CLUSTERING ANALYSIS OF LIGHTNING DISCHARGES USING WWLLN DATA AND FRIS-TAX METHOD

Assoc. Prof. Dr. Nikolay Baranovskiy[1]
Assoc. Prof. Dr. Svetlana Karanina[2]
Senior Lecturer Marina Belikova[3]
Assoc. Prof. Dr. Andrey Karanin[4]
Senior Lecturer Alena Glebova[5]
[1]Tomsk Polytechnic University, Russian Federation
[2, 3, 4, 5] Gorno-Altaysk State University, Russian Federation

ABSTRACT

The solution of the clustering problem is based on the concept of "similarity" ("proximity") of objects in the feature space. The formalization of this concept is the most consistent with human ideas is the function of competitive similarity (FRiS-function). The effectiveness and applicability of the FRiS function have been proven in solving problems of analyzing data of various types.

Clustering algorithms for space-time events or objects with spatial and temporal characteristics suggest searching for groups of events that occurred “close” from each other in space (in terms of the distance between points, calculated by geographic coordinates) and “close” in time of appearance. To determine the "proximity" of events can be used a priori threshold values. Note that for some events the “natural process” of group formation in time and space can be defined as follows: each next event in time initiates a new group of events or occurs close to an already existing group. At the same time, the spatial "proximity" of events can be considered in a competitive situation, that is, to which of the existing groups is closer a new event. Therefore, for the clustering of space-time events, it seems promising to develop an approach to the taxonomy of data based on the concept of competitive similarity. A modification of the FRiS-Tax algorithm for space-time events is proposed and an example of solving the problem of grouping data on lightning discharges using the proposed version of the algorithm is considered. This clustering method applied for clustering analysis of data on thunderstorm activity using data from the World Wide Lightning Location Network.

Keywords: *lightning activity, lightning discharge, clustering analysis, Republic of Buryatiya*

INTRODUCTION

Cluster analysis methods are widely used to mark out groups of lightning discharges corresponding to convective cells in thunderstorms. For example, the following algorithms are used: k-means, dbscan, grid algorithms, etc. [1],[2],[3],[4],[5],[6],[7]. One of the sufficiently new clustering algorithms that allow to select clusters of arbitrary shape and determine their hierarchical structure is the FRiS-Tax algorithm [8], [9], based on the use of the competitive similarity function (Function of Rival Similarity, FRiS function) [8].

This paper presents the results of automatic separation of thunderstorm cells based on a sample of lightning discharges registered by the World Wide Lightning Location Network (WWLLN) [10] in the Republic of Buryatiya. A modification of the FRiS-Tax algorithm for data with temporal and spatial characteristics is proposed, which allows determining the hierarchical structure of thunderstorm cells and its change over time.

METHODOLOGY

Competitive Similarity

Clustering algorithms involve the division of the sample into groups of close (similar) objects. One of the ways to estimate the proximity (similarity) of point objects with space-time reference is to compare the distance between objects with some a priori known threshold values [11]. So, in the case of grouping data on lightning discharges, two discharges are considered close (similar, belonging to the same cluster), if the distance between them, calculated by coordinates, does not exceed a certain value and the time difference of registration does not exceed a certain value. The average spatial-temporal characteristics of convective cells – area (from 20 to 100 km) and the average duration (from 20 minutes to one hour), respectively, can be used as a priori information to determine the threshold values and use.

The object's neighborhood can be defined as a set of objects, the distance from which to the object o is smaller R and the difference in registration time does not exceed value T. All objects that are included in the ε_{RT} - neighborhood of the object o will be considered "close" to the object o, objects that do not belong to the surrounding area of the object o will be considered as "far" from the object.

Let the objects of the two already "existing" clusters with numbers $C1$ and $C2$ be in the ε_{RT}-neighborhood of the object o. In this case, the object o is close to both the objects of the cluster $C1$ and the objects of the cluster $C2$. To determine whether an object o belongs to one of the clusters, it is necessary to determine which of the two clusters $C1$ or $C2$ of the object o is closer. Clusters $C1$ and $C2$ compete for the object o. Competitive similarity [8] (the function of competitive similarity, FRiS-function) of an object with the clusters under consideration can be calculated by the formula:

$$F_{1/2}(o) = \frac{r1 - r2}{r1 + r2},$$

where $r1$ is the distance from the object o to the cluster $C1$ and $r2$ is the distance to the competing cluster $C2$.

The distance from the object to the cluster can be defined as the distance from the object to the nearest cluster object, or the average distance from the object to all objects in the cluster, or the average distance from the object to several nearby cluster objects, etc.

Function values vary from -1 to +1. At the same time, positive values close to +1 indicate that the object *o* belongs to the cluster *C*1. Negative values close to -1 determine membership in a cluster *C*2. At distance values *r*1=*r*2, which indicates being on the border between the clusters.

It is shown, that the function of competitive similarity determined in this way can be successfully used for solving data analysis problems [8]. In particular, it allows one to obtain clustering results in spaces of low dimension, which are in good agreement with expert estimation. It is assumed that, being applied to solving the problem by grants, it will allow one to obtain solutions of better quality and interpretability of the results. Below is a description of the FRiS-Tax algorithm, which uses the FRiS-function in the process of building a taxonomy of lightning objects of a certain sample, and a modification of the FRiS-Tax algorithm for the taxonomy of a sample that represents lightning discharge data.

Fris-Tax Algorithm

The FRiS-Tax algorithm consists of two stages [8],[9]. At the first stage, called FRiS-Cluster, the centers of local densities of objects are searched throughout the sample under study, on the basis of which linearly separable clusters are selected in the sample. At the same time, clusters combine sample objects that are similar in the sense of competitive similarity with the center of the cluster (pillar).

Let the sample consisting of objects be considered $O = \left\langle o^1, o^2, ..., o^N \right\rangle$. At the initial stage, it is assumed that all objects in the sample belong to the same cluster described by some set of pillars $S = \{s1, s2, ..., sK\}$. The competing cluster is a virtual cluster, i.e. such a cluster is the closest distance from which to each sample object is equal to a constant r^*. The magnitude of the competitive similarity of an object o with the pillar s nearest to it in comparison with a virtual competing cluster is designated as follows:

$$F_S^*(o) = \frac{r(o,s) - r^*}{r(o,s) + r^*} \,. \quad (1)$$

The set of pillars S is sequentially selected from the sample objects O in such a way that the average value of the competitive similarity of each sample object O to the closest pillar from the set to it is maximum:

$$\overline{F}(S) = 1/N \sum_{o \in O} F_S^*(o) \to \max_S \,. \quad (2)$$

If the set of pillars S consists of objects that are centers of local clusters, then it has a maximum value.

The parameter of the FRiS-Cluster algorithm is the number K of pillars (clusters). When a given number of pillars are selected, the sample is divided into clusters as follows. For each object from the set O, a pillar is defined, the similarity with which is maximal. Many objects similar to a pillar form a cluster. At the last step of the algorithm, the position of the pillar for each cluster is refined.

At the second stage, called FRiS-Class, the close clusters are combined into classes of arbitrary shape. In this case, it is assumed that the clusters belonging to different classes are separated from each other by zones with a lower density of objects. Clusters belonging to the same class are separated by so-called competition zones (cluster boundaries), in which objects are distributed fairly evenly. Each pair of clusters is checked for belonging to one class as follows.

1. Each pair of clusters Ci and Cj check for the presence of objects that are located near the boundary separating them (in the competition area). The object a is considered to belong to the competition zone of clusters Ci and Cj, if the conditions are met:

a) The pillars of the clusters Ci and Cj are the two nearest pillars to it.

b) The absolute value of the FRiS function for a given object is less than a certain threshold F^*. Candidates for unification are those pairs of clusters whose competition zones are not empty.

2. The distance D_{ij} between the clusters Ci and Cj the minimum distance between two objects a and b those in the competition area and belonging to different clusters is taken.

3. For these objects a from Ci and b from Cj, the distances D_a and D_b from each of them to the nearest neighbor are determined.

4. Clusters Ci and Cj are considered to belong to the same class, if the value of three quantities D_{ij}, D_a and D_b differ little from each other. For example, the following condition can be checked:

$$\left(D_a < \alpha D_b\right) \wedge \left(D_b < \alpha D_a\right) \wedge \left(D_{ij} < \alpha (D_a + D_b)/2\right),\ \alpha > 1. \quad (3)$$

Fris-Tax Algorithm Modification

Let formed an ordered set $O = \left\langle o^{(1)}, o^{(2)}, ..., o^{(N)} \right\rangle$ corresponding to the selection of objects. Each object $o^{(i)}$ is described by a feature vector $o^{(i)} = \left(x^{(i)}, y^{(i)}, t^{(i)}\right)$, where $x^{(i)}$ is the object coordinate in the north-south direction, $y^{(i)}$ is the object coordinate in the west-east direction, $t^{(i)}$ is the object registration time, measured from a certain initial moment (depends on the specified time interval for the sample of discharges).

Two objects are considered close (similar, belonging to one cluster), if the distance between them, calculated by coordinates, does not exceed a certain value R and the difference in the registration time does not exceed a certain value T. These parameters describe the ε_{RT}-neighborhood of the object for which a decision is made about which cluster it belongs to.

Since the objects in question have a temporal characteristic, they can be ordered by the time of “appearance” and the rule for grouping objects is constructed using the following idea. Each next object (lightning discharge) “initiates” a new group of objects (cluster), if “a lot” of time has passed since the appearance of the previous object and/or the object appeared “far” from the existing groups (clusters), that is, there is no object in the ε_{RT}-neighborhood of other objects. An object can be assigned to an already existing group of objects, if in its ε_{RT}-neighborhood there are objects of other clusters. At the same time, if in the ε_{RT}-neighborhood of an object there are objects belonging to different clusters, then the FRiS function is used to make a decision on belonging to one of the clusters. The thunderstorm grouping algorithm based on the FRiS function consists of the following steps.

1. Set a priori parameters R and T.

2. The selection of objects $O=\left\langle o^{(1)}, o^{(2)}, ..., o^{(N)}\right\rangle$ is ordered by time. The first object $o^{(1)}$ is assigned a cluster label $C=1$.

3. The following objects are iterated sequentially. For each object $o^{(i)}$, where $i=2\,...\,N$, the cluster label is assigned according to the following rules:

3.1. If the $o^{(i)}$ object's ε_{RT}-neighborhood is empty, then a label $C=C+1$ is assigned to the object $o^{(i)}$. The cluster pillar is an object.

3.2. If the $o^{(i)}$ object's ε_{RT}-neighborhood is not empty and contains a certain subset of objects $\left\{o^{(j)}\right\}$, where $j\in\{2\ldots i-1\}$, assigned to only one cluster with a label C_s, then the object $o^{(i)}$ belongs to this cluster and receives the corresponding label C_s. The cluster pillar can be defined as follows. As a pillar, a cluster object o^* is selected, using which as a cluster pillar, the average FRiS function $\overline{F}(o^*)$ calculated by formula (2) has the maximum value.

3.3. If there are objects in a subset of objects $\left\{o^{(j)}\right\}$ belonging to several clusters, then the decision on which cluster the object belongs to is taken as follows.

3.3.1. Among all the objects of a subset $\left\{o^{(j)}\right\}$, two objects belonging to different clusters are determined, the distance from which to the object $o^{(i)}$ is minimal. Such objects $o^{(j1)}$ and $o^{(j2)}$ are the clusters to which they belong C_{s1} and C_{s2}, respectively. For these objects, the value $F_{C_{s1}/C_{s2}}\left(o^{(i)}\right)$ is calculated using formula (1). If $F_{C_{s1}/C_{s2}}\left(o^{(i)}\right) \neq 0$, then the object $o^{(i)}$ is assigned a label in accordance with the rule of interpretation of the value of the FRiS function.

3.3.2. In the cluster to which the object $o^{(i)}$ was added, the pillar can be redefined in accordance with clause 3.2.

3.3.3. To check the object $o^{(i)}$ of the cluster competition area C_{s1} and C_{s2} decide whether the clusters belong to the same class in accordance with the scheme of the FRiS-Class stage and formula (3).

RESULTS AND DISCUSSION

A computational experiment on the clustering of lightning discharges was carried out for data recorded by WWLLN in the territory of the Republic of Buryatiya and its environs. Lightning discharges registered on July 30-31, 2015 were chosen to demonstrate the capabilities of the FRIS-Tax algorithm, to allocate clusters corresponding to thunderstorm cells and combine them into classes corresponding to thunderstorm foci. These days were chosen in connection with the exacerbation of the fire danger situation this year in the study area in late July and early August. During this period, there were about 280 wildfires in remote mountainous areas, which were caused by lightning discharges [12].

On July 30-31, 2015, around 3900 lightning discharges were registered on the territory of Buryatiya by the WWLLN network. When clustering lightning discharges, the following a priori parameters were used: the minimum distance between two discharges is r = 50 km and the minimum time interval is t = 30 min. Clusters consisting of one and two discharges were excluded from the analysis of clustering results.

In total, 79 clusters were identified during the computational experiment. The average duration of a thunderstorm cell was 2 hours and 10 minutes, the median value was 1 hour and 38 minutes, the minimum duration was 8 minutes, and the maximum duration was 8 hours and 10 minutes. The area of thunderstorm cells was determined by the area of convex shells, built on the extreme thunderstorm discharges, selected clusters. The average area of thunderstorm cells was 3,994 square kilometers, the median is 2,186 square kilometers, the maximum area is 3,3454 square kilometers, and the minimum area is 1 square kilometers. At the same time, the largest area corresponds to the longest thunderstorm cells with the longest lifetime (from 5 to 8 hours). The intensity of lightning discharges for the selected clusters varies from 0.00002 to 0.11323 discharges/min per square kilometer; the average intensity is 0.00235 discharges/min per square kilometer, the median is 0.00013 discharges/ min per square kilometer. With an increase in the area of thunderstorm clusters, the intensity of lightning discharges in them decreases in inverse proportion to the area. The obtained characteristics of thunderstorm cells are comparable to the results of the clustering of lightning discharges conducted in the territory of Yakutia [7].

The highest lightning activity and the number of thunderstorm cells is characteristic of the southwestern and northeastern regions, which are characterized by mountainous terrain. The average height of the south-western regions of the republic is about 1500 m and increases to 2300 m in the axial parts of the Eastern Sayan and Khamar-Daban ranges. The average height of the north-eastern regions is about 1300 m above sea level, and the maximum heights are commensurate with the height of the Sayan ranges. Thus, on July 30-13, lightning activity was observed in the remote mountainous regions of the republic. Therefore, it could provoke, with other things being equal, the mass occurrence of forest fires. Using the clustering algorithm, 4 thunderstorm centers (classes) were identified, in which thunderstorm cells appeared quite close in space and time. These thunderstorm foci consisted of 2-5 and 8 thunderstorm cells. Figure 1 shows the thunderstorm cells and the two longest-lasting thunderstorm foci, which operated from 0:46 to 13:20 hours (UTC).

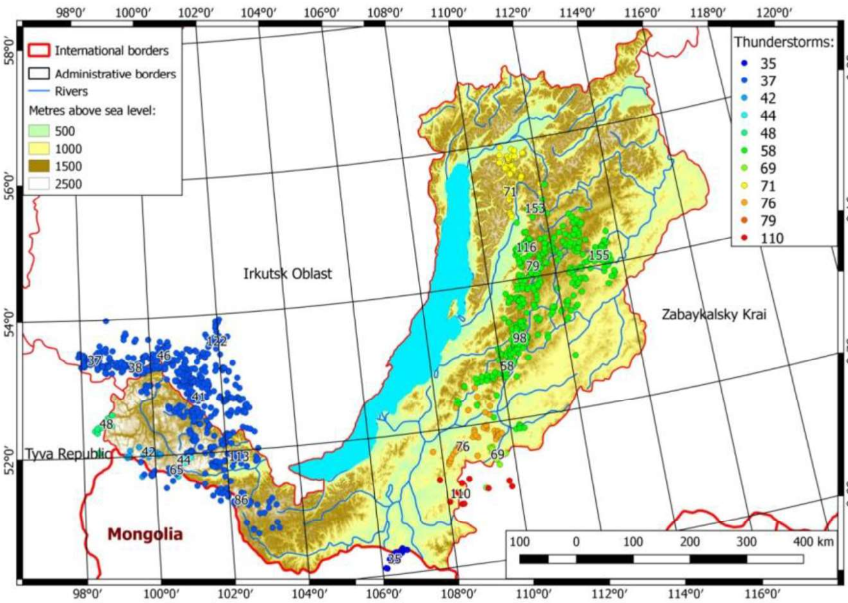

Figure 1. The storm cells and the thunderstorm hearth operating on the territory of the Republic of Buryatia on July 31 from 0:46 to 13:20 hours (UTC)

CONCLUSION

The proposed approach for clustering lightning discharges can be considered as a modified FRiS-Tax algorithm, which can be used to group space-time events [2], [11], [12] or objects that have a spatial reference and recording time, whose proximity is determined by a priori known threshold parameters in space and time. Unlike the original FRiS-Tax algorithm, in which the division into clusters is built on the entire initial sample of objects, the proposed algorithm allows us to construct a partition with successively "emerging" objects.

As an advantage, it is possible to point out that the proposed modification of the FRiS-Tax algorithm allows not only to distinguish clusters of space-time events, but also to combine nearby clusters into classes and thereby determine the hierarchical structure of clusters. In addition, compared with the k-means algorithm used to group lightning discharges [7], [9], the developed modification of the FRiS-Tax algorithm does not impose restrictions on the type of clusters, that is, the clusters may have a shape other than spherical or ellipsoidal.

With the help of the developed modification of the clustering algorithm based on the FRiS-Tax function, a grouping of lightning discharges registered by WWLLN in the Republic of Buryatia on July 30-31 was carried out. The characteristics of the identified thunderstorm clusters (cells) are in good agreement with the results of similar studies conducted for the territory of Yakutia [7]. The use of the FRiS-Tax algorithm allowed us to distinguish classes of thunderstorm cells that are close in space and time with each other and form thunderstorm foci.

ACKNOWLEDGEMENTS

The authors wish to thank the World Wide Lightning Location Network (http://wwlln.net), a collaboration among over 50 universities and institutions, for providing the lightning location data used in this paper.

This work is supported by Russian Foundation for Basic Researches. Scientific project N 17-29-05093.

REFERENCES

[1] Hutchins M. L., Holzworth R. H., Brundell J. B., Diurnal variation of the global electric circuit from clustered thunderstorms, Journal of Geophysical Research: Space Physics, Vol. 119. N 1, pp 620-629, 2014.

[2] Strauss C., Stephany S., Spatio-temporal clustering with a temporal sliding window for tracking of convective activity, XI Workshop de Computação Aplicada (WORCAP 2011 – Instituto Nacional de Pesquisas Espaciais (INPE), Spain, 2011.

[3] Moriah Kohn, Eli Galanti, Colin Price, Kostas Lagouvardos, Vassiliki Kotroni, Nowcasting thunderstorms in the Mediterranean region using lightning data, Atmospheric Research, Vol. 100, pp 489–502, 2011.

[4] Fuchs, B. R., E. C. Bruning, S. A. Rutledge,L. D. Carey, P. R. Krehbiel, and W. Rison, Climatological analyses of LMAdata with an open-source lightning flash-clustering algorithm, Journal of Geophysical Research Atmospheres, 2016, doi:10.1002/2015JD024663.

[5] Kononov I.I., Yusupov I.E., Cluster analysis of thunderstorm activity, Radioengineering and Electronics, Vol. 49, N 3, pp 283–291, 2004.

[6] Mezuman K., Price C., Galanti E., On the spatial and temporal distribution of global thunderstorm cells, Environmental Research Letters, Vol. 9, N 12, pp 124023, 2014.

[7] Shabaganova S.N, Karimov R.R, Kozlov V.I., Mullayarov V.A, Characteristics of storm cells from observations in Yakutia, Russian Meteorology and Hydrology, Vol. 37, N 11–12, pp 746–751, 2012.

[8] Zagoruiko N.G., Intelligent data analysis based on the function of competitive similarity, Avtometriya, Vol. 44, N 3, pp 31-40, 2008. (In Russian)

[9] Borisova I.A., The algorithm of taxonomy FRiS-Tax, Scientific Herald of NSTU, N 3, pp 3-12, 2007. (In Russian)

[10] Hutchins M.L., Holzworth R.H., Brundell J.B., Rodger C.J.,Relative detection efficiency of the World Wide Lightning Location Network, Radio Science, Vol. 47, N RS6005, 2012.

[11] Hadi Fanaee Tork, Spatio-temporal clustering methods classification, In Doctoral Symposium on Informatics Engineering, Norway, 2012, pp 199-209.

[12] Urbazaev Ch.B., Preliminary Results of the Forest Fire Season of 2015 in the Republic of Buryatia, Buryat state university report, Vol. 4, pp. 192-196, 2015. (In Russian)