УДК 551.58
Серга Э.Н., *к.геогр.н.*
Одесский государственный экологический университет

УНИВЕРСАЛЬНЫЙ ИТЕРАЦИОННЫЙ МЕТОД КЛАСТЕРИЗАЦИИ ДАННЫХ

Предлагается новый метод кластерного анализа, позволяющий производить разбиение массива данных на подмножества по принципу неоднородности. Численные эксперименты показали, что результаты кластеризации на примере температуры поверхности океана с помощью данного метода находят хорошее физическое обоснование.
Ключевые слова: *кластер, однородность, критерий, омега-квадрат, температура поверхности воды.*

Введение. Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster – гроздь, скопление. Впервые в 1939 был определен предмет кластерного анализа и сделано его описание исследователем Трионом [24]. Главное назначение кластерного анализа – разбиение множества исследуемых объектов и признаков на однородные в соответствующем понимании группы или кластеры. Это означает, что решается задача классификации данных и выявления соответствующей структуры в ней. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по количественному сходству. Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ в отличие от большинства математико-статистических методов не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассматривать множество исходных данных практически произвольной природы.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы различных данных, делать их компактными и наглядными. Важное значение кластерный анализ имеет применительно к совокупностям временных рядов, характеризующих физические процессы. Здесь можно выделять периоды, когда значения соответствующих показателей были достаточно близкими, а также определять группы временных рядов, динамика которых наиболее схожа.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл здесь может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения кластерного анализа. Этот процесс можно представить системой с обратной связью. В задачах прогнозирования весьма перспективно сочетание кластерного анализа с другими количественными методами (например, с регрессионным анализом). Как и любой другой метод, кластерный анализ имеет определенные недостатки и ограничения: вчастности, состав и количество кластеров зависит от выбираемых критериев разбиения. При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов за счет замены их характеристиками обобщенных значений параметров кластера. При проведении классификации объектов игнорируется очень часто возможность отсутствия в рассматриваемой совокупности каких-либо значений кластеров.

Существует множество литературы, где подробно описаны различные методы проведения кластерного анализа [3,5,6,21,22,23,24]. Однако идеального алгоритма не

существует и потенциально не может существовать [12]. Можно создать достаточно качественный алгоритм, который даст хорошо объяснимые результаты, например, в медицине, но при применении в других областях науки, полученные результаты могут быть сомнительными.

Материали и методы исследований. В 2003 году нами был разработан алгоритм УАИМКА [18], который хорошо показал себя при кластеризации территорий, соизмеримых с территорией Украины, но при применении его для более крупных пространств возникли некоторые трудности, связанные с выявлением небольшого количества крупных кластеров с потенциально неоднородными районами. Поэтому начальные кластеры приходилось «дробить», с помощью этого же метода, после чего можно было качественно обосновать, с физической точки зрения, полученную кластеризацию. Помимо всего, иногда проявляемые значимые коэффициенты корреляции, которые использовались в качестве критерия в УАИМКА, могли бы вызвать неоднозначное толкование. Кроме того, использование внутрикластерных и межкластерных дисперсий для определения различий или же критериев при кластеризации гидрометеорологических данных не всегда уместно, соответственно, и критерия Фишера (отношения суммарной дисперсии к уменьшаемой по ранжированному ряду или же отношения дисперсий двух случайных величин) в качестве единственного, определяющего те же самые различия, так как в зависимости от рассматриваемой характеристики, находящейся под влиянием одного глобального процесса, можно получить, например, один кластер, там где их несколько, что может вызвать ложные суждения о физике процесса.

В связи с вышесказанным, мы выбрали несколько иной подход к критериям кластеризации, акцентировав внимание на выявлении неоднородности кластеров, с помощью известных параметрических и непараметрических критериев Фишера, Крамера-Уэлча (при равенстве объёмов двух независимых выборок, он полностью совпадает с критерием Стьюдента для средних) и Лемана-Розеблатта, применение которых в математической статистике к независимым непрерывным случайным величинам с неизвестными законами распределения является наиболее аргументированным [8,9,11,13,14].

Известно [15,16], что наивысшая степень однородности достигается, если обе выборки взяты из одной генеральной совокупности, т.е. справедлива нулевая гипотеза $H_0 : F(x) = G(x)$ при всех x. Отсутствие однородности означает, что верна альтернативная гипотеза: $H_1 : F(x) \neq G(x)$, хотя бы при одном значении аргумента x. Если гипотеза H_0 принята, то выборки можно объединить в одну, если нет, то нельзя.

Прежде чем приступить более подробно к рассмотрению упомянутых критериев, необходимо сказать, что гидрометеорологические величины в большинстве случаев не подчиняются нормальному закону распределения, и являются независимыми непрерывными случайными величинами.

Критерии Фишера и Крамера-Уэлча (Критерий Стьюдента) достаточно хорошо рассмотрены в [4,22]. Поэтому приведём лишь конечные формулы их определяющие:

а) критерий Фишера: $F = \dfrac{S_x^2}{S_y^2}$;

б) критерий (статистика) Крамера-Уэлча: $T = \dfrac{\sqrt{mn}\left(\overline{x} - \overline{y}\right)}{\sqrt{nS_x^2 + mS_y^2}}$,

где S_x^2 - несмещённая дисперсия случайной величины x, S_y^2 - несмещённая дисперсия случайной величины y, $\overline{x}$ и $\overline{y}$ - средние значения случайных величин x, y, m и n - объёмы случайных выборок x и y.

Необходимо отметить, что применение критерия Крамера-Уэлча не менее обосновано, чем применение критерия Стьюдента. Дополнительное преимущество - не требуется равенства дисперсий S_x^2 и S_y^2 [11,13]. Поэтому, д*ля проверки однородности математических ожиданий (гипотеза H_0) целесообразно применять критерий Крамера-Уэлча* [11,15].

Критерии проверки гипотез о дисперсиях в отличии от гипотез о средних весьма чувствительны к любым отклонениям от предположений, в условиях которых они были получены. И также отсутствует или противоречива информация относительно мощности соответствующих критериев [8].

Неотклонение проверяемых гипотез о равенстве средних и (или) равенстве дисперсий еще не говорит о принадлежности выборок одной и той же генеральной совокупности. Это свидетельствует лишь о возможном равенстве числовых характеристик, но не законов распределения. Выбор же критериев проверки гипотез относительно законов распределения, соответствующих двум выборкам, более скромен. Как правило, на практике используется либо критерий Смирнова, либо критерий Лемана-Розенблатта [16]. Предпочтительность использования данных критериев для проверки однородности подробно обсуждалась в [14,16]. В [8,15] было показано, что для критерия типа омега-квадрат (ω^2) нет выраженного эффекта различия между номинальными и реальными уровнями значимости. Поэтому рекомендовано для проверки однородности функций распределения (гипотеза H_0) применять статистику A типа омега-квадрат, а при отсутствии методического, табличного или программного обеспечение для статистики Лемана-Розенблатта, рекомендовано использовать критерий Смирнова.

Рассмотрим критерий Лемана–Розенблатта, согласно тому, как он представлен в [8,15,16].

Статистика критерия типа омега-квадрат для проверки однородности двух независимых выборок имеет вид:

$$A = \frac{mn}{m+n} \int_{-\infty}^{\infty} \left(F_m(x) - G_m(x) \right)^2 dH_{m+n}(x),$$

где $H_{m+n}(x)$ – эмпирическая функция распределения, построенная по объединенной выборке. Легко видеть, что

$$H_{m+n}(x) = \frac{m}{m+n} F_m(x) + \frac{m}{m+n} G_n(x).$$

Согласно [14] значение статистики зависит лишь от рангов элементов выборки:

Статистика A типа омега-квадрат была предложена Э. Леманом [25] в 1951 г., изучена М. Розенблаттом [26] в 1952 г., а затем и другими исследователями. Она

зависит лишь от рангов элементов двух выборок в объединенной выборке. Пусть $x_1, x_2, x_3, ..., x_m$ - первая выборка, $x'_1 < x'_2 < x'_3 < ... x'_m$ - соответствующий вариационный ряд, $y_1, y_2, y_3, ..., y_n$ - вторая выборка, $y'_1 < y'_2 < y'_3 < ... y'_m$ - вариационный ряд, соответствующий второй выборке. Поскольку функции распределения независимых выборок непрерывны, то с вероятностью $p = 1$ все выборочные значения различны, совпадения отсутствуют. Статистика A представляется в виде [16]:

$$A = \omega^2 = \frac{1}{mn(m+n)}\left[m\sum_{i=1}^{m}(r_i - i)^2 + n\sum_{j=1}^{n}(s_j - j)^2 \right] - \frac{4mn-1}{6(m+n)},$$

где r_i - ранг x'_i и s_j - ранг y'_j в общем вариационном ряду, построенном по объединенной выборке.

Правила принятия решений при проверке однородности двух выборок на основе статистики типа омега-квадрат (ω^2), так же как и статистики Смирнова, на основе критических значений в зависимости от уровней значимости и объемов совокупностей случайных величин приведены в таблицах [2].

В разработанном нами методе (УИМКД), в качестве исходной информации выступает матрица $X = (x_{ij})_{m \times n}$, содержащая m векторов-строк мерности n, характеризующая статистические ряды объёмом n в m пунктах, которые и должны быть кластеризованы. В качестве априорной информации, в отличие от других методов, так же как и в УАИМКА, задается только минимальное количество векторов τ, которые могут составить кластер.

Итерационный процесс в алгоритме УИМКД состоит из ряда шагов:

1-й шаг: Рассчитывается квадратная матрица порядка m евклидовых расстояний между всеми векторами матрицы X

$$D = (D_{ij})_{m \times m}, \qquad D_{ij} = \sqrt{\sum_{s=1}^{n} (x_{js} - x_{is})^2}. \tag{1}$$

Матрица D является симметрической. На главной диагонали этой матрицы располагаются нули.

2-й шаг: В каждой строке матрицы D производится ранжирование её элементов, т.е. её элементы располагаются в возрастающем порядке. В результате получим матрицу D^1 вида

$$D^1 = \begin{pmatrix} 0 & d_{12}^{(p)} & d_{13}^{(p)} & d_{14}^{(p)} & \dots & d_{1m}^{(p)} \\ 0 & d_{22}^{(p)} & d_{23}^{(p)} & d_{24}^{(p)} & \dots & d_{2m}^{(p)} \\ 0 & d_{32}^{(p)} & d_{33}^{(p)} & d_{34}^{(p)} & \dots & d_{3m}^{(p)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & d_{m2}^{(p)} & d_{m3}^{(p)} & d_{m4}^{(p)} & \dots & d_{mm}^{(p)} \end{pmatrix}. \tag{2}$$

В ней производится перенумерация элементов каждой строки (p - номер столбца, в котором располагается элемент каждой строки матрицы (2) в матрице D $(p = \overline{1,m})$).

Пусть, предположим, мы условились, что минимальное число векторов, которые могут составить кластер равно τ. Тогда анализу подвергается блок матрицы (2), состоящий из первых $\tau = \vartheta$ столбцов (на первом этапе, например, $\vartheta = 3$).

3-й шаг: Производится сравнение евклидовых расстояний D_{lj} $(l, j = \overline{1,m})$ с евклидовыми расстояниями $d_{l3}^{(p)}$. Если $D_{lj} \le d_{l3}^{(p)}$ $(l = \overline{1,m})$, то j-й вектор может рассматриваться как потенциальный центр кластера, в который входит l-й вектор (с учетом значения индекса p).

4-й шаг: Для каждого такого j-го вектора определяется количество l-х векторов (число вхождений S_j), для которых он может являться центром кластера.

5-й шаг: Из общего числа j-х векторов выделяются те, для которых $S_j \ge \tau$ (в нашем примере $S_j \ge 3$). Остальные потенциальные центры кластеров, для которых это условие не выполняется, ликвидируются.

6-й шаг: Определяется число оставшихся j-х векторов как центров кластеров. Пусть их число равно r $(j = r)$.

7-й шаг: Из общего числа m-х векторов устанавливаются те, которые попали в s-й (V_s) и q-й (V_q) кластеры одновременно $(s, q = \overline{1,r};\ l \ne s, q)$ и производится их разведение по кластерам по решающему правилу: $X_l \in V_s$, если $D_{ls} < D_{lq}$, при этом $S_q = S_q - 1$.

8-й шаг: Находятся евклидовы расстояния $\tilde{D}_{sq}$ между центрами s-го (V_s) и q-го (V_q) кластеров.

9-й шаг: Находится максимальное из расстояний между векторами, входящими в s-й кластер D_{ts} и q-й кластер D_{fq} $(t = \overline{1,s_s};\ f = \overline{1,s_q})$. Пусть это будет D_{fq}.

10-й шаг: Евклидове расстояние D_{fq} сравнивается с расстояниями между центрами кластеров $\tilde{D}_{sq}$. Если $\tilde{D}_{sq} < D_{fq}$, то при $S_s \ge S_q$ ликвидируется q-й кластер . Если $S_q > S_s$, то ликвидируется s-й кластер.

11-й шаг: Производится формирование ряда предварительных центров кластеров z_j.

12-й шаг: Производится распределение по кластерам векторов исходной выборки в соответствии с решающим правилом $X \in V_j$ если $D_{xz_j} < D_{xz_k}$.

13-й шаг: Определяется количество векторов S_j, вошедших в каждый j- й кластер V_j.

14-й шаг: Рассчитываются средние векторы для каждого j-го кластера R_j.

15-й шаг: Для каждой пары s-го и q-го кластеров на основе R_j определяются значения критерия Фишера F_j и критерия Стьюдента t_j (статистика Крамера-Уэлча).

16-й шаг: Полученные значения критерия Фишера F_j и критерия Стьюдента t_j сравниваются с критическими значениями на уровне значимости $\alpha = 0{,}05$ $F_{кр}$ и $t_{кр}$.

17-й шаг: Если все j-е кластеры на данном шаге выявляются неоднородными по отношению к друг другу, производится дополнительная проверка на однородность с помощью расчёта критерия типа омега–квадрат (Лемана-Розенблатта) ω^2 и сравнения его с критическим значением $\omega^2_{кр}$ на уровне значимости $\alpha = 0{,}05$.

18-й шаг: При условии $\omega^2 > \omega^2_{кр}$ возвращаемся к шагу 15 и итерационная процедура продолжается вновь, только уже для изначальных центров полученных кластеров z_j вплоть до 17-го шага включительно.

19-й шаг: Возвращаемся к шагу 2 при условии $\vartheta = \vartheta + 1$, и итерационная процедура продолжается вновь вплоть до 18-го шага включительно.

20-й шаг: В случае выявления полной неоднородности средних векторов R_j и изначальных центров z_j кластеров, полученных на этапе ϑ и этапе $\vartheta - g$, $(g \geq 1)$ (согласно вышеупомянутым критериям), производится сравнение числа кластеров $N_\vartheta, N_{\vartheta-g}$ на соответствующих этапах, при условии $N_{\vartheta-g} \geq 2$ и $N_\vartheta \geq 2$. Если $N_\vartheta \geq N_{\vartheta-g}$, то ϑ этап считается основным, и количество неоднородных кластеров, полученных на последующих этапах, сравнивается с N_ϑ. Если $N_{\vartheta-g} > N_\vartheta$, то $\vartheta - g$ считается основным этапом.

21-й шаг: Процедура повторяется, начиная с шага 2 по 20 шаг до тех пор, пока $g \leq m$.

В качестве объекта для исследования методом кластеризации УИМКД были взяты поля среднемесячной температуры подстилающей поверхности Атлантического океана (данные ре-анализа ERA-40 [19]), заданные в узлах регулярной сетки точек 2,5º×2,5º в секторе, ограниченном по широте от 30º до 70º северной широты и по меридиану от 70º западной долготы до 10º восточной долготы, за период с 1958 по 2002 годы, зимние месяцы (декабрь, январь, февраль), так как с нашей точки зрения, наиболее ярко качество районирования может проявиться именно для температуры поверхности океана, которая тесно связана с известными типами течений и общей циркуляцией атмосферы. Таким образом, в каждом узле сетки был сформирован 34-х мерный вектор среднемесячных значений температуры для указанных месяцев. Множество этих векторов и было представлено алгоритму УИМКД для разбиения полей температуры поверхности воды на однородные кластеры. Выделенный район исследования представлен на рис.1, а результаты проведенной кластеризации - на рис.2-4. На них стрелками обозначены направления крупномасштабных океанических течений [1], а выделенные алгоритмом кластеры окрашены различными оттенками.

Прежде всего, следует отметить, что во все рассматриваемые месяцы хорошо проявляются в поле температуры кластеры циклонической и антициклональной циркуляции вод океана, кластер переменных течений, расположенный в субтропиках.

В Северной Атлантике под действием процессов, формирующих распределение поверхностной температуры с широтой (различное количество солнечной радиации, поглощаемое деятельным слоем океана и происходящих в нём процессов теплообмена, различные условия обмена с атмосферой теплом и количество движения, таяние арктических льдов, выносимых в Норвежское море) приводит к общему зональному распределению температуры воды. Это отчётливо проявляется в распределении кластеров в рассматриваемой акватории Северной Атлантики. Интересным является тот факт, что граница между двумя центральными кластерами практически совпадает с осью зоны дивергенции.

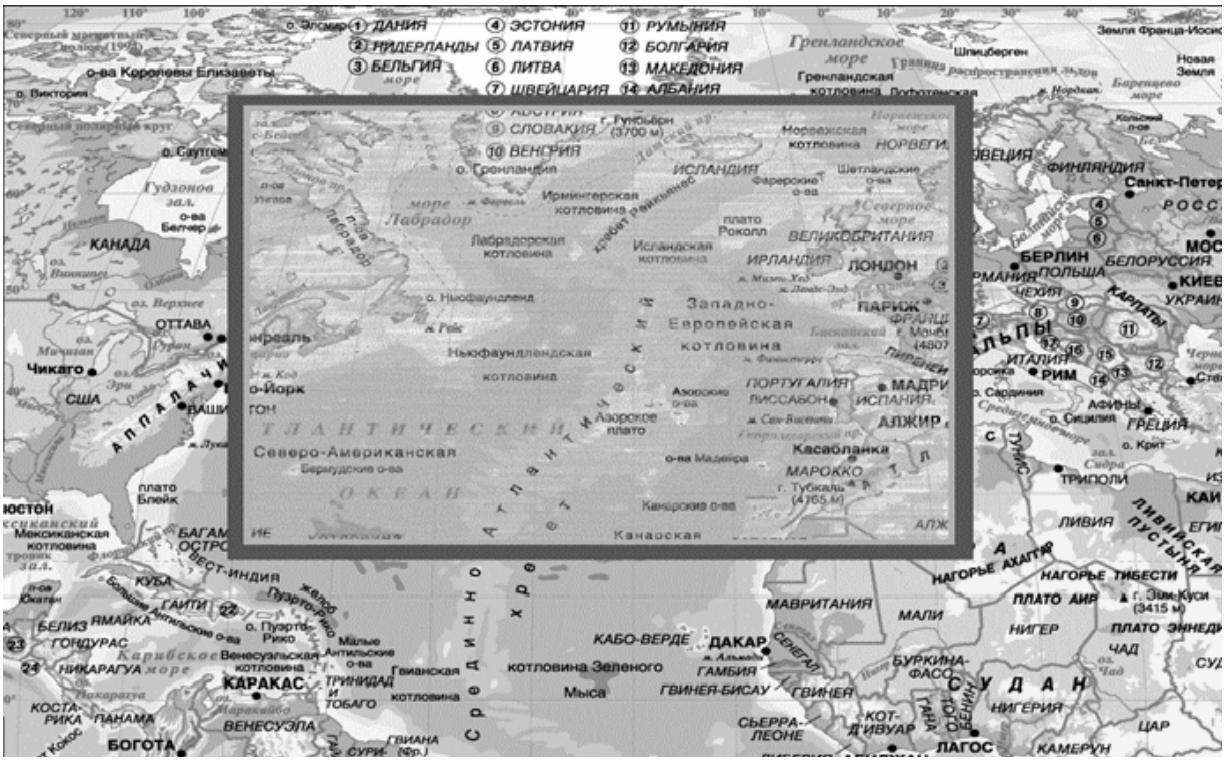

Рис. 1. Северо-Атлантический сектор подлежащий исследованию

Структура кластеров во все рассматриваемые месяцы отражает и распределение крупномасштабных течений в Северной Атлантике, таких как Гольфстрим, Северо-Атлантическое, Португальское, Восточно-Гренландское, Лабрадорское. Представляет большой интерес, что структура полученных кластеров в общем, а в некоторых местах океана в деталях, совпадает с однородными областями полей среднемесячных температур поверхности океана, полей осадков и испарений, затрат тепла на испарение, которые представлены в Атласе океанов [1].

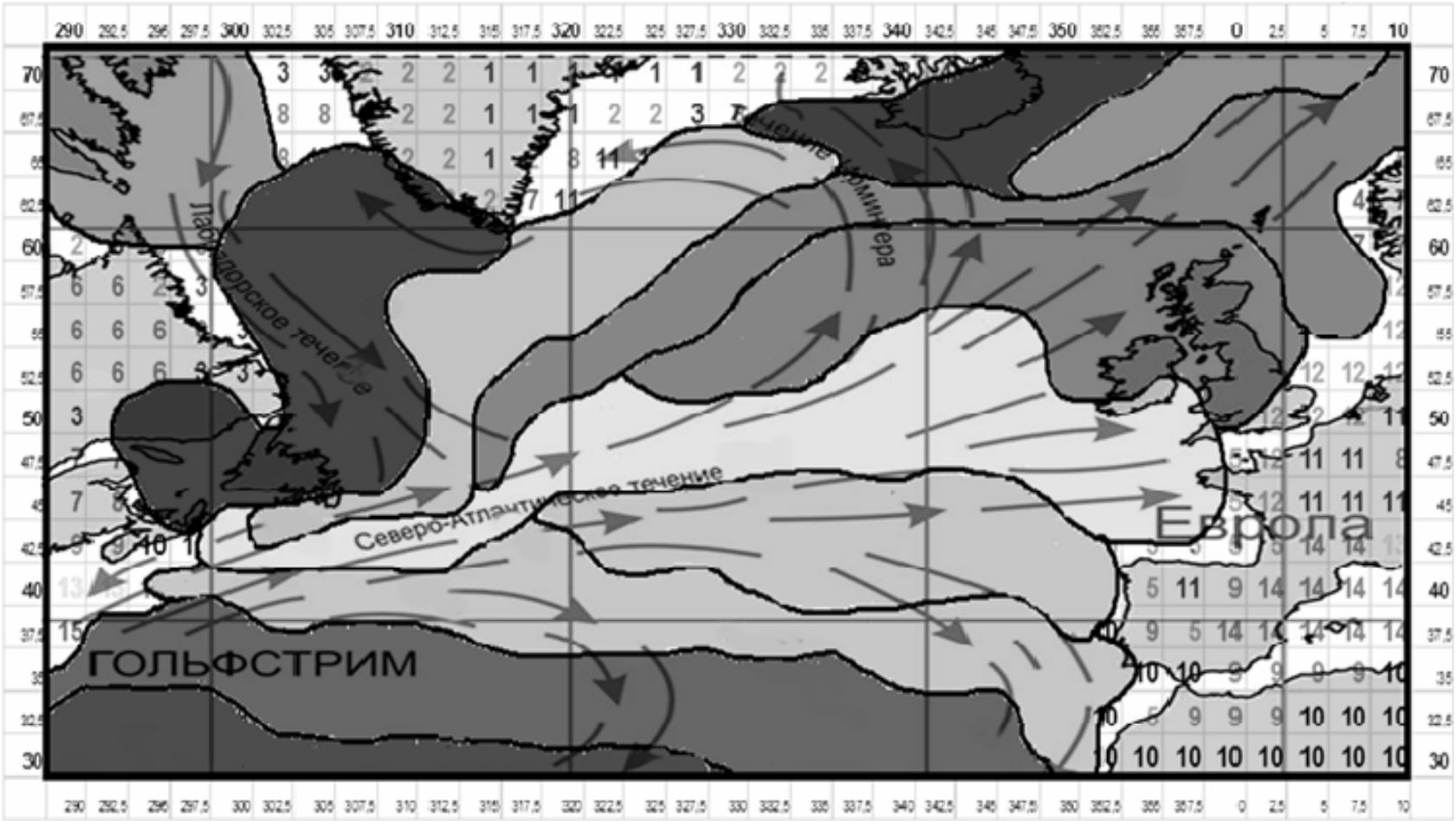

Рис. 2 Карта распределения кластеров среднемесячной температуры поверхности воды (декабрь)

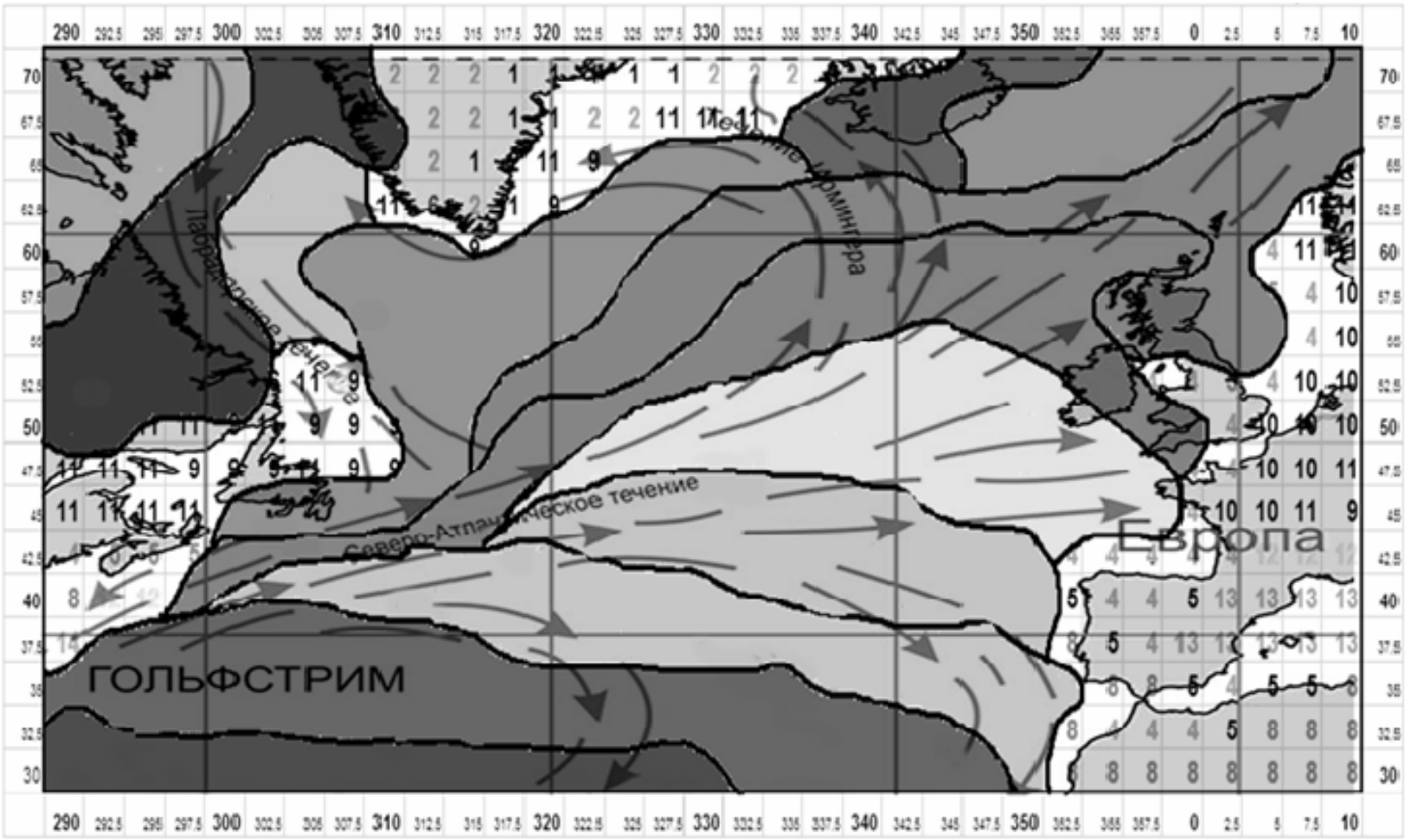

Рис. 3. Карта распределения кластеров среднемесячной температуры поверхности воды (январь)

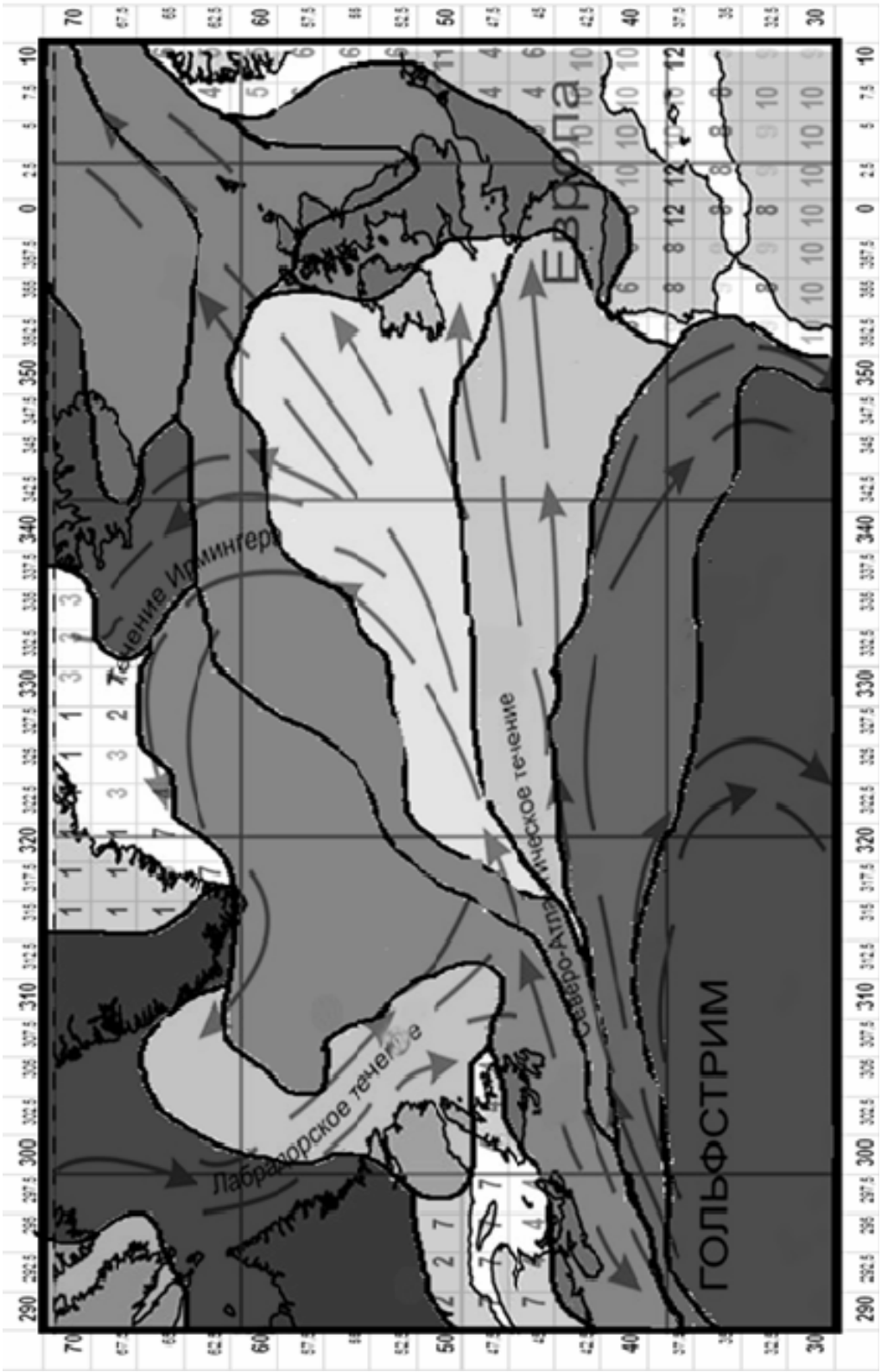

Рис. 4. Карта распределения кластеров среднемесячной температуры поверхности воды (февраль)

Вывод. Все приведенные факты особенности структуры распределения кластеров температуры поверхности воды океана, сходства различий кластеров в разные месяцы зимы свидетельствуют о том, что алгоритм УИМКД производит не только формальное распределение физических полей на однородные структуры, но и учитывает физические процессы, обуславливающие их формирование.

Список литературы

1. *Атлас океанов. Атлантический и Индийский океаны*/ Под ред. С.Г. Горшкова. – Л: Изд.ГУНИО, 1977.
2. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
3. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999. – 318 с.
4. *Крамер Г.* Математические методы статистики.: Пер. с англ.-2-е изд.- М,1975. – 325 с.
5. *Кулаичев А. П.* Методы и средства комплексного анализа данных. - М: ИНФРА- М, 2006. – 276 с.
6. *Лагутин М. Б.* Наглядная математическая статистика. - М.: П-центр, 2003. – 347 с.
7. *Лемешко Б. Ю., Горбунова А. А., Лемешко С. Б., Постовалов С. Н., Рогожников А. П., Чимитова Е. В*. Компьютерное моделирование и исследование вероятностных закономерностей // Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика. 2013. №1. С.74-85.
8. *Лемешко Б.Ю., Лемешко С.Б.* О сходимости распределений статистик и мощности критериев однородности Смирнова и Лемана-Розенблатта // Измерительная техника. 2005. № 12. С. 9-14.
9. *Лемешко Б.Ю., Лемешко С.Б., Миркин Е.П.* Исследование критериев проверки гипотез, используемых в задачах управления качеством // Материалы VII международной конференции “Актуальные проблемы электронного приборостроения” АПЭП-2004. Новосибирск, 2004. – Т. 6. – С. 269-272.
10. *Лемешко Б.Ю., Миркин Е.П.* Критерии Бартлетта и Кокрена в измерительных задачах при вероятностных законах, отличающихся от нормального // Измерительная техника. 2004. № 10. – С. 10-16.
11. *Лемешко Б.Ю., Помадин С.С.* Проверка гипотез о математических ожиданиях и дисперсиях в задачах метрологии и контроля качества при вероятностных законах, отличающихся от нормального // Метрология. 2004. – № 3.- С.3-15.
12. *Мандель И. Д.* Кластерный анализ. - М.: Финансы и Статистика, 1988. – 339 с.
13. *Орлов А.И.* О применении статистических методов в медико-биологических исследованиях. - М.: «Вестник Академии наук СССР», 1987.№2. С. 88-94.
14. *Орлов А.И.* О проверке однородности двух независимых выборок // Заводская лаборатория. – 2003. – Т.69. №.1. – С.55-60.
15. *Орлов А.И.* Прикладная статистика. - М.: «Экзамен», 2006. – 671 с.
16. *Орлов А.И.* Состоятельные критерии проверки абсолютной однородности независимых выборок // «Заводская лаборатория. Диагностика материалов».- 2012.Т.78. №.11. – С.66-70.
17. *Райзин Дж. Вэн.* Классификация и кластер. - М.: Мир, 1980. – 244 с.
18. *Серга Э.Н.* Универсальный адаптивный итерационный метод кластерного анализа // Міжвідомчий науковий зб. України: Метеорологія, кліматологія та гідрологія. – 2003. – Вип.47. – С.83-89.

19. Служба данных ECMWF ERA-40 [Електронний ресурс].- Режим доступа к журналу.: http://www.ecmwf.int/products/data.
20. *Смирнов Н.В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // Бюллетень МГУ, серия А. – 1939. – Т.2. №2. – С.3-14.
21. *Уиллиамс У.Т., Ланс Д.Н.* Методы иерархической классификации // Статистические методы для ЭВМ / Под ред. М. Б. Малютов. - М.: Наука, 1986. - С. 269–301.
22. *Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.* Обробка та аналіз гідрометеорологічної информації: Підручник.- К.: Міносвіти України, 1999. – 600 с.
23. *Jain A., Murty M., Flynn P.* Data clustering: A review // ACM Computing Surveys. - 1999. - Vol. 31, no. 3. - Pp. 264–323.
24. *Lance G. N., Willams W. T.* A general theory of classification sorting strategies. 1. hierarchical systems // Comp. J. - 1967. - no. 9. - Pp. 373–380.
25. *Lehmann E.L.* Consistency and unbiasedness of certain nonparametric tests / Ann. Math. Statist. – 1951. V.22. № 1. – P.165-179.
26. *Rosenblatt M.* Limit theorems associated with variants of the von Mises statistic // Ann. Math. Statist. – 1952. V.23. – P.617-623.

Універсальний ітераційний метод кластерізації даних.

Сєрга Е.М.

Пропонується новий метод кластерного аналізу, що дозволяє здійснювати розбиття масиву даних на підмножини за принципом неоднорідності. Чисельні експерименти показали, що результати кластеризації на прикладі температури поверхні океану за допомогою даного методу знаходять добре фізичне обгрунтування.

Ключові слова: *кластер, однорідність, критерій, омега-квадрат, температура поверхні води.*

The universal iterative method of clusterization data

Serga E.M.

A new method of cluster analysis that makes possibility to divide data set into multitudes in accordance with the heterogeneity principle. Numerical experiments show that the results of clusterization obtained for sea surface temperature, are physical explained.

Keywords: *cluster, homogeneity, criterion, omega-square, surface temperature.*