УДК 551.515.4

Методы и результаты кластеризации данных по грозовым разрядам*

М.Ю. Беликова[1], *С.Ю. Кречетова*[1], *А.А. Перелыгин*[2]

[1] Горно-Алтайский государственный университет (Горно-Алтайск, Россия)
[2] Алтайский государственный университет (Барнаул, Россия)

Methods and Results of Lightning Discharge Data Clustering

M.Y. Belikova[1], *S.Y. Krechetova*[1], *A.A. Perelygin*[2]

[1] Gorno-Altaisk State University (Gorno-Altaisk, Russia)
[2] Altai State University (Barnaul, Russia)

Изучение закономерностей пространственного распределения гроз является актуальной и практически значимой задачей как для решения фундаментальных проблем атмосферного электричества, так и для решения задач грозозащиты технических сооружений и грозовой пожарной опасности лесных массивов. Одним из источников данных о пространственном распределении гроз является Всемирная сеть локализации молний WWLLN (World Wide Lightning Location Network).

Обобщен опыт применения кластерного анализа (метод ближайшего соседа, алгоритм DBSCAN) к данным WWLLN. Выделены особенности данных, которые необходимо учитывать при выборе алгоритма кластеризации: наличие кластеров разной плотности, влияние на результат кластеризации времени регистрации разрядов. При этом результаты кластеризации должны быть сопоставимы с параметрами грозовой активности (например, средняя продолжительность грозы).

Показано, что алгоритм DBSCAN позволяет учитывать указанные особенности данных WWLLN, поэтому его использование для кластеризации данных о молниевых разрядах является наиболее целесообразным. Результаты кластеризации данных могут быть применены для уточнения и разработки новых методов диагноза и прогноза пространственного расположения и динамики развития гроз.

Ключевые слова: кластерный анализ, метод ближайшего соседа, DBSCAN, данные WWLLN.

DOI 10.14258/izvasu(2016)1-17

Investigation of spatial distribution of thunderstorms is complex and practically important task. It plays an important part in solving the fundamental problems of atmospheric electricity, as well as lightning protection of buildings and lightning fire protection of forests. One of the data sources of spatial distribution of thunderstorms is WWLLN (World Wide Lightning Location Network).

The paper provides the summary of WWLLN data cluster analysis (nearest neighbor algorithm, DBSCAN algorithm). Peculiar features to be taken into consideration are outlined for the data when a certain clustering algorithm is utilized: clusters of different densities, impact of lightning time on clustering results. At the same time, clustering results should be comparable to properties of thunderstorms (for example, thunderstorm mean duration).

It is shown that DBSCAN algorithm is capable of processing WWLLN data with peculiarities and thus is the most appropriate algorithm for data clustering. Clustering results can be utilized for further refining and developing of new prediction methods of spatial location and development of thunderstorm dynamics.

Key words: cluster analysis, K-means, DBSCAN, WWLLN data.

Введение. Молния — это мощный кратковременный электрический разряд, в котором каждая из стадий разряда вызывает характерные воз-

*Работа выполнена при финансовой поддержке РФФИ (13-05-98024 р_сибирь_а).

мущения электромагнитного поля Земли (атмосферики) [1]. Для регистрации атмосфериков создана Всемирная сеть локализации гроз World Wide Lightning Location Network, основной целью которой является решение одной из фундаментальных задач классической области исследований атмосферного электричества — оценки дневных вариаций глобальной электрической цепи [2].

Следует отметить, что для некоторых территорий Западной Сибири данные WWLLN являются единственным источником информации о грозовой активности. Такие наблюдения особенно актуальны, поскольку позволяют уточнять как временную, так и пространственную локализацию гроз для территорий с редкой сетью гидрометеостанций и отсутствием систем инструментальных наблюдений [3].

Количество и частота разрядов для одной грозы зависят от конкретной синоптической ситуации и варьируются на порядки: от десятков в секунду — до одного-двух за всю грозу. Так как атмосферик является характеристикой единичного грозового разряда, то «группа» атмосфериков, близких по местоположению и времени, может являться характеристикой одноячейкового или многоячейкового грозового облака или одной грозы [4,5]. Для исследования пространственных и временных закономерностей грозовой активности необходимо по данным WWLLN выделить группы атмосфериков таким образом, чтобы одна группа атмосфериков соответствовала одной и только одной грозе. Для этих целей используется кластерный анализ.

Кластерный анализ данных сети WWLLN. В качестве исследуемого множества объектов рассмотрим грозовые разряды, зарегистрированные международной сетью WWLLN. Сеть WWLLN предоставляет следующую информацию о грозовом разряде: дата, время, географические координаты, точность, количество станций, зарегистрировавших разряд.

Выделим следующие особенности, которые влияют на результат кластеризации: пространственный характер данных, наличие более плотных «сгустков» объектов и одиночных разрядов («шум»), отстоящих на некотором расстоянии от «сгустков» (рис. 1а). На определение близости объектов влияют не только географические координаты разрядов, но и время их регистрации (два разряда попадают в один кластер, если расстояние и разница по времени регистрации разрядов меньше заданных порогов).

Выбор переменных. Очевидно, что для решения поставленной задачи необходимы три характеристики грозовых разрядов: координаты разряда (x_1, x_2) и время регистрации разряда (x_3). Таким образом, грозовой разряд o может быть описан с помощью набора трех вещественных переменных $x(o) = (x_1, x_2, x_3)$. Множество $O = \{o^{(1)}, o^{(2)}, ..., o^{(N)}\}$, содержащее N грозовых разрядов, может быть представлено как множество N точек в трехмерном евклидовом пространстве E_3.

Расстояние между объектами. Переменные, описывающие характеристики (географические координаты и время) атмосферика, имеют разные единицы измерения. Определить расстояния между объектами только по пространственным характеристикам можно, используя евклидово расстояние. Для вычисления расстояния между объектами по географическим координатам и времени регистрации грозовых разрядов можно выделить три подхода.

Первый подход описан в [4] и заключается в разбиении множества всех объектов на группы по равным временным интервалам, например, часовым (в одну группу попадают атмосферики, зарегистрированные в течение часа). Расстояние между объектами вычисляется только с учетом географических координат между объектами одной группы. При использовании этого подхода возникают трудности в определении расстояния между объектами, находящимися на «границах» групп. Такие случаи требуют уточнения.

Второй подход был использован в работе [6]. Для каждой пары объектов (атмосфериков) $o^{(i)}$ и $o^{(l)}$, где $i, l = 1, ..., N$ и $i \neq l$, вычисляется расстояние

$$\rho_{E_2}(o^{(i)}, o^{(l)}) = \sqrt{\sum_{j=1}^{2}(x_j^{(i)} - x_j^{(l)})^2}, \qquad (1)$$

и разность во времени регистрации разрядов

$$\tau(o^{(i)}, o^{(l)}) = |x_3^{(i)} - x_3^{(l)}|. \qquad (2)$$

Здесь $(x_1^{(i)}, x_2^{(i)}), (x_1^{(l)}, x_2^{(l)})$ — географические координаты, а $x_3^{(i)}, x_3^{(l)}$ — время регистрации разрядов $o^{(i)}$ и $o^{(l)}$. При этом необходимо задать пороговые значения для ρ_{E_2} и τ.

Третий подход [5] основан на вычислении расстояния между двумя объектами по трем нормированным переменным. Расстояние между двумя объектами (атмосферики) $o^{(i)}$ и $o^{(l)}$ определим как

$$\rho_{E_3}(o^{(i)}, o^{(l)}) = \sqrt{\sum_{j=1}^{3}\left(\frac{x_j^{(i)} - x_j^{(l)}}{norm_j}\right)^2}, \qquad (3)$$

где $norm_j, j = 1, 2, 3$, — параметры нормировки, которые задаются в соответствии с физическими параметрами конвективной грозой ячейки (например, средняя площадь, время существования).

Для вычисления расстояния между двумя разрядами нами был выбран второй подход (см. (1), (2)).

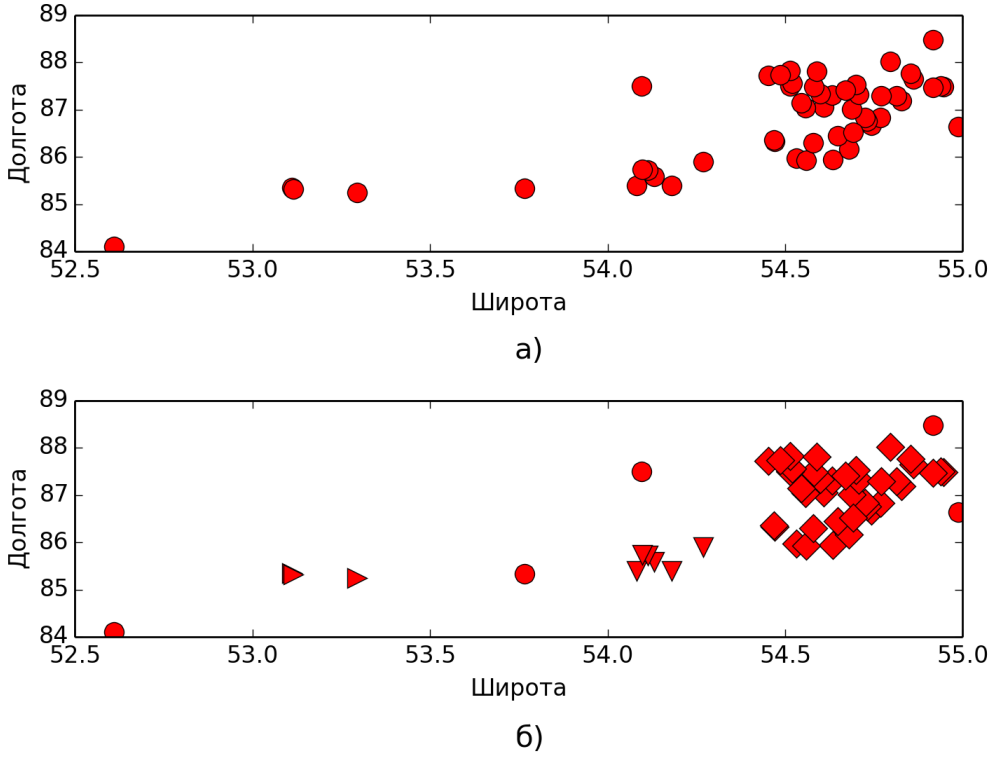

Рис. 1. Результаты кластеризации алгоритмом DBSCAN молниевых разрядов сети WWLLN за 15 мая 2013 года для территории Республики Алтай: а) расположение молниевых разрядов до кластеризации; б) результаты клатеризации (алгоритм DBSCAN), при этом кружками выделены молниевые разряды, отнесенные к «шуму», остальные разряды согласно различиям маркеров относятся к разным кластерам

Метод ближайшего соседа. Для группировки грозовых разрядов в [4, 5] используется подход, основанный на понятии ближайшего соседа. Учет влияния времени регистрации разрядов производится с помощью третьего подхода (3). Входными параметрами алгоритма кластеризации являются следующие параметры нормировки: $norm_1 = norm_2 = 50$ км, $norm_3 = 30$ мин. Таким образом, *грозовой кластер* — это множество разрядов, для которых расстояние между двумя разрядами меньше *50 км*, а временной интервал между регистрацией двух разрядов — меньше *30 мин*. Для выделения более «плотных» кластеров на фоне «шума» необходимо применять дополнительные вычислительные процедуры, например, в [4] на втором этапе кластеризации был использован модальный анализ.

Алгоритм DBSCAN. В работе [6] для кластеризации данных WWLLN применялся алгоритм DBSCAN (Density Based Spatiustering of Applications with Noise — плотностный алгоритм для кластеризации пространственных данных с присутствием «шума»). Учитывая специфику данных WWLLN, для их кластеризации было использовано три входных параметра (пороговых значений): $\varepsilon = 0.12^\circ$ — минимальное расстояние по координатам, в градусах; $\varepsilon_{time} = 18$ *мин* — минимальная разница по времени; $MinPts = 2$ — минимальное количество точек в кластере. Значение параметра ε было выбрано с учетом среднего числа гроз и средней площади конвективной грозовой ячейки для земного шара [6]. Значение ε_{time} должно быть меньше средней продолжительности грозы. Минимальное число разрядов для выделения кластера $MinPts = 2$, так как два распознанных сетью разряда подтверждают наличие грозы. Достоинство и преимущество алгоритма DBSCAN заключается в том, что он позволяет учесть указанные выше особенности данных без дополнительных вычислительных этапов.

Алгоритм был реализован на языке высокого уровня Python. Проведено тестирование алгоритма на модельных данных и данных WWLLN (летний период 2013 года) для территории Республики Алтай с указанными в описании алгоритма [6] входными параметрами (рис. 1б). В результате было выделено 2048 кластеров и средняя продолжительность по времени «грозовых кластеров» составила 33 мин., что не удовлетворяет требованию $\varepsilon < \varepsilon_{time}$. Для подбора оптимальных входных параметров, при которых данные кластеризации будут соответствовать региональным особенностям грозовой активности, требуется проведение дополнительных вычислительных экспериментов и привлечение формальных способов оценки результатов кластеризации.

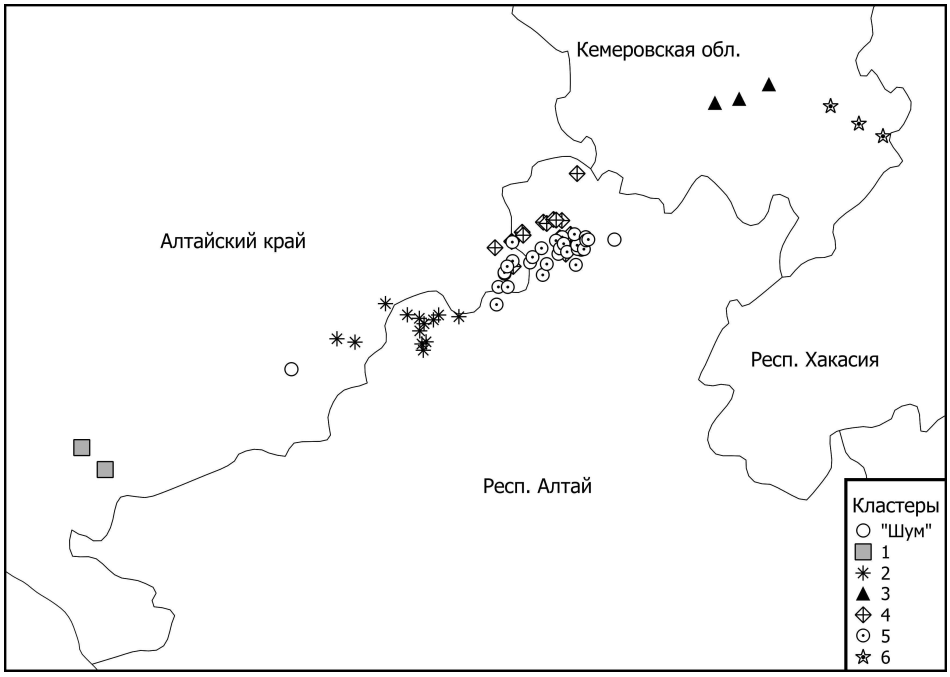

Рис. 2. Пример визуализации результатов кластеризации молниевых разрядов сети WWLLN за 15 мая 2013 года для территории Республики Алтай в программе QGIS

Учитывая, что данные WWLLN имеют пространственную привязку, для визуализации результатов кластеризации и анализа этой информации целесообразным является использование геоинформационных технологий (рис. 2).

Заключение. Обобщен опыт применения кластерного анализа к данным WWLLN. В работе выделены особенности данных о молниевых разрядах, регистрируемых сетью WWLLN, которые необходимо учитывать при выборе алгоритма кластеризации. Показано, что алгоритм DBSCAN позволяет учитывать эти особенности. При использовании алгоритма DBSCAN необходим подбор входных параметров для того, чтобы результаты кластеризации были сопоставимы с параметрами грозовой активности (например, средняя продолжительность гроз и средняя площадь грозового облака и/или грозовой ячейки в нем). Выделение кластеров грозовых разрядов, наиболее точно описывающих региональные особенности гроз, позволит проводить оценку исследуемых территорий по степени грозоопасности и уточнить результаты ранее проведенных исследований [3, 7].

Библиографический список

1. Альперт Я.Л. Распространение электромагнитных волн и ионосфера. — М., 1972.

2. Rodger C.J., Werner S., Brundell J.B., Lay E.H. et al. Detection efficiency of the VLF World-Wide Lightning Location Network (WWLLN): initial case study // Ann. Geophys. — 2006. — V. 24.

3. Дмитриев А.Н., Кречетова С.Ю., Кочеева Н.А. Грозы и лесные пожары от гроз на территории Республики Алтай. — Горно-Алтайск, 2011.

4. Козлов В.И., Шабаганова С.Н. Применение кластерного анализа для выделения грозовых очагов // Динамика сложных систем — 2010. — Т. 4, № 2.

5. Шабаганова С.Н., Каримов Р.Р., Козлов В.И., Муллаяров В.А. Характеристики грозовых ячеек по наблюдениям в Якутии // Метеорология и гидрология — 2012. — № 12.

6. Hutchins, Michael L., Robert H. Holzworth, Brundell J.B. Diurnal variation of the global electric circuit from clustered thunderstorms // Journal of Geophysical Research: Space Physics. — 2014.

7. Горбатенко В.П., Кречетова С.Ю., Беликова М.Ю., Разумова О.В. Идентификация мезомасштабной конвекции и гроз по данным MODIS и аэрологического зондирования // Вестник ТГУ. — 2012. — № 365.