mathematics

MDPI

Article

Processing Large Outliers in Arrays of Observations

Gurami Tsitsiashvili

Institute for Applied Mathematics, Far Eastern Branch of Russian Academy of Sciences, 690041 Vladivostok, Russia; guram@iam.dvo.ru; Tel.: +7-914-693-2749

Abstract: The interest in large or extreme outliers in arrays of empirical information is caused by the wishes of users (with whom the author worked): specialists in medical and zoo geography, mining, the application of meteorology in fishing tasks, etc. The following motives are important for these specialists: the substantial significance of large emissions, the fear of errors in the study of large emissions by standard and previously used methods, the speed of information processing and the ease of interpretation of the results obtained. To meet these requirements, interval pattern recognition algorithms and the accompanying auxiliary computational procedures have been developed. These algorithms were designed for specific samples provided by the users (short samples, the presence of rare events in them or difficulties in the construction of interpretation scenarios). They have the common property that the original optimization procedures are built for them or well-known optimization procedures are used. This paper presents a series of results on processing observations by allocating large outliers as in a time series in planar and spatial observations. The algorithms presented in this paper differ in speed and sufficient validity in terms of the specially selected indicators. The proposed algorithms were previously tested on specific measurements and were accompanied by meaningful interpretations. According to the author, this paper is more applied than theoretical. However, to work with the proposed material, it is required to use a more diverse mathematical tool kit than the one that is traditionally used in the listed applications.

Keywords: large outliers; arrays of observations; complex systems; digraphs

MSC: 93B07; 06A06

Citation: Tsitsiashvili, G. Processing Large Outliers in Arrays of Observations. *Mathematics* **2022**, *10*, 3399. https://doi.org/10.3390/math10183399

Academic Editor: Andrzej Sokołowski

Received: 18 August 2022
Accepted: 15 September 2022
Published: 19 September 2022

1. Introduction

This paper is devoted to the analysis of large outliers in data samples in medical and zoo geography, mining, an application of meteorology in fishing tasks, etc. The closest to this problem in probability theory, mathematical statistics, queuing theory and insurance is the analysis of heavy-tailed distributions [1–7].

It should be noted that recently, this topic has attracted the attention of a large number of data processing specialists from the fields of mathematical statistics [8,9], statistical methods in medicine [10,11] and physiological studies [12], as well as in the analysis of industrial processes [13,14]. Moreover, along with the statistical methods in this area, it requires the development of new algorithms and the application of graph theory elements, particularly in the study of protein networks [15].

However, in those applications with which the author had to work, it was necessary to shift the emphasis from estimates of heavy tails to the large outliers in empirical information. Apparently, this is due to the fact that we have to work with short samples or in the presence of rare events. However, the main reason is that there are no well-established theoretical models in these areas of application, and we have to work with data within the framework of a phenomenological approach. This circumstance required the development of original heuristic algorithms that allowed obtaining information useful and interesting to users who submitted their empirical results to the author. The novelty and significance of the algorithms constructed by the author were confirmed during last 20 years by the joint

results represented in [13,15–20]. Previously, for a long time, such tasks simply could not be solved by the author.

In the listed areas of application, the ability to consistently meet the user requirements plays a crucial role. The following motives are important for these specialists: the substantial significance of large emissions, the fear of errors in the study of large emissions by standard and previously used methods, the speed of information processing and the ease of interpretation of the results obtained. To meet these requirements, interval pattern recognition algorithms and the accompanying auxiliary computational procedures have been developed. These algorithms were designed for specific samples provided by the users. They have the common property of the original optimization procedures being built for them or well-known optimization procedures being used.

The emphasis on large outliers is due to the fact that their behavior usually obeys some asymptotic relations [21] and is therefore somewhat simplified. Such circumstances allow us to raise the question of increasing the reliability of the results of the processing arrays of observations and reducing the counting time. The latter plays an important role in the interdisciplinary interactions between domain specialists and mathematical programmers processing the arrays of observations. To carry out such work, it is advisable to identify the applied tasks in which such observation processing procedures may be implemented.

The considered samples of observations are defined by the number n of observations and the number m of their dimensions. The requirements of mathematical statistics [10] are such that it is desirable that the parameter n is large and the parameter m is small. However, in the arrays of observations with which we had to deal, the opposite situation was often observed, where the parameter n was small and the parameter m was large. For example, such a situation occurs in problems of medical geography [17] and in problems of meteorology and hydrology [18]. This circumstance forces one to look for sufficiently fast algorithms for processing short time series, and the accuracy of calculations determined in some way, on the contrary, increases with an increase in the parameter m.

At the same time, there are one-dimensional long time series ($m = 1$, and n is sufficiently large) in which not just rare but very rare events associated with large outliers are observed [13]. It is required to process these series in such a way that the length of the series and the number of large outliers in it do not create problems for either processing or interpretation of the results obtained.

Along with time series, which are not quite convenient for data processing, in various applications, there are large arrays of observations that require data compression and packaging and lead to extreme graph theory problems. These include disturbances in the rock according to the results of acoustic monitoring and the movement of animals in a territory. Despite the presence of well-known graph theory algorithms, special auxiliary algorithms, albeit simple, are designed well enough with the requirements of a particular subject area and are also required for processing such data.

This paper describes the methods of interval pattern recognition used in medical geography and meteorology recognition of rare outliers by a generalized indicator used in mining, studies of the vicinity of the extremes in the nodes of the square grid used in meteorology and hydrology and special classification methods used in the analysis of protein networks in zoo geography, mining and other subject areas.

2. Materials and Methods

The materials for constructing algorithms for processing empirical information are the following:

- Multidimensional short series of observations containing the main component and m accompanying components;
- Series of real observations equipped with Boolean variables indicating the presence or absence of critical events;
- An array of three-dimensional vectors characterizing the coordinates of sound sources;
- An array of one-dimensional characteristics of square lattice nodes;

- A scheme of the protein network in the form of a digraph;
- A map with a set of districts and a description of the presence or absence of borders between them.

The methods are as follows:

- The method of interval pattern recognition;
- The method for optimizing monotone piecewise constant functions;
- The method for converting a matrix of distances between points in three-dimensional space into an undirected graph;
- The method for difference approximation of first- and second-order partial derivatives for the functions of two variables;
- The method of sequentially allocating cyclic equivalence classes in a digraph and constructing a zero-one matrix of a partial order between these classes;
- The method of hierarchical classification of districts on a map with respect to the presence of common boundaries between them.

The following optimization problems are considered:

- When recognizing critical events from an array of one-dimensional observations, two optimization problems are considered. A connection between them is established, and it is shown how by reducing one task to another, the array of processed information may be significantly decreased.
- In determining the acoustic core, the connectivity component that contains the minimum number of vertices is selected from another connectivity component of a graph.
- An algorithm for approximating a level line of a smooth function, given at the nodes of a square lattice, in the form of an ellipse is constructed.
- In the hierarchical classification of districts on a map, for each district, the minimum number of borders, a crossing of which allows one to get out from this district to a common boundary, is determined.

All described methods are closely connected with the initial formulations of the applied problems and are adopted to real data processing. Moreover, in relation to each case, it is necessary to introduce some new element into the algorithm.

3. Interval Pattern Recognition Method and Related Algorithms

This section discusses the interval pattern recognition algorithm, which has found its application in the processing of time series in the problems of medical geography [17], as well as in meteorology, hydrology [18] and fishing [22,23].

3.1. Interval Pattern Recognition Method

Suppose that an array of observations is represented by a set of vectors with dimensions $m+1:\ X = \{(x_{01}, x_{11}, \ldots, x_{m1}), \ldots, (x_{n0}, x_{n1}, \ldots, x_{nm})\}$. Here, the components of vectors $x_{01}, \ldots, x_{0m}$ characterize the main features, and all other components of these vectors are related features. Let us say the element $(x_{k0}, x_{k1}, \ldots x_{km})$ corresponds to a larger outlier in the sample if the inequality $x_{k0} \geq x_0$ is satisfied at some critical level x_0 (selected by an expert) of the zero component value in the vector. Then, in the initial sample X, a set of elements with numbers $1 \leq k_1, \ldots, k_s \leq n$ is determined, for which the inequality $x_{k_j 0} \geq x_0,\ 1 \leq j \leq s$ is satisfied. All these elements are perceived as large outliers. We first calculate

$$x_i^+ = \max_{1 \leq j \leq s} x_{k_j i},\ x_i^- = \min_{1 \leq j \leq s} x_{k_j i}, \tag{1}$$

Then, a decisive rule is constructed according to which the sample element $(x_{k0}, x_{k1}, \ldots x_{km})$ is a large outlier if the following inequalities are met:

$$x_i^- \leq x_{ki} \leq x_i^+,\ 1 \leq i \leq m. \tag{2}$$

This decisive rule is defined as interval pattern recognition. Here, the image is understood as a large outlier determined by the value of the zero component of the sample element, and the decisive rule (2) is determined by the belonging of the components of the vector $(x_{k0}, x_{k1}, \dots x_{km})$ to the segments $[x_i^-, x_i^+]$, $1 \le i \le m$.

Let us now list the main properties of interval pattern recognition. For this, we denote S as the number of sample elements that are perceived by this decisive interval recognition rule as large outliers:

- All sample elements that are large outliers are perceived by interval recognition as large outliers. Therefore, the $S \ge s$ inequality is fulfilled. Then, the quality of interval recognition may be chosen by the ratio $s/S \le 1$.
- With an increase in the number m of associated features, the recognition quality of s/S increases and, for some samples of observations, may even approach unity.
- The number of arithmetic operations for the interval recognition procedure is proportional to the product nm and therefore depends linearly on the number n of sample elements X and on the number m of accompanying features.
- The solution of this problem in its initial version was tested with respect to s/S,, characterizing the quality of recognition for a given sample. Here, it is possible to increase the value 0.6 obtained by standard methods to 0.7 or more with an increase in the number m.

3.2. Investigation of the Extremum of a Function in the Nodes of a Square Lattice

The most important element of a structure of the pressure field at an altitude of 5 km above the Far East is a stable and extensive depression. The coordinates of this depression (which are usually associated with a square lattice node) and the pressure value H_{500} at an altitude of 5 km determine the nature of atmospheric circulation and significantly affect the weather [19]. This also includes observations represented by a finite number of points located at the nodes of a square lattice and characterizing a certain meteorological system. It is known from observations that the extremes of H_{500} at the nodes of such a grid largely determine the functioning of the meteorological system. If we assume that H_{500} is described by a smooth function defined on a rectangle and having a minimum at the lattice node, then by decomposing this function into a Taylor series and assuming the lattice step is small enough, we may approximate the level lines of this function with ellipses [19]. In turn, the direction of the major axis of the ellipse and its relation to the minor axis allow us to make meteorological forecasts concerning the behavior of anticyclones in the vicinity of the minimum point.

Suppose that the function $f(x, y)$, specifying H_{500}, is continuously differentiable twice in the domain $D = \{0 \le x \le Nh,\ 0 \le y \le Mh\}$, and at the point $(kh, lh), 0 < k < N$, $0 < l < m$, its first differential is zero, and its second differential $A(x - kh)^2 + B(y - lh)^2 + 2C(x - kh)(y - lh)$ is a positive definite quadratic form ($A = f_{x,x}(kh, lh)$, $B = f_{y,y}(kh, lh)$, $C = f_{x,y}(kh, lh)$). Then, the point (kh, lh) is the point of the local minimum of the functions f, and therefore, by virtue of the Sylvester criterion, the inequalities $A + B > 0$, $AB > C^2$ are fulfilled. The lines of the level of the function f in the vicinity of the point (kh, lh) are approximately ellipses. The angle of inclination of the major axis and the compression ratio of these ellipses determine the nature of the atmospheric circulation.

We denote $a = A + o(h)$, $b = B + o(h)$ and $c = C + o(h)$ as the finite difference approximations of the partial derivatives A, B and C. We approximate the function f by the function $\widehat{f}$ up to $o(h^2)$ in variables $X = \dfrac{x - kh}{h}$ and $Y = \dfrac{y - lh}{h}$:

$$\widehat{f}(x, y) = f(kh, lh) + \frac{1}{2}(aX^2 + bY^2 + 2cXY),\ a + b > 0,\ ab > c^2.$$

Therefore, for small h values, the quadratic form $aX^2 + bY^2 + 2kxy$ is also positively definite.

We reduce this form to a diagonal form by constructing a matrix $\mathcal{A} = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$ and writing out the characteristic equation $(a-\lambda)(b-\lambda) - c^2 = 0$, whose roots

$$\lambda_{\pm} = \frac{a+b}{2} \pm \sqrt{\left(\frac{a+b}{2}\right)^2 - ab + c^2} > 0,$$

are the eigenvalues of the matrix $\mathcal{A}$.

In the coordinate system (u_+, u_-) with an orthonormal basis $\overrightarrow{n}_+$, $\overrightarrow{n}_-$ from the eigenvectors of matrix $\mathcal{A}$, the quadratic form $aX^2 + bY^2 + 2cXY$ is represented by the sum of squares $\lambda_+ u_+^2 + \lambda_- u_-^2$ with level lines in the form of ellipses $\lambda_+ u_+^2 + \lambda_- u_-^2 = const > 0$, having a compression ratio $k = \sqrt{\lambda_+ / \lambda_-}$. The slopes of the major axis of the ellipses were found, and the compression ratio at the H_{500} level line allowed meteorologists to build a physical reconstruction of various processes occurring in the atmosphere. The lines of the level of the analyzed function H_{500}, constructed in the form of ellipses, were rechecked during the construction of a physical meteorological forecast in [19].

4. Recognition of Rare Outliers and Related Algorithms

Another type of observation may be time series in which $m = 1$ and the length of the series n is quite large, being to the order of several hundred. Such observations characterize important and therefore rare events in the system. These include the already described collapses in mine workings. The miners proposed to characterize the state of the system at some point in time by a generalized one-dimensional indicator ρ and a Boolean variable characterizing the presence or absence of a collapse in the system. The task is to recognize presence or absence of the collapse in the presented one-dimensional series of observations. An algorithm is proposed for constructing a recognition procedure for the presence or absence of the collapse, in which the amount of calculations is determined only by the number of important events N being much smaller than n. This algorithm is based on maximizing the frequency of correct recognition of the presence or absence of an event from the critical value ρ_*, determining the recognition result using the inequality $\rho \le \rho_*$.

Let us now turn to the consideration of long series of observations in which the number of large emissions is small (i.e., n is much larger than one, and N/n is much smaller than one). Such observations include, in particular, collapses in mine workings. There is a class of applied problems in which a certain generalized indicator is selected as a concomitant feature, formed by specialists of this subject area based on the results of numerous observations, such as mining specialists based on the results of acoustic monitoring of the rock strata [13,24,25].

4.1. Recognition of Rare Outliers by a Generalized Indicator

In this subsection, we assume that the initial sample is formed as follows. All generalized indicators form a sequence $\{x_{11}, \ldots, x_{n1}\}$, and the numbers $k_1, \ldots, k_s$ of the sample elements characterizing large outliers are given. It is required to build a recognition rule for determining emissions by this generalized (single) indicator. Let us place the sequence $\{x_{11}, \ldots, x_{n1}\}$ on the real line and mark it with crosses with the numbers $k_1, \ldots, k_s$ (see Figure 1). We are looking for a number x_* defining the following decisive rule: if $x_k \ge x_*$, then the sample element with the number k refers to large outliers. If $x_k < x_*$, then the sample element with the number k is not recognized as a large outlier.

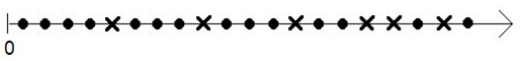

Figure 1. Representation of a training sample on a straight line by a set of characters $\times$, $\bullet$.

For each number c, we compare the frequency $\rho_\times(c)$ of correctly attributing a sample element to large outliers and the frequency $\rho_\bullet(c)$ of not correctly attributing a sample element to large outliers. The value ρ_* is introduced using an expert method, and it is

required that the solution corresponding to it satisfies the inequality $\rho_{\times}(c) \geq \rho_{*}$. Among all $c: \ \rho_{\times}(c) \geq \rho_{*}$, it is required to find one value that maximizes $\rho_{\bullet}(c)$. Here, $\rho_{\times}(c)$ characterizes the security of the decision being made, and $\rho_{\bullet}(c)$ characterizes its cost-effectiveness.

Since the function $\rho_{\times}(c)$ is stepwise and monotonically non-increasing by the argument c, being continuous to the left, and the function $\rho_{\bullet}(c)$ is stepwise and monotonically non-decreasing, being continuous on the right (see Figures 1–3), then this problem has many solutions that can be represented by some segment. In turn, the task of determining the maximum value of x_{*} at which $\rho_{\times}(c) \geq \rho_{*}$ has a unique solution, which is the right end of the segment specified above. It is natural, for security reasons, to determine the right end of the segment, which is the solution to the maximization problem $\rho_{\bullet}(c)$, under the condition $\rho_{\times}(c) \geq \rho_{*}$. Due to the specified property of this solution, it is sufficient to solve the problem for the maximum of the function $\rho_{\times}(c)$ under the condition $\rho_{\times}(c) \geq \rho_{*}$.

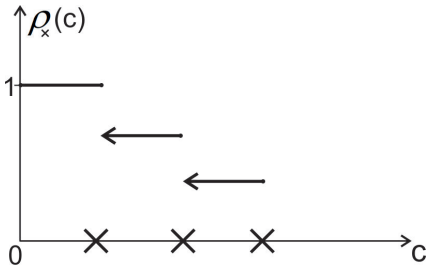

Figure 2. Type of function $\rho_{\times}(c)$.

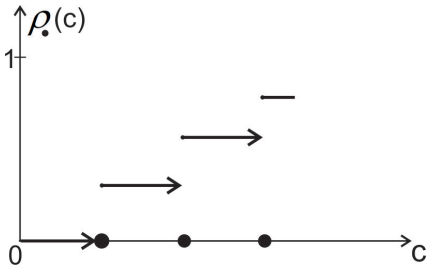

Figure 3. Type of function $\rho_{\bullet}(c)$.

The resulting solution to the problem of recognizing large outliers by sampling $\{x_{11}, \ldots, x_{n1}\}$ and numbers $k_1, \ldots, k_s$ requires only knowledge of the sequence $x_{k_1 1}, \ldots x_{k_s 1}$, which significantly reduces the amount of calculations, since s/n is much smaller than one.

Using the method of recognizing a large outburst (exceeding the generalized indicator of the critical level), the results were obtained for predicting collapses in the mine, which were confirmed by specialists in mining. Moreover, the frequency of correct recognition of a critical event (a collapse in a mining operation) constructed in solving this problem characterizes the safety factor of mining operations, and the frequency of correct recognition of an absence of a critical event characterizes the cost-effectiveness factor of mining operations. Therefore, when solving this problem, safety restrictions were first introduced, and under these restrictions, the efficiency indicator was optimized. The solution of the concrete problem considered in [13] was verified by comparing the optimization result c obtained by the author and the result independently obtained by mining specialists (which practically coincided). It was very important for the mining specialists to independently verify their own rather cumbersome calculations.

4.2. Clusters of Points in Space

When implementing an acoustic monitoring system, it becomes necessary to determine acoustically active zones and, on this basis, predict dangerous collapses in mining according to the generalized indicator introduced by mining specialists [24,25]. In the previous subsection, to construct a generalized indicator, it was necessary to determine

the acoustically active zone from a set of n points in three-dimensional space determined during acoustic monitoring of cod sources in the rock column [13].

In fact, we are talking about constructing a model of an acoustically active zone based on the available observations and an algorithm for determining it. This procedure is based on information about the matrix $||r_{ij}||_{i,j=1}^{n}$ of pairwise distances between the points detected during acoustic sounding and the critical distance r between them, as set by experts.

For the first step, the matrix $||r_{ij}||_{i,j=1}^{n}$ is converted to a zero-one matrix $||I(r_{ij} < r)||_{i,j=1}^{n}$.

In the second step, the constructed zero-one matrix is further considered as the adjacency matrix of an undirected graph, whose edges between the vertices i and j exist only under the condition $r_{ij} < r$.

In the third step, using the well-known methods of graph theory, in the set of $1, \ldots, n$ vertices of the graph G, a set of connectivity components is determined, among which the one with the maximum number of vertices is selected. This set of vertices is defined as an acoustically active zone (several zones are also possible).

In the forth step, the classification procedure is accelerated in the following way. Initially, the point 1 is taken, which is denoted by the first class. Let the vertex classes $I_1, \ldots, I_p$ from the set $\{1, \ldots, k\}$ be allocated in step k, and the point $k+1$ is connected by the edges to some of these classes. Then, a new class is formed from them and the point $k+1$, and the classes that are not included in this new class remain the same, together forming a set of classes in step $k+1$. In such an algorithm, information previously used is not lost at each step of the algorithm. The most significant step is the last step of this algorithm, in which it is proposed to preserve the classification of the connectivity components of the graph and not leave only one applicant for the formation of the final connectivity component.

The selection of clusters of points in the three-dimensional space detected during acoustic monitoring allows us to build generalized indicators by which critical events (collapses) in a mine are predicted. The solution to the problem considered in [13] was verified visually by mining specialists, who were interested in convenient computer algorithms for defining acoustically active zones.

5. Special Classification Algorithms

Classification algorithms allow us to identify some extreme modes in a complex system. In particular, with the help of classification algorithms, it is possible to determine the acoustically active zones. Of particular interest are hierarchical classification algorithms that identify objects, namely those that most influence the behavior of a complex system or objects that play the role of hubs through which numerous connections between elements pass. This section of the work is devoted to these issues.

5.1. Hierarchical Classification of Graph Vertices

This problem arose when analyzing a protein network presented by a complete digraph containing n vertices [26]. The vertices of such digraphs are proteins and the directed edges of the connection between them. The procedure of hierarchical classification in such a digraph is in some sense equivalent to the isolation of clots (aggregates of proteins close to each other).

Using Floyd's algorithm, we construct a matrix $||c_{ij}||_{i,j=1}^{n}$ of the lengths of the minimal paths between the vertices of the original digraph. We transform the matrix $||c_{ij}||_{i,j=1}^{n}$ into a symmetric matrix $||r_{ij}||_{i,j=1}^{n}$, $r_{ij} = c_{ij} + c_{ji}$. Thus, r_{ij} is the minimum length of a cycle connecting the vertices i and j. It is obvious that the minimum length of a cycle passing through a pair of vertices can be considered the distance between them, since it is nonnegative and satisfies the triangle inequality.

Let us construct a finite, monotonically increasing sequence of $R = \{r_1 < r_2 < \ldots < r_m\}$ nonzero elements of this matrix. Having chosen some critical level r, we transform the matrix $||r_{ij}||_{i,j=1}^{n}$ into a zero-one matrix $||I(r_{ij} \leq r)||_{i,j=1}^{n}$. Now, let us construct a graph G_r,

whose edges connect the vertices i and j provided that $r_{ij} \leq r$. Then, in the undirected graph G_r, the connectivity components may be distinguished using the classification algorithms described above. The parameter r may be selected in different ways, such as by assuming $r = r_1, \ldots, r_m$. In this case, with $1 \leq r_p < r_q \leq r_m$, the class defined with $r = r_q$ necessarily enters some class defined with $r = r_p$. Thus, the hierarchical classification of the set of vertices $1, \ldots, n$ is determined. However, the increasing sequence of values of the critical level r may be reduced at the choice of the users.

5.2. Allocation of Cyclic Equivalence Classes in a Digraph

The problem considered above requires for its formulation the allocation of cyclic equivalence classes (clusters) in the digraph. The cyclic equivalence relation between a pair of digraph vertices assumes the existence of a cycle containing this pair of vertices. Then, a partial order relation may be introduced between the cyclic equivalence classes in the digraph. There are different algorithms to define the cyclic equivalence classes and matrix of their partial order (see, for example, [27,28]).

In order to construct a sequential algorithm for solving this problem, it is required at each step to establish a partial order relation between the classes of cyclic equivalence. It is not enough to just allocate cyclic equivalence classes. It is also required to determine the zero-one matrix of the partial order relationship of clusters (a presence of a path from one cluster to another).

To accomplish this, at step 1, the vertex 1 is taken, and a cluster and a one-by-one unit matrix are formed from it. Let the clusters and the matrix of partial order relations between them be constructed at step $t-1$. We take the element t and select the following sets of clusters: B_1, B_2 and B. The set B_1 contains clusters, each of which has a path from the vertex t, and the set B_2 contains clusters from which there are paths to the vertex t. All other clusters fall into the set B, and from them, there can be paths only to the clusters of the set B_1, and paths can exist in them only from the set B_2. Then, at step t, a new cluster $[t]$ is built, consisting of the vertex t and the clusters of the set $B_1 \cap B_2$. The matrix of a partial order at step t is defined by rectangular sub matrices 0 consisting of only zeros, rectangular sub matrices 1 consisting of only ones and rectangular sub matrices repeating the corresponding submatrices of the matrix a at step $t-1$ (see Table 1).

Table 1. Algorithm of transition from step $t-1$ to step t for a matrix of partial order a.

<table>
<tr><th>Matrice Partial Order</th><th>Clusters Set A_1</th><th>Clusters Set $[t]$</th><th>Clusters Set A_2</th><th>Clusters Set B</th></tr>
<tr><td>set A_1 clusters</td><td>repeating step $(t-1)$</td><td>0</td><td colspan="2" rowspan="2">0</td></tr>
<tr><td>set $[t]$ clusters</td><td colspan="2" rowspan="2">1</td></tr>
<tr><td>set A_2 clusters</td><td colspan="2">repeating step $(t-1)$</td></tr>
<tr><td>set B clusters</td><td>repeating step $(t-1)$</td><td colspan="2">0</td><td>repeating step $(t-1)$</td></tr>
</table>

This method has been applied to the analysis of the thermal stability of some protein networks [16], and so far, requests have been received from various applied biological journals for the continuation of this topic.

5.3. Definition of Central Hub Areas on the Map

Another type of such observations may be maps divided into some areas and used to highlight areas associated with animal movements [29]. Let us assume that there is some bounded, connected territory with a set of U_0 singled-out, single-connection regions (administrative districts or hunting farms) on it. This territory is defined by a finite set of bounded regions on the plane. Everywhere else, without limiting generality, we assume that the boundaries between the regions are polylines. Our task is to compress information

about this map in order to use it further for studying the movement of rare animals in this area by traces of these animals found in these areas.

According to this division, it is necessary to build a hierarchical classification of internal (not touching the border of the map) districts in relation to their neighborhood. Such a hierarchical classification assumes the allocation on the map of a sequence of sets of districts $U_k,\ U_{k+1} \subseteq U_k,\ k \geq 1$ so that each district in the set U_{k+1} adjoins only the districts from the set U_k. It is shown that such a sequence is finite, and in real observations, the number of vertices at the end of the algorithm is usually significantly less than at the beginning. Thus, the final vertices allow us to compress the original information about the map.

This compression of map information is based on the "neighborhood" relationship between the specified areas. For this purpose, a map with the areas highlighted on it is represented as a planar graph, the faces of which are the regions, and the edges are the sections of the border between two neighboring regions.

This procedure can be continued in a recurrent way:

$$U_{k+1} = \{A \in U_k :\ S(A) \subseteq U_k\},\ k \geq 1 \tag{3}$$

This can continue up to some step n, at which point one of two equalities is fulfilled: $U_{n+1} = U_n$ or $U_{n+1} = \oslash$. Here, for $A \in U_0$, we define $S(A)$ as a set of regions bordering it.

The equality $U_{n+1} = U_n$ means that all regions of the set U_n border only on the regions of this set. However, due to the condition of the limitation of all areas of the map, the finiteness of the number of these areas and the presence of only polylines as boundaries, this condition cannot be fulfilled. In addition, since at each step k the strict inclusion of $U_{k+1} \subset U_k$ is performed, then the number of regions $N(U_k)$ in the set U_k satisfies the inequality $N(U_{k+1}) < N(U_k)$. This implies the inequality $n < N(U_0) < \infty$. Therefore, the algorithm in Equation (3) may be implemented in a finite number of steps n. In the second case, when $U_{n+1} = \oslash$, we have $N(U_{n+1}) = 0$, so no area from the set U_n may be completely surrounded by areas from the same set. This algorithm requires knowledge of the set of all inner regions U_1 and the sets $\{S(A) :\ A \in U_1\}$ of all regions bordering the inner regions (of the first kind). Thus, the implementation of the algorithm in Equaiton (3) is working with lists of the area numbers and not with their view on the plane, which greatly simplifies its implementation.

Denote $V_k = U_k \setminus U_{k+1},\ 1 \leq k < n,\ V_n = U_n$, and then the equalities are valid ($U_k = \bigcup_{j=k}^{n} V_j,\ 1 \leq k \leq n$), and any vertex of the set V_k is connected by an edge to some vertex of the set V_{k-1} where there are no edges connecting this vertex to the vertices of the sets $V_j,\ j < k-1$. Indeed, if the vertex is $v \in V_k$, then the inclusion of $v \in U_k$ is performed. However, a complete encirclement of a vertex v by vertices from the set U_k is impossible, because in this case, $v \in U_{k+1}$ means $v \in V_j$ for some $j \leq k-1$. Therefore, there is an edge connecting the vertex v with the set of vertices U_{k-1}. However, an edge connecting the vertex v to the set U_{k-2} is also impossible, because the vertex v is completely surrounded by the vertices of the set V_{k-1}. Finally, the vertex $v \in V_k$ may be connected with some vertices of this set also. Therefore, each region of the set $U_n = V_n$ may be considered some center on the map. Then, the set V_{n-1} consists of the areas bordering it and completely surrounding it, called its margin or periphery of the first kind. By attaching to the periphery of the first kind, with the regions bordering on the regions from this periphery, it is possible to build a periphery of the second kind, and so on. It follows from this construction that the minimum number of boundaries that the path from the vertex $v \in V_k$ to the total boundary of all districts crossed is equal to k, where $k = 1, \ldots, n$. The proposed algorithm was tested during the analysis of traces of the Amur tiger in the territory of Primorsky Krai with the help of ecologists and aroused their serious interest.

6. Discussions

What all the algorithms for processing large outliers given in this paper have in common is the fact that the algorithms themselves are fairly standard, but when applying

them to individual samples, it is necessary to select the correct combination of these algorithms. It is this combination that ensures the novelty of the results obtained. For example, when processing data on acoustic monitoring of the rock strata for an algorithm for predicting critical events (collapses), an algorithm for identifying acoustically active zones was required. In turn, when analyzing critical events in the climate system, it is necessary not only to highlight the moments of occurrence of these events but also their spatial localization and behavior in its vicinity.

In the practical application of the proposed algorithms, their computational complexity and computational speed play an important role. In some cases, for example, when processing data on animal movements over a certain territory, excessive requirements for data processing algorithms may encounter excessive computational complexity. This led to the construction and use of hierarchical classification algorithms, which at the top level of the hierarchy identify some central parts of the study area.

The final results of the proposed algorithms for processing observations are evaluated by experts from the subject area. Therefore, all elements of the proposed algorithms should be understood by these experts and allow them to be checked. Moreover, the proposed algorithms should be convenient to assist experts in constructing various scenarios of the behavior of the analyzed system. It should be noted that the results of processing large outliers tend to be some estimates that require estimates of their errors and the impact of the inaccuracies of the observations of them.

The experience of working with algorithms for processing large outliers shows that all the elements included in them should be selected as carefully as possible in order to ensure high quality and demand among specialists in the subject areas. It is also necessary to combine the proposed algorithms for processing large outliers with classical probabilistic models. For example, when processing data on animal tracks in a certain territory, it is convenient to use an inhomogeneous Poisson flow of points [30] as a model of animal tracks. Now, it is difficult to predict what new algorithms and models will have to be built to solve the problems discussed in the work. These tasks come from users and require additional mathematical processing, but it is already clear that various optimization procedures should play an important role in them.

When identifying flashes in a time series, some difficulties arise that require a set of different methods to overcome. For example, there are known time series of pink salmon yields, in which the harvest is small in even years and large in odd years. To analyze this phenomenon, it is necessary to distinguish stable cycles of a length of two in the Ricker model. These cycles appear when the growth coefficient of the model belongs to a certain interval. However, the noted phenomenon occurs only at the right end of the interval, and this can be detected only after additional and more detailed calculations.

7. Conclusions

This article presents an algorithm for constructing an interval pattern recognition procedure. The properties of this algorithm were investigated, and it was shown that with an increase in the dimension of observations, the recognition quality improves:

- An algorithm for recognizing a critical event from a one-dimensional series of observations was constructed by analyzing the (small) part of the series containing only critical events.
- An algorithm for determining the acoustically active zone by the coordinates of the sound source points was constructed. This algorithm is based on the transformation of an array of coordinates of sound source points into an undirected graph and the allocation of connectivity components in it.
- A sequential algorithm for determining cyclic equivalence classes and partial order relations between these classes in the digraph was constructed.
- A (fast) algorithm for the hierarchical classification of districts on the map based on the presence of common borders (neighborhood) between districts was constructed.

Therefore, their further development requires assessments of the stability of the results obtained with variations of these critical levels. In addition, an important role in the development of this topic should be played by estimates of the impact of observation errors on the results obtained in the work. If the array of observations of a system consists of parts of its elements' observations, then in the near future, it will be necessary to develop a procedure for comparing the results of processing these parts in order to determine the most sensitive part.

From the author's point of view, this paper is more applied than theoretical. However, to work with the proposed material, it is required to use a more diverse mathematical tool kit than the one that is traditionally used in the listed applications. In particular, when working with mining materials, this allows us to identify economic and safety indicators and significantly reduce the volume of the analyzed information.

The algorithms presented in this paper appeared as a result of long and rather unsuccessful computational experiments. Practice has shown that in order to obtain reasonable applied results, it is necessary to strictly follow the initial meaningful statement of the problem, but the algorithms proposed by the mathematicians themselves should be convenient in calculations and fast enough. Unfortunately, the consumers of these algorithms are often impatient users.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This paper has no processing of concrete data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Teugels, J.L. *The Class of Subexponential Distributions*; University of Louvain: Annals of Probability: Louvan, Belgium, 1975; Retrieved 7 April 2019.
2. Zolotarev, V.M. *One-Dimensional Stable Distributions*; American Mathematical Society: Providence, RI, USA, 1986.
3. Embrechts, P.; Klueppelberg, C.; Mikosch, T. *Modelling Extremal Events for Insurance and Finance*; Stochastic Modelling and Applied Probability; Springer: Berlin, Germany, 1997; Volume 33.
4. Asmussen, S.R. Steady-State Properties of GI/G/1. *Appl. Probab. Queues* **2003**, *51*, 266–301.
5. Foss, S.; Korshunov, D.; Zachary S. *An Introduction to Heavy-Tailed and Subexponential Distributions*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2013.
6. Novak, S.Y. *Extreme Value Methods with Applications to Finance*; CRC: London, UK, 2011.
7. Wierman, A. *Catastrophes, Conspiracies, and Subexponential Distributions (Part III). Rigor + Relevance Blog*; RSRG Caltech: Pasadena, CA, USA, 2014.
8. Siebert, C.F.; Siebert, D.C. *Data Analysis with Small Samples and Non-Normal Data Nonparametrics and Other Strategies*; Oxford University Press: Oxford, UK, 2017.
9. Chandrasekharan, S.; Sreedharan, J.; Gopakumar, A. Statistical Issues in Small and Large Sample: Need of Optimum Upper Bound for the Sample Size. *Int. J. Comput. Theor. Stat.* **2019**, *6*. [CrossRef]
10. Konietschke, F.; Schwab, K.; Pauly, M. Small sample sizes: A big data problem in high-dimensional data analysis. *Stat. Methods Med. Res.* **2020**, *30*, 687–701. [CrossRef] [PubMed]
11. Vasileiou, K.; Barnett, J.; Thorpe, S.; Young, T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Med. Res. Methodol.* **2018**, *18*, 148. [CrossRef] [PubMed]
12. Morgan, C.J. Use of proper statistical techniques for research studies with small samples. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **2017**, *313*, 873–877. [CrossRef] [PubMed]
13. Guzev, M.A.; Rasskazov, I.Y.; Tsitsiashvili, G.S. Algorithm of potentially burst-hazard zones dynamics Representation in massif of rocks by results of seismic-acoustic monitoring. *Procedia Eng.* **2017**, *191*, 36–42. [CrossRef]
14. Zhu, Q.X.; Chen, Z.S.; Zhang, X.H.; Rajabifard, A.; Xu, Y.; Chen, Y.Q. Dealing with small sample size problems in process industry using virtual sample generation: A Kriging-based approach. *Soft Comput.* **2020**, *24*, 6889–6902. [CrossRef]
15. Bulgakov, V.P.; Tsitsiashvili, G.S. Bioinformatics analysis of protein interaction networks: Statistics, topologies, and meeting the standards of experimental biologists. *Biochemistry* **2013**, *78*, 1098–1103. [CrossRef] [PubMed]

16. Tsitsiashvili, G.S.; Bulgakov, V.P.; Losev, A.S. Factorization of Directed Graph Describing Protein Network. *Appl. Math. Sci.* **2017**, *11*, 1925–1931. [CrossRef]
17. Bolotin, E.I.; Tsitsiashvili, G.S.; Golycheva, I.V. Some aspects and perspectives of factor prognosis for the epidemic manifestation of the Tick-Borne Encephalitis based on the multidimensional analysis of temporal rows. *Parazitology* **2002**, *36*, 89–95. (In Russian)
18. Shatilina, T.A.; Tsitsiashvili, G.S.; Radchenkova, T.V. Peculiarities of surface air temperature variations over the Far East regions in 1976–2005. *Russ. Meteorol. Hydrol.* **2010**, *35*, 740–743. [CrossRef]
19. Shatilina, T.A.; Tsitsiashvili, G.S.; Radchenkova, T.V. Okhotsk medium-tropospheric cyclone and its role in the formation of extreme air temperature in January in 1950–2019. *Hydrometeorol. Stud. Forecast.* **2021**, *3*, 64–79. (In Russian) [CrossRef]
20. Tsitsiashvili, G.S.; Shatilina, T.A.; Radchenkova, T.V. *Application of New Algorithms for Processing Meteorological Observations*; Publishing House "Buk": Kazan, Russia, 2022. (In Russian)
21. Lever, J.; Leemput, I.; Weinans, E.; Quax, R.; Dakos, V.; Nes, E.; Bascompte, J.; Scheffer, M. Foreseeing the future of mutualistic communities beyond collapse. *Ecol. Lett.* **2020**, *23*, 2–15. [CrossRef] [PubMed]
22. Radchenko, V. Abundance Dynamics of Pink Salmon, Oncorhynchus gorbuscha, as a Structured Process Determined by Many Factors. *NPAFC Tech. Rep.* **2011**, *8*, 14–18. [CrossRef]
23. Shuntov, V.P.; Temnikh, O.S. *Far Eastern Salmon Industry–2016: Good Results, Successes and Errors in Forecasts and the Traditional Failure of VNIRO on the Ways of Innovative Breakthroughs Announced by Him in Forecasting the Number and Catches of Fish. Study of Pacific salmon in the Far East*; TINRO-Center: Vladivostok, Russia, 2016; Volume 11, pp. 3–13. (In Russian)
24. Rasskazov, I.Y. *Control and Management of Rock Pressure in the Mines of the Far Eastern Region*; Gornaya Kniga Publ.: Moscow, Russia, 2008. (In Russian)
25. Rasskazov, I.Y.; Gladyr, A.V.; Anikin, P.A.; Svyatetsky, V.S.; Prosekin, V.A. Development and modernization of the control system of dynamic appearances of rock pressure on the mines of "Priargunsky Industrial Mining and Chemical Union". *JSC Gorn. Zhurnal (Min. J.)* **2013**, *8*, 9–14. (In Russian)
26. Tsitsiashvili, G.S.; Bulgakov, V.P.; Losev, A.S. Hierarchical classification of directed graph with cyclically equivalent nodes. *Appl. Math. Sci.* **2016**, *10*, 2529–2536.
27. Mezic, I.; Fonoberov, V.A.; Fonoberova, M.; Sahai, T. Spectral Complexity of Directed Graphs and Application to Structural Decomposition. *Complexity* **2019**, *2019*, 9610826. [CrossRef]
28. Tarjan, R. Depth-first Search and Linear Graph Algorithms. *SIAM J. Comput.* **1972**, *1*, 146–160. [CrossRef]
29. Pikunov, D.G.; Mikell, D.G.; Seredkin, I.V.; Nikolaev, I.G.; Dunishenko, Y.M. *Winter Tracking Records of the Amur Tiger in the Russian Far East (Methodology and History of Accounting)*; Dalnauka: Vladivostok, Russia, 2014. (In Russian)
30. Kingman, J.F.C. *Poisson Processes*; Oxford Studies in Probability-3; Clarendon Press: Oxford, UK, 1993.