# Sub-interval images. Big Data

Harin, Alexander

Modern University for the Humanities

22 November 2021

# Sub-interval images. Big Data

Alexander Harin
aaharin@gmail.com
Modern University for the Humanities

A systematic introduction to sub-interval images (or SI-images or S-IIs) is presented here. General outlook of possible use of the SI-analysis for Big Data is given.

Basic notions of S-IIs are formulated including cuboids of gravity and sub-interval copies of databases. Two concepts of SII-indexing are proposed for Big Data databases.

The S-IIs can be used in, e.g., search, and recognition in databases in, e.g., accounting and audit, micro- and macroeconomics, especially in Big Data databases.

Contents

# 1. Introduction

## 1.1. Preliminaries. Organization of the article

A sub-interval analysis (SI analysis or SI-analysis or S-I analysis or SIA or S-IA) was founded in 2011-2012 in reports and working papers of the author of the present article (see, e.g., Harin 2011a-c and 2012a-d).

The prerequisites of the sub-interval analysis are the needs for the tools for a number of possible fields of applications of the S-IA. They can include, e.g., decision theory, econometrics, accounting and audit, micro- and macroeconomics, databases, image analysis, search, and pre-recognition, theory of measurements, long-term and unfinished processes, etc. The realized considerations have confirmed the usefulness of the applications of the sub-interval analysis.

The systematic introduction to the basics of the S-IA (namely the basics of a SI-arithmetic) was started in Harin 2020. The present article develops it. In particular, sub-interval images or pictures (SI images or SI pictures or SI-pictures or S-I images or SI-images or SII or S-II) are introduced and considered here.

This draft version of the article is organized as follows. Section 1 presents its prehistory and a brief outlook of the SI-arithmetic. Section 2 presents initial notions of the S-IA. Section 3 presents cuboids of gravity. Section 4 presents original-corresponded, uniform, and mixed S-IIs. Section 5 presents computer-corresponded S-IIs. Section 6 presents sub-interval copies of databases. Section 7 presents a short consideration of sub-intervals and Big Data. Section 8 presents two concepts of S-II indexing for Big Data. Section 9 presents a combination of these two concepts for Big Data and hypothetical examples of such a combination. Section 10 presents conclusions. The article is finished by References.

## 1.2. General information on the sub-interval analysis. SI-arithmetic

Consider an interval $\boldsymbol{X} = [a, b] : 0 < (b-a) < \infty$. Consider a set of points $\{x_s\} : s = 0, 1, \ldots, S : 0 < S < \infty$, on this interval such that

$$\underline{a \equiv x_0 < x_1} < x_2 < \ldots < x_s < x_{s+1} < \ldots < x_S \equiv b .$$

This set of points divides the interval $\boldsymbol{X}$ into a set of $S$ adjacent sub-intervals $\{\boldsymbol{X}_s\} \equiv \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_S$. Due to this division, $\boldsymbol{X}$ may be denoted as $\boldsymbol{X}_{1..S}$.

The boundaries of $\{\boldsymbol{X}_s\}$ can be defined by various manners, for example by $\boldsymbol{X}_s \equiv [x_{s-1}, x_s)$ except of the far right sub-interval $\boldsymbol{X}_S \equiv [x_{S-1}, x_S] \equiv [x_{S-1}, b]$. The main condition of such definitions of the division is that any point of the interval should unambiguously belong to only one sub-interval.

So the interval $\boldsymbol{X}_{1..S}$ is divided into a set of adjacent sub-intervals $\{\boldsymbol{X}_s\}$.

The lengths of the sub-intervals may be denoted as $L_s$. They can be normalized by the length of the whole interval $L_{1..S} \equiv b-a$ and we have $l_s = L_s/L_{1..S}$ and $l_{1..S} \equiv 1$.

Suppose a set of quantities $\{W_s\} : W_s \geq 0,\ s = 1, 2, \ldots S : 1 < S < \infty$ and

$$\sum_{s=1}^{S} W_s = W \equiv W_{1..S} < \infty .$$

For the purposes of the SI-analysis, the quantities $\{W_s\}$ may be named as the weights of the sub-intervals and may be normalized by the whole weight $W_{1..S}$ as $w_s = W_s/W_{1..S}$ and $w_{1..S} \equiv 1$. The normalized (or relative) weights and also $W_{1..S} \equiv 1$ will be used here as a rule due to their convenience.

Generally $W_s$ may be assumed as, e.g., pointwise.

We can write an ensemble of the formulae named as a "Ring of formulae"

$$\Delta G_{1..S} = \sum_{n=1}^{S} w_n l_n =$$

$$= L_{1..S} - \sum_{m=1}^{S} w_m \sum_{q\in[1,S],|\,q\neq m} l_q = .$$

$$= L_{1..S} - \sum_{p=1}^{S} l_p \sum_{r\in[1,S],|\,r\neq p} w_r$$

When we know $P \le S$ lengths, $M \le S$ weights, and $N \le \min(M, P)$ both lengths and weights, we can write an ensemble of the inequalities that is named as a main "Chain of inequalities" or "Main chain of inequalities"

$$\sum_{n=1}^{N} w_n l_n \le \Delta G_{1..S} \le L_{1..S} - \max \begin{cases} \sum_{m=1}^{M} w_m \sum_{q\in[1,S],|\,q\neq m} l_q \\ \sum_{p=1}^{P} l_p \sum_{r\in[1,S],|\,r\neq p} w_r \end{cases}.$$

So these are the main features of the sub-interval arithmetic.

## 2. Initial notions of sub-interval images

A sub-interval image (or SI-image or S-II) will be referred to as a sub-interval representation of a set (in other words of an $n$-dimensional picture, where $n \ge 1$) or of some tracks or derivatives (see below) of this set and/or of its S-IIs.

The dimensionality of the set corresponds to real situations to be considered. The dimensionality of the S-II of this set corresponds to the goals of the consideration.

### 2.1. SII-approximations, SII-tracks, SII-derivatives

A sub-interval image of a set will be referred to as a SII-approximation of the set if this image can be considered as a sub-interval copy of this set and the resolution of at least one sub-interval is less than that of the set.

A sub-interval image of a set will be referred to as a SII-track of the set if this image represents only some features (at least one feature) of this set.

A sub-interval image of a set will be referred to as a SII-derivative of the set if this image represents a derivative or derivatives of this set and/or its track(s).

## 2.2. Uniform and non-uniform sub-interval images

A sub-interval image will be referred to as uniform for a dimension if for this dimension the lengths of the sub-intervals are equal to each other.

A sub-interval image will be referred to as a non-uniform one for a dimension if for this dimension the lengths of the sub-intervals are not equal to each other.

A uniform sub-interval image will be referred to as a binary-corresponded one for a dimension if for this dimension the lengths of the sub-intervals can be obtained by the division of the length of the total interval by two or by its powers.

A non-uniform sub-interval image such that the sub-intervals of this S-II or their boundaries correspond to some characteristic features of the original set will be referred to as an original-corresponded sub-interval image.

## 2.3. S-IIs of texts. Black-and-white pictures

One-dimensional sets and sub-intervals, for example, S-IIs of texts can be considered as well. The chapters, subchapters,.. paragraphs and sentences can play the role of sub-intervals. The symbols (only letters and/or letters with spaces between the letters) can play the role of the weight. Pictures, formulae, tables, etc. as another/additional objects can be also accounted in such sub-intervals.

Two-dimensional sets (black-and-white pictures) and their sub-interval images will be considered here as a rule for the sake of simplicity and clarity.

## 2.4. About the exactness of S-IIs

The exactness of S-IIs is evidently limited by two sources. The first source is a limited exactness of the measurements. The second source can arise if an analyzed point (or a value) coincides with a border between sub-intervals.

In both these cases the one more adjacent sub-interval (or the two adjacent sub-intervals) should be used to overlap the problematic area of the limited exactness of the measurements or of the coincidence points.

If there are some coincidences of some points (or values) with the boundaries of sub-intervals then the sequential numbers of the records and sub-intervals, and information on the coincidences can be stored in a separate array.

## 3. Cuboids of gravity

### 3.1. Main notions

An $n$-dimensional sub-interval will be referred to as the $p$-percent cuboid of gravity if it contains the center of gravity of the total $n$-dimensional interval (at least at the boundary) and $p$ percent of the weight of the interval.

If the weights of the rest (left and right) parts of the total interval are equal to $(100\% - p\%)/2$ for each dimension then it will be referred to as a regular cuboid of gravity and by default as simply a cuboid of gravity.

In the case of the unequal rest parts it will be referred to as an irregular one.

If the percentage of the weight is not the same for each dimension (if the percentages are different for different dimensions) then the cuboids of gravity can be named as correspondingly regular or irregular parallelepipeds of gravity.

**An important note.** The regular cuboids of gravity have equal percentage of the weight in each dimension, but from the point of their geometrical sizes they are often parallelepipeds except of symmetric distributions of the weight.

The centers of gravity may be denoted as **CoGs** or **CroGs**.

The parallelepipeds of gravity may be denoted as **PoGs**.

The cuboids of gravity may be denoted as **CoGs** or generally, especially when centers of gravity are also used in the same text, as **CboGs**.

Evidently a cuboid of gravity as an $n$-dimensional sub-interval can be composed of some smaller sub-sub-intervals.

The 50% (or 25%-50%-25%) CboG will be referred to as the standard one.

One-dimensional cuboids of gravity can be also named simply as sub-intervals of gravity. A $p$-percent sub-interval of gravity contains $p$ percent of the weight of the total interval and the weights of the rest (left and right) parts of the total interval are equal to $(100\% - p\%)/2$.

Two-dimensional cuboids of gravity can be also named as quadrats of gravity or **QoG**. A $p$-percent quadrat of gravity contains $p$ percent of the weight of the total interval and the weights of the rest neighbor parts of the total interval are equal to $(100\% - p\%)/2$ for each of the two dimensions. If the percentage of the weight is not the same for each of the two dimension then the quadrats of gravity can be named as rectangles of gravity.

#### 3.1.1. Concentric cuboids of gravity

Systems of concentric cuboids of gravity may be used in sub-interval images. For example, 10% : 30% : 50% : 70% (beginning from the middle of the total interval) systems may be used. Such systems may be also named due to their lateral parts as 45% : 35% : 25% : 15% systems. Systems 1/8 : 1/4 : 2/4 : 3/4 and 0 : 1/4 : 2/4 : 3/4 (where "0" represents the center of gravity and minimal cuboid of gravity) can be also used.

3.2. Examples of CboGs

Let us choose any two-dimensional picture, e.g. a picture of a digit "4" or "5" (see Figure 10) for a postal envelope as an example for explanation of CboGs.

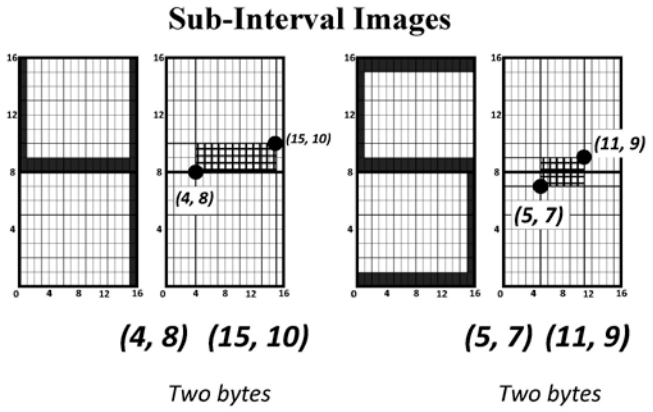


Figure 1. Digits "4" and "5" and their sub-interval images
namely their standard cuboids of gravity

This or any other two-dimensional picture is divided by the grid 16x16 and by 3x3 sub-intervals. Then sub-interval images are formed (see Figures 2-7). (In this particular case, due to the character of the postal envelope digits, the grids are elongated two times in the vertical direction)

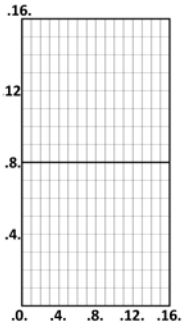


Figure 2. The grid 16×16

For the sake of simplicity and obviousness let us demonstrate the process of formation of the standard cuboid of gravity (CboG) on the uniform picture.

Let us form the 1/4 : 1/2 : 1/4 image. Let us measure 1/4 of the weight of the picture in the vertical direction from the bottom and from the top (see Figure 8).

Let us measure 1/4 of the weight of the picture in the vertical direction from the bottom and from the top (see Figure 3) .

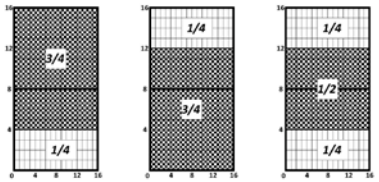


Figure 3. The formation of three vertical sub-intervals namely of the standard CboG of the uniform picture. The sub-interval weights equal to 1/2, 1/4, 1/8

Let us measure 1/4 of the weight of the picture in the horizontal direction from the left and from the right (see Figure 4)

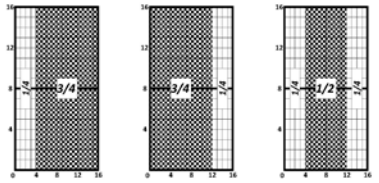


Figure 4. The formation of three horizontal sub-intervals for the standard CboG 1/4-1/2-1/4 on the uniform picture

So we obtain the sub-interval image of the uniform picture. The weight of the central sub-interval is 1/2. This image is depicted on Figure 5. The images with the weights of the concentric central sub-intervals 1/4 and 1/8 are depicted also.

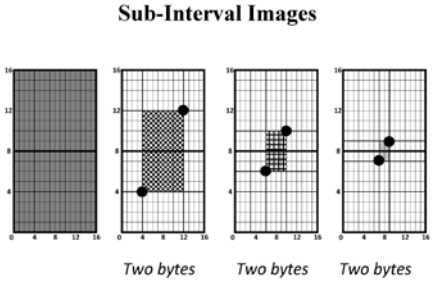


Figure 5. The concentric CboG sub-interval images of the uniform picture. The central concentric sub-interval weights equal to 1/2, 1/4, 1/8

We may obtain the sub-interval image of, e.g., a picture of a digit "4" (see Figure 11) for a postal envelope in a similar way.

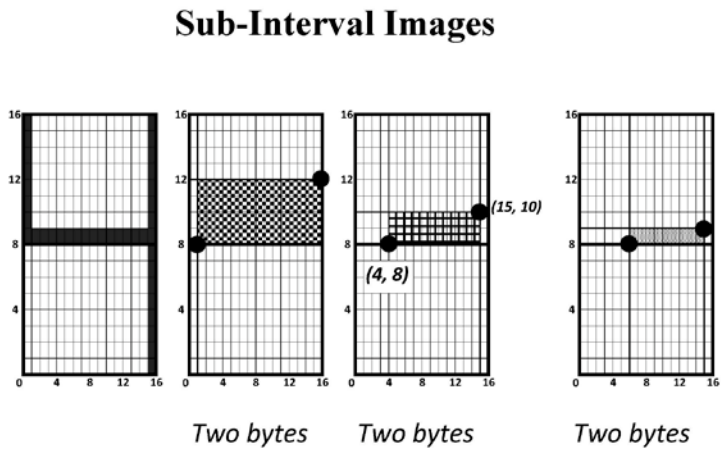


Figure 6. The CboG concentric sub-interval images of the digit "4".
The concentric central sub-interval weights equal to 1/2, 1/4, 1/8

We may obtain the sub-interval image of, e.g., a picture of a digit "5" (see Figure 12) for a postal envelope in a similar way.

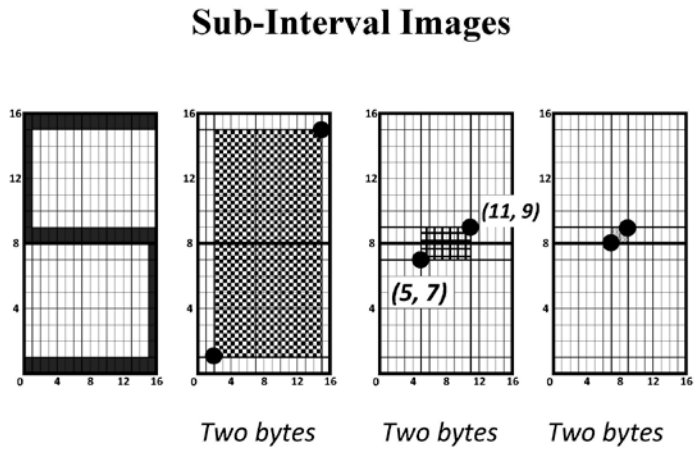


Figure 7. The concentric CboG sub-interval images of the digit "5".
The concentric central sub-interval weights equal to 1/2, 1/4, 1/8

### 3.3. About analytical methods of formation of CboGs

Cuboids of gravity may be obtained in a general form by purely analytical methods. For example, let us suppose an $N$-dimensional box $[a_1, b_1; \ldots a_N, b_N]$ and a local weight density which has a form of a nonnegative integrable $N$-dimensional function $w(x_1, \ldots x_N)$ such as

$$x_n \in [a_n, b_n]$$

and

$$w(x_1, ..., x_N) \geq 0$$

and the total weight $W$ of the function $w(x_1, \ldots x_N)$ is finite

$$\int_{a_1}^{b_1} dx_1 ... \int_{a_N}^{b_N} w(x_1, ..., x_N) dx_N \equiv W < \infty .$$

Such a function may represent the density of an $N$-dimensional picture.

For example, the $(N-1)$-dimensional planes $X_{n.a4}$ and $X_{n.b4}$ for a 1/4 : 1/2 : 1/4 CboG may be found from the integral equations

$$\int_{a_1}^{b_1} dx_1 ... \int_{a_n}^{X_{n.a4}} dx_n ... \int_{a_N}^{b_N} w(x_1, ..., x_N) dx_N = \frac{W}{4}$$

and

$$\int_{a_1}^{b_1} dx_1 ... \int_{X_{n.b4}}^{b_n} dx_n ... \int_{a_N}^{b_N} w(x_1, ..., x_N) dx_N = \frac{W}{4} .$$

In practice, such planes for any picture may be simply found by means of a photo-element by measuring 1/4 of the total optical weight of the picture from the border to the middle of the picture in the direction of the $n$-th coordinate.

If there is an $N$-dimensional grid, for example $16^N$, then the planes should be averaged to nearest grid lines.

So, CboGs may be formed in general case for any $N$-dimensional picture, which density may be represented by abovementioned function $w(x_1, \ldots x_N)$.

Note, the cuboids of gravity are evidently more integrated spatial characteristics than sub-interval compressions of pictures.

### 3.4. To reduce the sizes of the minimal CboG

In the scope of the sub-interval images the CroGs are really represented as the minimal (or elementary) cuboids of gravity that include the exact CoGs.

One of the main goals of the S-IA and S-IIs is to reduce the sizes of these minimal cuboids of gravity.

### 3.5. To analyze parameters of typical cuboids

Another main goal of the S-IIs is to analyze parameters of typical cuboids.

This goal can include, for example, considerations of the standard cuboids of gravity: their coordinates, lengths of their sides (edges) and hypotenuses, mutual

ratios of their lengths, ratios of the sizes of the S-IIs to the lengths and hypotenuses of the standard CboGs.

Note that the mutual ratios of the lengths of CboGs are invariant to rotations by 90 degrees, reflections, shifts, uniform compressions, etc.

This goal can include also considerations of concentric CboGs including, for example, considerations and mutual comparison of parameters of different cuboids, e.g., of ¾ to ¼ cuboids.

Note that the mutual ratios of a number of parameters of different cuboids from concentric CboGs are invariant to some rotations, reflections, shifts, uniform compressions, etc.

## 4. Original-corresponded, uniform, and mixed S-IIs

### 4.1. Original-corresponded sub-interval images

Original-corresponded sub-interval images are non-uniform sub-interval images that represent characteristic features of the original pictures (sets).

#### 4.1.1. Plateaus and valleys. Slopes and cliffs

Original-corresponded sub-interval images represent the features of the original pictures (sets) such as parts, areas where the values are approximately constant or changed insignificantly. Such “flat” areas can seem as plateaus and valleys and the sub-intervals should correspond to them.

Original-corresponded sub-interval images should represent the features of the original pictures (sets) such as parts where the values are crucially or at least essentially changed. Every such slope and cliff should be represented by at least one boundary (or by at least one sub-interval) between the two adjoining sub-intervals.

#### 4.1.2. Possible general cases

In the general case the boundaries of the sub-intervals of an original-corresponded S-II may be non-straight lines and their intersections may be non-rectangular.

### 4.2. Uniform and mixed S-IIs

Uniform S-IIs are evidently more simple, regular, and often more suitable than non-uniform ones. This allows their natural use in computers (see below).

Suppose the sizes of the sub-intervals are much less than the sizes of the characteristic features of the original pictures. In this case, original-corresponded S-IIs can evidently be superimposed on the grid of the uniform S-IIs.

## 5. Computer-corresponded sub-interval images

### 5.1. Binary, and computer-corresponded S-IIs. Sub-intervals and computers. Natural conformity

Binary-corresponded S-IIs constitute an important category of the uniform sub-interval images. Probably the most important category of uniform binary sub-interval images is a sub-category of binary S-IIs that is harmonized with computers.

In particular, the number of sub-intervals in the total interval should exactly correspond to the sizes of the machine words (usually to the bytes). In this case the natural consistency between sub-interval images and their processing by computers can be reached.

Let us consider such computer-corresponded sub-interval images.

### 5.2. About the structure of computer-corresponded S-IIs. Primary and secondary collections

Let us consider a SII-database of two-dimensional data, e.g., of black-and-white photos and paintings. This database can be arranged as the combination of, at least, primary and secondary collections.

The primary collection represents the initial data. The secondary collections represent sub-interval images of the database records. The tertiary collections can exist that can represent pre-indexing of the database.

The primary collection consist of three parts.

The first part of the primary collection includes the **sequence numbers** (**SNs**) of the records (The records can be arranged e.g. chronologically in the order of their recording in the database). Since the sizes of the records can differ from each other, the second part includes the addresses of the beginnings and ends of the records. The third part includes the records (the original pictures) themselves.

Note 1. The first part can be omitted if the structure of computer memory allows to determine these SNs without any additional memory volume.

Note 2. If the records are recorded continuously and do not be modified, then the second part can include only the addresses of the beginnings of the records. At that the beginning of the next record can determine the end of the preceding one. So the addresses of the ends of the records can be omitted as well.

The secondary collections can consist of a number of parts.

The secondary collections can include SII-approximations, SII-tracks, SII-derivatives, blocks of the sequence numbers presorted in a proper manner, pre-assessed S-IIs, and pre-equalized S-IIs (see below).

## 5.4. Indexing. Pre-sorting. Pre-assessment. Pre-limitation

Indexing is an effective general tool to accelerate search and recognition.

One can distinguish some types of a preliminary indexing of the S-IIs. They include pre-sorting, pre-assessment, and pre-equalization (see below).

### 5.4.1. Pre-sorting

The sub-interval images can be used for a preliminary indexing of databases, especially of Big Data databases.

One can name the preliminary indexing of sub-interval images as pre-sorting by some parameters when every combination of these parameters corresponds to a set of sequence numbers (SNs) of the records.

Probably the simplest type of pre-sorting is a pre-sorting by means of SII-approximations. One can estimate the needs and possibilities for various S-IIs.

4×4 elementary sub-intervals with 2 grades (1/8 of a byte) of the weight for every elementary sub-interval need 16/8 = 2 bytes for every record. These S-IIs enable $(2^1)^{16} = 2^{16}$ combinations of the parameters.

4×4 elementary sub-intervals with 4 grades (1/4 of a byte) of the weight (white-light-dark-black) for every elementary sub-interval need 16/4 = 4 bytes for every record. These S-IIs enable $(2^2)^{16} = 2^{32} = 4$ Gigabits combinations of the parameters. This case can be considered, e.g., for the pre-sorting in distributed networks with low-power and low-cost terminals.

4×4 elementary sub-intervals with 16 grades (1/2 of a byte) of the weight for every elementary sub-interval need 16/2 = 8 bytes for every record. These S-IIs enable $(2^4)^{16} = 2^{64}$ combinations of the parameters.

8×8 elementary sub-intervals with 2 grades (1/8 of a byte) of the weight for every elementary sub-interval need 64/8 = 8 bytes $(2^8)$ for every record. These S-IIs enable $(2^1)^{64} = 2^{64}$ combinations of the parameters.

8×8 elementary sub-intervals with 4 grades (1/4 of a byte) of the weight for every elementary sub-interval need 64/4 = 16 bytes $(2^8)$ for every record. These S-IIs enable $(2^2)^{64} = 2^{128}$ combinations of the parameters.

…

64×64 elementary sub-intervals with 4 grades (1/4 of a byte) of the weight for every elementary sub-interval need 4096/4 = 1024 bytes $(2^{14})$ for every record. These S-IIs enable $(2^2)^{4096} = 2^{8192}$ combinations.

256×256 elementary sub-intervals with 4 grades (1/4 of a byte) of the weight for every elementary sub-interval need 4096/4 = 1024 bytes $(2^{10})$ for every record. These S-IIs enable $(2^2)^{65536} = 2^{131072}$ combinations.

### 5.4.2. Pre-assessment

Instead of, for example, the above $2^{131072}$ combinations of the pre-sorting, one cannot exclude that an elementary sub-interval will include more (and even much more) than one record. This is especially actual when the quantity of the combinations is less than that of the records. So the problem is to continue the search and recognition among this set.

Let us consider how are the records distributed among the values of a parameter. One can choose, for example, the grades of a sub-interval, the coordinates of vertices of CboGs, the mutual ratio of the sides of CboGs, etc.

For further pre-indexing of such elementary sub-intervals, one can use a type of pre-indexing that can be named as a pre-assessment. One can name pre-indexing of sub-interval images as a pre-assessment by a parameter (or a combination of some parameters) when this parameter (or this combination) of the pre-assessment corresponds to a certain amount of the records and this amount is stored as the quantity in corresponding memory array of this pre-assessment.

In other words, the pre-assessment of a parameter for a certain value of this parameter shows how many records possess this value (of this parameter).

That is, the pre-sorting determines both the sequence numbers of the records and their amount but the pre-assessment determines only the amount of the records. This allows to diminish the volume of memory by the addresses of the SNs of the records. And, naturally and mainly, this excludes the blocks of the memory to store these sequence numbers.

In many cases one can suppose the following. The less records correspond to a value of a parameter, the less records should be checked by more than this parameter. The less records should be checked, the faster is the search and recognition. So, the aim of the pre-assessment is to facilitate the continuing of pre-search and pre-recognition after the pre-sorting.

### 5.4.3. Pre-limitation

There can be reasons to limit the maximal quantity of the records in elementary sub-intervals. In this case an additional pre-sorting can be relevant in the elementary S-Is those exceed the limit quantity.

This additional pre-sorting can be performed both by an enhancement of the initial pre-sorting and by another types of pre-sorting.

## 5.5. Indexing. From the center to periphery. From uniform to non-uniform. Examples

Two assumptions are made here:

1) Usually the most valuable and interesting features are placed in the middles of photos and paintings. The first steps of the indexing are the most valuable ones.

Hence the first S-Is in the indexing should be probably determined not in the corners but in the center of the image.

For example, concentric clockwise circles can be determined from the center of the image to its border. The beginnings of the circles can be determined at the first top-right elementary sub-interval that corresponds to the first hour on the clock face or (preferably, due to simplicity) from the top-right diagonal.

2) The uniform image is chosen as the initial image for the indexing.

That is the grades of S-Is (determined analogously to the above cases) are considered from their uniform to peripheral values.

### 5.5.2. An example of 8×8. Pre-sorting from the top right corner and mean grades

Let us consider a hypothetical example of approximation S-IIs of 8×8 and four grades, i.e. 0, 1, 2, 3 (white, light, dark, black).

#### Enumeration of S-Is from the left bottom corner

The first elementary sub-interval of the first circle is 5:5 (grade 2) followed by 5:4, 4:4, 4:5 (all sub-intervals have grade 2).

The second circle begins from 6:6, followed by 6:5, 6:4, 6:3, 5:3, 4:3, 3:3, 3:4, 3:5, 3:6, 4:6, 5:6 (all grades are 2).

The third circle begins from 7:7 followed by 7:6, 7:5, 7:4, 7:3, 7:2, 6:2, 5:2, 4:2, 3:2, 2:2, 2:3, 2:4, 2:5, 2:6, 2:7, 3:7, 4:7, 5:7, 6:7 (all grades are one unit less than the preceding middle ones, that is 1).

The last, fourth, circle begins from 8:8 followed by 8:7, 8:6, 8:5, 8:4, 8:3, 8:2, 8:1, 7:1, 6:1, 5:1, 4:1, 3:1, 2:1, 1:1, 1:2, 1:3, 1:4, 1:5, 1:6, 1:7, 1:8, 2:8, 3:8, 4:8, 5:8, 6:8, 7:8 (all grade 1).

Then the grades of the last sub-interval are modified as 7:8 (grade 2), 7:8 (grade 0), 7:8 (grade 3).

Then the grades of the pre-last and last sub-intervals are modified as 6:8 (grade 2), 7:8 (grade 1), 7:8 (grade 2), 7:8 (grade 0), 7:8 (grade 3). Then 6:8 (grade 0), 7:8 (grade 1), 7:8 (grade 2), 7:8 (grade 0), 7:8 (grade 3).

And so on till changing the grades of the first elementary sub-interval 5:5.

#### Enumeration of S-Is from the center of the entire interval

If the sub-intervals are enumerated from the center of the total interval then we have a bit different numbers of the elementary sub-intervals (the grades are omitted for the sake of brevity):

The first elementary sub-interval is 1:1 followed by 1:-1, -1:-1, -1:1.

The second circle begins from 2:2 followed by 2:1, 2:-1, 2:-2, 1:-2, -1:-2, -2:-2, -2:-1, -2:1, -2:2, -1:2, 1:2.

The third circle begins from 3:3 followed by 3:2, 3:1, 3:-1, 3:-2, 3:-3, 2:-3, 1:-3, -1:-3, -2:-3, -3:-3, -3:-2, -3:-1, -3:1, -3:2, -3:3, -2:3, -1:3, 1:3, 2:3.

The last, fourth, circle begins from 4:4 followed by 4:3, 4:2, 4:1, 4:-1, 4:-2, 4:-3, 4:-4, 3:-4, 2:-4, 1:-4, -1:-4, -2:-4, -3:-4, -4:-4, -4:-3, -4:-2, -4:-1, -4:1, -4:2, -4:3, -4:4, -3:4, -2:4, -1:4, 1:4, 2:4, 3:4. And so on as above.

### 5.5.3. An example.
### Pre-sorting by the standard four concentric cuboids of gravity

Consider concentric cuboids of gravity, for example 1/8, 1/4, 1/2, 3/4 (beginning from the middle of the total interval) with the S-I grid 256×256. The uniform image is chosen as the initial image for the indexing. The last (from the center to periphery) elementary sub-interval is used to mark the corners of the cuboids of gravity.

Enumeration of S-Is is from the left bottom corner.

The first CboG borders are determined as 128-16=112 and 128+16=144. That is the coordinates of the first CboG are 112:112 and 144:144.

The coordinates of the second CboG are 96:96 and 160:160.

The coordinates of the third CboG are 64:64 and 192:192.

The coordinates of the fourth CboG are 32:32 and 224:224.

The first elementary sub-interval is 112:112 followed by 144:144, 112:112, 144:144, 96:96, 160:160, 64:64, 192:192, 32:32, 224:224.

Then the coordinates of the last sub-interval are modified as, for example, 224:225, 225:225, 225:224, 225:223, 224:223, 223:223, 223:224, 223:225, and so on.

Then the coordinates of the second-to-last are modified once as, for example, 32:31. Then the coordinates of the last sub-interval are modified in its total spectrum as, for example, 224:224, 224:225, 225:225, 225:224, 225:223, 224:223, 223:223, 223:224, 223:225, and so on.

Then the coordinates of the second-to-last are modified once more as, for example, 31:31. Then the coordinates of the last sub-interval are modified in its total spectrum as, for example, 224:224, 224:225, 225:225, 225:224, 225:223, 224:223, 223:223, 223:224, 223:225, and so on.

And so on. And so on.

## 6. Sub-interval copies of databases (DB S-ICs, DBS-ICs)

Sub-interval copies of databases (DB S-ICs, DBS-ICs) can include some parts of the above secondary collections.

Evidently such sub-interval copies of databases can be much smaller than the initial databases and the search and analysis in these DBS-ICs can be simplified due to the constant sizes of the DBS-ICs, i.e., due to the absence of additional addressing (as in the case of the varied sizes of the original records).

Each SII-copy of a record can include, for example:

- the numbers of the elementary elements in the record and the maximum of their grades;
- the hypotenuse of the maximal relative deviation of the CroG from the center of the total two-dimensional interval;
- the ratio of the (larger to smaller) lengths of the standard CboG of the record;
- the CroG or some minimal (e.g., 1/8) CboG (one or two coordinates);
- the standard CboG of the record (two coordinates) that can be represented as a small picture;
- the standard concentric CboG of the record (eight coordinates) that can be represented as a small picture;
- 8×8 four grades elementary sub-interval approximation of the record (8×2=16 bytes) that can be represented as a small picture;
- 64×64 four grades elementary sub-interval approximation of the record (64×16=1024 bytes) that can be represented as a small picture.

So DB S-ICs can be used for preliminary analysis of databases.

## 7. Sub-intervals and Big Data. A short outlook

### 7.1. General approach

Suppose there is a quantity (or a continuum) of some elements. Suppose these elements are located within an interval. Let us divide the interval by (in the general case) another quantity of sub-intervals.

Generally, the essence of the sub-interval analysis is:

**"An analysis of the set of the sub-intervals instead of an analysis of the set of the elements."**

### 7.2. General advantages of sub-intervals

Generally, there are two basic advantages of sub-intervals:

1) Proper quantities
2) Proper structures

The term "proper quantities" means that one can choose or determine the quantities of the sub-intervals in the accordance with the goals of the analysis. This can be made by divisions and mergers of the sub-intervals.

The term "proper structures" means in particular the following.

a) The lengths (and sometimes the weights) of the sub-intervals can be equal to each other, i.e. uniform.

b) The lengths (and sometimes the weights) of the sub-intervals can represent main features of the interval.

c) The sub-intervals can be organized as a regular hierarchical structure.

### 7.3. Basic advantage of sub-intervals for Big Data

The evident feature of Big Data is the vast quantity of the data. Such a vast quantity makes it difficult to analyze these data.

So from the perspective of the Big Data, the main advantage of sub-intervals is the possibility to radically diminish the quantity of these S-Is down to quantities that can be analyzed easily and conveniently. This diminishing can be evidently reached by, e.g., simple merging of some neighbor sub-intervals.

## 8. Big Data. Two concepts of S-II indexing

### 8.1. Comparison of the quantities

One of the main goals of the sub-interval analysis and sub-interval images is to assist in pre-search and pre-recognition in databases. Its evident sub-goal is to diminish the density of records (i.e. the quantity of the records) per a combination of the parameters of sub-intervals.

In this light the more combinations the less the density. And the above examples demonstrate that sub-interval images provide the advantage to enhance the quantity of such combinations to the quantities that are much more than the usual volume of Big Data databases.

Nevertheless the considered question is a dialectical one: if the capacity of the database is constant then the more the quantity of the combinations the more empty combinations (that is the combinations that correspond to no records). However one should store these empty combinations, sometimes count them, etc. This is another side and disadvantage of increasing the quantity of combinations of the parameters of sub-intervals.

So the question is to compare the quantities of the records per an elementary S-II and the quantities of the empty elementary S-IIs, and to balance the above advantage and disadvantage.

### 8.2. Concept of under-sorting and uninterrupted homogeneous indexing

The essence of the concept of under-sorting and uninterrupted homogeneous indexing is to provide a uniform array or arrays of combinations of the parameters of the sub-intervals.

The goal and advantage of the concept is to omit the numbers of combinations of the parameters of sub-intervals. This is possible due to the well-ordered uniform structure where the sequence number of a combination can be unambiguously determined by its position in the array and vice versa.

The disadvantage of this concept consists in empty combinations. That is it consists in the combinations that does not correspond to any record of the database. The more combinations the more empty combinations. Therefore the upper limit of this concept is the point or area where the disadvantage overweights the advantage. To successfully use this concept, the database should be in a sense under-sorted.

### 8.3. Concept of over-sorting and interrupted non-homogeneous selective indexing

The essence of the concept of over-sorting is to omit the empty sub-intervals.

The array includes the information only on non-empty S-IIs. Namely every line in the array includes the sequence number of the non-empty S-II, the quantity of the records that belong to this S-II, and the address of the start (and maybe the address of the end) of the section where the SNs of these records are stored.

If the quantity of the records that belong to this S-II exceeds a certain maximal value then, after these SNs, the address of the refinement to this S-II is written.

## 9. Big Data. Combination of the two concepts

Natural solutions for Big Data can be some combinations of the above two concepts: of under-sorting and over-sorting.

Suppose the above two assumptions are true. Namely

1) Usually the most valuable and interesting features are placed in the middles of photos and paintings.

2) The uniform image is chosen as the initial image for the indexing.

Then one can use the under-sorting in the middle circles and middle (1 and 2) grades. The over-sorting can be used for the periphery grades of the middle circles and for periphery circles.

## 10. Conclusions

The systematic introduction to the sub-interval analysis is developed (see Harin 2020a, 2020b) in the present draft version of the article. Namely the sub-interval images (S-IIs) are considered here.

Basic concepts, terms, and types of S-IIs are outlined. The consideration is concentrated on the computer-corresponded S-IIs and Big Data.

The features that are specific for S-IIs are considered, in particular, cuboids of gravity, sub-interval copies of databases, pre-sorting, pre-assessment, and pre-equalization.

The mutually reinforcing concepts of under-sorting and over-sorting are proposed.

Some hypothetical examples are reviewed.

The sub-interval images can be used among others tools in dealing with databases, especially with Big Data, including economic information.

## References


Harin, A. (2011a) About possible additions to interval arithmetic, *X International conference on Financial and Actuarial Mathematics and Eventoconvergence of Technologies*, Krasnoyarsk, (2011).

Harin, A. (2011b) Ruptures in the probability scale. Interval analysis. *International conference "Modern problems of applied mathematics and mechanics" devoted to 90th Anniversary from the birthday of academician N.N.Yanenko*, Novosibirsk, (2011).

Harin, A. (2011c) Interval analysis of distributions. Interval images of text, music, representations and video information, *54-th Scientific conference of MIPT "Modern problems of fundamental and applied sciences"*, Moscow, (2011).

Harin, A. (2012a) Subinterval analysis. First results. *15th International Symposium on Scientific Computing, Computer Arithmetic and Verified Numerical Computations*, Novosibirsk, (2012).

Harin, A. (2012b) Sub-interval analysis and possibilities of its use, *55-th Scientific conference of MIPT "Modern problems of fundamental and applied sciences"*, Moscow, (2012).

Harin, A. (2012c) Interval pictures and images. Their use for preliminary analysis and recognition, *XIX International conference "Mathematics. Computer. Education"*, Dubna, (2012).

Harin, A. (2012d) About global optimization in subinterval analysis at analog of Lipschitz's condition, *XX International conference "Mathematics. Economics. Education"*, Rostov-na-Donu, (2012).

Harin, A. (2020a) Introduction to sub-interval analysis. Estimations for the centers of gravity, MPRA paper 100496.

Harin, A. (2020b) Introduction to sub-interval analysis. Sub-interval arithmetic, *Applied Mathematical Sciences*, Vol. 14, no. 12, 607-620.