

## RESEARCH PAPER

# The Study of Time Series Using the DMA Methods and Geophysical Applications

Sergey Agayan, Shamil Bogoutdinov, Anatoly Soloviev and Roman Sidorov

Geophysical Center of the Russian Academy of Sciences, Moscow, Russia

Corresponding author: Anatoly Soloviev (a.soloviev@gcras.ru)

The discrete mathematical analysis (DMA) is a series of algorithms aimed at the solution of basic problems of data analysis: clustering and tracing in multidimensional arrays, morphological analysis of reliefs, search for anomalies and trends in records etc. All the DMA algorithms are of universal nature, joined by the same formal foundation, based, in its turn, on fuzzy logic (FL) and fuzzy mathematics (FM).

The current study finalizes the search for the anomalies in one-dimensional time series within the scope of DMA: here the initial concept of an interpreter's logic gets its additional development. First, the formal expert's opinions are more fully expressed, and this is realized with the more complex measures of activity (the concept of straightenings (Gvishiani et al. 2003; Gvishiani et al. 2004; Zlotnicki et al. 2005) is replaced by the measures of activity which come to the fore); second, for the junction of anomalies, a recently created DPS (Discrete Perfect Sets) algorithm is used DPS (Discrete Perfect Sets) (Agayan et al. 2011; Agayan et al. 2014).

**Keywords:** time series; anomalies; fuzzy logic; discrete mathematics; data mining; geoinformatics; geomagnetism

## A short summary of the study

We will now describe, in a semi-formal way, what the study is dedicated to, and what will be carried out in it.

Let there be a time series  $f$ , defined on a finite, in common case, irregular set of nodes  $T = \{t_1 < \dots < t_N\}$ . The time series  $f$  is analyzed by an expert  $\varepsilon$ .

It is supposed that the expert  $\varepsilon$  has a point of view on the dynamic pattern of  $f$  in every node  $t \in T$ , and a number of questions is posed at this time. Let us formulate the first two questions:

Question 1: To what extent  $\mu_{\varepsilon_f}(t) \in [0, 1]$  is the dynamic pattern of the time series  $f$  the expert  $\varepsilon$  is interested in, expressed (how active is it) in the node  $t$ ?

Question 2: How is the dynamic pattern of the time series  $f$ , the  $\varepsilon$  is interested in, alternating at the time moment  $t$  is the measure  $\mu_{\varepsilon_f}$  increasing (decreasing) in the node  $t$  or passing through an extremal transition?

The answers for this questions form the basis of what is called a monitoring of activity of the time series  $f$  with the expert  $\varepsilon$  point of view. The measure of activity  $\mu_{\varepsilon_f}$  gives the key to the spot  $\varepsilon$ -anomalies:  $\varepsilon$ -anomalous moments of  $f$  are the moments  $t$  where  $\mu_{\varepsilon_f}(t) \geq 0.75$  (if some feature is expressed with the scale of  $[0, 1]$ , then it's a matter of course to consider its occurrences within the interval  $[0.75, 1]$  as anomalous).

Let us denote  $LA(\mu_{\varepsilon_f})$  the set of  $\varepsilon$ -anomalous nodes:

$$LA(\mu_{\varepsilon_f}) = \{t \in T : \mu_{\varepsilon_f}(t) \geq 0.75\}$$

Forming of  $LA(\mu_{\varepsilon_f})$  is also included into the problem of activity monitoring.

Due to the randomness of the time series  $f$ , its local anomalous nodes  $LA$  can comprise a subset in  $T$  which demands extra work (preliminary topological filtering (Gvishiani et al. 2003) and further clustering

(Mikhailov et al. 2003; Agayan et al. 2004; Soloviev et al. 2005)). The point is that the connected side-by-side and sufficiently large excesses of the measure  $\mu_{\mathcal{E}f}$  over the anomaly level 0.75 are of the main interest as it is them what should be considered expressions (traces) of global anomalies of  $f$  from the perspective of  $\mathcal{E}$ .

Let us denote  $A(\mu_{\mathcal{E}f})$  their plurality:

$$A(\mu_{\mathcal{E}f}) = \{A_1, \dots, A_L\}$$

The transition  $LA(\mu_{\mathcal{E}f}) \rightarrow A(\mu_{\mathcal{E}f})$  comes with the help of the DPS. algorithm. This is how the answer is obtained for the question 3.

Question 3: Which fragments of the time series  $f$  are considered interesting from the perspective of the expert  $\mathcal{E}$ ?

The last question is formulated this way:

Question 4: How to compare and rank the  $\mathcal{E}$ -anomalies of  $f$ ?

DMA includes a corresponding implement which makes it possible to compare the fragments to each other and rank them (Gvishiani et al. 2008; Gvishiani et al. 2008), it will be told about that further. Here we shall list a couple of examples for the quantitative indicators of the anomalies  $A_i$ , inducing their ranking.

**Example 1** Massiveness (continuity)  $mA_i$  of an anomaly  $A_i$ :  $mA_i = |\text{Supp } A_i|$  – number of nodes in the carrier of  $A_i$ .

**Example 2** Activity  $\mu A_i$  of an anomaly  $A_i$ :  $\mu A_i = \frac{\mu_{\mathcal{E}f}(t) : t \in \text{Supp } A_i}{|\text{Supp } A_i|}$

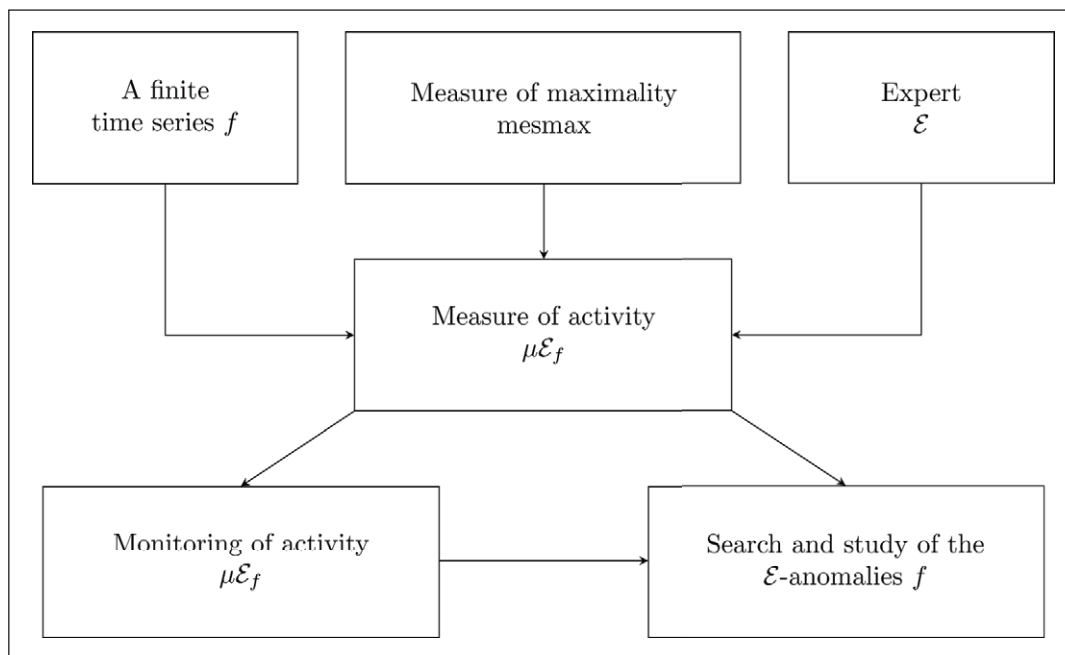
The summary will follow the scheme displayed on **Fig. 1**, in the simplest suppositions, however, fully keeping the thoroughness of the analysis of a time series announced above.

### A finite time series $f$ (initial data)

Period of observation  $T$  – the finite regular number of nodes

$$T = \{t_1 < \dots < t_N\}, t_{i+1} - t_i = h, i = 1, \dots, N - 1$$

Function  $f$  – a time series defined on  $T$



**Figure 1:** The scheme of the analysis of a time series using the FL methods.

$$f : T \rightarrow \mathbb{R}, f \sim (f_i = f(t_i))_i^N$$

Localization of space  $T$  in the node  $t$  is realized with the fuzzy structures  $\delta_t$  on  $T$ , normalized at  $t$  and decreasing with the increase of the distance from  $t$ :

$$(\delta_t(t) = 1) \wedge (|\bar{t} - t| > |\bar{t} - t| \Rightarrow \delta_t(\bar{t}) \leq \delta_t(t))$$

The  $\delta_t$  structures are called the *measures of closeness* on  $T$  to the node  $t$ . They center  $T$  around  $t$  by assigning the weight functions  $\delta_t(\bar{t})$  to the nodes  $\bar{t}$ , act as surroundings for  $t$  in  $T$ , and formalize the view of the expert  $\varepsilon$  on  $T$  at  $t$  (Gvishiani et al. 2003; Gvishiani et al. 2004).

In the examples given below,  $p$  stand for a non-negative parameter responsible for the scale of  $\delta$ -localization: the higher  $p$  is, the stronger the localization is: in particular,  $p = 0$  results in the global point of view on  $T$ , the same for each node.

**Example 3**

1. The global measure of closeness (Fig. 2)

$$\delta_t(\bar{t}) = \left( 1 - \frac{|\bar{t} - t|}{\max(t_N - t, t - t_1) + h} \right)^p$$

2. The local measure of closeness (Fig. 3)

$$\delta_t(\bar{t}) = \begin{cases} \left( 1 - \frac{|\bar{t} - t|}{r} \right)^p & , \text{ if } |\bar{t} - t| \leq r \\ 0 & , \text{ if } |\bar{t} - t| > r \end{cases}$$

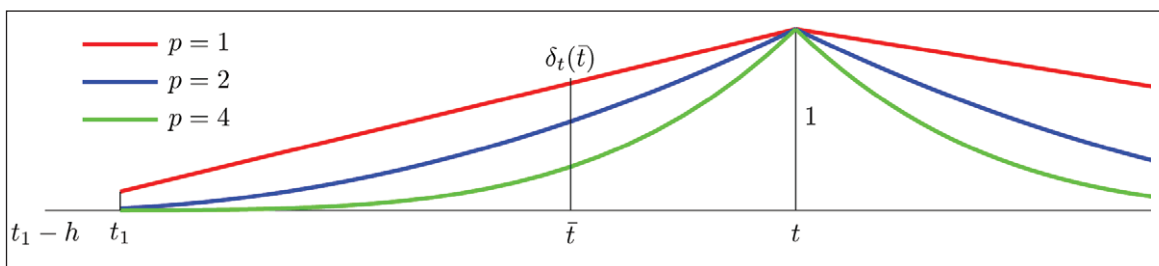
where  $r \geq h$  – observation radius.

**An expert  $\varepsilon$  (formalization)**

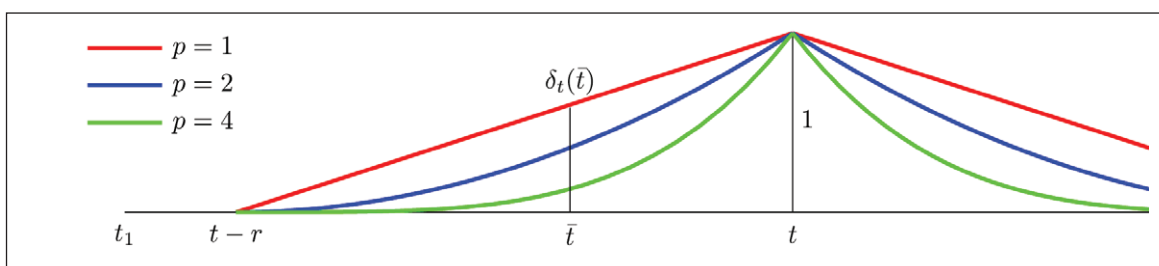
Dynamical indicator  $D$  – a non-negative functional on  $T$ , parametrized by  $T$

$$D : F(T) \times T \rightarrow \mathbb{R}^+$$

where  $\mathfrak{F}(T)$  is a function space on  $T$ .



**Figure 2:** The global measure of closeness.



**Figure 3:** The local measure of closeness.

The values of the function  $D(f, t)$  are denoted  $F_f(t)$  and regarded as quantitative estimates of the behavior of the function  $f$  in a node  $t \in T$  according to the approach  $D$  to its dynamics.

There exists an open to supplementation set of basic indicators which showed a good performance in practice. Let us list some of them.

**Example 4**

1. Energy

$$E_f(t|\delta) = \frac{\sum_{\bar{t} \in T} |f(\bar{t}) - M_f(t|\delta)| \delta_t(\bar{t})}{\sum_{\bar{t} \in T} \delta_t(\bar{t})} \tag{1}$$

where

$$M_f(t|\delta) = \frac{\sum_{\bar{t} \in T} f(\bar{t}) \delta_t(\bar{t})}{\sum_{\bar{t} \in T} \delta_t(\bar{t})}$$

2. Jaggedness

$$L_f(t|\delta) = \frac{\sum_{\bar{t} \in T} |f'_g(\bar{t})| \delta_t(\bar{t})}{\sum_{\bar{t} \in T} \delta_t(\bar{t})} \tag{2}$$

where  $f'_g(\bar{t})$  is the  $f$  regression derivative with respect to the closeness  $\delta$

$$f'_g(\bar{t}) = \frac{\left| \frac{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) f(\tilde{t}) \tilde{t}}{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}} - \frac{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}^2}{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}} \right|}{\left| \frac{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}^2}{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}} - \frac{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t}) \tilde{t}}{\sum_{\tilde{t} \in T} \delta_{\tilde{t}}(\tilde{t})} \right|}$$

3. Scatterness

$$O_f(t|\delta, q) = \left( \frac{\sum_{\bar{t}, \tilde{t} \in T} |f(\bar{t}) - f(\tilde{t})|^q \delta_t(\bar{t}) \delta_t(\tilde{t})}{\sum_{\bar{t}, \tilde{t} \in T} \delta_t(\bar{t}) \delta_t(\tilde{t})} \right)^{\frac{1}{q}} \tag{3}$$

where  $q > 0$ .

The dynamical indicators are related to the very natural and important but the simplest points of view of the expert  $\varepsilon$  on the dynamics of the time series  $f$ . The general case is combined from the simple ones, plus, the composition can't be straight and mechanical. The point is that, initially, the dynamical indicators  $D_f$  for different  $D$  are not comparable to each other and their junction additional efforts by means of the measures of maximality or dynamical activity of the indicator  $D_f$ .

**Measure of maximality mesmax (constructions)**

Measure of dynamical activity  $\mu D_f$  is a function membership to the fuzzy concept of the "activity of the function  $f$  in the node  $t$  from the perspective of the indicator  $D$ " on  $T$  (Gvishiani et al. 2008; Gvishiani et al. 2008; Gvishiani et al. 2008c). The measure of activity  $\mu D_f(t)$  results from the  $D_f(t)$  indicator using one or another construction of the measure of maximality mesmax.

The measure of maximality  $\text{mesmax}_B(x)$  is a fuzzy structure on  $\mathbb{R}$ , which answers the following question for a finite weighted arrangement  $B = \{(b_k, \omega_k) |_{k=1}^M, \omega_k > 0\}$ : "To what degree the number  $x$  is big modulo  $B$ ?" In DMA, for a mesmax measure there are four constructions: "fuzzy comparisons", "Kolmogorov averages", "fuzzy means", "iteration bounds" (Gvishiani et al. 2004; Zlotnicki et al. 2005; Gvishiani et al. 2008).

For a measure  $\mu D_f(t)$ , the  $(\text{Im } D_f, \delta_t)$ , which is an image of  $D_f$  centered in the node  $t$  with the help of the measure of closeness  $\delta_t$ , acts as B:

$$(\text{Im } D_f, \delta_t) = \{ (D_f(\bar{t}), \delta_t(\bar{t}), \bar{t} \in T \}$$

and

$$\mu D_f(t) = \text{mesmax}_{(\text{Im } D_f, \delta_t)} D_f(t) \tag{4}$$

**Example 5**

1. The "fuzzy comparison" construction: binary version.

$$\mu D_f(t) = \left( \sum_{\bar{t} \in T} \frac{D_f(t) - D_f(\bar{t})}{D_f(t) + D_f(\bar{t})} \delta_t(\bar{t}) \right) \left( \sum_{\bar{t} \in T} \delta_t(\bar{t}) \right)^{-1}$$

2. The "fuzzy comparison" construction: gravitational version.

$$\mu D_f(t) = \frac{D_f(t) - \overline{D_f(t)}}{D_f(t) + D_f(t)}$$

where

$$\overline{D_f(t)} = \frac{\sum_{\bar{t} \in T} D_f(\bar{t}) \delta_t(\bar{t})}{\sum_{\bar{t} \in T} \delta_t(\bar{t})}$$

3. The "fuzzy comparison" construction: sigma version.

$$\mu D_f(t) = \frac{\sigma^r(D_f(t)) - \sigma^l(D_f(t))}{\sigma^r(D_f(t)) + \sigma^l(D_f(t))}$$

where

$$\begin{aligned} \sigma^l(D_f(t)) &= \sum_{\bar{t} \in T} (D_f(t) - D_f(\bar{t})) \delta_t(\bar{t}) : D_f(\bar{t}) < D_f(t) \\ \sigma^r(D_f(t)) &= \sum_{\bar{t} \in T} (D_f(\bar{t}) - D_f(t)) \delta_t(\bar{t}) : D_f(\bar{t}) < D_f(t) \end{aligned}$$

4. The "Kolmogorov averages" construction.

$$\mu D_f(t) = \frac{4}{\pi} \arctan p^* - 1$$

where  $p^*$  is the solution of the equation

$$\left( \frac{\sum_{\bar{t} \in T} f(\bar{t})^{p^*} \delta_t(\bar{t})}{\sum_{\bar{t} \in T} \delta_t(\bar{t})} \right)^{1/p^*} = D_f(t)$$

**Remark 1** The listed measures of activity  $\mu D_f$  possess the values within  $[-1, 1]$ . The proper measures of activity possess the values within  $[0, 1]$  and result from the transformation  $\mu D_f \rightarrow \frac{\mu D_f + 1}{2}$ .

**Measure of activity  $\mu \epsilon_f$  (constructions)**

The transition  $D_f \rightarrow \mu D_f$  translates the function  $f$  analysis into the fuzzy logic and fuzzy mathematics language (Gvishiani et al. 2003): the measures of activity  $\mu D_f$  for different  $D$  possess the values within the unified scale of the interval  $[0, 1]$  and can be combined in any proportions and quantities using numerous

FL operations and averagings of any type which will be denoted \*. It becomes possible to instill the content to the complicated activity of  $f$  by the plurality of dynamical indicators  $\mathfrak{D}$  in the node  $t \in T$ :

$$\mu D_f(t) = *_{D \in \mathfrak{D}}(\mu D_f(t)) \tag{5}$$

It is the construction (1) by which the view of expert  $\varepsilon$  on the dynamics of  $f$  is commonly modeled (Gvishiani et al. 2004). Such modeling is an art in many respects, and it involves selection of basic dynamical indicators  $\mathfrak{D}(\varepsilon)$  and their appropriate junction  $*(\varepsilon)$ :

$$\mu e_f(t) = *(e)_{D \in \mathfrak{D}(e)}(\mu D_f(t)) \tag{6}$$

**Example 6**

1. Junctions \* using the fuzzy conjunction  $\top$  make it possible to consider all the basis dynamics from  $\mathfrak{D}(\varepsilon)$  simultaneously:

$$\mu e_f(t) = \top_{D \in \mathfrak{D}(e)}(\mu D_f(t)) \tag{7}$$

2. Junctions \* using the fuzzy disjunction  $\perp$  make it possible for every basic dynamic from  $\mathfrak{D}(\varepsilon)$  to approve itself:

$$\mu e_f(t) = \perp_{D \in \mathfrak{D}(e)}(\mu D_f(t)) \tag{8}$$

3. Let us suppose that the expert  $\varepsilon$  gives initial precedence on the basic dynamics from  $\mathfrak{D}(E)$  in the form of weights  $\omega(D)$ , then one of possible junctions \*, able to consider this fact will be the weighted averaging

$$\mu e_f(t) = \frac{\sum_{D \in \mathfrak{D}(e)} \omega(D) \mu D_f(t)}{\sum_{D \in \mathfrak{D}(e)} \omega(D)} \tag{9}$$

**Monitoring of activity  $\mu \mathcal{E}_f$  (concept)**

Analysis of the measure  $\mu \mathcal{E}_f$  and its basic components  $\mu D_f^m, D^m \in \mathfrak{D}(\varepsilon)$  is the essence of the monitoring of  $f$  by an expert  $\varepsilon$ . It includes:

1. Direct analysis of the  $\mu \mathcal{E}_f$  activity which makes it possible to understand to what extent the dynamic of interest for  $\varepsilon$  is expressed on  $f$  (**Fig. 4**) for every moment  $t \in T$ . The result is the disjunctive partition of  $T$  in three pieces

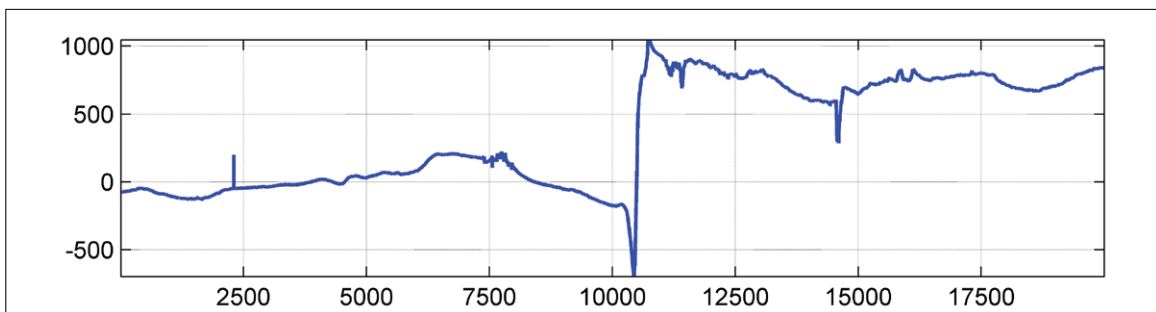
$$T = LB(\mu \mathcal{E}_f) \vee LPA(\mu \mathcal{E}_f) \vee LA(\mu \mathcal{E}_f) \tag{10}$$

- (a)  $LB(\mu \mathcal{E}_f)$ : locally background nodes

$$LB(\mu \mathcal{E}_f) = \{t \in T : \mu \mathcal{E}_f(t) \in [0; 0.5]\} \tag{10.1}$$

- (b)  $LPA(\mu \mathcal{E}_f)$ : locally potentially anomalous nodes

$$LPA(\mu \mathcal{E}_f) = \{t \in T : \mu \mathcal{E}_f(t) \in [0.5; 0.75]\} \tag{10.2}$$



**Figure 4:** The function under examination.

(c)  $LA(\mu\varepsilon_f)$ : locally anomalous nodes

$$LA(\mu\varepsilon_f) = \{t \in T: \mu\varepsilon_f(t) \in [0.75; 1]\} \tag{10.3}$$

2. Trend analysis of the activity of  $\mu\varepsilon_f$  based on regression derivatives: it helps to understand how is the  $f$  dynamics of  $\varepsilon$ 's interest  $t$  alternating at every moment  $t \in T$ : is it increasing, decreasing or passing through an extremal transition.
3. statistical analysis of the total of the basic activities  $\{\mu D_f^m, D^m \in \mathcal{D}(\varepsilon)\}$  using mean and covariance characteristics  $\bar{\mu D_f^m}$  and  $Cov(\mu D_f^{m_1}, \mu D_f^{m_2})$ . The result is the coding of the behavior of  $f$  on  $T$  from the perspective of  $\varepsilon$  and the possibility of comparison with the behavior of another time series  $\bar{f}$  on, generally speaking, another period  $\bar{T}$  from the perspective of the same  $\varepsilon$ .

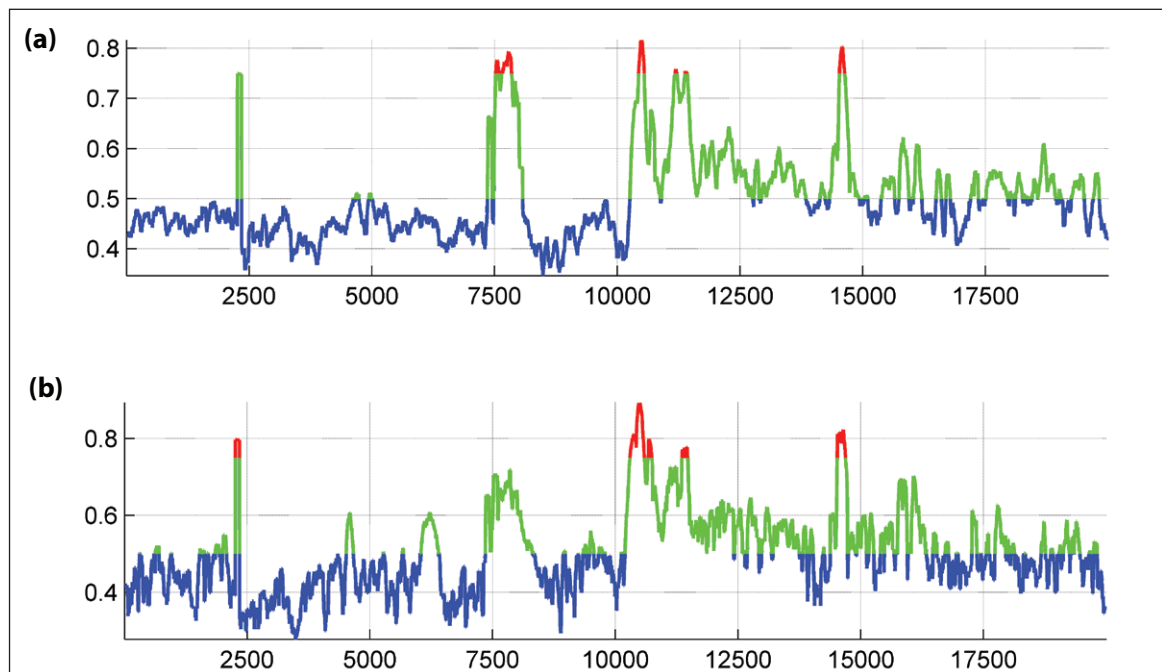
**Remark 2** Real partitions (10) have close thresholds but not always equal, to ideal thresholds 0.5 and 0.75. It is not principal for further studies and therefore in the present paper is convenient to consider monitoring of activity to be ideal.

### Measure and monitoring of activity (examples)

For a record  $f$  (Fig. 4), the simple measures of activity  $\mu L_f$  (Fig. 2) and  $\mu O_f$  (Fig. 3), were constructed. They correspond to the dynamical indicators of jaggedness (2) and scatterness (3). Afterwards, they are joined in the complex measures of activity using three approaches (7)–(9). On the constructed measures (Figs. 5 and 7) and the initial record (Fig. 6 and 8), a partition is shown (10): blue color for the background nodes (10.1), green for the potentially anomalous nodes (10.2), and red for the anomalous nodes (10.3).

On the simple measures of activity  $\mu L_f$  (Fig. 2) and  $\mu O_f$  (Fig. 3) respectively, 6 and 5 anomalous areas were identified and marked red. On the basis of simple measures, three complex measures of activity were constructed: using the fuzzy conjunction (7) (Fig. 7a), using the fuzzy disjunction (8) (Fig. 7b), using the weighted average (9) (Fig. 7c). Correspondingly, 3, 8 and 5 anomalous areas were identified, respectively, whose configuration and location are different from the anomalous areas identified by the simple measures of activity. This proves the disjunctive, conjunctive and compromise character of the complex measures of activity (7–9) (Zlotnicki et al. 2005).

On Fig. 9a and Fig. 10a, the trend analysis of the  $\mu L_f$  are shown, based on the regression derivatives with respect to the sufficiently global measure of closeness  $\delta$ . The result of the analysis is the partition of the  $\mu L_f$  in increasing (red), decreasing (green) and flat areas (blue). As a result, a morphological separation of anomalies (Fig. 6a) trend components was carried out (Figs. 9b, 10b) (Gvishiani et al. 2008).



**Figure 5:** The measures of activity  $\mu D_f(t)$ : **a)** based on (2), **b)** based on (3).

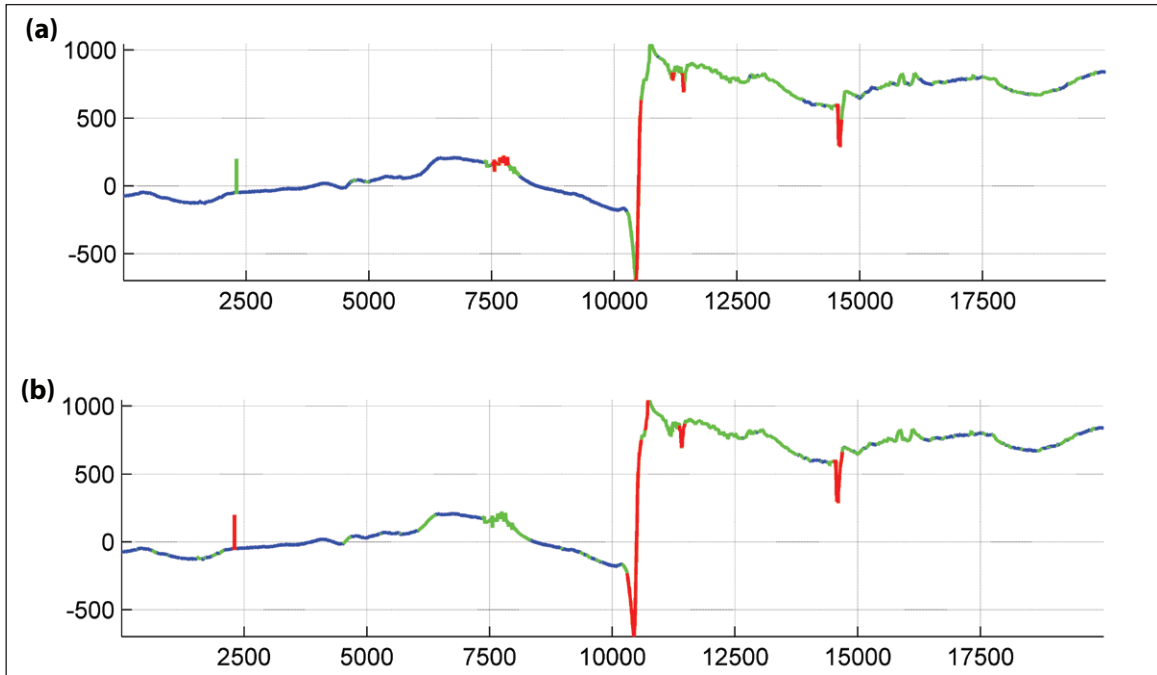


Figure 6: Anomalies marked out using the measures of activity (Fig. 5).

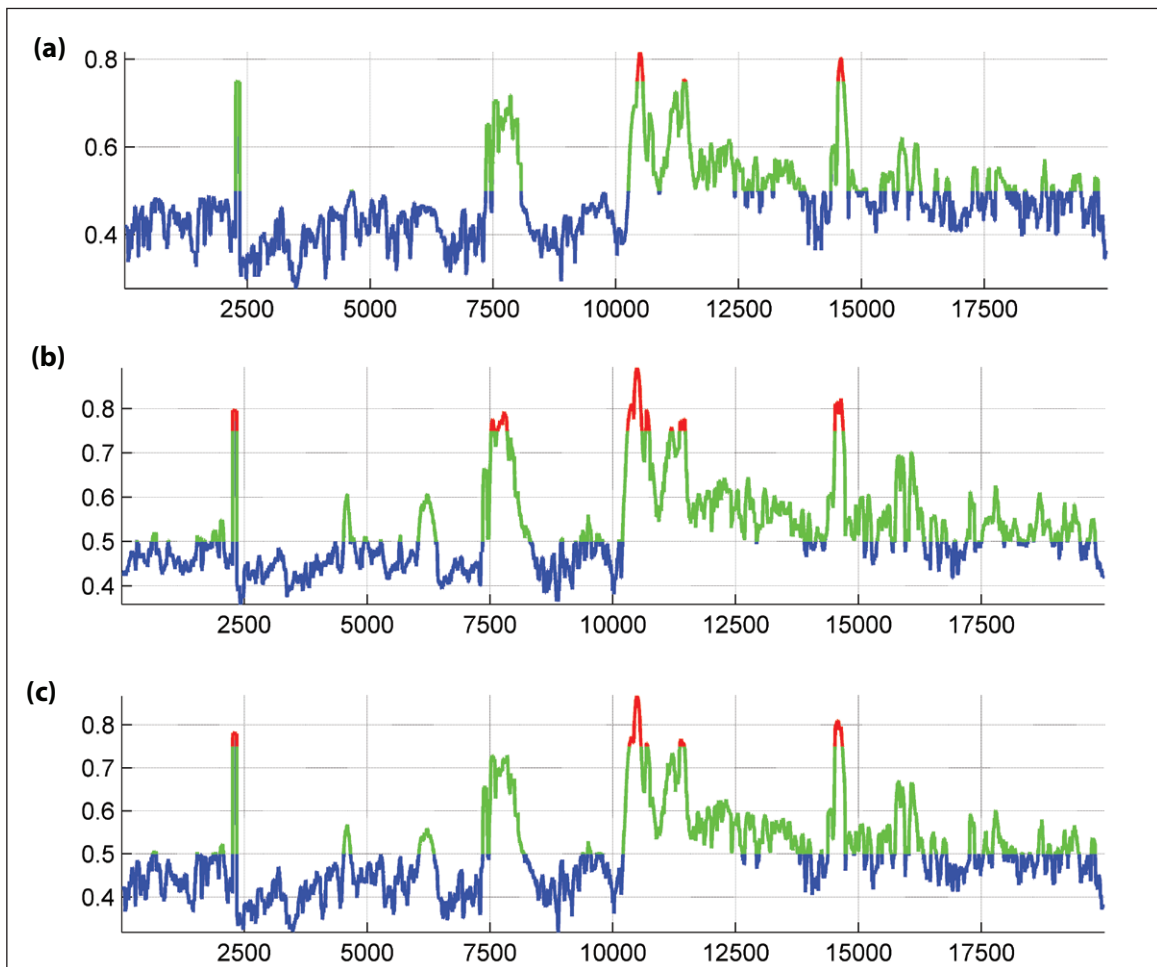
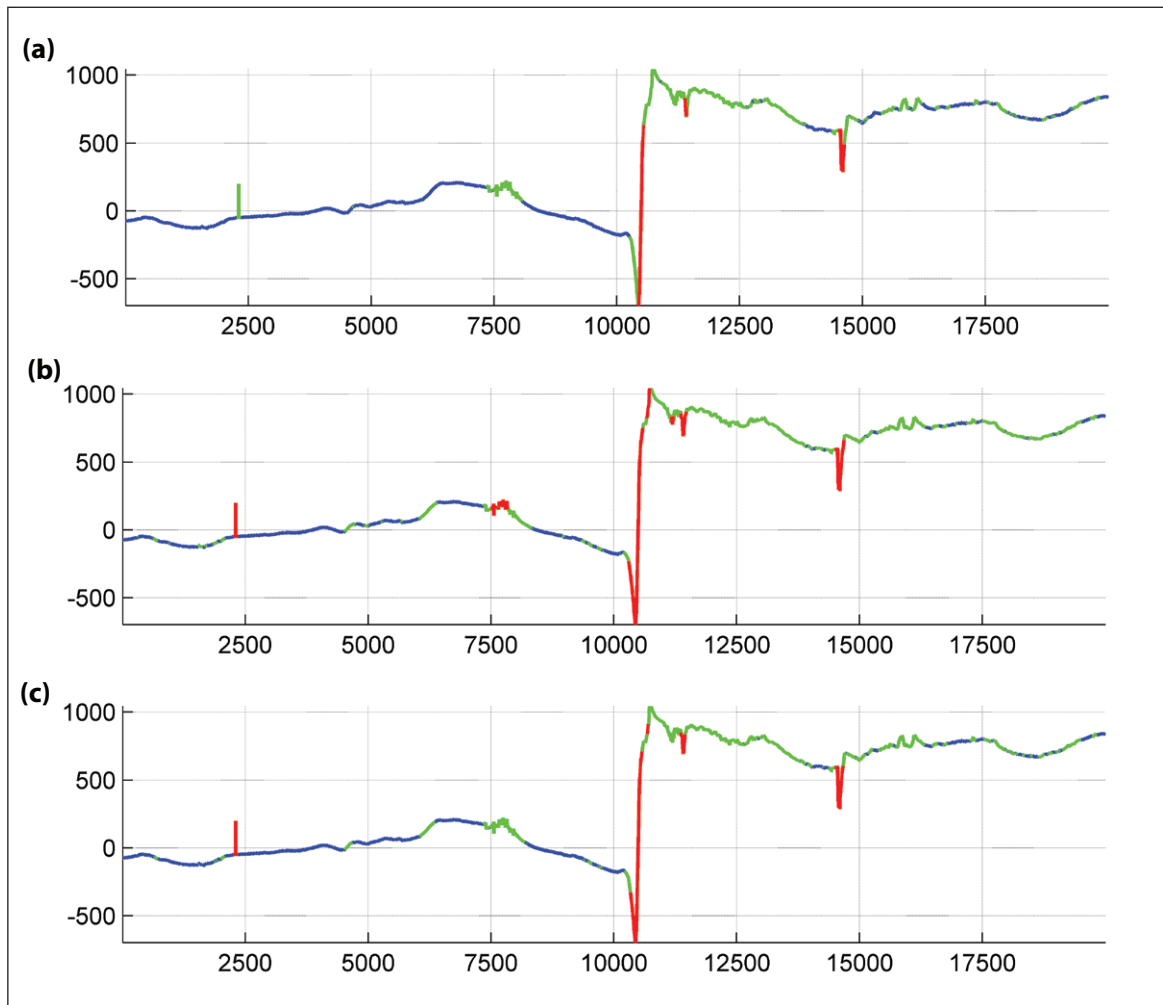
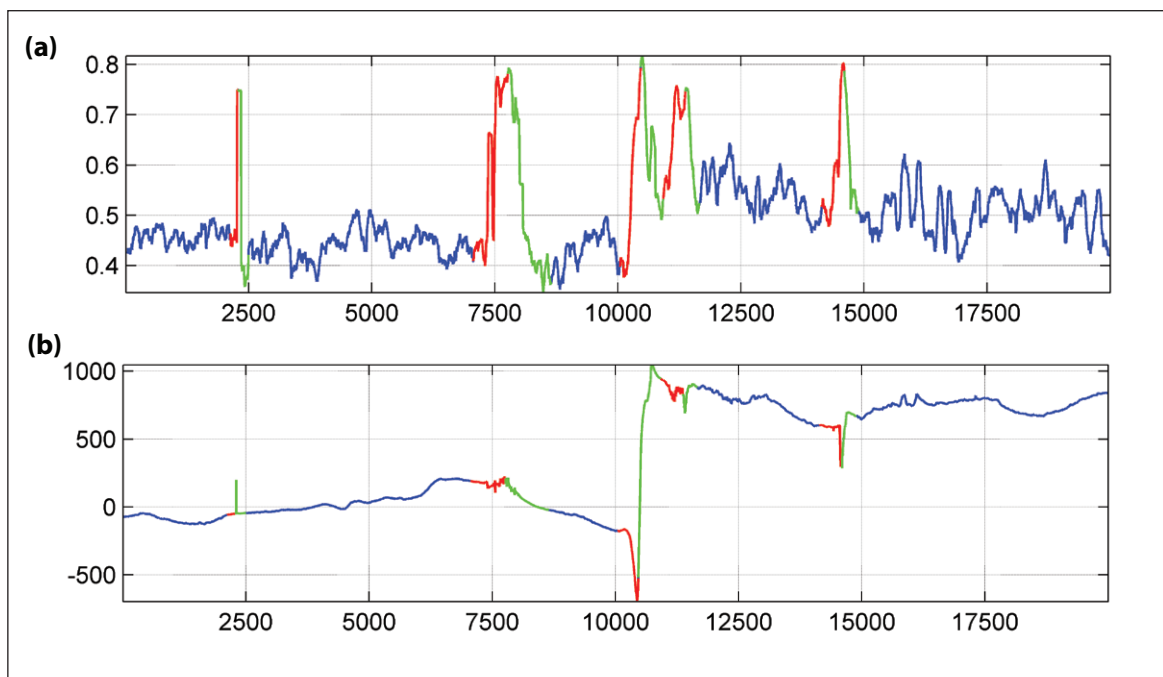


Figure 7: The measures of activity  $\mu\varepsilon_f(t)$ : **a)** based on (7), **b)** based on (8), **c)** based on (9).

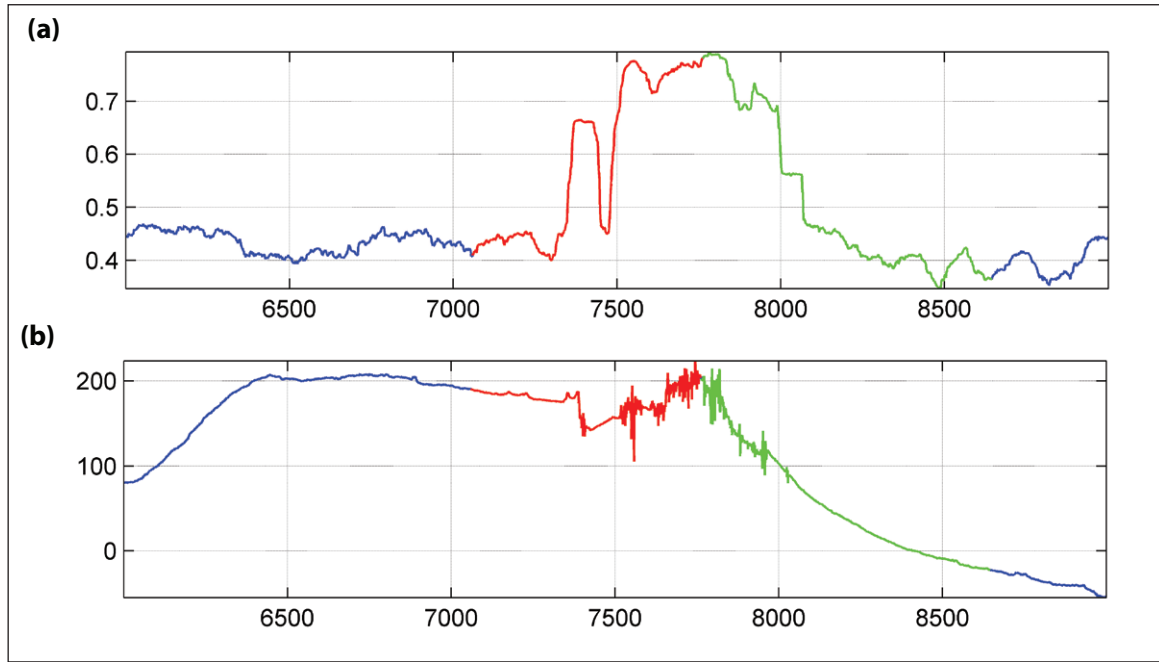




**Figure 8:** Anomalies marked out using the measures of activity (Fig. 7).



**Figure 9:** **a)** Areas of monotonicity (red is for increase, green is for decrease); **b)** Reduction to the initial record.



**Figure 10:** A fragment of initial record (Fig. 9): **a)** Areas of monotoneness (red is for increase, green is for decrease); **b)** Reduction to the initial record.

**Global  $\varepsilon$ -anomalies (conception)**

Globally anomalous fragments of a time series  $f$  from the perspective of the expert  $\varepsilon$  ( $\varepsilon$ -anomalies) are the uplifts of the relief  $\mu\varepsilon_f$

$$e - \text{anomalies of } f \equiv \text{uplifts of } \mu\varepsilon_f.$$

The second part of the study is dedicated to the search for uplifts on an arbitrary non-negative relief. Their search is realized using the DPS algorithm in two stages.

**The DPS algorithm**

Let  $X$  be a finite set, and  $A, B, \dots$  and  $x, y, \dots$  are the subsets and the points in it, respectively.

The *density*  $P$  on a subset  $X$  is a mapping of  $2^X \times X$  into the interval  $[0, 1]$ , increasing with respect to the first argument:

$$P(A, x) = P_A(x)$$

$$\forall x \in X, A \subset B \Rightarrow P_A(x) \leq P_B(x)$$

The value of  $P_A(x)$  is defined as the density of the subset  $A$  in the point  $x$ .

The subset  $A$  is called the  $\alpha$ -perfect in  $X$  for  $\alpha \in [0, 1]$ , if

$$A = \{x \in X : P_A(x) \geq \alpha\}$$

The process of construction of the maximal  $\alpha$ -perfect subset  $X(\alpha)$  in  $X$  is called the algorithm DPS (Agayan et al. 2011; Agayan et al. 2014):

$$X(\alpha) = \text{DPS}(X|P, \alpha)$$

The subset  $X(\alpha)$  results from the intersection

$$X(\alpha) = \bigcap_{k=1}^{\infty} X^k(\alpha)$$

where

$$X^{k+1}(\alpha) = \{x \in X^k(\alpha) : P_{X^k(\alpha)}(x) \geq \alpha\}$$

The intersection  $\bigcap_{k=1}^{\infty}$  is necessarily achieved because it's always finite due to finiteness of  $X$  and the multiplicity of  $X^{k+1}(\alpha)$  into  $X^k(\alpha)$  for all  $k = 0, 1, \dots$

### The search for $\mathcal{E}$ -anomalies: stage 1 – functional clustering

The search for  $\mathcal{E}$ -anomalies will be illustrated for the function  $f$  (Fig. 11) and the simple measure of activity  $\mu_{L_f}$  (Fig. 12) corresponding to the dynamical indicator of jaggedness (2).

Now we define the density  $P$  on  $T$  by the measure of dynamic activity  $\mu_{D_f}$ : let  $\Delta > h$  be the local observation parameter,  $A$  is a subset in  $T$ ,  $t$  – node from  $T$ , then

$$P_A(t) = \sum \frac{\mu_{e_f}(\bar{t})}{|T|} : \bar{t} \in [t - \Delta, t + \Delta] \cap A$$

The result of the DPS algorithm operation with such density on  $T$  with a fixed  $\alpha$  is the first part of the search for anomalous nodes for the function  $f$  on the period of observation  $T$  from the perspective of the expert  $\mathcal{E}$  (Fig. 13):

$$T(\alpha) = \text{DPS}(T|P, \alpha, \mu_{\mathcal{E}_f})$$

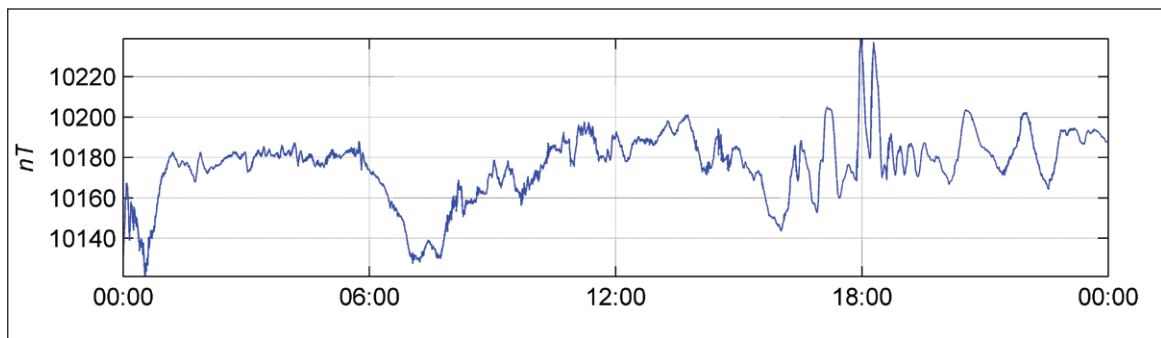


Figure 11: The initial time series.

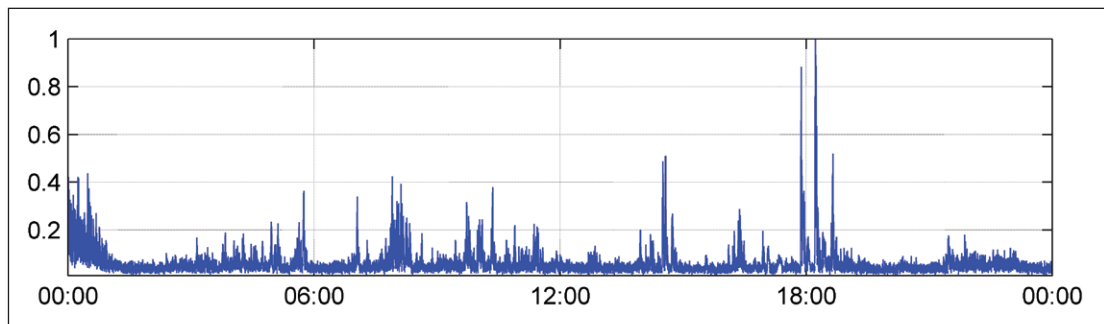


Figure 12: The corresponding activity.

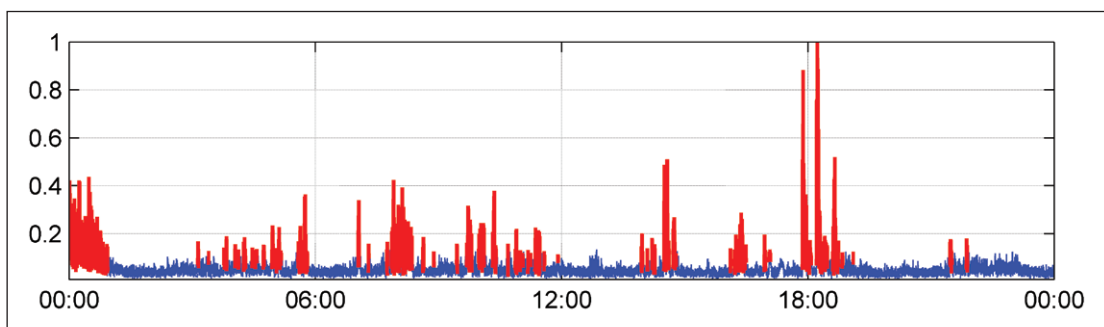


Figure 13: DPS: step 1. Results of identification of anomalies on the indicator of activity.

The set  $T(\alpha)$  is assumed as a disjunctive union of the intervals  $[a_k, b_k], k = 1, \dots, m$ . Every  $[a_k, b_k]$  is located in  $T$  and has no gaps inside (**Fig. 13**):

$$T(\alpha) = \bigvee_{k=1}^m [a_k, b_k]$$

**The search for  $\varepsilon$ -anomalies: stage 2 – interval clustering**

Let  $X(\alpha)$  be the totality of the intervals  $\{[a_k, b_k]_{k=1}^m\}$ , and let us introduce the measure of closeness  $\rho$  between two intervals from  $X(\alpha)$ :

$$\rho([a_{k_1}, b_{k_1}], [a_{k_2}, b_{k_2}]) = \frac{|[a_{k_1}, b_{k_1}]| + |[a_{k_2}, b_{k_2}]|}{|\min(a_{k_1}, a_{k_2}), \max(b_{k_1}, b_{k_2})|}$$

The closeness  $\rho$  helps to construct the density  $\mathcal{P}$  on  $X(\alpha)$ : if  $S \subset X(\alpha), [a_k, b_k] \in X(\alpha)$ , then

$$\mathcal{P}_S = \max_{[a_{\bar{k}}, b_{\bar{k}}] \in S \setminus [a_k, b_k]} \rho([a_k, b_k], [a_{\bar{k}}, b_{\bar{k}}])$$

The result of the algorithm operation DPS with such density on  $X(\alpha)$  with a fixed  $\beta$  is the second part of the search for anomalous points for the function  $f$  on the period of observation  $T$  from the perspective of the expert  $\varepsilon$  (**Figs. 14, 15**)

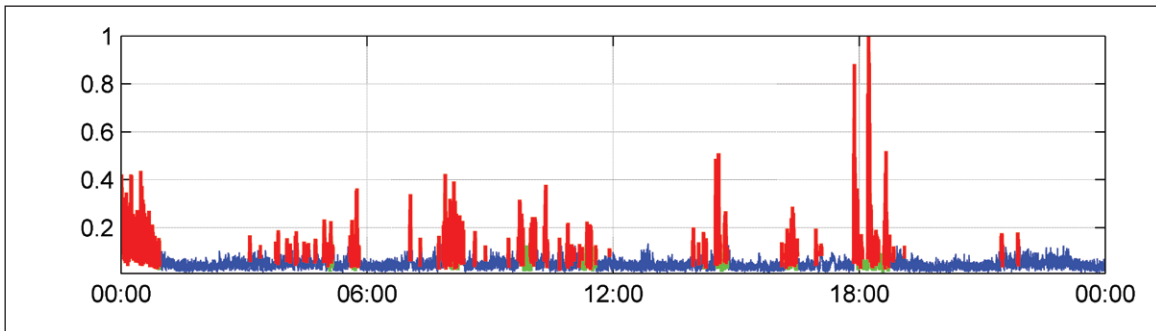
$$X(\alpha)(\beta) = \text{DPS}(X(\alpha)|\mathcal{P}, \beta, \mu\varepsilon_f)$$

On the second step, the DPS algorithm indicates which intervals in  $X(\alpha)$  should additionally be joined into one anomaly – these are exactly the intervals from  $X(\alpha)(\beta)$  (Agayan et al. 2014).

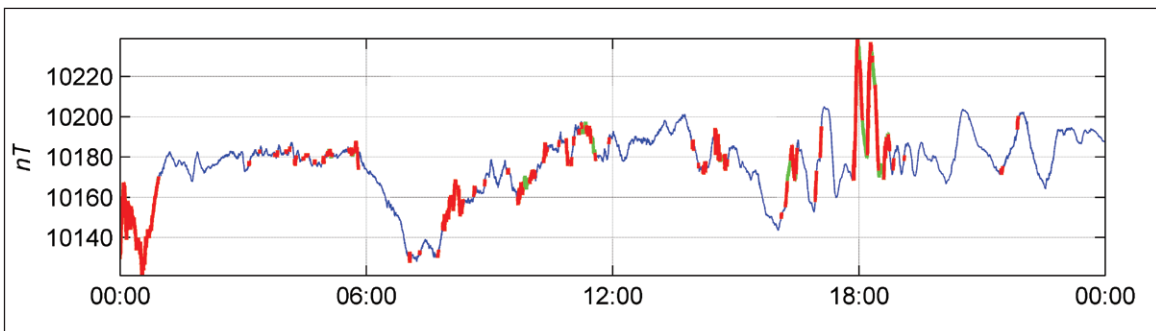
Thus, the finally anomalous intervals for the function  $f$ , from the perspective of the expert  $\varepsilon$ , would be the fairly isolated ones from  $X(\alpha)$  which are not included into  $X(\alpha)(\beta)$ , and also the new intervals which result from uniting the intervals from  $X(\alpha)(\beta)$ .

Let us denote the totality of all the anomalies  $\mathcal{A} = \mathcal{A}(\mu\varepsilon_f)$ :

$$\mathcal{A} = \{A_i = A_i(\mu\varepsilon_f)\}_{i=1}^l$$



**Figure 14:** DPS: step 2. Clustering of the intervals.



**Figure 15:** DPS: final result. Reduction to the initial function.

### Studying the $\varepsilon$ -anomalies

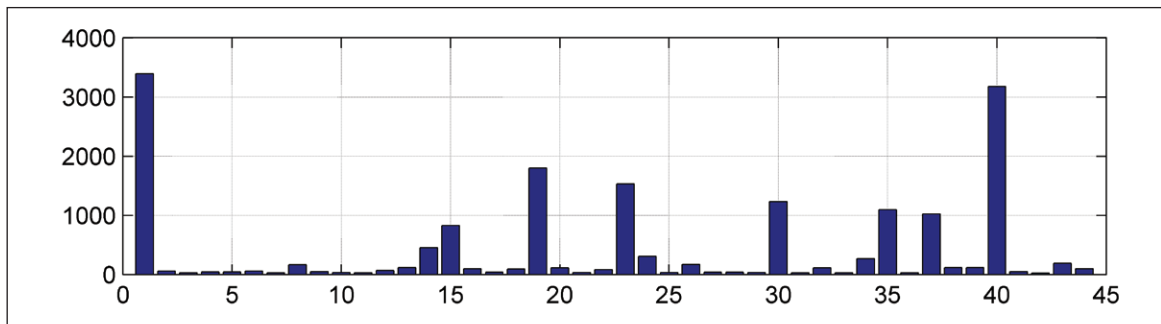
Let us denote  $\mathcal{A}(\mu\varepsilon_f) = \{A_i |_{t=i}^t\}$  the totality of all the  $\varepsilon$ -anomalies detected by DPS. Any  $\varepsilon$ -anomaly  $A_i$  defines the fragment  $f_i = f|_{A_i}$  of the time series  $f$  and could be studied according to the program listed above with respect to the measure  $\mu\varepsilon_f$  (the internal  $A_i$ -monitoring of activity, superanomalies inside  $A_i$ , and so on). It is natural to call such study of  $A_i$  the internal study. As an ambivalent an external study of the anomaly  $A_i$  is supposed, characterizing  $A_i$  among the other anomalies from  $\mathcal{A}(\mu\varepsilon_f)$ . Let's talk about it more in detail.

There are many characteristics of  $\varepsilon$ -anomalies, each of them ranks the  $\varepsilon$ -anomalies in  $\mathcal{A}(\mu\varepsilon_f)$  in its own way. The characteristics can be simple and complex (an analogy to the simple and complex measures of activity). Keeping in mind a compromise between simplicity and thoroughness, mentioned in the beginning, we shall list five natural characteristics and display how to make the complex ones from them.

The DPS algorithm identified 44 anomalies on the record  $f$  (**Fig. 15**). Let us construct the simple characteristics from them:

1. Massiveness (continuity) of the anomaly  $A_i$  (**Fig. 16**):

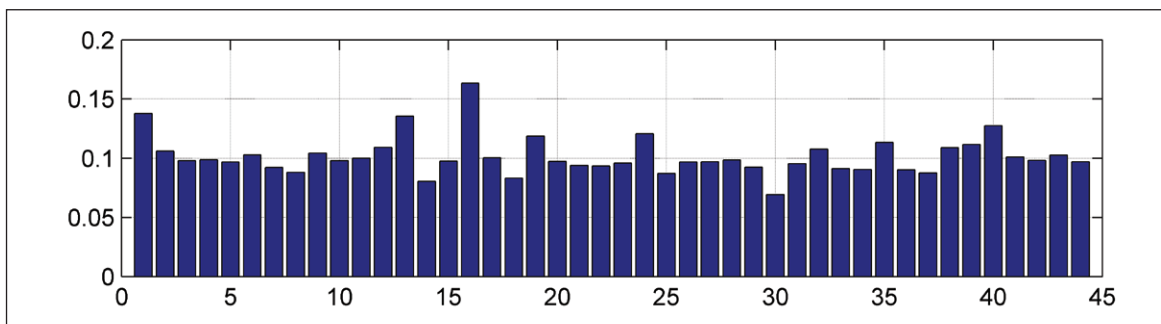
$$mA_i = |A_i|$$



**Figure 16:** The massiveness of anomalies.

2. Mean activity of the anomaly  $A_i$  (**Fig. 17**):

$$\mu A_i = \frac{\sum \mu\varepsilon_f(t) : t \in A_i}{|A_i|}$$



**Figure 17:** Mean activity of the anomalies.

3. Mean energy of the anomaly  $A_i$  (**Fig. 18**):

$$E A_i = \frac{\sum E_f(t) : t \in A_i}{|A_i|}$$

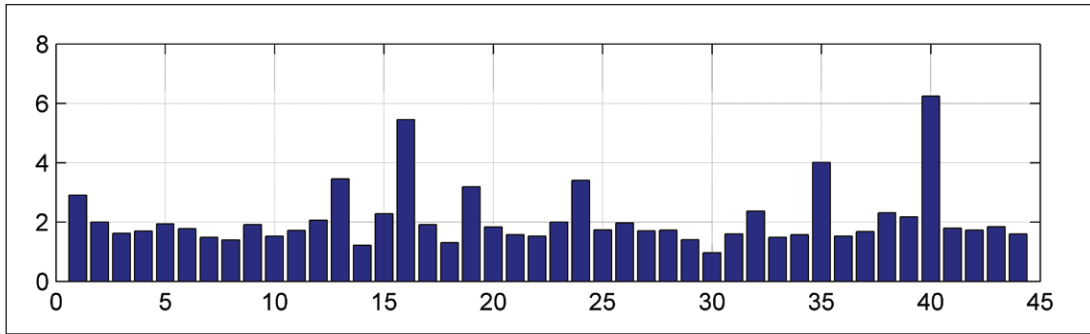


Figure 18: Mean energy of the anomalies.

4. Mean jaggedness of the anomaly  $A_t$ (Fig. 19):

$$LA_t = \frac{\sum L_f(t) : t \in A_t}{|A_t|}$$

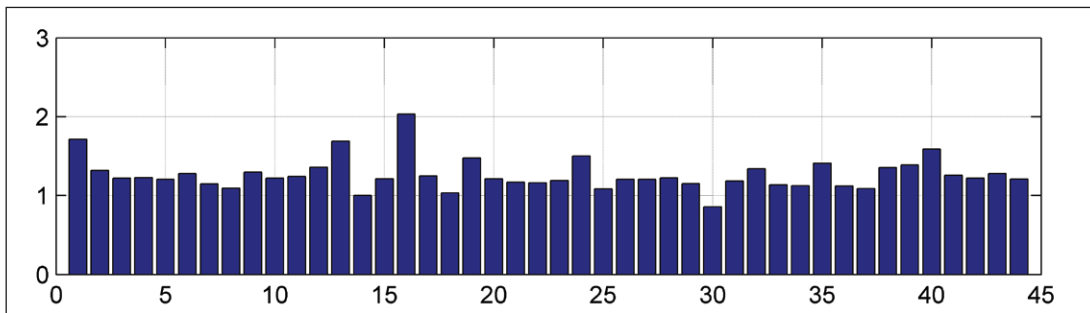


Figure 19: Mean jaggedness of the anomalies.

5. Mean scatterness of the anomaly  $A_t$ (Fig. 20):

$$OA_t = \frac{\sum O_f(t) : t \in A_t}{|A_t|}$$

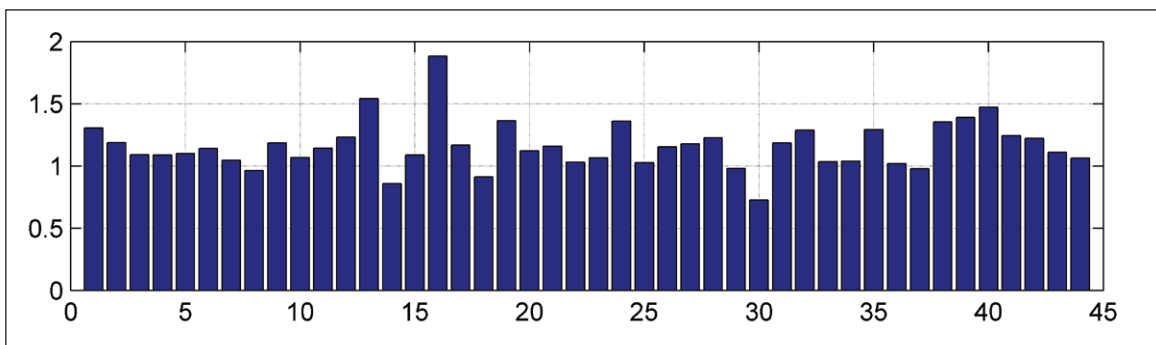


Figure 20: Mean scatterness of the anomalies.

As a result, five totalities occur:

$$mA(\mu_{\varepsilon_f}) = \{mA_t |_{t=1}^L\}; \mu A(\mu_{\varepsilon_f}) = \{\mu A_t |_{t=1}^L\};$$

$$EA(\mu_{\varepsilon_f}) = \{EA_t |_{t=1}^L\}; LA(\mu_{\varepsilon_f}) = \{LA_t |_{t=1}^L\}; OA(\mu_{\varepsilon_f}) = \{OA_t |_{t=1}^L\}$$

The comparison of every anomaly  $A$  with the corresponding indicators of other anomalies from  $\mathcal{A}(\mu\varepsilon_f)$  based on the measure of maximality  $\text{mesmax}$  gives five indices of the anomaly  $A$ , similar to dynamical indicators:

1.  $\text{ind}_m A = \text{mesmax}_{m\mathcal{A}(\mu\varepsilon_f)} mA$  (Fig. 21)

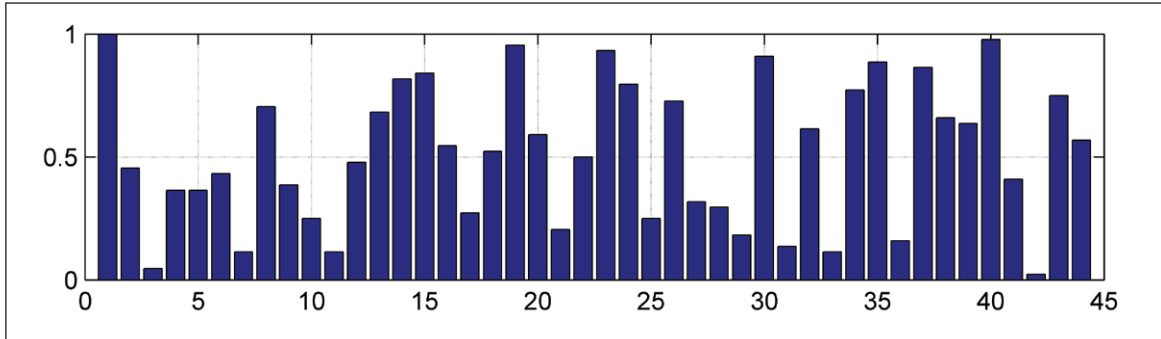


Figure 21: Index of anomalies based on massiveness.

2.  $\text{ind}_\mu A = \text{mesmax}_{\mu\mathcal{A}(\mu\varepsilon_f)} \mu A$  (Fig. 22)

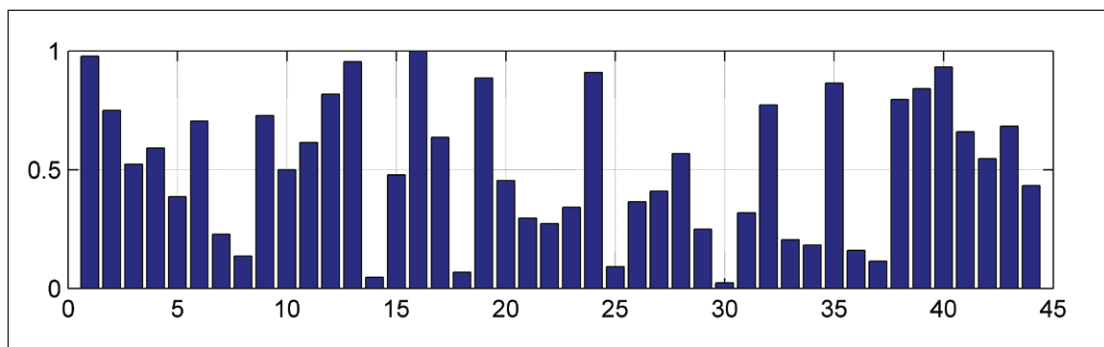


Figure 22: Index of anomalies based on mean activity.

3.  $\text{ind}_E A = \text{mesmax}_{E\mathcal{A}(\mu\varepsilon_f)} EA$  (Fig. 23)

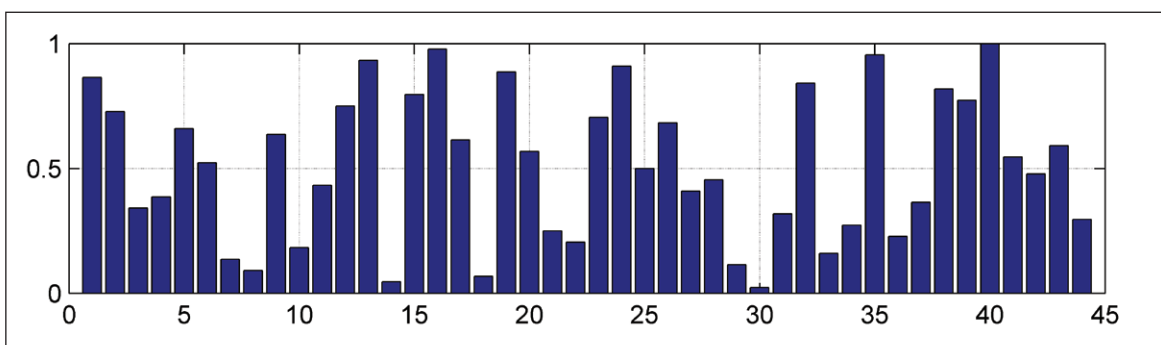


Figure 23: Index of anomalies based on mean energy.

4.  $\text{ind}_L A = \text{mesmax}_{L\mathcal{A}(\mu\varepsilon_f)} LA$  (Fig. 24)

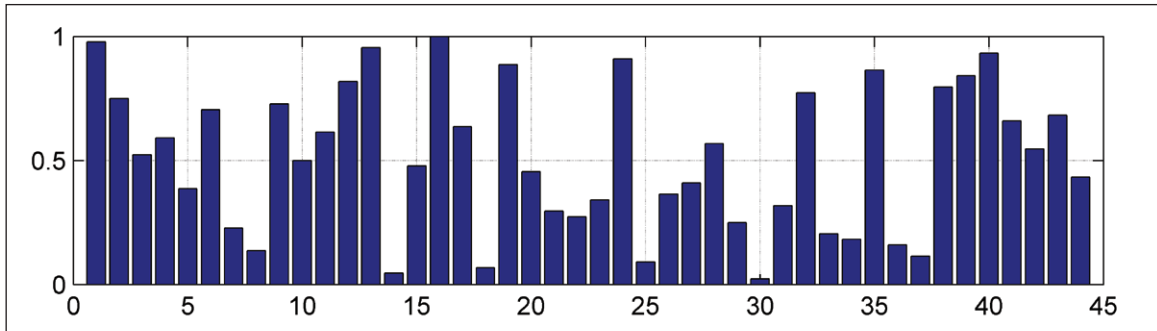


Figure 24: Index of anomalies based on mean jaggedness.

5.  $ind_{\circ}A = \text{mesmax}_{OA(\mu \varepsilon_j)} OA$  (Fig. 25)

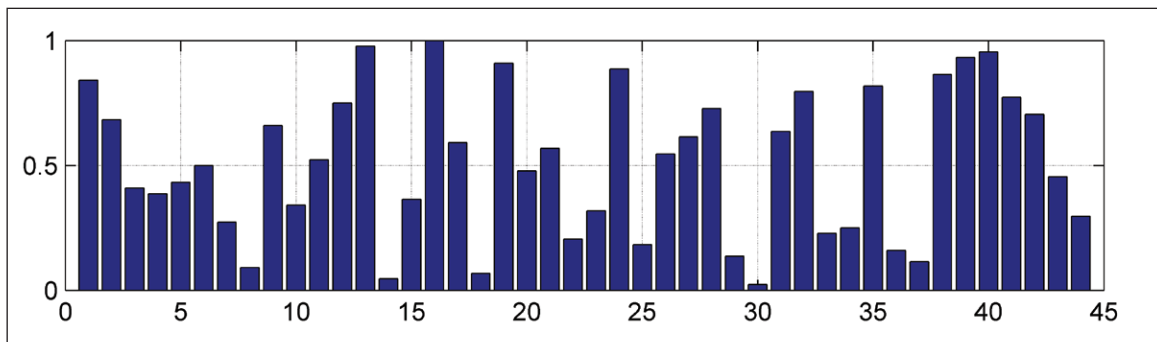


Figure 25: Index of anomalies based on mean Scatterness.

Simple indices, using the FL \* operations, produce the new complex integral indices of anomalies  $\widetilde{ind}$ .

Example 7 A complex index based on two simple indices (Fig. 26)

$$\widetilde{ind}A = \frac{ind_L A + ind_{\circ}A}{2} \tag{11}$$

Example 8 On Fig. 27a, five of 44 anomalies identified using the  $ind_L A$  index are displayed, on Fig. 27b anomalies identified using the  $ind_{\circ}A$  are displayed, and on Fig. 27c anomalies marked out using the  $\widetilde{ind}A$  index (11) are shown.

### Measure and monitoring of geomagnetic activity

Let us analyze the geomagnetic monitoring (10–10.3) on the basis of the measure of activity  $\mu = \mu L_f$  (4) constructed according to the dynamical indicator of irregularity  $L_f(2)$ . The application of this measure to the ground-based observation data from the global geomagnetic observatory network brought an opportunity

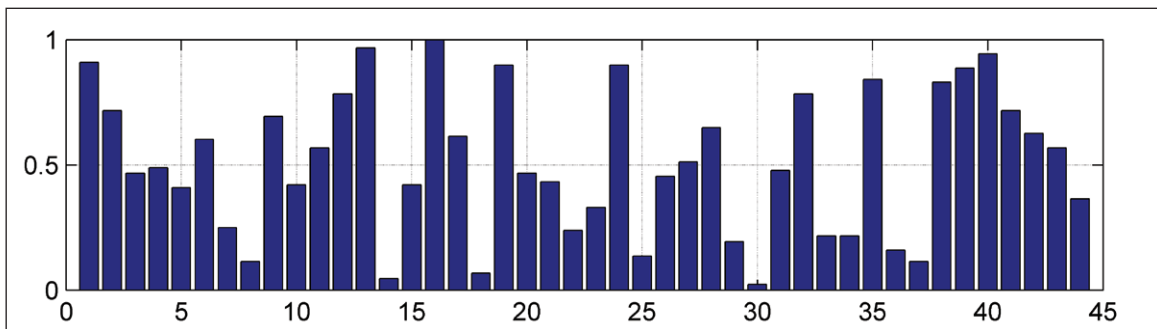
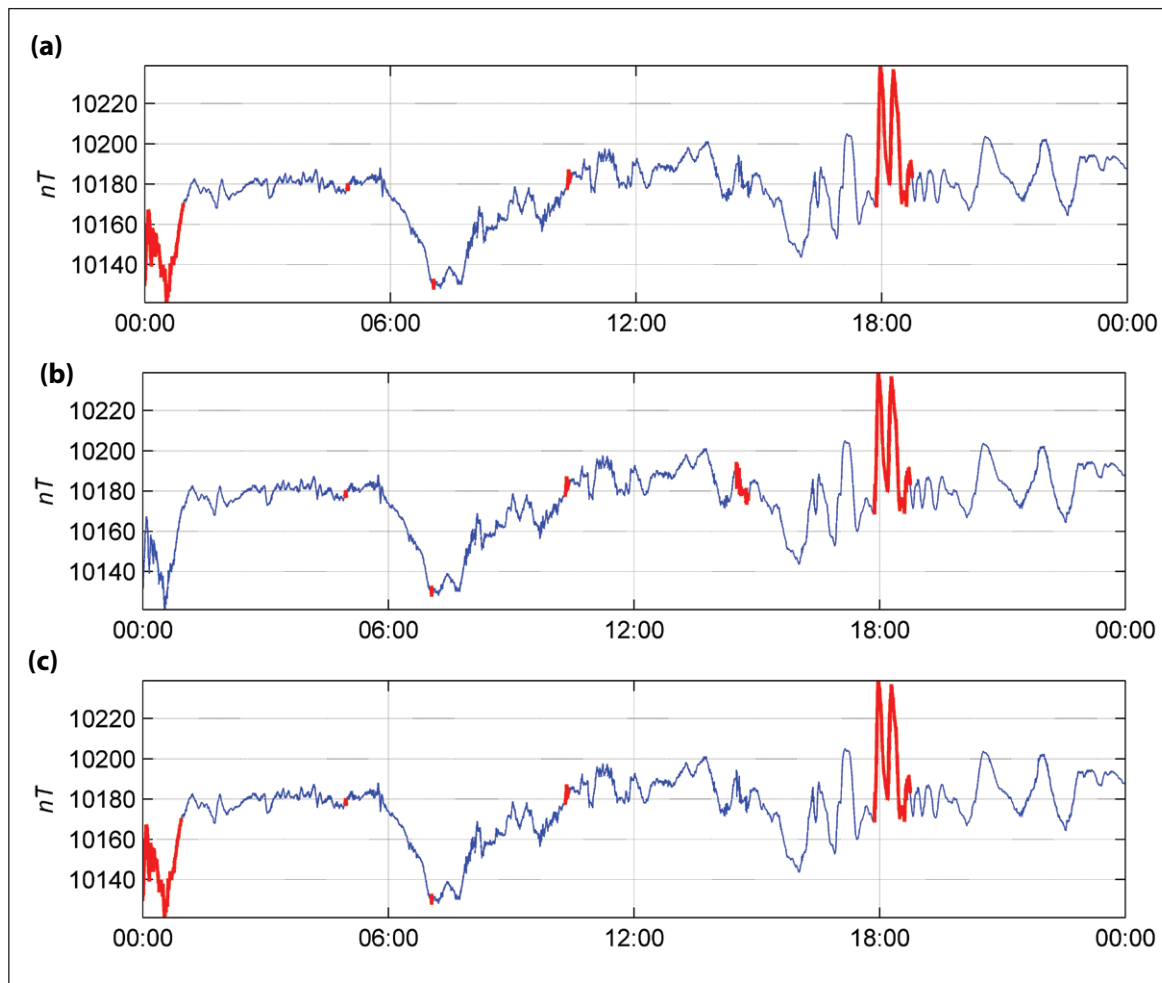


Figure 26: Complex index of anomalies.





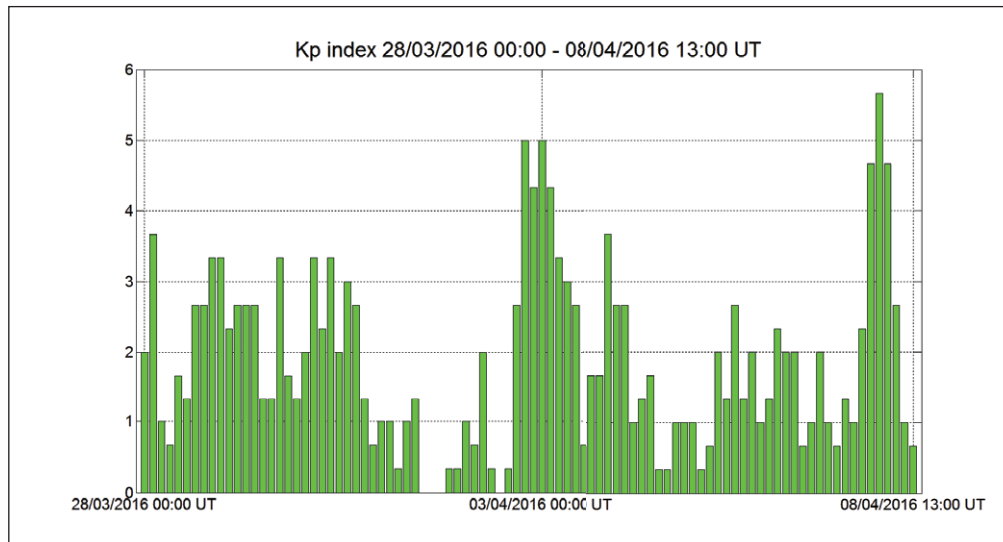
**Figure 27:** Anomalies marked out using: **a)**  $ind_L A$ ; **b)**  $ind_O A$ ; **c)**  $\widetilde{ind} A$ .

of spatial-temporal analysis of geomagnetic activity and the estimation of the dynamics of magnetic storm propagation in realtime (Soloviev et al. 2013; Gvishiani et al. 2014; Soloviev et al. 2016). The usage of the measure makes it possible to estimate the geomagnetic activity in various regions of the Earth within a single scale, taking its regional specification into account. The developed measure allows the analysis of the internal fine structure of magnetic storms, the dynamics of their development both around the globe and within particular regions. When estimating the geomagnetic activity, the whole range of initial data is efficiently considered in, as the temporal resolution of the measure matches the initial data sampling frequency.

The example of application of the developed measure of anomaly to the geophysical data analysis was the study of magnetic activity using the time series from the magnetic observatory “Saint Petersburg” (IAGA code SPG). The initial observatory data is available at the Russian-Ukrainian Geomagnetic Data Center website (<http://geomag.gcras.ru>). For illustration purposes, the data for the period March 28 – April 8, 2016 was taken, as it contains the fragments of both increased (April 2–3, April 7–8) and decreased (March 31 – April 1, April 5) magnetic activity.

According to the planetary  $Kp$  index data (**Fig. 28**), during the given time period the magnetic conditions varied from calm (the  $Kp$  index values were from 0 to 2) and slightly disturbed ( $Kp = 2 \dots 3$ ) to disturbed (the  $Kp$  index values were up to 5+). The disturbed data fragments corresponded to the magnetic storms which occurred from April 7, 2016, 17:00 UT, to April 8, 2016, 02:00 UT, and the increased magnetic activity on April 2–3, 2016.

The magnetic disturbance, which lasted from April 2, 2016, 15:00 UT to April 3, 2016, 03:00 UT, was characterized first by the increase of the amplitude of variations of the components  $X$ ,  $Z$  and the total intensity  $F$  by 70–80 nT, after which the values of these components smoothly decreased to  $-180 \dots -200$  nT from



**Figure 28:** Planetary  $K_p$  index values for the period from 28.03.2016 to 08.04.2016.

their initial levels. This minimum took place approximately at midnight on April 2–3, next the values of the components reached the initial values by 03:00 UT, April 3, 2016. The magnetic storm which took place on April 7–8 was characterized by rapid evolution. During the sudden commencement period (about 17:11 UT) the amplitude of the variations of the  $X$  component at the Saint Petersburg magnetic observatory increased respectively to its background value from 14565 to 14580 nT. During the main phase of the magnetic storm, which evolved approximately from 19:00 to 21:53 UT, the decrease by about 150 nT for the  $X$  component, 180 nT for the  $Y$  component, and more than 200 nT for the  $Z$  component took place. According to the data records of several Russian magnetic observatories located in the same latitudinal belt, the main phase was accompanied by the decrease, which varied between 100–150 and 250 nT for the  $X$  and  $Y$  components, and was up to 300 nT for the  $Z$  component. The recovery phase of the magnetic storm lasted from about April 7, 2016, 22:00 UT to April 8, 2016, 02:00–02:30 UT. During this phase, some oscillations in the magnetic field intensity with a period of about 1 hour and the maximal amplitude of about 140–150 nT were observed. By the end of the recovery phase, the magnetic field at the observatory reached the values of the same level, which was registered before the storm onset.

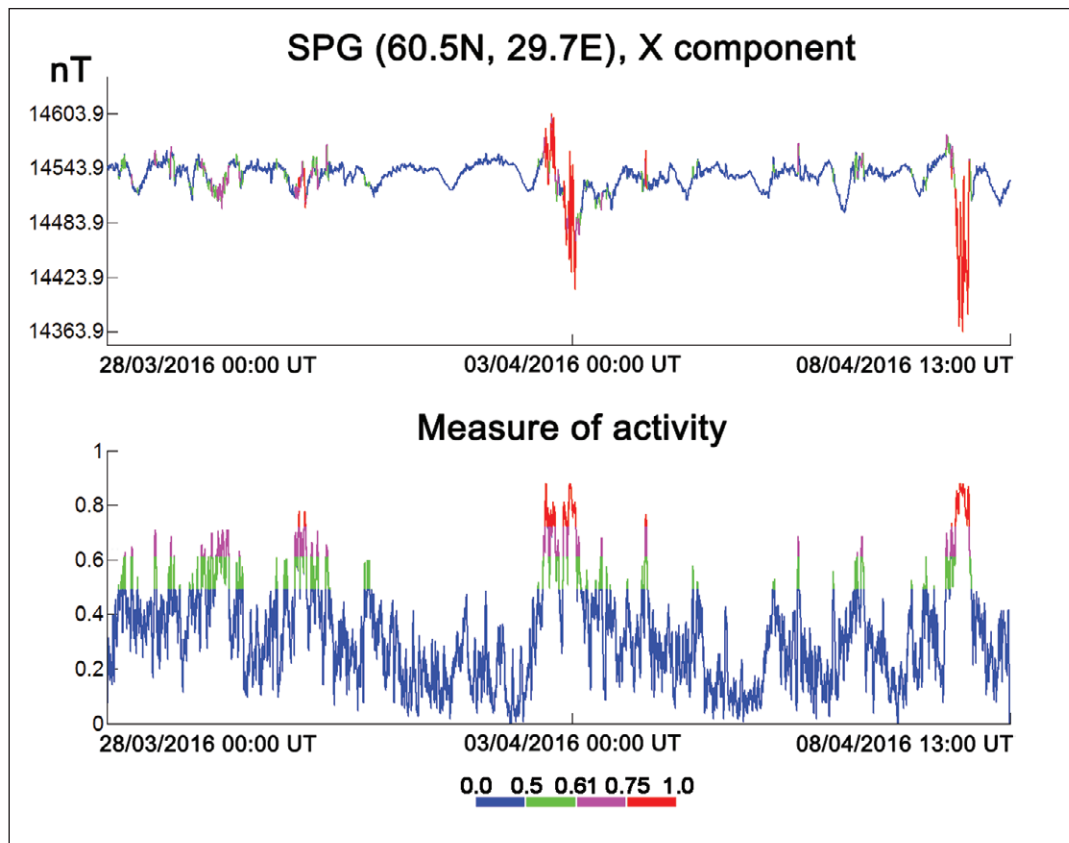
The monitoring of geomagnetic activity for its Northern component  $X$  was performed using the measure of activity  $\mu = \mu L_f$ . Its results are displayed on **Fig. 29**. The geomagnetic events listed above were considered anomalous according to the measure  $\mu$  and marked red on the plot. The data registered during the period of less disturbed magnetic field (e.g., for the period of 28–29 March), were recognized as potentially anomalous and marked green and purple on the plot (the potentially anomalous fragments are divided into two parts by the threshold value 0.61). Finally, for the period of quiet magnetic field, March 31–April 1, the values of the measure  $\mu$  practically do not exceed the threshold value 0.5, and no data fragment for this period was classified as anomalous.

## Discussion

The comparison of the result of monitoring with the planetary  $K_p$  index data demonstrated that the classification of magnetic activity using the described measure does not contradict the classical methods for its estimation. At that, the principal advantages of the measure of anomaly should be mentioned:

1. the matching of the temporal resolution of the measure to the initial sampling frequency of the time series;
2. the estimation of anomaly using the records from different observatories in a unified and normalized scale.

The first advantage gives an opportunity of application of the measure to the monitoring of magnetic disturbance with minimal temporal delay in realtime mode. At that, the estimation of the magnetic activity is performed less raw than using the traditional 3-hour  $K$  index. The second advantage makes it possible



**Figure 29:** The results of classification of initial data into background (blue), weakly anomalous (green), anomalous (purple) and strongly anomalous (red) fragments using the measure of anomaly. On the upper plot the initial record is shown, on the bottom plot the measure of anomaly is displayed, calculated for this record.

to study the magnetic disturbance regardless of latitudinal location of the observatory, taking the regional specifications into account.

## Conclusions

An exponential growth of digital observations of the Earth’s environment acts as a challenge for the development of adequate mathematical methods of data mining and systems analysis of input data. It is the usage of these methods that provides an opportunity operational extraction of useful information, such as specific geomagnetic events of different origin, from the large volumes of continuously collected data. The approach to the studying of time series based on fuzzy logic, proposed in the paper, is largely focused on this challenge and finds its practical application. In particular, several algorithms based on DMA (Soloviev et al. 2009; Bogoutdinov et al. 2010; Soloviev et al. 2012; Sidorov et al. 2012; Zelinskiy et al. 2014) are integrated into Russian-Ukrainian Geomagnetic Data Center (<http://geomag.gcras.ru>) and used for continuous automated detection of anthropogenic disturbances and geomagnetic pulsations in observatory data.

Further research plans include, first of all, the expansion of analysis from a single time series  $f$  to a network  $S$  of time series  $f_s(t)$ ,  $s \in S$ . This will allow to define the monitoring of activity in network, anomalous events in network, and more.

The results presented in this paper rely on data collected at magnetic observatories. We thank the national institutes that support them and INTERMAGNET for promoting high standards of magnetic observatory practice ([www.intermagnet.org](http://www.intermagnet.org)).

The research has been carried out in the framework of the Federal target program of the Ministry of Education and Science of the Russian Federation, agreement No. 14.607.21.0058, project unique identifier RFMEFI60714X0058, and is a part of research aimed at the creation of the experimental prototype of the hardware and software complex for monitoring and detection of extreme geomagnetic phenomena using ground-based and satellite data.

## Competing Interests

The authors have no competing interests to declare.

## References


- Agayan, S M, Bogoutdinov, Sh R and Dobrovolsky, M N** 2011 On one algorithm for searching the dense areas and its geophysical applications Materials of 15th Russian national workshop “Mathematical methods of pattern recognition, MMRO-15”, pp. 543–546 (in Russian).
- Agayan, S M, Bogoutdinov, Sh R and Dobrovolsky, M N** 2014 Discrete Perfect Sets and Their Application in Cluster Analysis. *Cybernetics and Systems Analysis*, 50(2): 176–190. DOI: <http://dx.doi.org/10.1007/s10559-014-9605-9>
- Agayan, S M and Soloviev, A A** 2004 Recognition of dense areas in metric spaces basing on crystallization. *System Research and Information Technologies*, 2: 7–23.
- Bogoutdinov, Sh R, Gvishiani, A D, Agayan, S M, Solovyev, A A and Kihn, E** 2010 Recognition of disturbances with specified morphology in time series. part 1: spikes on magnetograms of the worldwide INTERMAGNET network. *Izv. Phys. Solid Earth*, 46(11): 1004–1016. DOI: <http://dx.doi.org/10.1134/S1069351310110091>
- Gvishiani, A D, Agayan, S M and Bogoutdinov, Sh R** 2008 Fuzzy recognition of anomalies in time series. *Dokl. Earth Sci.*, 421(5): 838–842. DOI: <http://dx.doi.org/10.1134/S1028334X08050292>
- Gvishiani, A D, Agayan, S M, Bogoutdinov, Sh R, Graeva, E M, Zlotnicki, J and Bonnin, J** 2008 Recognition of anomalies from time series by fuzzy logic methods. *Russian Journal of Earth Sciences*, 10: ES1001. DOI: <http://dx.doi.org/10.2205/2007ES000278>
- Gvishiani, A D, Agayan, S M, Bogoutdinov, Sh R, Ledenev, A V, Zlotnicki, J and Bonnin, J** 2003 Mathematical methods of geoinformatics. II. Fuzzy-logic algorithms in the problems of abnormality separation in time series. *Cybern Syst. Anal.*, 39(4): 555–563. DOI: <http://dx.doi.org/10.1023/B:CASA.0000003505.56410.4f>
- Gvishiani, A D, Agayan, S M, Bogoutdinov, Sh R, Tikhotsky, S A, Hinderer, J, Bonnin, J and Diament, M** 2004 Algorithm FLARS and recognition of time series anomalies. *System Research & Information Technologies*, 3: 7–16.
- Gvishiani, A D, Agayan, S M, Bogoutdinov, Sh R, Zlotnicki, J and Bonnin, J** 2008 Mathematical methods of geoinformatics. III. Fuzzy comparisons and recognition of anomalies in time series. *Cybern. Syst. Anal.*, 44(3): 309–323. DOI: <http://dx.doi.org/10.1007/s10559-008-9009-9>
- Gvishiani, A, Lukianova, R, Soloviev, A and Khokhlov, A** 2014 Survey of Geomagnetic Observations Made in the Northern Sector of Russia and New Methods for Analysing Them. *Surveys in Geophysics*, 35: 1123–1154. DOI: <http://dx.doi.org/10.1007/s10712-014-9297-8>
- Mikhailov, V, Galdeano, A, Diament, M, Gvishiani, A, Agayan, S, Bogoutdinov, Sh, Graeva, E and Sailhac, P** 2003 Application of artificial intelligence for Euler solutions clustering. *Geophysics*, 68(1): 168–180. DOI: <http://dx.doi.org/10.1190/1.1543204>
- Sidorov, R V, Soloviev A A and Bogoutdinov, Sh R** 2012 Application of the SP algorithm to the INTERMAGNET magnetograms of the disturbed geomagnetic field. *Izvestiya Physics of the Solid Earth*, 48(5): 410–414. DOI: <http://dx.doi.org/10.1134/S1069351312040088>
- Soloviev, A, Agayan, S and Bogoutdinov, S** 2016 (in press) Estimation of geomagnetic activity using measure of anomalousness. *Annals of Geophysics*, 59(6). DOI: <http://dx.doi.org/10.4401/ag-7116>
- Soloviev, A A, Agayan, S M, Gvishiani, A D, Bogoutdinov, S R and Chulliat, A** 2012 Recognition of disturbances with specified morphology in time series. Part 2: Spikes on 1-second magnetograms. *Izv. Phys. Solid Earth*, 48(5): 395–409. DOI: <http://dx.doi.org/10.1134/S106935131204009X>
- Soloviev, A A, Bogoutdinov, Sh R, Agayan, S M, Gvishiani, A D and Kihn, E** 2009 Detection of hardware failures at INTERMAGNET observatories: Application of artificial intelligence techniques to geomagnetic records study. *Russ. J. Earth Sci.*, 11: ES2006. DOI: <http://dx.doi.org/10.2205/2009ES000387>
- Soloviev, A, Bogoutdinov, S, Gvishiani, A, Kulchinskiy, R and Zlotnicki, J** 2013 Mathematical Tools for Geomagnetic Data Monitoring and the INTER-MAGNET Russian Segment. *Data Science Journal*, 12: WDS114–WDS119. DOI: <http://dx.doi.org/10.2481/dsj.WDS-019>
- Soloviev, A A, Yu Shur, D, Gvishiani, A D, Mikhailov, V O and Tikhotskii, S A** 2005 Determination of the magnetic moment vector using cluster analysis of the local linear pseudoinversion of Delta T anomalies. *Doklady earth sciences*, 404(7): 1068–1071.

- Zelinskiy, N R, Kleimenova, N G, Kozyreva, O V, Agayan, S M, Bogoutdinov, Sh R and Soloviev, A A** 2014 Algorithm for recognizing Pc3 geomagnetic pulsations in 1-s data from INTERMAGNET equatorial observatories. *Izvestia-Physics of the Solid Earth*, 50(2): 240–248. DOI: <http://dx.doi.org/10.1134/S106935131402013X>
- Zlotnicki, J, LeMouel, J-L, Gvishiani, A, Agayan, S, Mikhailov, V and Bogoutdinov, Sh** 2005 Automatic fuzzy-logic recognition of anomalous activity on long geophysical records. Application to electric signals associated with the volcanic activity of la Fournaise volcano (Reunion Island). *Earth and Planetary Science Letters*, 234: 261–278. DOI: <http://dx.doi.org/10.1016/j.epsl.2005.01.040>

**How to cite this article:** Agayan, S, Bogoutdinov, S, Soloviev, A and Sidorov, R 2016 The Study of Time Series Using the DMA Methods and Geophysical Applications. *Data Science Journal*, 15: 16, pp.1–21, DOI: <http://dx.doi.org/10.5334/dsj-2016-016>

**Submitted:** 20 June 2016    **Accepted:** 10 October 2016    **Published:** 06 December 2016

**Copyright:** © 2016 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 