

УДК 53.05, 519.6, 004.02

ОЦЕНКА УСТОЙЧИВОСТИ АЛГОРИТМА ДИНАМИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ТЕМПЕРАТУРНЫХ СИГНАЛОВ

© Ю. В. Волков

*Институт мониторинга климатических и экологических систем СО РАН,
634055, г. Томск, просп. Академический, 10/3
E-mail: yvvolkov@mail.ru*

Рассмотрен алгоритм динамической кластеризации температурных сигналов, используемый для решения задачи выделения климатических областей на поверхности Земли, характеризующихся определённым однородным в границах области типом климата. Дано описание численного эксперимента, реализованного в целях оценки его устойчивости. Модель эталонного сигнала, использованная в численном эксперименте, представлена в виде суммы гармонических составляющих. Аддитивные составляющие шума формировались в частотной области. В численном эксперименте изменялись величина шума и размер выборки. Приведены оценки среднеквадратических отклонений и устойчивости алгоритма динамической кластеризации.

Ключевые слова: анализ сигналов, математическая модель, приземная температура, классификация климата, кластеризация, устойчивость алгоритма.

DOI: 10.15372/AUT20200604

Введение. Информация о современной среде обитания, экологическое состояние которой характеризуется отдельными компонентами геосистемы планеты, востребована для эффективного социально-экономического планирования. В последние десятилетия особое внимание уделяется глобальным климатическим изменениям, сопровождающимся трансформацией полей отдельных характеристик климата, что может стать причиной изменения существующих биоклиматических и сельскохозяйственных зон [1]. Актуальная задача климатического районирования расширяется до необходимости получения классификаций климата в оперативном режиме на основе постоянного объективного анализа изменений. Для решения этой задачи требуется разработка новых подходов динамической классификации климата.

Существуют различные классификации, определяющие границы распространения отдельных типов и видов климата в соответствии с выделенными природно-климатическими закономерностями. Классы в одной из наиболее используемой в мире классификации климата В. П. Кепена [2] выделены на основе количественных критериев тепловлажностного режима с учётом ландшафтных особенностей территории. Классификация климатов Л. С. Берга [3] основана на учёте ландшафтно-географических зон суши. Границы климатических зон в генетической классификации Б. П. Алисова [4] определяются по среднему положению климатических фронтов. В основе этой классификации лежит учёт условий формирования климата в зависимости от типов воздушных масс и их циркуляции.

Классические подходы к районированию климата [2–4] получены для некоторых средних климатических условий в период доказанных климатических изменений, являются сложными для реализации и не могут дать оперативную оценку изменений, что является актуальной задачей в связи с существующими тенденциями к усилению климатических изменений [5–10]. Таким образом, повышается научная и практическая значимость любой климатической структуризации, реализующей формализованные подходы и позволяющей исследовать в динамике региональные особенности изменения глобального климата.

В работах [11–14], развивающих классификации климата по оценкам измеренных климатических параметров, применяются методы, пригодные для организации оперативного мониторинга климатического районирования. Алгоритмы используют параметры температурных сигналов, полученных на метеостанциях. К недостаткам нужно отнести отсутствие оценок устойчивости работы и точности предлагаемых алгоритмов классификации в условиях глобальных климатических изменений.

Целью данной работы является получение оценок устойчивости работы алгоритма [14, 15], использующего синхронность как основное свойство климатических кластеров, и оценок среднеквадратических отклонений (СКО) полученных результатов.

Алгоритм динамической кластеризации. В качестве исходных данных для алгоритма динамической кластеризации (АДК) используются характеристики температурных сигналов, представленные огибающими или функциями фаз. Данные характеристики образуют множество $M_{\tau,k}^l$, номера элементов которого соответствуют метеостанциям с номерами $k = 1, 2, \dots, N$ или $j = 1, 2, \dots, N$, $N = 818$, где τ — время, l — порядковый номер итерации. Процедура кластеризации состоит в вычислении корреляционной матрицы $C_{j,k}$ элементов одинаковых множеств $M_{\tau,j}^l$ и $M_{\tau,k}^l$.

Расчёт значений взаимной корреляции Пирсона можно представить в следующем виде:

$$C_{j,k} = \frac{\sum_{\tau=1}^T \left(M_{\tau,j} - \frac{1}{T} \sum_{\tau=1}^T M_{\tau,j} \right) \left(M_{\tau,k} - \frac{1}{T} \sum_{\tau=1}^T M_{\tau,k} \right)}{\left[\sum_{\tau=1}^T \left(M_{\tau,j} - \frac{1}{T} \sum_{\tau=1}^T M_{\tau,j} \right)^2 \left(\sum_{\tau=1}^T M_{\tau,k} - \frac{1}{T} \sum_{\tau=1}^T M_{\tau,k} \right)^2 \right]^{1/2}},$$

где T — количество временных отсчётов.

Далее производится разбиение множества $M_{\tau,j}^l$ на кластеры. Каждый кластер формируется из элементов, соответствующих условию $C_{j,k} = \langle M_j^l, M_k^l \rangle \geq r$, где $r = \text{const}$ — требуемый уровень взаимной корреляции. В соответствии с заданным условием создаётся бинарный массив значений элементов

$$S_{j,k} = \begin{cases} 0, & \text{если } C_{j,k} < r, \\ 1, & \text{если } C_{j,k} \geq r. \end{cases}$$

Для каждого элемента множества $M_{\tau,j}^l$, попавшего в один из кластеров $S_{j,k}$, определяется текущая характеристика, которая вычисляется как средняя оценка всех элементов, вошедших в кластер:

$$M_{\tau,j}^{l+1} = \sum_{k=1}^N (M_{\tau,k}^l S_{j,k}) / \sum_{k=1}^N (S_{j,k}) \quad \text{при} \quad \sum_{k=1}^N (S_{j,k}) > 0.$$

Формируется новое множество $M_{\tau,j}^{l+1}$, элементами которого являются рассчитанные текущие характеристики.

Далее определяется уровень отличия между массивами путём расчёта абсолютной разницы между элементами массивов $M_{\tau,j}^l$ и $M_{\tau,j}^{l+1}$. Итерационный процесс останавливается при выполнении следующего условия:

$$\max |M_{\tau,j}^l - M_{\tau,j}^{l+1}| \leq \varepsilon, \quad (1)$$

где ε — величина отклонения, равная 0,001.

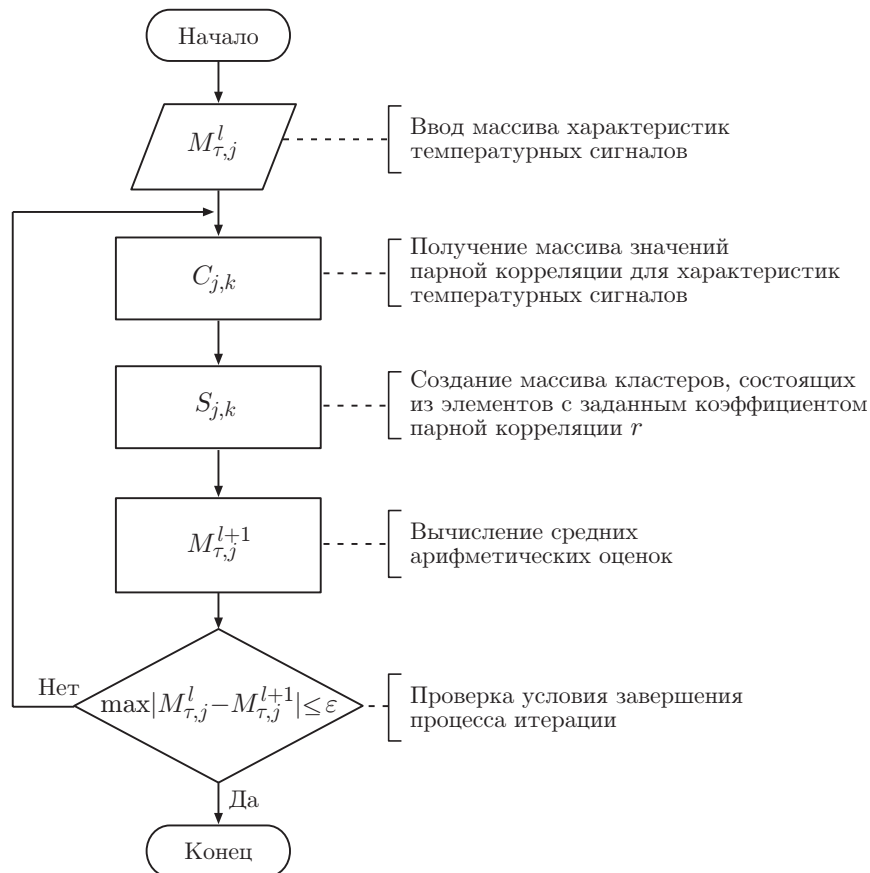


Рис. 1. Блок-схема алгоритма динамической кластеризации

В случае превышения максимального значения абсолютной разницы между элементами массивов $M_{\tau,j}^l$ и $M_{\tau,j}^{l+1}$ величины отклонения ϵ повторяется итерационный процесс, начинающийся с расчёта вычисления корреляционной матрицы $C_{j,k}$ для элементов массива $M_{\tau,j}^{l+1}$. При выполнении условия (1) итерационный процесс завершается, выделяются типовые функции, характеризующие усреднённые климатические закономерности. Элементы массива $M_{\tau,j}^l$, структурированные вокруг полученных типовых закономерностей в соответствии с заданным уровнем корреляции r , формируют климатические кластеры. Разработанный алгоритм (рис. 1) можно использовать для проведения оценки регионального климата в разных временных и пространственных масштабах. На его основе может быть построена интерактивная система оперативного мониторинга климатических зон.

Численный эксперимент. Оценка устойчивости работы алгоритма динамической кластеризации проводится при изменении объёма выборки, состоящей из характеристик температурных сигналов, и при изменении уровней шумов, накладываемых на эталонные сигналы.

Изучаемые процессы, подлежащие исследованию, формируются под воздействием большого числа независимых случайных факторов при незначительном воздействии каждого из них. Основными задачами алгоритма динамической кластеризации являются выделение сигналов в отдельные кластеры по признаку синхронного поведения и восстановление исходных закономерностей, лежащих в основе изучаемых процессов.

В численном эксперименте создана имитация влияния независимых случайных факторов, накладываемых в виде шумов на эталонный сигнал, с помощью которой можно

получить оценки, характеризующие работоспособность численного алгоритма кластеризации. Реализация численного эксперимента позволяет сделать выводы об эффективности, устойчивости и о наличии ограничений для применения предложенного алгоритма.

Исследование проводится по фазовым характеристикам температурных сигналов. Модельный сигнал фазовой функции состоит из несущего сигнала и имеет модуляцию фазы. В большинстве случаев модулирующие сигналы являются сложными функциями времени, их можно представить в виде конечной или бесконечной суммы гармонических составляющих. В качестве модели сигнала полной фазы используется функция вида

$$\Phi(\tau) = \varphi_n(\tau) + \sum_{i=1}^N I_i^\Phi \varphi_i(\tau) = \varphi_n(\tau) + \sum_{i=1}^N I_i^\Phi \cos(2\pi\nu_i\tau), \quad (2)$$

где τ — время, $\varphi_n(\tau)$ — фаза несущего колебания, I_i^Φ — индекс фазовой модуляции (определяет максимальное отклонение фазы), $\varphi_i(\tau)$ — сигнал, модулирующий фазу, ν_i — частота сигнала, модулирующего фазу, i — порядковый номер гармоники, $i = 1, 2, \dots, N$, N — количество модулирующих гармоник.

Для реализации алгоритма кластеризации при построении модели эталонного сигнала фазы из полной фазовой функции (2) вычитается фаза несущего колебания

$$\Phi'(\tau) = \Phi(\tau) - \varphi_n(\tau) = \sum_{i=1}^N I_i^\Phi \cos(2\pi\nu_i\tau).$$

Форма эталонного модулирующего сигнала варьируется в частотной области путём изменения количества гармоник N . Для эксперимента сформированы эталонные сигналы со следующим количеством гармоник: $N = 4, 6, 12, 24$. Частоты гармоник заданы в интервале от 5 до 30 частотных полос. Частотный интервал выбран исходя из анализа спектра огибающих и фаз характерных температурных сигналов [14].

Сигнал, соответствующий шуму, сформирован также в частотной области в интервале от 5 до 30 частотных полос. Интервал частот наложения шума выбран равным интервалу частот основных гармонических составляющих эталонного сигнала. Для генерации шума применён датчик случайных чисел программы Mathcad, который создаёт вектор из заранее заданного количества случайных чисел, имеющих равномерное распределение. Сигнал с шумом сформирован как исходный сигнал, содержащий аддитивную составляющую шума. Уровень шума определяется как отношение СКО шума к СКО исходного сигнала:

$$P_\varphi = \left[\sum_{i=1}^N (S_i - \bar{S})^2 / \sum_{i=1}^N (\Phi'_i - \bar{\Phi}')^2 \right]^{1/2} \cdot 100 \%,$$

где S — сигнал-шум, \bar{S} — среднеарифметическое значение для сигнала-шума, Φ' — эталонная фаза, $\bar{\Phi}'$ — среднеарифметическое значение эталонной фазы.

Количество эталонных сигналов для одновременной реализации алгоритма кластеризации выбрано равным 4. Для каждого эталонного сигнала сформирована выборка из 25 эталонных сигналов с шумом. Таким образом, общий массив данных сформирован из 100 сигналов с шумом.

Алгоритм динамической кластеризации применён одновременно ко всему общему массиву сигналов. В результате реализации АДК получены восстановленные сигналы — вос-

становленные функции фаз, для которых рассчитываются оценки нормированных среднеквадратических отклонений

$$E = \left[\sum_{i=1}^N (\Delta\Phi_i - \bar{\Delta\Phi})^2 / \sum_{i=1}^N (\Phi'_i - \bar{\Phi}')^2 \right]^{1/2} \cdot 100 \%,$$

где $\Delta\Phi$ — разность между эталонной и восстановленной фазами: $\Delta\Phi = \Phi' - \Phi^T$, Φ^T — восстановленная фаза; $\bar{\Delta\Phi}$ — среднеарифметическое значение разности фаз, $\bar{\Phi}$ — среднеарифметическое значение эталонной фазы, N — количество отсчётов.

В численном эксперименте изменяемыми параметрами являются уровень шума (в процентах от эталонного сигнала) и размер выборки (количество сигналов с шумом, приходящееся на один эталонный сигнал). Постоянные параметры эксперимента: ширина спектрального окна шума в 25 частот в интервале от 5 до 30, количество эталонных сигналов $K_i = 4$. Эталонные сигналы сформированы таким образом, чтобы максимальное значение коэффициента корреляции между ними не превышало 0,1. Значение коэффициента корреляции для реализации алгоритма динамической кластеризации задано $r \geq 0,8$.

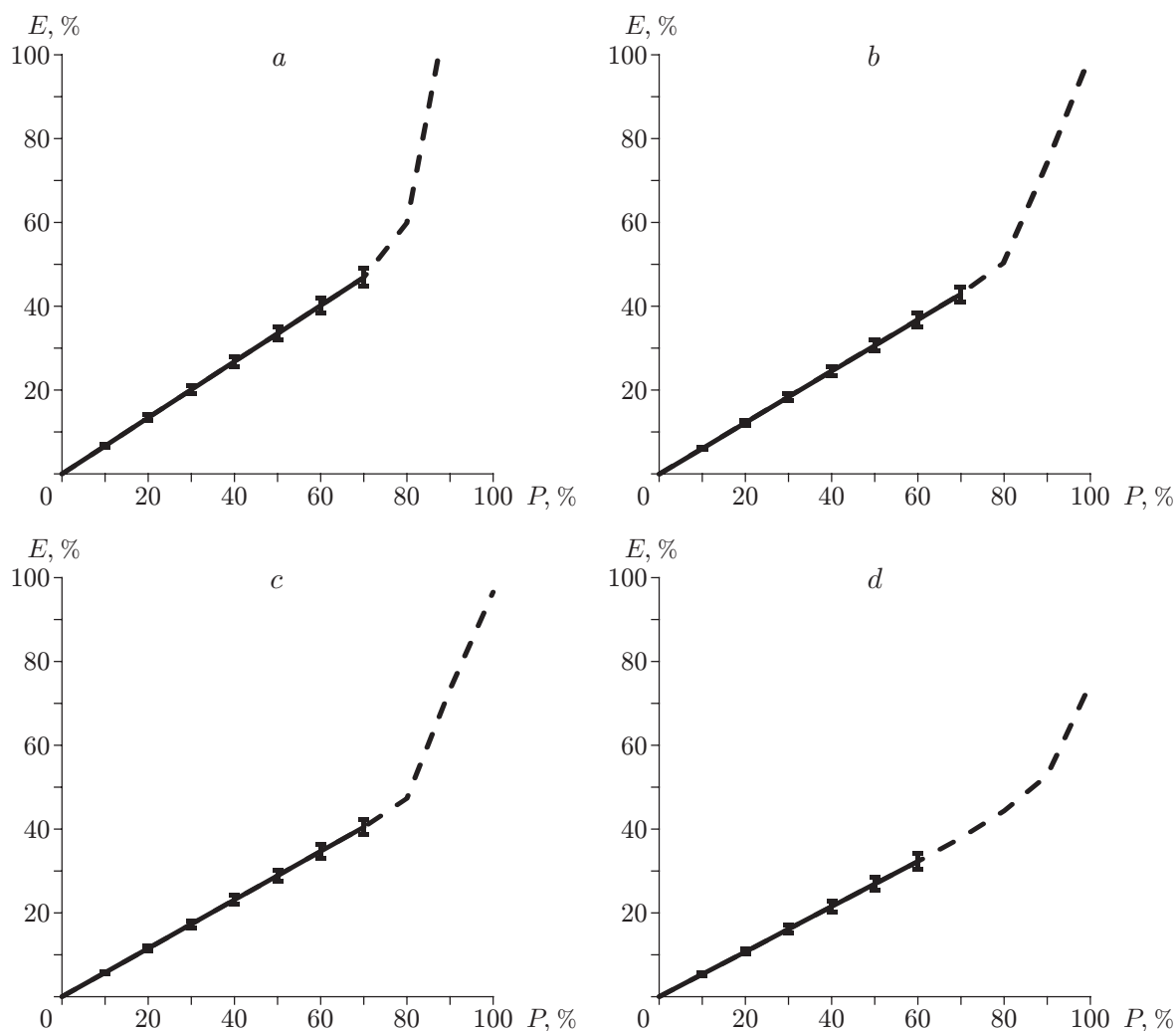


Рис. 2. Оценки нормированной среднеквадратической ошибки восстановления фаз алгоритмом динамической кластеризации при разном уровне шума и количестве сигналов: a — 20, b — 40, c — 60, d — 100

Результаты численного эксперимента. В численном эксперименте выборка сформирована таким образом, что на один эталонный сигнал приходится 5, 10, 15, 25 сигналов с шумом. Таким образом, размер выборки имеет значения 20, 40, 60, 100 сигналов с шумом. Отношение шума к сигналу варьируется от 0 до 100 %.

В результате применения алгоритма динамической кластеризации к сформированным массивам данных получены результаты (рис. 2), из которых видно, что при всех размерах выборки с увеличением уровня шума растёт среднее значение нормированного среднеквадратического отклонения. При этом чем больше сигналов в выборке, тем ниже среднее значение нормированного среднеквадратического отклонения, вычисленное при устойчивой работе алгоритма. Условием устойчивой работы алгоритма является равенство между количеством восстановленных K_r и количеством заданных эталонных K_i фазовых функций при малом значении ошибки E . Критерии устойчивой работы алгоритма: $E \leq 1,1P$ и $S = K_r/K_i = 1$. В зависимости от величины шума в соответствии с критериями устойчивости определяется диапазон устойчивой работы алгоритма. На графиках (см. рис. 2) диапазон устойчивой работы алгоритма обозначен сплошными линиями. Результат, при котором количество восстановленных сигналов не равно количеству эталонных, не может быть отнесён к диапазону устойчивой работы алгоритма, так как в этом случае возникает неоднозначность при выборе действительных сигналов. В данной работе определение действительных сигналов реализовано по наибольшему значению коэффициента корреляции между эталонными и восстановленными сигналами. На графиках диапазон неустойчивой работы алгоритма обозначен штриховыми линиями.

Устойчивая работа алгоритма сохраняется при размере выборки 20, 40, 60 сигналов до отношения шума к сигналу 70 %. При размере выборки, равной 100 сигналам, алгоритм устойчив до отношения шума к сигналу 60 %. Значение среднего нормированного СКО при устойчивой работе алгоритма не превышает 50 %.

Заключение. В результате исследования показано, что алгоритм динамической кластеризации, использующий свойство согласованного поведения процессов внутри кластеров, обеспечивает сходимость процесса кластеризации к конечному числу кластеров. Алгоритм динамической кластеризации не требует предварительного определения количества кластеров. Результаты численного эксперимента показывают, что разработанный алгоритм устойчив до уровня отношения шума к сигналу 60 % при величине выборки, состоящей из 100, 60, 40, 20 сигналов с шумом, при значении среднего нормированного СКО, не превышающем 50 %. В результате численного эксперимента установлено, что при увеличении размера выборки значение среднего нормированного СКО снижается. Полученные результаты позволяют применить разработанный алгоритм для анализа температурных сигналов, измеренных на метеорологических станциях, в целях выделения климатических кластеров в оперативном режиме и для организации системы оперативного мониторинга изменения в климатических зонах, формируемых на поверхности Земли.

СПИСОК ЛИТЕРАТУРЫ

1. Stocker T. F., Qin D., Plattner G.-K. et al. The physical science basis // Proc. of the Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, 2013. 1535 p.
2. Köppen W., Geiger G., Borntraeger G. Das geographische System der Klimate. Handbuch der Klimatologie. Berlin: Verlag von Gebrüder Borntraeger, 1936. P. 1–44.
3. Григорьев А. А., Будыко М. И. Классификация климатов СССР // Изв. АН СССР. Сер. Геогр. 1959. № 3. С. 3–19.
4. Алисов Б. П., Полтараус Б. В. Климатология. М.: Изд-во Московского ун-та, 1974. 300 с.

5. **Kottek M., Grieser J., Beck C. et al.** World map of the Köppen — Geiger climate classification updated // *Meteorologische Zeitschrift*. 2006. 15 (3). P. 259–263.
6. **Peel M. C., Finlayson B. L., McMahon T. A.** Updated world map of the Köppen — Geiger climate classification // *Hydrology and Earth Syst. Sci. Discuss.* 2007. 4, Iss. 2. P. 439–473.
7. **Rohli R. V., Joyner T. A., Reynolds St. J. et al.** Globally extended Köppen — Geiger climate classification and temporal shifts in terrestrial climatic types // *Phys. Geographi.* 2015. 36, Iss. 2. P. 142–157.
8. **Stern H., Hoedt G. D., Ernst J.** Objective classification of Australian climates // *Aust. Meteorol. Mag.* 2000. 49 (2). P. 87–96.
9. **Zhang X., Yan X.** Spatiotemporal change in geographical distribution of global climate types in the context of climate warming // *Clim. Dyn.* 2014. 43, Iss. 3–4. P. 595–605.
10. **Franziska H., Körper J., Spangehl T., Cubasch U.** Shifts of climate zones in multi-model climate change experiments using the Köppen climate classification // *Meteorologische Zeitschrift*. 2012. 21, N 2. P. 111–123.
11. **Салугашвили Р. С.** Колебание климата на территории первого естественного синоптического района и климатическое районирование // *Уч. зап. Казанского университета. Сер. Естественные науки*. 2012. 154, кн. 3. С. 216–227.
12. **DeGaetano A. T.** Spatial grouping of United States climate stations using a hybrid clustering approach // *Intern. Journ. Climatology*. 2001. 21. P. 791–807.
13. **Unal Y., Kindap T., Karaca M.** Redefining the climate zones of Turkey using cluster analysis // *Intern. Journ. Climatology*. 2003. 23, Iss. 9. P. 1045–1055.
14. **Волков Ю. В.** Анализ температурных сигналов и алгоритм их кластеризации // *Автоматрия*. 2019. 55, № 3. С. 38–44. DOI: 10.15372/AUT20190305.
15. **Cheredko N. N., Tartakovsky V. A., Krutikov V. A., Volkov Yu. V.** Climate classification in the Northern hemisphere using phases of temperature signal // *Atmospheric and Oceanic Optics*. 2017. 30, N 1. P. 63–69.

Поступила в редакцию 22.04.2020

После доработки 19.06.2020

Принята к публикации 22.06.2020
