DOI 10.1007/s11018-023-02180-2
*Measurement Techniques, Vol. 65, No. 12, March, 2023*

# GENERALIZATION OF JACCARD INDEX FOR INTERVAL DATA ANALYSIS

**A. N. Bazhenov[1,2] and A. Yu. Telnova[1]** UDC 519.612

*Data samples with interval uncertainty are analyzed. It is proposed to use the Jaccard measure (index), which is widely used when comparing sets in various problem areas, as a measure (functional) of the consistency of interval values and their samples. Information about interval analysis, classical and complete (Kaucher) interval arithmetic is presented. For interval quantities, the necessary concepts and definitions of operations are introduced, in particular, generalizations of the concepts of intersection and union of sets. The Jaccard measure is generalized to the case of data with interval uncertainty and samples of interval data. Various variants of interval relations are described in detail — from their coincidence to incompatible cases. Various definitions of the Jaccard measure are given, both symmetric and nonsymmetric with respect to the operands. The connections of the proposed measure with the interval mode and the results of calculations with tweens are considered. A practical example of finding the information set of an interval problem using a new measure is given. Two areas of application of both symmetric and asymmetric measures are presented — computational processes (for characterizing iterative computational processes) and data analysis (for characterizing measurement workspaces and classifying data by a set of features).*
***Keywords:*** *Jaccard coefficient, t-norm, consistency measure, interval analysis, covering and noncovering measurements and samples, information set, interval mode, interval tweens.*

**Introduction.** One of the tasks of data analysis is to find out the degree of similarity of the sample elements among themselves or the degree of similarity of different samples. To characterize the degree of similarity, it is necessary to introduce a quantitative measure (similarity coefficient). Let us define the measure of similarity formally. The measure of similarity is a two-place function $S(A, B) \to [0, 1]$ — a real function on the set of pairs of objects $(A, B)$, where $A$, $B$ are arbitrary finite sets. Formally, membership of an operation to similarity measures is determined by a system with the following properties:

- boundedness $0 \le S(A, B) \le 1$;
- symmetry $S(A, B) = S(A, B) \ge 0$;
- reflexivity $S(A, B) = 1 \Leftrightarrow A = B$;
- transitivity $A \subseteq B \subseteq C \Rightarrow S(A, B) \ge S(A, C)$.

There are other systems of similarity axioms. The listed properties, also called the *t*-norm, can be implemented in a variety of ways. Measures of inclusion and similarity have been widely used in practice since the second half of the last century. In the publication [1], various similarity measures and their mutual relationships are considered in detail. The first such measure was proposed by Jaccard in 1901 in relation to biology. The Jaccard coefficient $J$ (index) for sets $A$, $B$ has the form

$$J\left(A, B\right) = \frac{n\left(A \cap B\right)}{n\left(A \cup B\right)}, \tag{1}$$

where $n$ is the cardinality (number of representatives) of the sets.

[1]Loffe Institute, St. Peterburg, Russia; e-mail: bazhenov_an@spbstu.ru; anna.telnova@mail.ioffe.ru.
[2]Peter the Great St. Petersburg Polytechnic University, Institute of Applied Mathematics and Mechanics, St. Petersburg, Russia.

Translated from Izmeritel'naya Tekhnika, No. 12, pp. 15–22, December, 2022. Original article submitted June 2, 2022. Accepted October 20, 2022.

 

In computer applications (image processing, machine learning), the measure of similarity of sets is often referred to as IoU (intersection over union).

The purpose of this work is to give a mathematical definition of the proposed option for calculating the binary measure of similarity of sets, to consider various options for its calculation and to study its properties. The measure is generalized to work with data samples with interval uncertainty. The article gives a number of examples and presents various directions for further research.

**Interval arithmetic.** In this publication, similarity measures are applicable to mathematical objects with interval uncertainty. The modern notation system for interval objects used in this article is given in [2]. Here and throughout the article, interval objects are indicated in bold italic type, as is customary in interval arithmetic.

Classical interval arithmetic $\mathbb{IR}$ is an algebraic system formed by intervals $x = [\underline{x}, \overline{x}] \subset \mathbb{R}$ so that for any arithmetic operation "$*$" from the set $\{+, -, \cdot, /\}$ the result of the operation between intervals is defined as $\boldsymbol{x}*\boldsymbol{y} = \{x*y | x \in \boldsymbol{x}, y \in \boldsymbol{y}\}$, where $x$, $y$ are real numbers; $\boldsymbol{x}$, $\boldsymbol{y}$ are intervals. The symbol $\mathbb{I}$ in the notation $\mathbb{IR}$ is used to distinguish interval and real arithmetic, with $\mathbb{IR} \supseteq \mathbb{R}$.

We introduce the characteristics of the interval $\boldsymbol{a}$:

$$\text{middle mid } \boldsymbol{a} = (\overline{\boldsymbol{a}} + \underline{\boldsymbol{a}})/2;$$

$$\text{radius rad } \boldsymbol{a} = (\overline{\boldsymbol{a}} - \underline{\boldsymbol{a}})/2;$$

$$\text{width wid } \boldsymbol{a} = \overline{\boldsymbol{a}} - \underline{\boldsymbol{a}}, \tag{2}$$

where $\overline{\boldsymbol{a}}$, $\underline{\boldsymbol{a}}$ are the right and left (or upper and lower) boundaries of the interval $\boldsymbol{a}$.

In a more general setting, the problem of interval analysis can be solved in the so-called complete interval arithmetic or Kaucher arithmetic $\mathbb{KR}$ [3]. This arithmetic is obtained by adding improper intervals $[\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}]$, $\underline{\boldsymbol{x}} > \overline{\boldsymbol{x}}$ to the set $\mathbb{IR}$. The introduction of Kaucher's interval arithmetic is motivated by a number of limitations of classical interval arithmetic: the absence of inverse elements relative to addition and multiplication operations; the impossibility in the general case of taking the minimum by inclusion; the complexity with the formulation of minimax problems. The use of improper intervals allows you to build original methods for solving problems with interval objects.

**Generalization of the Jaccard coefficient.** To analyze data, it is necessary to compare interval objects in a universal way, regardless of the property of the covering. Let us introduce a numerical characteristic of the degree of consistency of two intervals $\boldsymbol{x}, \boldsymbol{y} \subset \mathbb{IR}$ in the form $JK(\boldsymbol{x}, \boldsymbol{y})$ ($K$ indicates the relation to Kaucher's interval arithmetic):

$$JK\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\text{wid}\left(\boldsymbol{x} \wedge \boldsymbol{y}\right)}{\text{wid}\left(\boldsymbol{x} \vee \boldsymbol{y}\right)}, \tag{3}$$

where the measure for the Jaccard coefficient, in contrast to (1), is the width of the interval (2); and instead of the operations of intersection and union of sets, we have the operations of taking the minimum $\wedge$ and maximum $\vee$, respectively, by including two quantities in the complete interval arithmetic (Kaucher), described by

$$\boldsymbol{a} \wedge \boldsymbol{b} = \left[\max\left\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\right\}, \min\left\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\right\}\right];$$

$$\boldsymbol{a} \vee \boldsymbol{b} = \left[\min\left\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\right\}, \max\left\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\right\}\right].$$

In the general case, the inclusion minimum in expression (3) may be an improper interval of negative width. To further generalize measure (3) to sets of intervals, we write its numerical expression

$$JK\left(\boldsymbol{x}, \boldsymbol{y}\right) = \frac{\min\left\{\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}\right\} - \max\left\{\underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}\right\}}{\max\left\{\overline{\boldsymbol{x}}, \overline{\boldsymbol{y}}\right\} - \min\left\{\underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}\right\}}. \tag{4}$$

In writing formula (4), instead of interval widths, we use explicit expressions for taking the minimum and maximum by inclusion, which provide the universal character of (4), regardless of whether the result of the operation of taking the minimum by inclusion $\wedge$ is a proper or improper interval.

Let us illustrate various measures of binary consistency. Consider different cases of the relative position of intervals $\boldsymbol{x}$, $\boldsymbol{y}$ and various measures of consistency. Suppose

$$\min\left\{\underline{x},\ \underline{y}\right\} = 1; \quad \max\left\{\overline{x},\ \overline{y}\right\} = 6\ .$$

In Fig. 1 we show the possible options for the relative position of the intervals $\boldsymbol{x}$, $\boldsymbol{y}$ and the corresponding values of the measure $JK$ (cf. formula (4)). Summing up the indicated options, we obtain the inequality

$$-1 \le JK(\boldsymbol{x}, \boldsymbol{y}) \le 1\ .$$

For non-empty intersections of intervals, the classical Jaccard measure (1) (not shown in Fig. 1) and the new measure $JK(\boldsymbol{x}, \boldsymbol{y})$ (3), (4) give the same results. For empty intersections of intervals, the classical Jaccard measure is equal to zero, and $JK(\boldsymbol{x}, \boldsymbol{y})$ numerically describes the measure of inconsistency from 0 to –1.

**Asymmetric variants of the measure $JK(\boldsymbol{x}, \boldsymbol{y})$.** Measure (3) is symmetric with respect to its arguments, which may be inconvenient in case of a large difference in the widths of the arguments. It is desirable that, in this case, the sensitivity of the measure also differ significantly when it is taken according to different arguments. In this case, asymmetric options are useful:

$$JK_x\left(\boldsymbol{x},\ \boldsymbol{y}\right) = \frac{\operatorname{wid}\left(\boldsymbol{x} \wedge \boldsymbol{y}\right)}{\operatorname{wid}\boldsymbol{x}}, \tag{5}$$

$$JK_y\left(\boldsymbol{x},\ \boldsymbol{y}\right) = \frac{\operatorname{wid}\left(\boldsymbol{x} \wedge \boldsymbol{y}\right)}{\operatorname{wid}\boldsymbol{y}}\ . \tag{6}$$

Similar expressions in the literature are called similarity measures (similarity, overlapping ratio [4, 5]). Expressions (5), (6) differ from similar ones given in [4, 5] in that they allow the interval to be improper when taking the minimum by inclusion.

*Example*. Let $\boldsymbol{x} = [1; 2]$, $\boldsymbol{y} = [3; 7]$. Calculate the measures (5), (6):

$$JK_{\boldsymbol{x}}(\boldsymbol{x},\ \boldsymbol{y}) = \frac{2-3}{1} = -1\,, \tag{7}$$

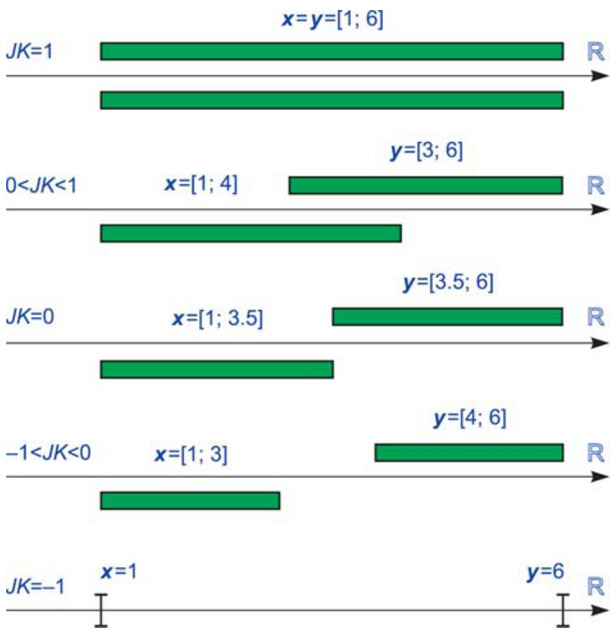


Fig. 1. Variants of relations between the positions of two intervals and the value of the Jaccard measure $JK$.

$$JK_y(\boldsymbol{x},\ \boldsymbol{y}) = \frac{2-3}{4} = \frac{-1}{4}. \tag{8}$$

The absolute values of expressions (7), (8) reflect the measure of the "dissimilarity" of the $\boldsymbol{x}$, $\boldsymbol{y}$ intervals relative to their widths. In this case, the value of expression (8) is significantly less. This fact corresponds to the difference in the widths of the intervals $\boldsymbol{x}$, $\boldsymbol{y}$ and makes it possible to build meaningful constructions on this basis, for example, in regularization procedures.

**Jaccard measure for sampling interval data.** The consistency measure introduced for two intervals in the form (3) admits a natural generalization to the case of interval sampling. Let there be an interval sample $\boldsymbol{X} = \{\boldsymbol{x}_i\}$, $i = 1, 2, ..., n$. Let us define the measure $JK(\boldsymbol{X})$ as

$$JK\ (\boldsymbol{X}) = \frac{\text{wid}\ (\Lambda_i \boldsymbol{x}_i)}{\text{wid}\ (\mathrm{V}_i \boldsymbol{x}_i)}, \tag{9}$$

where

$$\bigvee_{1\le i\le n} \boldsymbol{x}_i = \left[\min_{1\le i\le n} \underline{\boldsymbol{x}}_i,\ \max_{1\le i\le n} \overline{\boldsymbol{x}}_i\right]; \quad \bigwedge_{1\le i\le n} \boldsymbol{x}_i = \left[\max_{1\le i\le n} \underline{\boldsymbol{x}}_i,\ \min_{1\le i\le n} \overline{\boldsymbol{x}}_i\right].$$

It is important that in the case of an interval sample of two elements, expression (9) becomes expression (3).

Consistency measures for interval values are considered in publications [4–6], but only for consistent samples.

*Example*. Let us calculate the Jaccard coefficient (9) for consistent $\boldsymbol{X}_1 = \{[0; 8], [2; 9], [3; 10]\}$ and inconsistent $\boldsymbol{X}_2 = \{[9; 10], [10; 11], [11; 12]\}$ samples:

$$JK\ (\boldsymbol{X}_1) = \frac{8-3}{10-0} = \frac{5}{10} = 0.5,$$

$$JK\ (\boldsymbol{X}_2) = \frac{10-11}{12-9} = \frac{-1}{3} = -0.33(3).$$

**Jaccard measure and interval mode.** Let us consider the connection of the *JK* measure with other characteristics of interval samples, for example, with the interval mode of the sample. In the book [7], following the ideas of publication [8], this object (interval) is introduced as follows: "The interval sample mode $\boldsymbol{X}$ is the set of intersection intervals of the largest consistent subsamples of the considered sample." The greatest length of consistent subsamples of a given sample is called the mode frequency.

Let us consider the simplest example illustrating the relationship between the Jaccard measure and the interval mode. In Fig. 2 we show options for calculating the interval mode and the mode consistency measure for a sample of two intervals $\boldsymbol{X} = \{\boldsymbol{x}, \boldsymbol{y}\}$, from which follows the relationship between the interval mode and the value of the Jaccard measure. In this case, the interval mode is a complex object that provides additional information about the sample. The interval mode is generally a multi-interval [7]. The nature of the Jaccard measure *JK* is different and has a numerical character. The numerator of expression (5) contains the inclusion minimum, and its position in multimode distributions may not belong to any sampling interval. In this case, additional expressions are required for the numerical description of the interval mode.

Preliminarily, it can be noted that the value of $JK(\text{mode}\ \boldsymbol{X})$ in the case of a multi-interval mode is nonpositive. In this case, the absolute value of $JK(\text{mode}\ \boldsymbol{X})$ depends both on the distance between the multisampling intervals and on the width of their inclusion maximum.

*Example*. For a noncovering sample $\boldsymbol{X} = \{[1; 4], [5; 9], [1.5; 4.5], [6; 9]\}$, the interval mode is the union of two intervals (multi-interval) mode $\boldsymbol{X} = [1.5; 4] \cup [6; 9]$ and the measure of consistency is defined as

$$JK\ (\text{mode}\ \boldsymbol{X}) = \frac{4-6}{9-1} = -0.25.$$

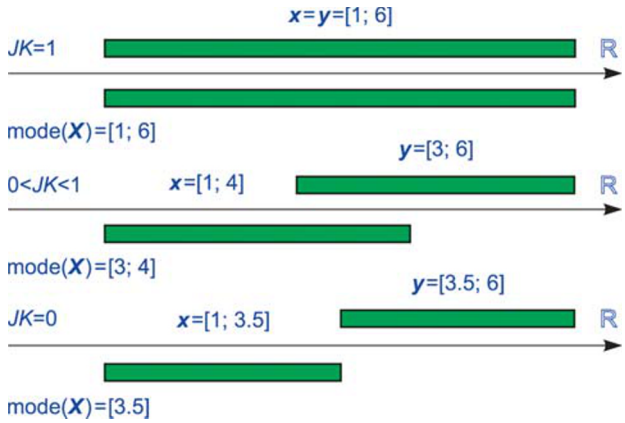


Fig. 2. Variants of relations between the position of intervals for the covering sample, interval mode and *JK* values.

In this case, the inclusion minimum $\bigwedge\limits_{1\le i\le n} x_i = [6;4]$ for the sample elements $\boldsymbol{X}$ in the numerator of expression (4) is an improper interval, which in the given case is also the inclusion minimum for the interval mode multi-interval elements.

**Application of *JK* to tween arithmetic and presentation of measurement data with errors of several kinds.** In practice, the ends of the intervals representing the results of measurements may not be known exactly, so that it becomes necessary to work with intervals that have interval ends. Such objects are known in interval analysis and are called tweens (English tween is an abbreviation for twice interval, double interval).

Tween, as an "interval of intervals" or an interval with interval ends, is represented as

$$\boldsymbol{X} \quad [\boldsymbol{a}, \boldsymbol{b}] = \left[\left[\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}\right], \left[\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}\right]\right]. \tag{10}$$

An alternative consideration of tweens as objects that give internal and external estimates of interval computations is proposed in [9]:

$$\boldsymbol{X} = \left[\boldsymbol{X}^{\text{in}}, \boldsymbol{X}^{\text{out}}\right] = \left[\left[\underline{\boldsymbol{X}}^{\text{in}}, \overline{\boldsymbol{X}}^{\text{in}}\right], \left[\underline{\boldsymbol{X}}^{\text{out}}, \overline{\boldsymbol{X}}^{\text{out}}\right]\right], \tag{11}$$

where $\boldsymbol{X}^{\text{in}}$ and $\boldsymbol{X}^{\text{out}}$ are intervals of internal and external estimates of interval calculations.

In definitions (10), (11), a tween is defined by four values. The functional *JK* also depends on four values — the ends of two intervals $\boldsymbol{x}$, $\boldsymbol{y}$. If we take the internal and external components of the tween as arguments of *JK*, we obtain a relative value characterizing the relation of the internal and external estimates of the tween $\boldsymbol{X}$ in the form (11):

$$\delta\boldsymbol{X} = JK\left(\boldsymbol{X}^{\text{in}}, \boldsymbol{X}^{\text{out}}\right). \tag{12}$$

*Example*. Given a temperature tween in the form (10):

$$\boldsymbol{T} = \left[[19.15;19.85], [20.15;20.85]\right]\ ^{\text{o}}\text{C}.$$

We write it in the Nesterov form:

$$\boldsymbol{T} = \left[[19.85;20.15], [19.15;20.85]\right]\ ^{\text{o}}\text{C}.$$

In accordance with construction (12), we have the "relative tween width"

$$\delta\boldsymbol{T} = JK\left(\boldsymbol{T}^{\text{in}}, \boldsymbol{T}^{\text{out}}\right) = 0.18.$$

The positive value of $\delta \boldsymbol{T}$ reflects the consistency of the internal and external estimtes of the interval object, and the absolute value is the "reliability margin" of the estimate. A negative value of $\delta \boldsymbol{T}$ corresponds to the nonempty intersection of the intervals of the lower and upper estimates of the interval $\boldsymbol{T}$.

**Analysis of data with interval uncertainty.** Analysis of interval data emerged in the last decades of the 20th century as an alternative to traditional "probabilistic statistics" based on the methods of probability theory. Many works are devoted to various aspects of the analysis of interval data, in particular [10–19]. Analysis of data with interval uncertainty (statistics of interval data, analysis of interval data) is considered in the book [8], where the system of concepts and terms related to the processing of the specified data is presented. One of the most important concepts of interval statistics is a covering measurement (measurement) — an interval measurement result that is guaranteed to contain the true value of the measured value [7]. A measurement for which it is impossible to assert that it contains the true value of the measured quantity will be called noncovering. A covering measurement is a guaranteed two-sided "fork" of the values of the measured quantity, while for a noncovering measurement nothing of the kind can be asserted. This gives the covering measurements a fundamentally higher status and makes it possible to build more meaningful constructions on their basis.

Further, we will call a covering sample a set of measurements, the dominant part (most, etc.) of the measurements (observations) of which are covering. On the contrary, a sample is called noncovering if the predominant part of the measurements included in it are noncovering. Possible alternative terms are "inclusive dimension," "encompassing dimension" (their negation is "noninclusive", "nonencompassing"). Suggested English equivalents are enclosing measurement, covering measurement. Such a definition is not strict, which will avoid overly strict and rarely achievable sampling requirements.

When formulating new definitions, we follow the principle of correspondence in the methodology of science — any new scientific theory should include the old theory and its results as a special limiting case. As applied to the measure of consistency, further more general constructions contain (as special cases) the previously defined constructions. Symbolically, this can be expressed as

$$\mathbb{KR} \supseteq \mathbb{IR} \supseteq \mathbb{R} .$$

It is also important that the measures introduced for more general objects, namely, on samples of interval values and multidimensional interval objects, naturally reduce to constructions for simpler objects.

Thus, the measure $JK(\boldsymbol{X})$ in the form (3) can be defined both for noncovering samples in Kaucher arithmetic and for covering samples in classical interval arithmetic, as well as for real "point" samples.

*Example*. We will demonstrate the use of the Jaccard index for data processing. The subject area refers to the physics of semiconductors — studies of the photoelectric characteristics of the sensor under test, carried out by specialists from the A. F. Ioffe Laboratory of Photoelectric Converters of the Physico-Technical Institute. There are two data samples with interval uncertainty.[1] Samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, respectively, refer to sensors (photodetector, PD) PD1, PD2 (standard). The number of readings in the samples is 200. The PD1 sensor is calibrated according to the PD2 standard. The dependence between quantum efficiencies of the sensors $QE_{\mathrm{PD1}}$, $QE_{\mathrm{PD2}}$ is assumed constant for each pair of measurement sets:

$$QE_{\mathrm{PD2}} = \frac{I_{\mathrm{PD2}}}{I_{\mathrm{PD1}}} QE_{\mathrm{PD1}} , \tag{13}$$

where $I_{\mathrm{PD1}}$, $I_{\mathrm{PD2}}$ are the measured currents of the detectors.

Let us present the data in such a way as to apply the concepts of data statistics with interval uncertainty. One of the common ways to obtain interval results in primary measurements is the "intervaluing" of point values, when an error interval $\varepsilon$ is added to the point (base) value that is read from the indicators of the measuring device:

$$\boldsymbol{x} = \dot{x} + \varepsilon .$$

We set the error interval as a balanced interval

$$\varepsilon = [-\varepsilon, \varepsilon] ,$$

[1] M. Z. Schwartz, working materials. URL: https://github.com/AlexanderBazhenov/Solar-Data (date of access: 11/18/2022).

for a specific procedure of measurements $\varepsilon = 10^{-4}$ mV.

Due to a significant drift of the measuring equipment, the data is pre-processed and the drift is subtracted. After calculations, the data have a nonequivariant form, shown in Fig. 3.

According to the terminology of interval analysis, the sample under consideration is a vector of intervals, or an interval vector $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$. In the case of estimating a single physical quantity from a sample of interval data, the information set will also be an information interval — an interval containing the values of the estimated quantity that are consistent with the measurements of the sample (consistent with the data of these measurements).

We calculate $JK(\boldsymbol{X})$ for a set composed of the union of the values of the original samples $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2\}$. Using programs for calculations in interval arithmetic (cf. https://github.com/szhilin/octave-interval-examples), we obtain $JK(\boldsymbol{X}) = -0.65$. The negative value of $JK(\boldsymbol{X})$ characterizes the inconsistency of the $\boldsymbol{X}$ sample.

We use the model (13) of the connection between samples for different measurements ($\boldsymbol{R}$ is an unknown factor):

$$\boldsymbol{X}_1 = \boldsymbol{R}\boldsymbol{X}_2 .$$

In accordance with the tasks of researching sensors, it is necessary to give point and interval estimates of $\boldsymbol{R}$. We optimize the parameter $\boldsymbol{R}$ as follows. A point estimate for the ratio of quantum efficiencies is

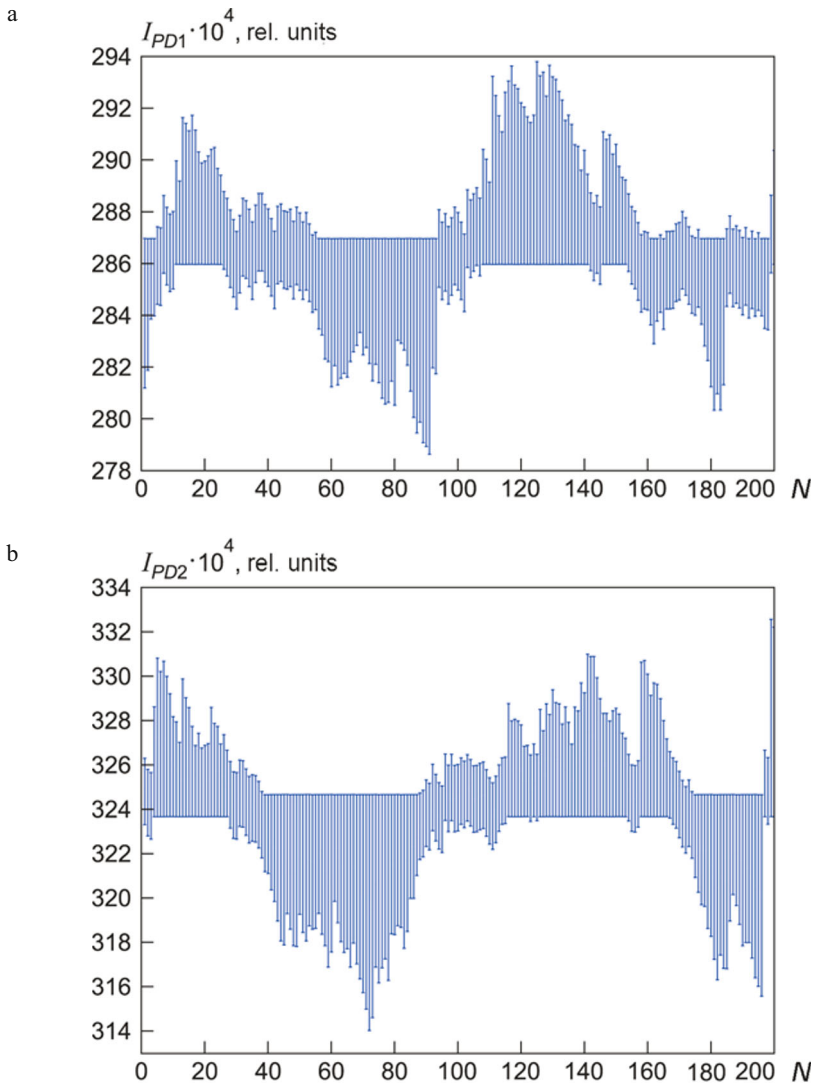


Fig. 3. Quantum efficiency measurement data samples (rel. units) for the studied (a) and reference (b) sensors.

$$\boldsymbol{R}_{\text{opt}} = \arg \max_{\boldsymbol{R}} JK\left(\boldsymbol{X}'\right), \tag{14}$$

where

$$\boldsymbol{X}' = \left\{\boldsymbol{R}\boldsymbol{X}_1, \boldsymbol{X}_2\right\}.$$

The interval estimate $\boldsymbol{R}$ corresponds to the information set of the problem

$$\boldsymbol{R}{:}JK\left(\boldsymbol{X}'\right) \geq 0. \tag{15}$$

Let us carry out calculations using formula (14) and plot $JK(\boldsymbol{R})$ (Fig. 4), where the area corresponding to condition (15) is highlighted. The optimal value in formulation (14) $\boldsymbol{R}_{\text{opt}} = 1.132$, the information set of values (15) $\boldsymbol{R} = [1.12; 1.14]$.

The histogram of values of the $\boldsymbol{X}'$ combined sample at $\boldsymbol{R}_{\text{opt}}$ is shown in Fig. 5. The frequency distribution of histograms is close to symmetrical and unimodal, which positively indicates the correctness of the optimization procedure.

**Applications of measure of consistency.** For computational processes, two directions of application of the proposed measure can be proposed.

The first direction is the characterization of computational processes. When iteratively solving interval systems of equations, it is necessary to monitor the intersection of sets of successive operations. For this, a convenient means is an asymmetric measure equal to the ratio of the intersection of the sets of successive operations to the width of the iterative operator (Newton, Kravchik, etc.). The second direction is the solution of interval systems of linear algebraic equations (ISLAE). Traditionally, the solvability of ISLAE (in other words, the nonemptiness of solution sets) is studied using the technique of recognizing functionals. Recognizing functionals are constructed in such a way that their sign indicates that the solution sets are nonempty. For the main types of problem setting for solving ISLAE — finding a combined, admissible, and controlled decision sets — it is possible to build recognizing functionals based on symmetric and non-symmetric consistency measures.

For data analysis, the proposed measure can be used to characterize the range of a variable. When measuring any value with a set of sensors, it is necessary to allocate operating domains of parameters or measurement time ranges. A measure of consistency is naturally suitable for this. To select a reliable measurement range, it is necessary to set its practically reasonable threshold. A similar situation arises with multiple measurements of any quantity. For example, in the case of forward and reverse motion in the system under study, a hysteresis type of data may occur, including inconsistent ones. For characterization (identification of hysteresis), a measure of consistency can be used.

In clustering (factorization) problems, the working tool is the classification of data according to a set of features. A popular example is histogramming. If the data is essentially interval, one measurement may fall into more than one interval (bin) of the set according to which the classification is carried out. In this case, an asymmetric measure of consistency is well suited: the choice is made according to its maximum value.

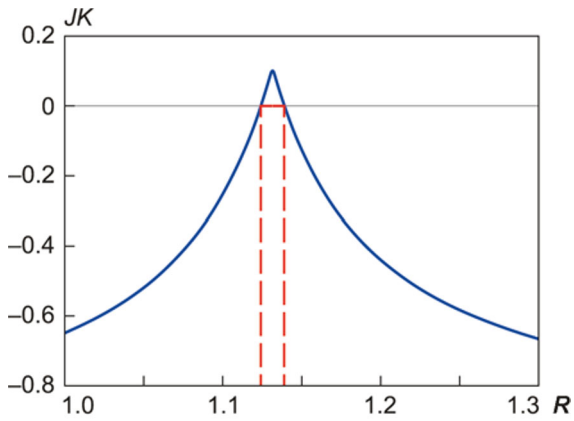


Fig. 4. Jaccard coefficient dependence on $\boldsymbol{R}$.

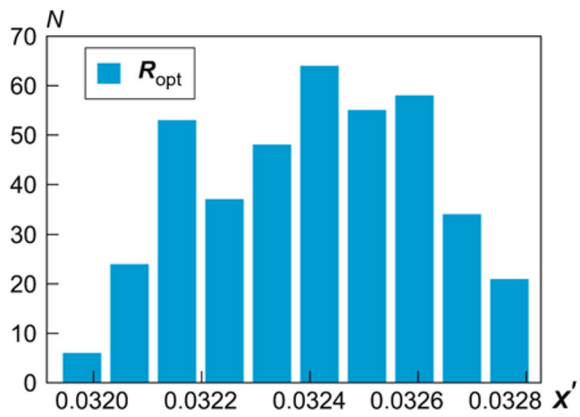


Fig. 5. Histogram of values of the $\boldsymbol{X}'$ combined sample at $\boldsymbol{R}_{\text{opt}}$.

**Conclusion.** The proposed measure made it possible to generalize the concept of similarity of sets to interval objects: pairs of intervals and data samples with interval uncertainty. It is essential that the measure provides meaningful information for non-joint data samples as well. The above example demonstrated the possibility of obtaining an interval estimate of an unknown model parameter in the problem of dependency recovery. Directions for further research on the application of the proposed measure to characterize computational processes and analyze data with interval uncertainty are outlined.

**Acknowledgment.** The authors are grateful to the participants of the All-Russian webinar on interval analysis S. I. Zhilin, S. I. Kumkov, A. V. Prolubnikov, E. V. Chausova, and S. P. Shary for creative and constructive cooperation in the field of data analysis with interval uncertainty and also to V. M. Nesterov for discussion of issues related to twin arithmetic.

Works presented in the sections "Jaccard measure and interval mode", "Application *JK* to twin arithmetic and presentation of measurement data with several types of error," "Data analysis with interval uncertainty", supported by the Russian Science Foundation, project No. 21-72-20007. The works presented in the sections "Interval arithmetic", "Generalization of the Jaccard coefficient" and "Asymmetric variants of the $JK(\boldsymbol{x}, \boldsymbol{y})$ measure" were supported by the Ioffe Institute RAS (within the framework of the RF state task 0034-2019-0001 and 0040-2019-0023).

## REFERENCES

1. B. I. Semkin, On the Relation Between Mean Values of Two Measures of Inclusion and Measures of Similarity, *Byull. Botanicheskogo sada-instituta DVO RAS*, No. 3, 91–101 (2009).
2. R. B. Kearfott, M. T. Nakao, A. Neumaier, S. M. Rump, S. P. Shary, and P. van Hentenryck, Standardized Notation in Interval Analysis, *Comput. Technol.*, **15**, No. 1, 7–13 (2010).
3. S. Shary, *Numerical Computation of Formal Solutions to Interval Linear Systems of Equations,* arXiv:1903.10272v1 [math.NA], https://doi.org/10.48550/arXiv.1903.10272.
4. S. Kabir, C. Wagner, T. C. Havens, D. T. Anderson, and U. Aickelin, *IEEE Int. Conf. on Fuzzy Systems* (*FUZZ-IEEE 2017*), 2017, https://doi.org/10.1109/FUZZ-IEEE.2017.8015623.
5. T. Wilkin and G. Beliakov, *IEEE Int. Conf. on Fuzzy Systems* (*FUZZ-IEEE 2019*), 1–6 (2019), https://doi.org/10.1109/FUZZ-IEEE.2019.8858850.
6. S. Kabir, C. Wagner, and Z. Ellerby, *Towards Handling Uncertainty-at-Source in AI — A Review and Next Steps for Interval Regression*, arXiv:2104.07245 [cs.LG], https://doi.org/10.48550/arXiv.2104.07245.
7. A. N. Bazhenov, S. I. Zhilin, S. I. Kumkov, and S. P. Sharyj, *Processing and Analysis of Data with Interval Uncertainty*, 2022, available at: http://www.nsc.ru/interval/Library/ApplBooks/InteData Processing.pdf (accessed: 10.11.2022).
8. C. Hu and Z. H. Hu, On Statistics, Probability, and Entropy of Interval-Valued Datasets, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science*, M. J. Lesot et al. Eds., Cham, Springer, **1239**, 2020, https://doi.org/10.1007/978-3-030-50153-2_31.
9. V. M. Nesterov, Tween Arithmetics and Their Application in Methods and Algorithms of Two-Sided Interval Estimation, *Diss. Doct. Phis. Math. Sci. St. Petersburg (St. Petersburg Institute of Informatics and Automation of the Russian Academy of Sciences),* 1999.
10. S. Shary, *Comput. Technol.*, **2**, No. 2, 150–172 (2017), http://dx.doi.org/10.14529/mmph170105.
11. S. Shary, *J. Comput. Syst. Sci. Int.*, **56**, No. 6, 897–913 (2017), https://doi.org/10.7868/S0002338817060014.
12. S. Shary, Identification of Outliers in the Maximum Mat-Ching Method in the Analysis of Interval Data, *Proc. All-Russian Conf. on Mathematics Int. Participation "MAC-2018,"* Barnaul, AltGU Publishing House, 2018, pp. 215–218.
13. S. Shary, On a Variability Measure for Estimates of Parameters in the Statistics of Interval Data, *Comput. Techol.*, **24**, No. 5, 90–108 (2019), https://doi.org/10.25743/ICT.2019.24.5.008.
14. S. P. Shary, Data Fitting Problem under Interval Uncertainty in Data, *Ind. Lab. Diagn. Mater.*, **86**, No. 1, 62–74, (2020), https://doi.org/10.26896/1028-6861-2020-86-1-62-74.
15. S. I. Zhilin, *Reliab. Comput.*, **11**, 433–442 (2005), https://doi.org/10.1007/s11155-005-0050-3.
16. S. I. Zhilin, *Chemometr. Intell. Lab. Syst.*, **88**, No. 1, 60–68 (2007), https://doi.org/10.1016/j.chemolab.2006.10.004.
17. S. I. Kumkov, Processing of Experimental Data on Ionic Conductivity of Molten Electrolyte by Methods of Interval Analysis, *Russian Metallurgy* (*METALLY*), No. 3, 79–89 (2010).
18. S. I. Kumkov and Yu. V. Mikushina, *Reliab. Comput.*, **19**, 197–214 (2013).
19. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty. Applications to Computer Science and Engineering*, Springer, Berlin-Heidelberg, 2012, https://doi.org/10.1007/978-3-642-24905-1.