УДК 551.501

ОБ ОДНОМ ПОДХОДЕ К КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ

Ботыгин Игорь Александрович,
кандидат технических наук, доцент кафедры информатики и проектирования систем Института кибернетики Национального исследовательского Томского политехнического университета, Россия, 634050, г. Томск, пр. Ленина, 30. E-mail: bia@tpu.ru

Катаев Сергей Григорьевич,
доктор технических наук, профессор кафедры общей физики физико-математического факультета Томского государственного педагогического университета, Россия, 634061, г. Томск, ул. Киевская, 60. E-mail: sgkataev@sibmail.com

Тартаковский Валерий Абрамович,
доктор физико-математических наук, заведующий лабораторией биоинформационных технологий Института мониторинга климатических и экологических систем СО РАН, Россия, 634055, г. Томск, пр. Академический, 10/3. E-mail: trtk@list.ru

Шерстнёва Анна Игоревна,
кандидат физико-математических наук, доцент кафедры высшей математики физико-технического института Национального исследовательского Томского политехнического университета, Россия, 634050, г. Томск, пр. Ленина, д. 30. E-mail: sherstneva@tpu.ru

Актуальность работы *обусловлена необходимостью разработки универсальных информационно-аналитических подходов для извлечения знаний из бурно растущего объема геофизических данных. Одна из основных задач в обработке геофизических данных заключается в выявлении в них объективно существующих закономерностей, на которых можно построить разнообразные, в том числе и прогнозные, модели поведения выделенных параметров геофизических полей. И именно технологии кластеризации данных являются начальным базисом для создания программного обеспечения подобных информационных систем анализа неструктурированных данных.*

Цель работы. *Разработка методики кластеризации экспериментальных данных геофизической природы на основе выделения структур для решения задач анализа неструктурированной информации при изучении и контроле состояний сложных систем.*

Методы исследования. *Классические и современные методы и алгоритмы кластеризации, алгоритмы теории графов, контрольный пример кластеризации геофизического поля метеорологических параметров с территории северной части Евразии.*

Результаты. *Разработан новый алгоритм выделения структур в исходном геофизическом поле, позволяющий по пространственным характеристикам декомпозировать исследуемое пространство на области с похожим поведением исследуемых параметров. Алгоритм основывается на структуризации различных разложений (сезон, аномалия и т. д.) геофизических полей и обеспечивает получение разнообразной информации об исследуемом объекте в виде наборов параметров выделенных структур. Подобная информация вместе с сопутствующими эмпирическими зависимостями между параметрами рассматривается как обобщенная экспериментальная характеристика исследуемого объекта и служит основой для формирования гипотез и моделей его поведения. Кроме того, построенная таким образом структурная модель пространства метеорологического параметра обеспечивает возможность сжатия первичной информации без существенной потери семантической значимости исследуемого геофизического поля.*

Ключевые слова:
Геофизическое поле, кластеризация, теория графов, структура, метод выделения структур, метеонаблюдения, временной ряд.

Введение

Решение задач развития цивилизации однозначно предполагает наличие пространственно-временных систематических наблюдений за изменениями природных и антропогенно-техногенных систем. Получаемые таким образом данные, в соответствии с определением из [1], интерпретируются как геофизические поля (ГП), или, другими словами, как «множество значений физических величин (параметров), характеризующих естественное или искусственно созданное в Земле физическое поле (или его отдельные элементы) в пределах определенной территории или области Земли». Интерес к изучению геофизических полей как природного, так и техногенного происхождения, связан, прежде всего, с необходимостью оценки их влияния на изменение окружающей среды и геоэкологической обстановки, так как именно физические поля коренным образом влияют на энергетический обмен между живой и неживой приро-

дой и, следовательно, на качество жизни не только отдельных экологических систем, но и всей биосферы. Интерес к изучению геофизических полей связан и с изменением климата, однозначно влияющим на экологическую обстановку на планете.

Анализ структуры климатических полей (климатическое районирование, выделение классов и т. п.), с одной стороны, направлен на анализ закономерностей формирования различных типов климата в глобальной климатической системе. С другой стороны, определение территориальных границ типов климата, различных по своим свойствам, позволяет преобразовать огромное количество информации о климатических параметрах в гораздо меньшее число информационных структур с целью использования полученных результатов в хозяйственно-экономических мероприятиях и при моделировании климатических ситуаций. Таким образом, научная и практическая значимость любой климатической структуризации бесспорна. Чем больший региональный уклон имеют подобные исследования, тем более высокую социально-экономическую эффективность может нести полученная информация для конкретных отраслей. Если задача решается для исследования генезиса климата местности, основываясь на всем комплексе климатических условий соответствующих ландшафтных зон, то говорят о климатической классификации. Если выделение структур в полях элементов климата проводят для прикладных целей, то данную процедуру называют климатическим районированием.

Широко известным классическим классификациям климата присуща значительная доля субъективизма. При выделении климатических типов, зон, районов, помимо непосредственно температурно-влажностных характеристик, учитывается преобладание над территорией соответствующих типов воздушных масс по сезонам года и особенности их циркуляции, степень континентальности климата, характер подстилающей поверхности. Так, классы в одной из наиболее используемой в мире классификации климата В.П. Кеппена [2] выделяются на основе количественных критериев тепловлажностного режима с учетом ландшафтных особенностей территории. Классификация климата Л.С. Берга [3] основана на учете ландшафтно-географических зон суши. Границы климатических зон в генетической классификации Б.П. Алисова [3] определяются по среднему положению климатических фронтов, то есть в основе этой классификации лежит учет условий формирования климата в зависимости от типов воздушных масс и их циркуляции. На основе градаций характеристик тепло- и влагообеспеченности приземного воздуха и учета параметров теплового баланса деятельной поверхности построена классификация климатических режимов в работе [4]. Современные классификации предлагают более формализованные подходы, основанные, например, на учете вклада каждого влияющего фактора, но ранжированного своим весовым коэффициентом [5].

В работах, посвященных нахождению закономерностей поведения во времени полей метеопараметров (временных рядов, заданных в определенных пространственных точках) или иных характеристик на территориях разного масштаба, обычно исследуется поведение тренда в выбранном ареале, поскольку именно тенденция изменения временного ряда дает возможность осуществлять прогнозирование. В этих задачах широко применяются все методы анализа многомерных геофизических данных, направленные на поиск в этих данных тех или иных регулярностей, проявляющихся в существовании явных или неявных структур. К классическим методам многомерной статистики обычно относят: метод главных компонентов, факторный и корреляционный анализ, дискриминантный и кластерный анализ, многомерное шкалирование [6–13]. Выбор сочетания методов исследования зависит от целей исследования, природы данных и наличия априорной информации о возможных связях.

В ряде работ дополнительно исследуется структура временного ряда, под которой понимаются его характерные особенности, сформулированные в сжатом виде. Так, в работах [14–17] при исследовании периодических рядов среднемесячных температур (общего содержания озона, осадков и др.) в качестве параметров, оценивающих характерные особенности, использовались и среднее значение, и дисперсия, и фаза, и параметры тренда, и др. Используя этот набор данных, можно с той точностью, которую допускают данные, описать временной ряд и дать статистический прогноз его поведения.

Пространственно-временные связи между различными полями метеопараметров обычно изучают с использованием корреляционного анализа, который позволяет определить и временные показатели запаздывания или опережения событий и явлений. Канонический корреляционный анализ в линейной и нелинейной формах широко используется в климатологии, в частности для сезонного прогноза [18], анализа структуры колебаний Эль-Ниньо [19], определения среднеширотного атмосферного отклика на вариации приповерхностной температуры Тихого океана [19, 20].

При решении задачи прогнозирования климата также используются разнообразные подходы, основанные на применении классических методов регрессионного анализа [21–27], главных компонентов (эмпирических ортогональных функций) [28–35].

Среди методов многомерной статистики именно кластеризация чаще всего применяется для обработки данных во многих природно-климатических исследованиях. Заметим, что результат применения кластерного анализа – набор пространственных областей (кластеров), обладающих похожим поведением изучаемого параметра или целого набора параметров. Например, в [36] для систематизации полей гидротермического коэффициента Селянинова были использованы алгоритмы иерар-

хического кластерного анализа. Итогом явилось сжатие рассматриваемых массивов данных. Иерархическая кластеризация использовалась также при решении задачи прогноза облачности в [37]. Эксперименты проводились с разным числом кластеров и разными признаками. В работе [38] иерархическая кластеризация использовалась для решения задачи климатического районирования. Исходные данные были взяты для 35 метеостанций по 22 климатическим показателям. В исследованиях по дистанционному зондированию Земли [39] также был успешно использован иерархический (дивизимный) алгоритм кластеризации.

Кластеризация по методу *k*-средних была применена для решения задачи пространственной декомпозиции метеорологических полей Северного полушария [40]. Для исследования были использованы глобальные агроклиматические данные из базы FAOCLIM-2, где для каждой из почти 32000 метеостанций мира указаны до 14-ти наблюдаемых и вычисляемых параметров за длительные периоды наблюдений. Используемые для кластеризации итоговые данные составили порядка 100000 значений годовых осадков и около 5000 среднегодовых температур. В [41] для исследования была использована разновидность метода *k*-средних – метод динамических ядер. С его помощью были выделены 4 кластера – географо-климатических варианта светлохвойных травяных лесов (подтайги). Алгоритм кластеризации, известный как Shared Nearest Neighbor (SNN), был использован в [42] для исследования окружающей природной среды. Полный объем использованного архива составил 2.9 Терабайта. Архив содержит более 80 различных переменных (включая атмосферное давление на уровне моря, влажность воздуха, солнечную радиацию) в нескольких координатных системах с 1948 г. по нынешнее время.

Таким образом, кластеризация является методом, интенсивно применяемым при анализе природно-климатических данных. Существует ещё много разновидностей методов кластеризации числовых данных, которые обладают как своими достоинствами, так и недостатками [43]. Тем не менее, у большинства из них есть общая черта – необходимость задавать количество кластеров до начала кластеризации, что существенно затрудняет обработку «сырых» данных и сужает возможности кластеризации.

В настоящей работе описывается разработанный авторами метод выделения структур, относящийся к методам кластерного анализа и позволяющий решить две задачи для рядов наблюдений:

- осуществить распределение анализируемых данных по структурам;
- выявить наличие в анализируемых данных заранее заданных структур.

Структуры в настоящем контексте – это математические объекты, отражающие совокупное проявление скрытых связей между исследуемыми данными, которые не всегда являются очевидными.

Формальное описание

Рассмотрим конечное множество объектов $A=\{a_1,a_2,...,a_m\}$. Пусть каждый из объектов идентифицируется некоторой парой значений из пространства основных $X=\{x_1,x_2,...,x_n\}$ и вспомогательных $\Omega=\{\omega_1,\omega_2,...,\omega_k\}$ признаков. Общее пространство признаков обозначим как $X_\Omega=\{X,\Omega\}$ Построим матрицу расстояний $D=\|d_x\|_{m\times m}$ между объектами $A=\{a_1,a_2,...,a_m\}$, используя только элементы пространства основных признаков $X=\{x_1,x_2,...,x_n\}$. Причем вид функции, определяющей расстояние между любой парой объектов в построенном метрическом пространстве, может быть любым. Сформулируем некоторые абстрактные наборы требований к объектам $A=\{a_1,a_2,...,a_m\}$. При каждом конкретном анализе такие ограничения (аксиомы) определяются целью проводимых исследований. Пусть A_X – набор ограничений, сформулированных в терминах пространства основных признаков $X=\{x_1,x_2,...,x_n\}$. Набор ограничений, сформулированных в пространстве вспомогательных признаков $\Omega=\{\omega_1,\omega_2,...,\omega_k\}$, обозначим как A_Ω. Набор же ограничений, сформулированных только в терминах матрицы расстояний $D=\|d_{a_i}d_{a_j}\|_{m\times m}$ между объектами $A=\{a_1,a_2,...,a_m\}$, обозначим как A_D. В принятых обозначениях полный набор всех возможных ограничений можно представить упорядоченной тройкой аксиом $R=\{A_X,A_\Omega,A_D\}$.

Таким образом, задачу выделения структур можно сформулировать как задачу поиска на множестве объектов $A=\{a_1,a_2,...,a_m\}$ подмножеств (или структур), удовлетворяющих заданному набору ограничений из совокупности аксиом $R=\{A_X,A_\Omega,A_D\}$. В рассматриваемом варианте и $R=\{A_X,A_\Omega,A_D\}$ естественным образом формируются 8 классов наборов ограничений (типов аксиом):

$$\{0,0,0\},\{0,0,1\},\{0,1,0\},\{0,1,1\},\{1,0,0\},\ \{1,0,1\},\{1,1,0\},\{1,1,1\},$$

где цифры 0 и 1 указывают, на отсутствие или наличие соответствующих наборов ограничений из $R=\{A_X,A_\Omega,A_D\}$. Так, например, отсутствие каких-либо наборов ограничений $\{0,0,0\}$ определяет любое выделенное подмножество объектов из $A=\{a_1,a_2,...,a_m\}$ структурой. В случае же $\{0,0,1\}$ для выделения структур используются ограничения, сформулированные в терминах расстояний между объектами. Например, аксиома из такого класса $A_D^\circ=\{\max d_{a_ia_j}\le d_0,(a_i,a_j)\in A\}$ выделяет структуры объектов, в которых максимальное расстояние между любыми парами объектов $\max d_{a_ia_j}$ не превышает заданную норму d_0. Выделенные таким образом структуры находятся внутри гиперповерхности с радиусом d_0.

Аксиомы, сформулированные в терминах пространства основных признаков (A_X), а также в терминах пространства вспомогательных признаков (A_Ω), являются, по сути, абстрактными координатами гиперсфер, разбивающих множество объектов $A=\{a_1,a_2,...,a_m\}$ на несвязные подмножества $\{U_1,U_2,...,U_k\}$. При этом элементы несвязных подмножеств принадлежат множеству объектов A:

$U_1,U_2,...,U_k \in \{a_1,a_2,...,a_m\}$, пересечение несвязных подмножеств – пустое подмножество: $U_1 \cap U_2 \cap ... \cap U_k = \varnothing$, а их объединение есть само множество A: $U_1 \cup U_2 \cup ... \cup U_k = A$.

Алгоритм выделения структур

Для описания алгоритма выделения структур воспользуемся формализмами теории графов. Поскольку решаемая задача выделения структур хорошо отождествляется с задачей нахождения всех компонентов связности в неориентированном графе, в которой необходимо разбить вершины графа на несколько групп так, чтобы внутри одной группы можно было дойти от одной вершины до любой другой, а между разными группами такого бы пути не существовало.

Сопоставим с каждым объектом a_i, $i=1,...,m$ из множества $A=\{a_1,a_2,...,a_m\}$ вершину v_i некоторого графа $G=(V,E)$. С каждым ребром $e_{i,j}=(e_i,e_j)\in E: i,j=1,...,m$, связывающим вершины v_i и v_j, сопоставим абстрактную стоимость $c_{i,j}$, равную расстоянию $d_{a_i}d_{a_j}$ между соответствующими объектами a_i и a_j в матрице расстояний $D=\|d_{a_i}d_{a_j}\|_{m\times m}$. Для начального состояния, с которого алгоритм выделения структур начинает работать, возможны два варианта – начальное состояние есть полный граф или начальное состояние есть пустой граф.

В случае, когда множество объектов определяют начальное состояние в виде полного графа $G=(V,E)$, для каждой пары вершин $(v_i,v_j)\in V: i,j=1,...,m$ существует ребро $e_{i,j}$, инцидентное вершине v_i и инцидентное вершине v_j (все вершины графа соединены ребрами между собой). Последовательное удаление из графа ребер $e_{i,j}$, значения которых больше заданного порога d_p^t, приведет к получению на каждом шаге графа уровня t, равного $G^t=(V^t,E^t)$, для множества ребер которого справедливо $E^t=\{e_{i,j}\in E: c_{i,j}\le d_p^t\}$. Процесс уменьшения порога d_p^t приведет к ситуации, когда при некотором значении d_p^{t*} исходный граф $G=(V,E)$ превратится в несвязный и появится связанное с d_p^{t*} некоторое количество k_t подграфов $G_1^t,G_2^t,...G_{k_t}^t$ – компонентов связности. Достаточно очевидно, что в общем случае число вершин у подграфов $G_1^t,G_2^t,...G_{k_t}^t$ – различное, вплоть до наличия подграфов с одной вершиной, а продолжение процесса уменьшения порога d_p^t будет приводить к увеличению компонентов связности исходного граф $G=(V, E)$, т. е. к увеличению количества связных подграфов.

В случае, когда множество объектов определяют начальное состояние в виде вполне несвязанного графа (пустого графа) $G=(V)$, последовательное добавление в этот граф ребер $e_{i,j}$, значения которых меньше заданного порога d_p^t, также приведет к получению на каждом шаге графа уровня t, равного $G^t=(V^t,E^t)$, для множества ребер которого справедливо $E^t=\{e_{i,j}\in E: c_{i,j}\le d_p^t\}$. А процесс увеличения порога d_p^t приведет к ситуации, когда при некотором значении d_p^{t**} исходный граф $G=(V)$ превратится в связный и появится связанное с d_p^{t**} некоторое количество k_t подграфов $G_1^t,G_2^t,...G_{k_t}^t$ – компонентов связности. Дальнейшее увеличение порога d_p^t приведет к уменьшению компонентов связности $G_1^t,G_2^t,...G_{k_t}^t$ и числа изолированных вершин и при некотором значении порога d_p^{t*} исходный пустой граф $G=(V)$ превратится в связный граф $G=(V,E)$.

Определение. Компонент связности, появляющийся на некотором шаге t и удовлетворяющий априори заданному набору ограничений из совокупности $R=\{A_X,A_\Omega,A_D\}$, называется структурой $S=\{v: v=V_\alpha^t, G_\alpha^t=(V_\alpha^t, E_\alpha^t), c_\alpha^t\le d_p^t\}$.

Каждая структура характеризуется максимальным $d_{\max}^k$ и минимальным $d_{\min}^k$ пороговыми значениями расстояния. Минимальное значение порога $d_{\min}^k$ определяет условия, при которых из компонента связности выходит одна или несколько вершин, т. е. структура $S=\{v: v=V_\alpha^t, G_\alpha^t=(V_\alpha^t,E_\alpha^t), c_\alpha^t\le d_p^t\}$ разрушается. Таким образом, $d_{\min}^k$ характеризует компактность подграфов $G_1^t,G_2^t,...G_{k_t}^t$, причем, чем меньше величина $d_{\min}^k$, тем существеннее в некотором смысле связь между компонентами связности. Максимальное же значение порога $d_{\max}^k$ определяет условия, характеризующие степень связанности k-й структуры с оставшейся частью исходного графа $G=(V,E)$. В частности, по величине $d_{\max}^k=\max(d_{\max}^k)$, $k=1,2,...$ определяются условия выделения первого подграфа из $G=(V,E)$.

Таким образом, разность $d_{\max}^k-d_{\min}^k$ определяет степень изолированности структуры $S=\{v: v=V_\alpha^t, G_\alpha^t=(V_\alpha^t,E_\alpha^t), c_\alpha^t\le d_p^t\}$ от остальной части графа $G=(V,E)$. Причем выбор метрики расстояния d не является ключевым, так как в общем случае априори определить свойства исследуемого пространства достаточно сложно. По крайней мере, метрики Евклида и Хемминга приводили к одинаковым результатам. А с вычислительной точки зрения евклидово расстояние – самое простое.

Результаты исследования

Для иллюстрации работы предложенного алгоритма кластеризации были привлечены средние месячные температуры воздуха 249 станций на территории северной части Евразии за период 1955–2010 гг. Сеть станций – неравномерная, более плотная на юге и западе исследуемой территории и менее плотная в Сибири и на севере. Районирование территории с использованием временных рядов средней месячной температуры имеет своей основной целью нахождение естественных структур, т. е. районов, обладающих похожим поведением температуры. Привязка полученных таким образом структур-классов к ландшафту позволяет приблизиться к оценке причин формирования и выделения однородных зон.

Ниже приведены результаты начальной (рис. 1), промежуточной (рис. 2) и конечной (рис. 3) стадий выделения структур с аномальной составляющей среднемесячной температуры по 90 станциям, находящимся в Сибирской части Северной Евразии.

Рис. 1 иллюстрирует результаты выделения структур при d_p=0,5. Состояние – близкое к начальному.

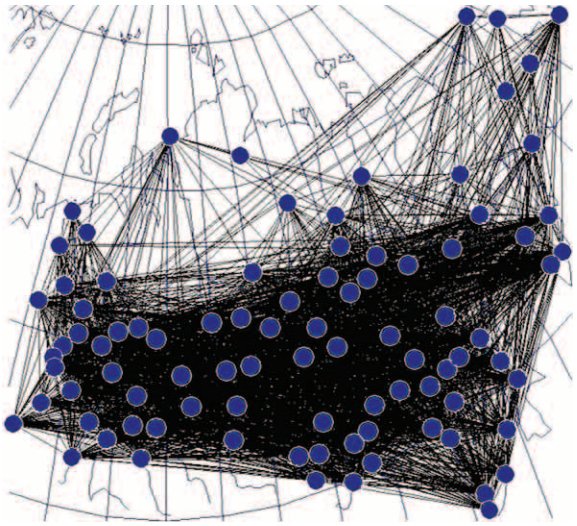


Рис. 1. *Кластеризация при d_p=0,5*

Fig. 1. *Clustering at d_p=0,5*

Дальнейшее уменьшение порогового значения расстояния иллюстрируется на рис. 2: в левой части – выделение в исходном графе двух компонентов связности, в правой части – выделение уже шести компонентов связности.

В финальной части выделения структур (рис. 3) присутствует 20 компонентов связности – 11 подграфов и 9 изолированных точек.

Заключение

Разработанный алгоритм кластеризации, примененный для классификации специфического геофизического поля средних месячных температур воздуха, позволяет выделить в исследуемом пространстве отдельные автономные области и, следовательно, проводить более детальный анализ влияния ландшафта и подстилающей поверхности на их климат.

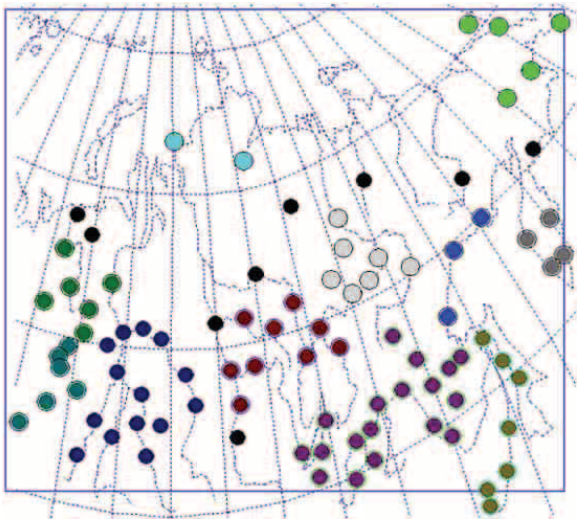
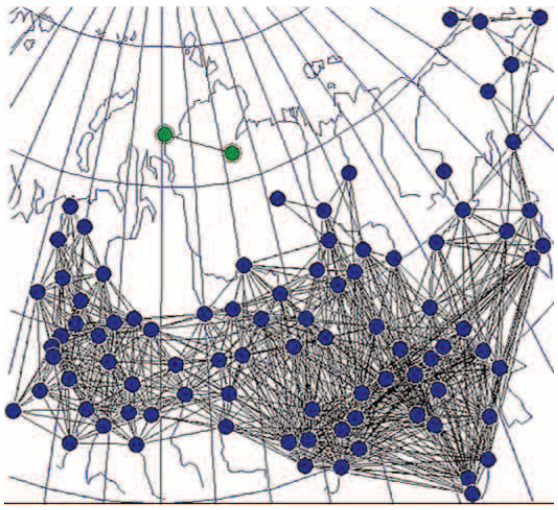


Рис. 3. *Финальная стадия классификации*

Fig. 3. *Final stage of classification*

Кроме того, использование предложенного алгоритма позволяет сформировать области, в которых изменение температуры похожее, а следовательно, и определить тенденции изменения температурных рядов (моменты изменения климата) не только в региональном, но и глобальном масштабах. Вычисление статистических характеристик температур в выделенных структурах (например, дисперсии и др.), дополнение их определенными универсальными характеристиками (например, зависящими от длины связи) дает возможность и количественно сравнивать отдельные структуры между собой.



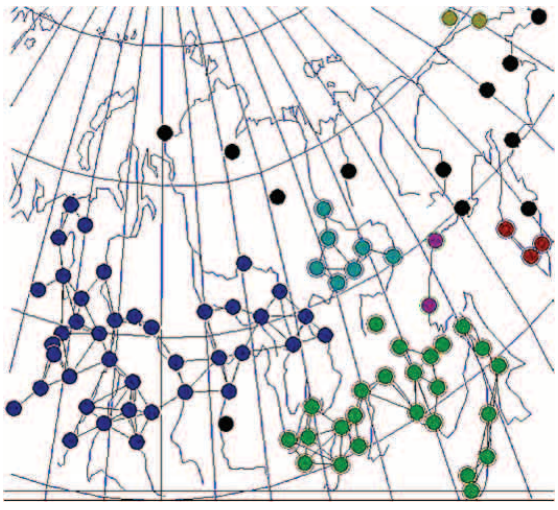


Рис. 2. *Промежуточные стадии классификации*

Fig. 2. *Intermediate stages of classification*

СПИСОК ЛИТЕРАТУРЫ

1. Геологический словарь / под ред. К.Н. Паффенгольца. – М.: Недра, 1978. – Т. 2. – 456 с.
2. Köppen W. Das geographische system der klimate. – Berlin: Verlag von Gebrüder Borntträger, 1936. – 44 p.
3. Хромов С.П., Петросянц М.А. Метеорология и климатология. – М.: Изд-во МГУ, 2004. – 582 с.
4. Григорьев А.А., Будыко М.И. Классификация климатов СССР // Известия АН СССР. Серия геогр. – 1959. – № 3. – С. 58–70.
5. Коробов В.Б., Васильев Л.Ю. Климатическое районирование территорий экспертно-статистическими методами. Постановка задачи // Метеорология и гидрология. – 2004. – № 6. – С. 38–48.
6. Fovell R., Fovell M.-Y. Climate zones of the conterminous United States defined using cluster analysis // American Meteorological Society. – 1993. – № 6. – P. 2103–2135.
7. Mining and heavy metal pollution: assessment of aquatic environments in Tarkwa (Ghana) using multivariate statistical analysis / F.A. Armah, S. Obiri, D.O. Yawson, A.N.M. Pappoe, B. Akoto // Journal of Environmental Statistics. – 2010. – V. 1. – № 4. – P. 1–13.
8. Mrutu A., Luilo G.B. Data mining using multivariate statistical analysis. The case of heavy metals in sediments of the Msimbazi Creek mangrove wetland // Environmental Skeptics and Critics. – 2013. – V. 2 (4). – P. 153–163.
9. Boyles R., Raman S. Analysis of climate trends in North Carolina (1949–1998) // Environment International. – 2003. – V. 29. – P. 263–275.
10. A local search approximation algorithm for k-means clustering / T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu. // Computational Geometry: Theory and Applications. – 2004. – № 28. – P. 89–112.
11. Kalkstein L.S., Tan G., Skindlov J.A. An evaluation of three clustering procedures for use in synoptic climatological classification // Journal of Climate and Applied Meteorology. – 1987. – V. 26. – P. 717–730.
12. Bretherton C.S., Smith C., Wallace J.M. An intercomparison of methods for finding coupled patterns in climate data // J. of Climate. – 1992. – V. 5. – P. 541–560.
13. Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988 / F. D'Andrea, S. Tibaldi, M. Blackburn, G. Boer, M. Deque, M.R. Dix, B. Dugas, L. Ferranti, T. Iwasaki, A. Kitoh, V. Pope, D. Randall, E. Roeckner, D. Straus, W. Stern, H. Van Den Dool, D.L. Williamson // Climate Dynamics. – 1998. – V. 14. – P. 385–407.
14. Катаев С.Г., Кусков А.И. Исследование озонных полей над территорией России и сопредельных государств. I. Составляющие полей озона // Вестник ТГПУ. Сер.: Естественные и точные науки. – 1998. – Вып. 5. – С. 3–9.
15. Катаев С.Г., Кусков А.И. Исследование озоновых полей над территорией России и сопредельных государств. II. Классификация составляющих полей озона // Вестник ТГПУ. Сер.: Естественные и точные науки. – 1998. – Вып. 1. – С. 10–17.
16. Кусков А.И., Катаев С.Г. Закономерности современных изменений теплового поля в приземном слое атмосферы Сибири и на Дальнем Востоке // Известия вузов. Физика. – 2004. – № 11. – С. 81–92.
17. Кусков А.И., Катаев С.Г. Структура и динамика приземного температурного поля над азиатской территорией России. – Томск: Изд-во ТГПУ, 2006. – 176 с.
18. Wu A.M., Hsieh W.W., Zwiers F.W. Nonlinear modes of North American winter climate variability derived from a general circulation model simulation // J. of Climate. – 2003. – V. 16. – P. 2325–2339.
19. Wu A.M., Hsieh W.W. Nonlinear interdecadal changes of the El Nino-Southern oscillation // Climate Dynamics. – 2003. – V. 21. – P. 719–730.
20. Wu A.M., Hsieh W.W. Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature // Climate Dynamics. – 2002. – V. 19. – P. 713–722.
21. Barnett T.P., Preisendorfer R. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis // Monthly Weather Review. – 1987. – V. 115 (9). – P. 1825–1850.
22. Roswintiarti O., Niyogi D.S., Raman S. Tele-connections between tropical Pacific sea surface temperature anomalies and North Carolina precipitation anomalies during El Niño events // Geophys. Res. Lett. – 1998. – V. 25. – P. 4201–4204.
23. Correlation study of time-varying multivariate climate data sets / J. Sukharev, C. Wang, K.L. Ma, A.T. Wittenberg // Proceeding of IEEE VGTC Pacific Visualization Symposium. – Beijing, 2009. – P. 161–168.
24. Cannon A.J., Hsieh W.W. Robust nonlinear canonical correlation analysis: application to seasonal climate forecasting Nonlin // Processes Geophys. – 2008. – V. 15. – P. 221–232.
25. Livezey R.E., Smith T.M. Considerations for use of the Barnett and Preisendorfer algorithm for canonical correlation analysis of climate variations // J. of Climate. – 1999. – V. 12. – P. 303–305.
26. Barnett T.P., Preisendorfer R. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis // Mon. Weather Rev. – 1987. – V. 115. – P. 1825–1850.
27. Barnston A.G., Ropelewski C.F. Prediction of ENSO episodes using canonical correlation-analysis // J. of Climate. – 1992. – V. 5. – P. 1316–1345.
28. The NCEP climate forecast system / S. Saha, S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H.M. van den Dool, H.-L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Pena, S. Lord, G. White, W. Ebisuzaki, P. Peng, P. Xie // J. of Climate. – 2006. – V. 19. – P. 3483–3517.
29. Seasonal climate: forecasting and managing risk / A. Troccoli, M. Harrison, D.L.T. Anderson, S.J. Mason. – Dordrecht: Springer Science, 2008. – 461 p.
30. Greenfield R.S., Fisher G.M. Improving responses to climate predictions – an introduction // Bull. Amer. Meteorol. Soc. – 2003. – V. 84. – P. 1685–1685.
31. Harrison M.S. The development of seasonal and interannual climate forecasting // Climatic Change. – 2005. – V. 70. – P. 201–220.
32. Multimodel ensemble forecasts for weather and seasonal climate / T.N. Krishnamurti, C.M. Kishtawal, Zh. Zhang, T. Larow, D. Bachiochi, E. Williford // J. of Climate. – 2000. – V. 13. – P. 4196–4216.
33. Kim K.Y., North G.R. EOF-based linear prediction algorithm: examples // J. of Climate. – 1999. – V. 12. – P. 2076–2092.
34. Kim K.Y., Wu Q. A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics // J. of Climate. – 1999. – V. 12. – P. 185–199.
35. Forecasting ENSO events: a neural network extended EOF approach / F.T. Tangang, B.Y. Tang, A.H. Monahan, W.W. Hsieh // J. of Climate. – 1998. – V. 11. – P. 29–41.
36. Поляков Д.В., Кужевская И.В. Применение кластерного анализа для оценки температурно-влажностных условий в период активной вегетации на территории юга Западной Сибири и его связь с гидротермическим коэффициентом Т.Г. Селянинова // Вестник Том. гос. ун-та. – 2012. – № 360. – С. 188–192.
37. Ветрова Е.И., Скриптунова Е.Н., Шакина Н.П. Прогноз низкой облачности на аэродромах европейской территории бывшего СССР // Метеорология и гидрология. – 2013. – № 1. – С. 12–31.
38. Овечкин С.В., Майнашева Г.М. Опыт использования кластерного анализа при климатическом районировании Московской области // Вестник МГПУ. Серия: Естественные науки. – 2010. – № 2 (6). – С. 65–74.

39. Вершовский Е.А. Разработка методов и алгоритмов кластеризации мультиспектральных данных дистанционного зондирования Земли: автореф. дисс. ... канд. техн. наук. – Таганрог, 2010. – 17 с.
40. Кирста Ю.Б., Курепина Н.Ю., Ловцкая О.В. Пространственная декомпозиция метеорологических полей Евразии: разделение воздействий растительности и антропогенной деятельности // Фундаментальные исследования. – 2014. – № 5. – С. 1030–1036.
41. Дробушевская О.В., Царегородцев В.Г. Географо-климатические варианты светлохвойных травяных лесов Сибири // Сибирский экологический журнал. – 2007. – № 2. – С. 211–219.
42. Родригес Залепинос Р.А. Данные и методы интеллектуального анализа данных для исследования окружающей природной среды // Системный анализ и информационные технологии в науках о природе и обществе. – 2011. – Вып. 1. – С. 94–107.
43. Нейский И.М. Классификация и сравнение методов кластеризации // Научно-образовательный кластер CLAIM. URL: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (дата обращения: 03.06.2015).

Поступила 17.11.2015 г.

UDC 551.501

APPROACH TO CLUSTERING OBJECTS

Igor A. Botygin,
National Research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, 634050, Russia. E-mail: bia@tpu.ru

Sergey G. Kataev,
Tomsk State Pedagogical University, 60, Kievskaya Street, Tomsk, 634061, Russia. E-mail: sgkataev@sibmail.com

Valeriy A. Tartakovskiy,
Institute of Monitoring of Climatic and Ecological Systems SB RAS, 10/3, Academichesky Avenue, Tomsk, 634055, Russia. E-mail: trtk@list.ru

Anna I. Sherstneva,
National Research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, 634050, Russia. E-mail: sherstneva@tpu.ru

Relevance of the work *is due to the need to develop universal information-analytical approaches to extract knowledge from the rapidly growing volume of geophysical data. One of the main problems in the processing of geophysical data is to find in it objectively existing laws which could become the basis for diverse, including forward-looking, behavior models of selected parameters of geophysical fields. And that data clustering technologies are the foundation for the software development of similar information systems analysis of unstructured data.*
The main aim of the study *is to develop a method of experimental data clustering of geophysical nature on the basis of allocation of structures for solving problems of the analysis of unstructured information when studying and controlling complex systems.*
The methods used in the study: *classical and modern methods and clustering algorithms, graph theory algorithms, test case of clustering of a geophysical field of meteorological parameters from the territory of the northern part of Eurasia.*
The results. *The authors developed a new algorithm of structures allocation in the initial geophysical field, which allows decomposing the test space into fields with the same behavior of the studied parameters based on spatial characteristics. The algorithm is based on the structuring of the various expansions of geophysical fields (season, anomaly, etc.) and provides a wide range of information on the object in the form of sets of parameters of the selected structures. This information, along with the accompanying empirical relationship between the parameters is considered as a generalization of the experimental characterization of the object and is the basis for the formation of hypotheses and behavior models. In addition, a structural model of the space of a meteorological parameter provides the ability to compress primary data without significant loss of semantic value of the target geophysical field.*

Key words:
Geophysical field, clustering, graph theory, structure, method of allocation of structures, weather observations, time series.

REFERENCES

1. *Geologicheskiy slovar* [Geological dictionary]. Ed. by K.N. Paffengolts. Moscow, Nedra Publ., 1978, vol. 2. 456 p.
2. Köppen W. *Das geographische system der klimate*. Berlin, Verlag von Gebrüder Bornträger, 1936. 44 p.
3. Khromov S.P., Petrosyants M.A. *Meteorologiya i klimatologiya* [Meteorology and climatology]. Moscow, Moscow State University Publ., 2004. 582 p.
4. Grigorev A.A., Budyko M.I. Klassifikatsiya klimatov SSSR [Climate classification of USSR]. *Izvestiya AN SSSR. Seriya geogr.*, 1959, no. 3, pp. 58–70.
5. Korobov V.B., Vasilev L.Yu. Klimaticheskoe rayonirovanie territoriy ekspertno-statisticheskimi metodami. Postanovka zadachi [Climatic zoning of territories by expert and statistical methods. Formulation of the problem]. *Russian meteorology and hydrology*, 2004, no. 6, pp. 38–48.
6. Fovell R., Fovell M.-Y. Climate zones of the conterminous United States defined using cluster analysis. *American Meteorological Society*, 1993, no. 6, pp. 2103–2135.
7. Armah F.A., Obiri S., Yawson D.O., Pappoe A.N.M., Akoto B. Mining and heavy metal pollution. Assessment of aquatic environments in Tarkwa (Ghana) using multivariate statistical analysis. *Journal of Environmental Statistics*, 2010, vol. 1, no. 4, pp. 1–13.
8. Mrutu A., Luilo G.B. Data mining using multivariate statistical analysis. The case of heavy metals in sediments of the Msimbazi Creek mangrove wetland. *Environmental Skeptics and Critics*, 2013, vol. 2 (4), pp. 153–163.
9. Boyles R., Raman S. Analysis of climate trends in North Carolina (1949–1998). *Environment International*, 2003, vol. 29, pp. 263–275.
10. Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., Silverman R., Wu A.Y. A local search approximation algorithm for k-means clustering. *Computational Geometry: Theory and Applications*, 2004, no. 28, pp. 89–112.
11. Kalkstein L.S., Tan G., Skindlov J.A. An evaluation of three clustering procedures for use in synoptic climatological classification. *Journal of Climate and Applied Meteorology*, 1987, vol. 26, pp. 717–730.
12. Bretherton C.S., Smith C., Wallace J.M. An intercomparison of methods for finding coupled patterns in climate data. *J. of Climate*, 1992, vol. 5, pp. 541–560.
13. Andrea F.D., Tibaldi S., Blackburn M., Boer G., Deque M., Dix M.R., Dugas B., Ferranti L., Iwasaki T., Kitoh A., Pope V., Randall D., Roeckner E., Straus D., Stern W., Van den Dool H., Williamson D.L. Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979–1988. *Climate Dynamics*, 1998, vol. 14, pp. 385–407.
14. Kataev S.G., Kuskov A.I. Issledovanie ozonnykh poley nad territoriey Rossii i sopredelnykh gosudarstv. I. Sostavlyayushchie poley ozona [Ozone research fields over the territory of Russia and neighboring countries. I. Components of ozone fields]. *Tomsk State Pedagogical University Bulletin. Natural Sciences*, 1998, iss. 5, pp. 3–9.
15. Kataev S.G., Kuskov A.I. Issledovanie ozonovykh poley nad territoriey Rossii i sopredelnykh gosudarstv. II. Klassifikatsiya sostavlyayushchikh poley ozona [Ozone research fields over the territory of Russia and neighboring countries. II. Classification of components of ozone fields]. *Tomsk State Pedagogical University Bulletin. Natural Sciences*, 1998, iss. 1, pp. 10–17.
16. Kuskov A.I., Kataev S.G. Zakonomernosti sovremennykh izmeneniy teplovogo polya v prizemnom sloe atmosfery Sibiri i na Dalnem Vostoke [Laws of modern changes of the thermal field in the surface layer the atmosphere of Siberia and the Far East]. *Russian Physics Journal*, 2004, no. 11, pp. 81–92.
17. Kuskov A.I., Kataev S.G. *Struktura i dinamika prizemnogo temperaturnogo polya nad aziatskoy territoriey Rossii* [Structure and dynamics of the surface temperature field over the Asian territory of Russia]. Tomsk, TSPU Publ., 2006. 176 p.
18. Wu A.M., Hsieh W.W., Zwiers F.W. Nonlinear modes of North American winter climate variability derived from a general circulation model simulation. *J. of Climate*, 2003, vol. 16, pp. 2325–2339.
19. Wu A.M., Hsieh W.W. Nonlinear interdecadal changes of the El Nino-Southern oscillation. *Climate Dynamics*, 2003, vol. 21, pp. 719–730.
20. Wu A.M., Hsieh W.W. Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature. *Climate Dynamics*, 2002, vol. 19, pp. 713–722.
21. Barnett T.P., Preisendorfer R. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, 1987, vol. 115 (9), pp. 1825–1850.
22. Roswintiarti O., Niyogi D.S., Raman S. Tele-connections between tropical Pacific sea surface temperature anomalies and North Carolina precipitation anomalies during El Niño events. *Geophys. Res. Lett.*, 1998, vol. 25, pp. 4201–4204.
23. Sukharev J., Wang C., Ma K.L., Wittenberg A.T. Correlation study of time-varying multivariate climate data sets. *Proceeding of IEEE VGTC Pacific Visualization Symposium*, Beijing, 2009. pp. 161–168.
24. Cannon A.J., Hsieh W.W. Robust nonlinear canonical correlation analysis: application to seasonal climate forecasting Nonlin. *Processes Geophys.*, 2008, vol. 15, pp. 221–232.
25. Livezey R.E., Smith T.M. Considerations for use of the Barnett and Preisendorfer algorithm for canonical correlation analysis of climate variations. *J. of Climate*, 1999, vol. 12, pp. 303–305.
26. Barnett T.P., Preisendorfer R. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.*, 1987, vol. 115, pp. 1825–1850.
27. Barnston A.G., Ropelewski C.F. Prediction of ENSO episodes using canonical correlation-analysis. *J. of Climate*, 1992, vol. 5, pp. 1316–1345.
28. Saha S., Nadiga S., Thiaw C., Wang J., Wang W., Zhang Q., van den Dool H.M., Pan H.L., Moorthi S., Behringer D., Stokes D., Pena M., Lord S., White G., Ebisuzaki W., Peng P., Xie P. The NCEP climate forecast system *J. of Climate*, 2006, vol. 19, pp. 3483–3517.
29. Troccoli A., Harrison M., Anderson D.L.T., Mason S.J. *Seasonal climate: forecasting and managing risk*. Dordrecht, Springer Science, 2008. 461 p.
30. Greenfield R.S., Fisher G.M. Improving responses to climate predictions an introduction. *Bull. Amer. Meteorol. Soc.*, 2003, vol. 84, pp. 1685–1685.
31. Harrison M.S. The development of seasonal and interannual climate forecasting. *Climatic Change*, 2005, vol. 70, pp. 201–220.
32. Krishnamurti T.N., Kishtawal C.M., Zhang Z., Larow T., Bachiochi D., Williford E. Multimodel ensemble forecasts for weather and seasonal climate. *J. of Climate*, 2000, vol. 13, pp. 4196–4216.
33. Kim K.Y., North G.R. EOF-based linear prediction algorithm: examples. *J. of Climate*, 1999, vol. 12, pp. 2076–2092.
34. Kim K.Y., Wu Q. A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics. *J. of Climate*, 1999, vol. 12, pp. 185–199.
35. Tangang F.T., Tang B.Y., Monahan A.H., Hsieh W.W. Forecasting ENSO events: a neural network extended EOF approach. *J. of Climate*, 1998, vol. 11, pp. 29–41.
36. Polyakov D.V., Kuzhevskaya I.V. Primenenie klasternogo analiza dlya otsenki temperaturno-vlazhnostnykh usloviy v period aktivnoy vegetatsii na territorii yuga Zapadnoy Sibiri i ego svyaz s gidrotermicheskim koeffitsientom T.G. Selyaninova [Application of cluster analysis to estimate the temperature and humidity con-

ditions during the active growing season in the south of Western Siberia and its relation to hydrothermal coefficient of T.G. Selyaninov]. *Bulletin of Tomsk State University*, 2012, no. 360, pp. 188–192.

37. Vetrova E.I., Skriptunova E.N., Shakina N.P. Prognoz nizkoy oblachnosti na aerodromakh evropeyskoy territorii byvshego SSSR [The forecast of low cloud at aerodromes of European territory of the former USSR]. *Russian meteorology and hydrology*, 2013, no. 1, pp. 12–31.
38. Ovechkin S.V., Maynasheva G.M. Opyt ispolzovaniya klasternogo analiza pri klimaticheskom rayonirovanii Moskovskoy oblasti [Experience in the use of cluster analysis in the climatic zone of the Moscow region]. *Bulletin of Moscow State Pedagogical University. Natural Sciences*, 2010, no. 2 (6), pp. 65–74.
39. Vershovskiy E.A. *Razrabotka metodov i algoritmov klasterizatsii multispektralnykh dannykh distantsionnogo zondirovaniya Zemli*. Avtoref. Kand. nauk [Development of methods and algorithms for clustering multispectral data of remote sensing of Earth. Author's abstract Cand. Diss.]. Taganrog, 2010. 17 p.
40. Kirsta Yu.B., Kurepina N.Yu., Lovtskaya O.V. Prostranstvennaya dekompozitsiya meteorologicheskikh poley Evrazii: razdelenie vozdeystviy rastitelnosti i antropogennoy deyatelnosti [The spatial decomposition of meteorological fields in Eurasia: the separation effects of vegetation and human activities]. *Fundamental Research*, 2014, no. 5, pp. 1030–1036.
41. Drobushevskaya O.V., Tsaregorodtsev V.G. Geografo-klimaticheskie varianty svetlokhvoynykh travyanykh lesov Sibiri [Geographic and climatic variations of herbal coniferous forests of Siberia]. *Contemporary Problems of Ecology*, 2007, no. 2, pp. 211–219.
42. Rodriges Zalepinos R.A. Dannye i metody intellektualnogo analiza dannykh dlya issledovaniya okruzhayushchey prirodnoy sredy [Data and data mining techniques for the study of the natural environment]. *System analysis and information technologies in the sciences of nature and society*, 2011, iss. 1, pp. 94–107.
43. Neyskiy I.M. Klassifikatsiya i sravnenie metodov klasterizatsii [Classification and comparison of clustering methods]. *Nauchno-obrazovatenyy klaster CLAIM*. Available at: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (accessed 3 June 2015).

Received: 17 November 2015.