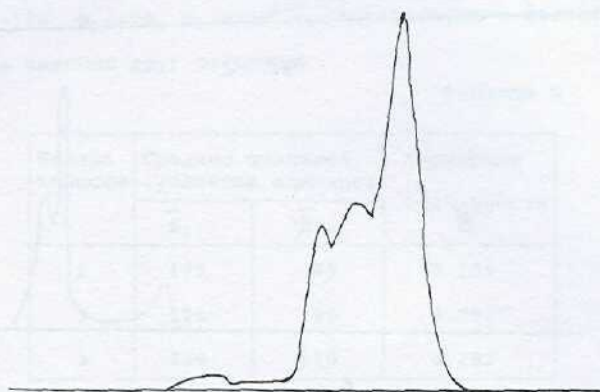


Рис. 6

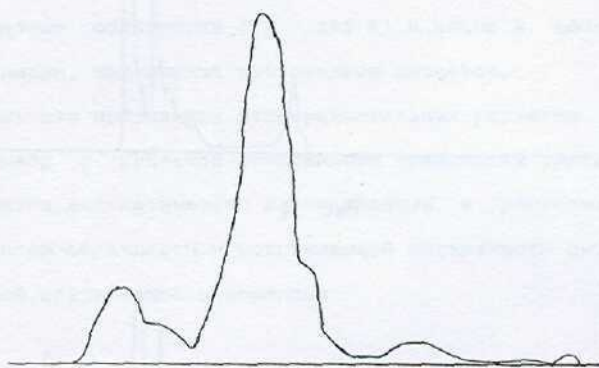


Рис. 7

Литература

1. Бакст Л.А., Федорова Н.Н. Исследование облачности по многогоспектральным данным AVHRR ИСЗ NOAA для целей синоптического анализа// Исследование Земли из Космоса, 1994, No 4. С. 3-8.
2. Федорова Н.Н., Баясгалан М., Эрдэнтуяа М. К вопросу классификации облачности вв интерактивном режиме// Информационное письмо: Опыт работы научно-исследовательских и сетевых подразделений Госкомгидромета СССР. Ч. 1. М.: Гидрометеиздат, 1991. С. 10-13.
3. Соловьева И.С., Сонечкин Д.М., Харитонов В.Ф. Обработка и анализ телевизионных изображений облачности на ЭВМ// Тр. Гидрометцентра СССР. Вып. 73, 1971. С. 64-74.
4. Tokuno Masami, Tsuchiya Kijoshi. Classification of cloud types based on data of multiple satellite sensors// Adv. Space Res., 1994, v. 14, No. 3. P.P. 199-206.
5. Strabala K.I., Ackerman S.A., Menzel W.P. Cloud properties inferred from 8-12 μm data// J. Appl. Meteorology, 1994, v. 33, No. 2. P.P. 212-219.
6. Bankert R.L. Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network// J. Appl. Meteorology, 1994, v. 33, No. 8. P.P. 909-918.

7. Segmentation of Satellite imagery using hierarchial thresholding and neural networks// J. Appl. Meteorology, 1994, v. 33, No. 5. P.P. 605-616.
8. Апраушева Н.Н., Раджабова М.Б. Классификация хлопка-сырца по статистическому алгоритму. М.: ВЦ АН СССР, 1990.
9. Апраушева Н.Н., Козельцева В.Ф., Толстиков Э.Е. Об опыте применения алгоритма Дзя к распознаванию метеорологических однородных полей// Тр. Гидрометцентра СССР. Вып. 192, 1977. С. 60-70.
10. Волошин Г.Я., Бурлаков И.А., Косенкова С.Т. Статистические методы решения задач распознавания, основанные на аппроксимационном подходе// Владивосток: 1992, ч.1, 155 с.
11. Day N.E. Estimating the components of a mixture of normal distributions// Biometrika, 1969. V. 56, No. 2.
12. Андерсон Т. Введение в многомерный статистический анализ. М.: Физматгиз, 1963.
13. Уилкс С. Математическая статистика. М.: Наука, 1973.
14. Апраушева Н.Н. Алгоритм расщепления смеси нормальных классов// Программы и алгоритмы. М.: ЦЭМИ АН СССР, 1976, N 68. 26 с.
15. Апраушева Н.Н. Определение числа классов в задачах классификации// Изв. АН СССР. Сер. Техн. Кибернетика, 1981, N 3. С. 71-77.
16. Апраушева Н.Н. Новый подход к обнаружению кластеров. М.: ВЦ РАН, 1993.

Н.Н. Апраушева, Л.А. Бакст, И.А. Горлач,
А.А. Желнин, С.В. Сорокин, Н.Н. Федорова

О распознавании типов
фронтальной облачности
по данным искусственных спутников Земли

Корректурa авторов

Подписано в печать 19.03.96

Формат бумаги 60x84 1/16

Уч.-изд.л. 1. Усл.-печ.л. 1,25

Тираж 120 экз. Заказ 17. Цена договорная

Отпечатано на ротaпpинтах в ВЦ РАН
117333, Москва, ул. Вавилова, 40