ACTA IMEKO
ISSN: 2221-870X
April 2017, Volume 6, Number 1, 20-26

Numerical experimental investigation of comparison data evaluation method using preference aggregation

Sergey V. Muravyov, Irina A. Marinushkina, Diana D. Garif

National Research Tomsk Polytechnic University, Pr. Lenina 30, 634050 Tomsk, Russia

ABSTRACT
An integrated software for experimental testing preference aggregation method for interlaboratory comparison data processing is presented. The data can be obtained by a Monte-Carlo simulation and/or taken from real comparisons. Numerical experimental investigations with the software have shown that, as against traditional techniques of interlaboratory comparison data processing, the preference aggregation method provides a robust comparison reference value to be closer to a nominal value.

Section: RESEARCH PAPER

Keywords: interlaboratory comparisons; reference value; largest consistent subset; preference aggregation; robust method

Citation: Sergey Muravyov, Irina Marinushkina, Diana Garif, Numerical experimental investigation of comparison data evaluation method using preference aggregation, Acta IMEKO, vol. 6, no. 1, article 4, April 2017, identifier: IMEKO-ACTA-06 (2017)-01-04

Editor: Paolo Carbone, University of Perugia, Italy

Received July 4, 2016; **In final form** October 13, 2016; **Published** April 2017

Funding: This work was supported by the Ministry of Education and Science of Russian Federation

Corresponding author: Sergey Muravyov, e-mail: muravyov@tpu.ru

1. INTRODUCTION

Interlaboratory comparisons (IC) are now a quite common and important metrological procedure that is used under key comparisons [1], measurement laboratories proficiency testing [2], etc. The procedure consists in arrangement and implementation of assessment of measurement quality of a given object characteristic by means of several different laboratories in accordance with definite prescribed rules.

Main task of any kind of interlaboratory comparisons is establishing a *reference value of measured quantity* x_{ref} that characterizes a largest subset of consistent (reliable) measurement results, i.e. so called *largest consistent subset* (LCS) [3]. For this aim, participating in comparisons laboratories estimates the same *nominal value* x_{nom} of the measured quantity. Laboratories having unreliable measurement results do not participate in establishing the final reference value.

It should be noticed that, in contrast to proficiency testing, an official procedure of key comparisons (KCs) of the MRA [1] does not allow to discard any of the participant results, even though a result looks like unreliable or outlying. In this paper we will adhere to a hypothetical position that the two types of ICs can be treated to be similar actions tolerating exclusion of outliers, understanding the resulting reference value can be biased in the sense that some participants were excluded from its computation.

There are different approaches to check consistency of laboratory measurement results and to find the reference value x_{ref}, see, for example [3]-[7]. The choice of a particular consistency test method depends on a kind of travelling standard, measurement conditions and number of participating laboratories. Widely used methods are statistical ones characterizing IC participant competences to carry out measurements based on, for example, calculation of the difference of laboratory measurement results and assigned by comparison providers, percent differences, percentiles, or ranks [8]. However, these methods usually impose limitations on a feasible IC participating laboratories number. Moreover, statistical methods may evince low discriminating ability, that is the capacity to differ truly unreliable laboratories from laboratories providing results to be trusted.

In [4]-[5], a rather widely known so called Procedure A was presented. The procedure uses the weighted mean value y:

$$y = \sum_{i=1}^{m} x_i u^{-2}(x_i) / \sum_{i=1}^{m} u^{-2}(x_i), \tag{1}$$

where x_i is the nominal value estimate provided by the i-th laboratory; $u(x_i)$ are corresponding standard uncertainties; m is the number of IC participating laboratories. The standard uncertainty of value y has the view:

$$u^2(y) = 1/\sum_{i=1}^{m} u^{-2}(x_i). \quad (2)$$

In this procedure the weighted average value y is accepted as the reference value x_{ref} only if its consistency with IC participating laboratories data is confirmed in accordance to the criterion χ^2.

If the consistency test is not satisfied, it is proposed in [3] to use a strategy of successive exclusion of outliers, that is, results which are not consistent with the others by limits of claimed uncertainties. A result is deemed to be inconsistent if $|E_n| > 2$, where

$$E_n = \frac{x_i - y}{\sqrt{u^2(x_i) \pm u^2(y)}}, \qquad i = 1,...,m. \quad (3)$$

The process of exclusion of one inconsistent result is repeated until a consistency of results by the criterion χ^2 is achieved. For LCS obtained in this way, the reference value is determined by (1), where instead of m number of reliable laboratories m' is used.

Procedure A can be reasonably applied if measurement results provided by participating laboratories are characterized by a normal probability distribution. That is why there is a need to develop robust methods for interlaboratory comparison data processing that are well-behaved in cases where the law of laboratory measurement results distribution differs from normal or unknown.

For example, in paper [9] Nielsen proposed the method, the successful application of which has been described in [10]. The method offers to consider the uncertainty range $u(x_i)$ as the rectangular distribution and to deem that each participant gives one vote to each value within its uncertainty range and no votes for values outside this range. This produces a robust algorithm for the reference value x_{ref} determination that is insensitive to outliers, i.e. results with an uncertainty considerably lower than those of other participants.

This paper is devoted to software implementation of the comparison reference value determination method presented in terms of preference aggregation [11]-[13]. In Section 2 a way is considered to transform uncertainty intervals provided by participating laboratories into rankings of measured quantity values. Then the obtained rankings, constituting an initial preference profile, can serve as input data for determination of consensus ranking by the Kemeny rule that allows to find the reference value of the measurand and to assess an ability of participating laboratories to provide reliable measurement results. In Section 3 specially developed software is discussed to carry out numerical experimental researches of IC methods including Procedure A, the Nielsen algorithm and the proposed preference aggregation method. In Section 4 processing of real comparison data by the preferences aggregation method is presented.

2. IC DATA PROCESSING ON THE BASE OF PREFERENCE AGGREGATION

Define the procedure of transformation of uncertainty intervals provided by laboratories into rankings. For this aim, designate an uncertainty interval gained by the i-th laboratory through $u(x_i) = [u_l(x_i), u_u(x_i)]$.

Define A, a *range of actual values* (RAV), of the measurand for converting uncertainty intervals of m laboratories to rankings. The initial value a_1 of A is chosen to be equal to a least lower bound of uncertainty intervals $a_1 = \min\{u_1(x_i) \mid i = 1,...,m\}$ provided by laboratories. The finite value a_n of A is chosen to be equal to a largest upper bound of laboratories uncertainty intervals $a_n = \max\{u_u(x_i) \mid i = 1,...,m\}$.

Divide A into $n - 1$ equal intervals (divisions) in such a way that their amount guarantees a necessary and sufficient accuracy of the measurand values representation. Then there will be n values of the measurand $A = \{a_1, a_2, \ldots, a_n\}$ corresponding to boundaries of the division intervals (marks), see Figure 1. Details on the proper selection of a particular value of n can be found in [14].

Compose a *preference profile* Λ of m rankings representing the uncertainty intervals of laboratories. Each i-th ranking, $i = 1, \ldots, m$, is a union of binary relations of strict order and equivalence possessing the following properties at $k = 1, \ldots, m$ and $i, j = 1, \ldots, n$:

a) $a_i \succ a_j$ if $a_i \in u(x_k) \wedge a_j \notin u(x_k)$;

b) $a_i \sim a_j$ if $a_i, a_j \in u(x_k) \vee a_i, a_j \notin u(x_k)$;

c) $a_i \prec a_j$ if $a_i \notin u(x_k) \wedge a_j \in u(x_k)$.

Then the measurement result indicated by some laboratories is represented by a ranking of the measurand values where one or more equivalent values which belong to the uncertainty interval of the laboratory are more preferable. All other values of A in this ranking are less preferable and equivalent to each other. Thus, each ranking includes a single symbol of strict order $\succ$ and $n - 1$ symbols of equivalence $\sim$.

To aggregate the m ranking means to determine a single preference relation β ensuring a best compromise between them. Such a ranking β is called *consensus ranking*.

In the authors' works [12], [15], [16] it was shown that the Kemeny median can be used in the capacity of consensus ranking. One of the possible algorithms is based on the branch and bound technique and described in [12].

As soon as a consensus ranking β is found, a value ranked first in it can be selected as the reference value x_{ref} of the measurand.

The LCS consists of laboratories whose uncertainty intervals include the revealed reference value x_{ref}. Laboratories that do not contain the reference value are ignored when forming the largest consistent subset.

A standard uncertainty of the obtained reference value for the LCS is defined as the smallest of the two values, i.e. from the maximum lower bound $u_l(x_i) \le x_{\text{ref}}$ and the minimum upper bound $u_u(x_i) \ge x_{\text{ref}}$ of the uncertainty intervals of laboratories.

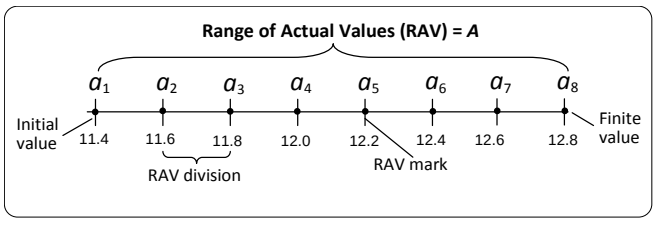

Figure 1. An example of shaping a range of actual values A.

3. EXPERIMENTAL INVESTIGATIONS OF IC DATA PROCESSING METHODS

To investigate experimentally the proposed method for IC data processing on the base of preference aggregation special software was developed called INTERLABCOM in the environment Microsoft Visual C#. The software has a user-friendly interface and, in its current version, implements the following three IC data processing methods: the proposed preference aggregation method (PAM), Procedure A and the Nielsen algorithm.

Measurement results provided by laboratories can be real and/or simulated by means of a program pseudo-random numbers generator that provides an opportunity to realize various modifications of the Monte-Carlo method when conducting numerical computing experiments. There is a possibility to choose a uniform or a normal distributions of generated measurement results. Uniformly distributed comparison data x_i and $u(x_i)$ can be generated at a given value x_{nom} using the standard library function `rand()`. Normally distributed data of comparison results are obtained from uniformly distributed data using the well-known Box–Muller transform [17].

When preparing an experiment, in a special window, one can preset a nominal measurand value x_{nom}, the number of participating laboratories m, and the number of the measurand values n. By pushing the button "Generation" the generated measurement result x_i and its uncertainty $u(x_i)$ are displayed on a monitor screen. The uncertainty $u(x_i)$ is represented as the couple of upper and lower bounds. A graph of the initial generated IC data is indicated in a special window (Figure 2). Uncertainty intervals are shown in a two-dimensional graph with dimensions “Measurand” (vertical axis) and “Laboratories” (horizontal axis).

The software allows to indicate IC data processing of each method in a separate window including a table with initial comparison data (measurand values and corresponding uncertainty intervals), a graph of comparison processed data and conclusion on consistency of each participating laboratory results.

All the IC data processing results by means of different methods are reduced to a summary table and graph. An inconsistent result is labelled by a special mark and the corresponding data are removed from the processed set. The graph and final data of comparison can be saved in Microsoft Excel format for further processing.

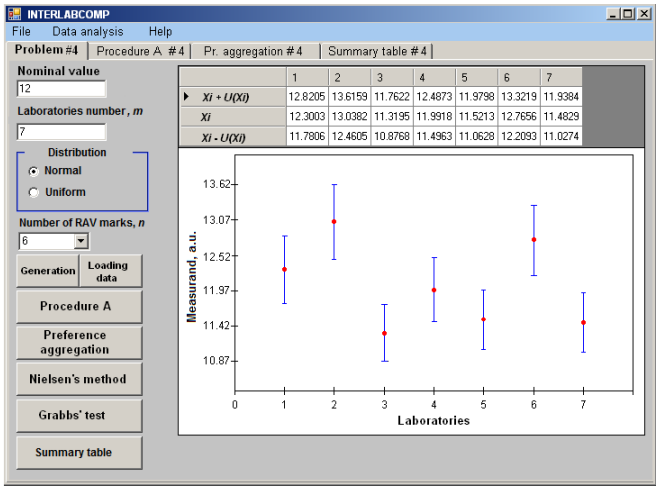

Figure 2. One of the software user interface windows.

In order to demonstrate the developed software tool operation, some IC measurement data for 7 participating laboratories are shown in Figure 3. In this case the RAV with lower and upper bounds 11.43 and 12.73 is divided into 5 equal divisions, bounds of which define 6 values a of the measurand.

The appropriate preference profile Λ, constructed as described in Section 2, has the following view:

λ_1: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5 \sim a_6$
λ_2: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5 \sim a_6$
λ_3: $a_3 \sim a_4 \sim a_5 \sim a_6 \succ a_1 \sim a_2$
λ_4: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5 \sim a_6$
λ_5: $a_3 \sim a_4 \sim a_5 \succ a_1 \sim a_2 \sim a_6$
λ_6: $a_2 \sim a_3 \sim a_4 \succ a_1 \sim a_5 \sim a_6$
λ_7: $a_1 \sim a_2 \succ a_3 \sim a_4 \sim a_5 \sim a_6$

For this profile two optimal consensus rankings exist:

$a_3 \succ a_2 \succ a_4 \succ a_5 \succ a_6 \succ a_1$
$a_3 \succ a_2 \succ a_4 \succ a_5 \succ a_1 \succ a_6$,

from where the final consensus ranking is:

$$\beta = \{\underline{a_3} \succ a_2 \succ a_4 \succ a_5 \succ a_6 \sim a_1\},$$

where the first position is occupied by the value $a_3 = 11.95$. This value is accepted as the measurand reference value x_{ref}.

Our hypothesis consists in that, as ordinal data are used in the PAM, a reference value obtained by means of this method should not significantly depend on the particular probability distribution law of measurement results.

For experimental investigations of this hypothesis normally distributed data for 100 individual problems have been generated that were distinguished from each other by random uncertainty intervals; the laboratories number $m = 15$; $x_{\text{nom}} = 3$. These data were processed by PAM, Procedure A and the Nielsen algorithm. The same steps under similar conditions were undertaken for uniformly distributed generated data.

In Table 1 and Table 2 the numerical experimental investigation results of PAM as compared with Procedure A and the Nielsen algorithm are reduced. The fact that the program model allows to assign and know a nominal value beforehand, gives a possibility to assess a quality of method M intended for IC data processing by means of calculation of the difference

$$\xi = \left| x_{\text{ref}}(M) - x_{\text{nom}} \right|. \tag{4}$$

Thus, Table 1 includes x_{ref} and ξ for each individual problem solved by each of the three methods obtained for normal distribution and Table 2 includes the values acquired for uniform distribution.

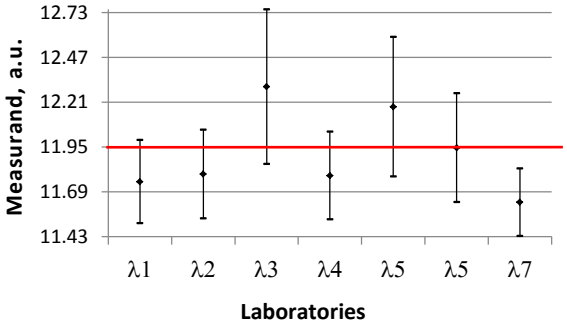

Figure 3. Example of IC measurement results.

Table 1. A fragment comparison generated data procession results for x_{nom} = 3.0 arbitrary units (a.u.); normal distribution.

Problem number	PAM		Procedure A		Nielsen algorithm	
	x_{ref}	ξ	x_{ref}	ξ	x_{ref}	ξ
1	2.97	0.03	2.92	0.08	2.95	0.05
2	2.91	0.09	2.90	0.10	2.93	0.07
3	2.95	0.05	2.91	0.09	2.91	0.09
4	2.98	0.02	2.98	0.02	2.90	0.12
5	3.05	0.05	2.90	0.10	2.96	0.04
6	2.89	0.11	2.89	0.11	2.86	0.14
7	2.98	0.02	3.00	0.00	2.79	0.21
8	2.93	0.07	2.98	0.02	3.10	0.10
9	2.98	0.02	2.86	0.14	2.91	0.09
10	2.97	0.03	2.97	0.03	2.68	0.32
11	2.98	0.02	2.95	0.05	3.02	0.02
12	2.92	0.08	2.99	0.01	2.85	0.15
13	2.99	0.01	2.97	0.03	2.92	0.08
14	2.96	0.04	2.99	0.01	2.92	0.08
15	2.93	0.07	2.99	0.01	2.99	0.01
...						
86	3.03	0.03	2.90	0.11	2.94	0.06
87	2.99	0.01	2.97	0.03	2.85	0.15
88	2.94	0.06	2.97	0.03	2.83	0.17
89	2.98	0.02	2.94	0.06	2,74	0.26
90	2.91	0.09	2.94	0.06	2.88	0.12
91	2.93	0.07	2.97	0.03	2.92	0.08
92	2.98	0.02	2.90	0.10	2.93	0.07
93	2.97	0.03	2.81	0.19	2.70	0.30
94	2.99	0.01	2.99	0.01	2.94	0.06
95	2.99	0.01	2.78	0.22	2.87	0.13
96	2.96	0.04	2.99	0.01	2.83	0.17
97	2.98	0.02	2.97	0.03	3.05	0.05
98	2.97	0.03	2.84	0.16	2.93	0.07
99	2.99	0.01	2.91	0.09	3.11	0.11
100	3.01	0.01	3.01	0.01	2.85	0.15

Table 2. A fragment comparison generated data procession results for x_{nom} = 3.0 a.u.; uniform distribution.

Problem number	PAM		Procedure A		Nielsen algorithm	
	x_{ref}	ξ	x_{ref}	ξ	x_{ref}	ξ
1	3.01	0.01	2.92	0.08	2.95	0.05
2	2.97	0.03	2.92	0.08	2.95	0.05
3	3.12	0.03	2.43	0.57	3.25	0.25
4	3.04	0.04	2.69	0.31	2.67	0.33
5	2.98	0.02	2.65	0.35	2.46	0.54
6	2.98	0.02	2.16	0.84	2.86	0.14
7	2.89	0.11	2.54	0.46	2.86	0.14
8	2.81	0.19	2.57	0.43	2.54	0.46
9	2.91	0.09	2.49	0.51	2.74	0.26
10	3.10	0.10	3.00	0.00	2.71	0.29
11	2.96	0.04	2.62	0.38	3.20	0.20
12	3.04	0.04	2.97	0.03	3.37	0.37
13	3.14	0.14	2.69	0.31	2.73	0.27
14	2.98	0.02	2.90	0.10	3.00	0.00
15	2.90	0.10	2.54	0.46	3.06	0.06
...						
86	2.99	0.01	3.01	0.01	2.95	0.05
87	2.84	0.17	2.66	0.34	2.90	0.10
88	3.03	0.03	2.88	0.12	2.85	0.15
89	2.94	0.06	2.77	0.23	2.85	0.15
90	2.86	0.14	2.40	0.60	3.09	0.09
91	2.98	0.02	2.90	0.10	2.79	0.21
92	3.11	0.11	2.38	0.62	3.27	0.27
93	2.97	0.03	2.88	0.12	2.75	0.25
94	2.73	0.27	1.90	1.10	2.69	0.31
95	3.00	0.00	2.49	0.51	3.11	0.11
96	2.96	0.04	2.95	0.05	3.11	0.11
97	2.97	0.03	2.90	0.10	3.12	0.12
98	2.95	0.05	2.62	0.38	3.12	0.12
99	2.98	0.02	2.85	0.15	2.89	0.11
100	3.08	0.08	3.01	0.01	2.93	0.07

The experimental data were used to plot curves illustrating how values ξ are changed from problem to problem for each comparison method. Values ξ were taken for every 100 individual problems and organized in ascending order.

Figure 4 shows the graph of deviations ξ obtained by the proposed PAM compared to Procedure A for uniform (U) and normal (N) distributions of comparison data. It should be noticed that Procedure A is not intended to be applied for data distributed by laws other than normal.

Therefore, the experimental results obtained under the uniform law are given here in order to demonstrate the non-robust method behaviour compared to the robust ones over the same data. One can see in Figure 4 that a particular kind of measured results probability distribution practically does not influence the PAM (curves 3 and 4) performance. It means that the PAM is a robust procedure. Over the same data, the Procedure A (curves 1 and 2) has shown a considerable increase of ξ when passing from normally to uniformly distributed measurements.

Figure 5 represents a graph of deviations ξ obtained by the

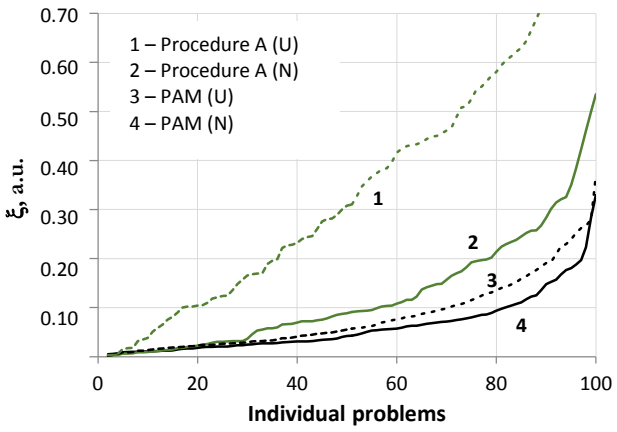

Figure 4. Deviations ξ obtained by PAM and Procedure A for uniform (U) and normal (N) distributions of comparison data.

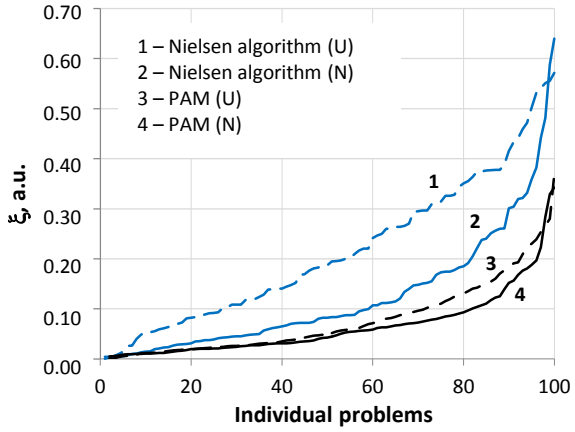

Figure 5. Deviations ξ obtained by PAM and Nielsen algorithm for uniform (U) and normal (N) distributions of comparison data.

proposed PAM compared to the Nielsen algorithm for uniform (U) and normal (N) distributions of comparison data. It can be seen from Figure 5 that the PAM provides estimates of x_{ref} closer to the nominal value x_{nom} than the Nielsen algorithm. At the same time, the latter method (curves 1 and 2) shows a discrepancy between normally and uniformly distributed data of about 0.18 which is more than twice bigger than PAM with a discrepancy 0.08.

4. REAL COMPARISONS DATA PROCESSING BY THE METHOD OF PREFERENCES AGGREGATION

Let us demonstrate the applicability of the PAM to real world examples of comparison data taken from open sources [10], [18].

4.1. The key comparison on high frequency power

Participating national metrology institutes (NMI) in key comparison (KC) CIPM CCEM.RF-K25.W [18] determined the effective efficiency and the calibration factor reference of two waveguide thermistor power sensors in the frequency range from 33 to 50 GHz. The effective efficiency of the travelling standard was determined by formula:

$$\eta_{\text{eff}} = \frac{P_{\text{DC,sub}}}{P_{\text{RF,abs}}}, \tag{5}$$

where $P_{\text{DC,sub}}$ is the substituted DC power and $P_{\text{RF,abs}}$ is the total absorbed RF power. Participants of the comparisons calculated also the calibration factor η_{cal} according to equation:

$$\eta_{\text{cal}} = (1-\Gamma^2)\eta_{\text{eff}} \tag{6}$$

where Γ is the input reflection coefficient of the travelling standard which was measured as a complex quantity stated as magnitude and phase at the measuring frequencies.

The median absolute deviation was used to identify an outlier:

$$\sigma \approx S(\text{MAD}) \equiv k_1 \text{median}\{[\eta_i - \eta_{\text{med}}]\}, \tag{7}$$

where k_1 is a multiplier determined by simulation; η_{med} is the median value of measurement results $\{\eta\}$.

The value of η_i which differed from the median by more than $2.5 \cdot S(\text{MAD})$ has been regarded as an outlier. It has been excluded from the calculation of the reference value. This criterion was used to check each measurement result:

$$\left|\eta_i - \eta_{\text{med}}\right| > 2.5 \cdot S(\text{MAD}). \tag{8}$$

The reference value of KC was determined in accordance with section 8 of the technical report [18] on the basis of the unweighted mean value:

$$\eta_{\text{eff,ref}} = \frac{1}{m'}\sum_{i=1}^{m'} \eta_{\text{eff},i}. \tag{9}$$

The standard uncertainties were calculated:

$$u(\eta_{\text{eff,ref}}) = \sqrt{\frac{1}{m'}\sum_{i=1}^{m'} u^2(\eta_{\text{eff},i})}. \tag{10}$$

KC data treatment in accordance with CCEM.RF-K25.W.

The results of the comparison on effective efficiency at 36 GHz are reduced to Table 3.

The comparison reference value $\eta_{\text{eff,ref}} = 0.9161$ was determined for the effective efficiency η_{eff} with the uncertainty $u(\eta_{\text{eff,ref}}) = 0.0027$. NIM, MNIA and NRC have not participated

Table 3. Key comparison data on effective efficiency.

m	NMI	Effective efficiency, η_{eff}	
		$\eta_{\text{eff},i}$	$u(\eta_{\text{eff},i})$
1	PTB	0.9153	0.0031
2	NPL	0.9167	0.0060
3	NIST	0.9184	0.0064
4	LNE	0.9157	0.0018
5	KRISS	0.9143	0.0104
6	VNIIFTRI	0.9160	0.0079
7	NIM	0.8360	0.0072
8	MNIA	0.9174	0.0071
9	NRC	0.9375	0.0130

in the reference value determination as NIM and NRC measurement results were considered to be outliers in accordance with criterion (8). The result of MNIA was proved to be traceable to the results of other participants. A graphic illustration of the comparison results and the reference value are shown in Figure 6.

The results of the comparison on the calibration factor are reduced to Table 4.

The reference value for the calibration factor $\eta_{\text{cal,ref}} = 0.7942$ was determined with the uncertainty $u(\eta_{\text{cal,ref}}) = 0.0024$ (Figure 7). Results of VNIIFTRI and NRC were recognized as outliers. Result of NMIA turned out to be traceable to the results of other participants.

KC data treatment by the PAM.

Data of Table 3 were processed using PAM at $n = 8$. Then the RAV was divided into $n - 1 = 7$ equal divisions. Bounds of divisions corresponded to eight values a of the measurand: $a_1 = 0.8288$, $a_2 = 0.8462$, $a_3 = 0.8636$, $a_4 = 0.8809$, $a_5 = 0.8983$, $a_6 = 0.9157$, $a_7 = 0.9331$, and $a_8 = 0.9505$. The preference profile consisted of nine rankings describing the uncertainty intervals

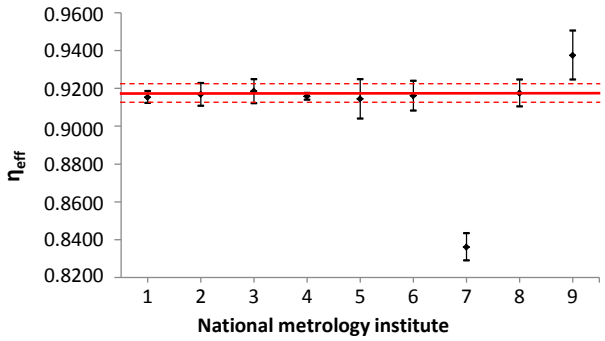

Figure 6. Uncertainty intervals of effective efficiency value provided by NMIs.

Table 4. Key comparisons data on the calibration factor.

m	NMI	Calibration factor, η_{cal}	
		$\eta_{\text{cal},i}$	$u(\eta_{\text{cal},i})$
1	PTB	0.7954	0.0036
2	NPL	0.7937	0.0067
3	NIST	0.7976	0.0070
4	LNE	0.7914	0.0046
5	KRISS	0.7935	0.0079
6	NIM	0.7936	0.0031
7	VNIIFTRI	0.7820	0.0105
8	MNIA	0.7972	0.0073
9	NRC	0.8140	0.0130

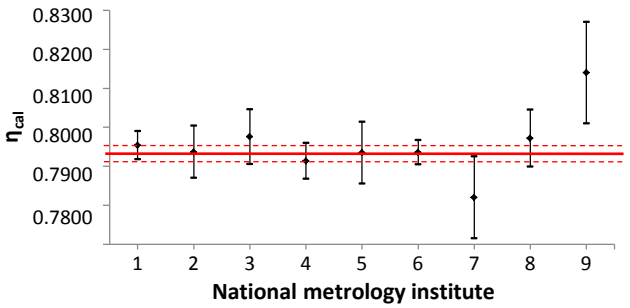

Figure 7. Uncertainty intervals of calibration factor value provided by NMIs.

of the appropriate NMIs:

λ_1: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_2: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_3: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_4: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_5: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_6: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_7: $a_1 \succ a_2 \sim a_3 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_8: $a_6 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6 \sim a_7 \sim a_8$
λ_9: $a_7 \sim a_8 \succ a_1 \sim a_2 \sim a_3 \sim a_4 \sim a_5 \sim a_6$.

The final consensus ranking was determined as $\beta_{fin} = \{\underline{a_6} \succ a_1 \sim a_7 \sim a_8 \succ a_2 \sim a_3 \sim a_4 \sim a_5\}$. The comparison reference value of $a_6 = \eta_{eff,ref} = 0.9157$ was obtained with uncertainty $u(\eta_{eff,ref}) = 0.0018$. The cardinality of LCS was $m' = 7$ as measurement results of NIM and NRC were recognized to be outliers because of not containing the obtained reference value (Figure 8).

Data of Table 5 were also processed, using the PAM, at $n = 6$. The RAV was divided into five equal divisions. Bounds of the intervals corresponded to six values of the measurand: $a_1 = 0.7715$, $a_2 = 0.7826$, $a_3 = 0.7937$, $a_4 = 0.8048$, $a_5 = 0.8159$, $a_6 = 0.8270$ (Figure 9).

The preference profile was shaped of 9 rankings:

λ_1: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_2: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_3: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_4: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_5: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_6: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_7: $a_1 \sim a_2 \succ a_3 \sim a_4 \sim a_5 \sim a_6$
λ_8: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5 \sim a_6$
λ_9: $a_4 \sim a_5 \succ a_1 \sim a_2 \sim a_3 \sim a_6$.

The final consensus ranking was determined as $\beta_{fin} = \{\underline{a_3} \succ a_2 \sim a_4 \sim a_5 \sim a_6 \succ a_1\}$. The reference value of comparison $\eta_{cal,ref} = 0.7937$ was obtained with uncertainty $u(\eta_{cal,ref}) = 0.0019$. The cardinality of LCS was $m' = 7$ (Figure 9).

Measurement results of VNIIFTRI and NRC were recognized to be outliers.

4.2. Interlaboratory power comparison in the microwave region

In [5], results of interlaboratory power comparisons in the microwave region (50 MHz–26.5 GHz) on the project SIT.AF-01 were reviewed. They were organized by the INRiM (Istituto Nazionale di Ricerca Metrologica, Italy) in Turin. A Hewlett Packard power meter model 438A as a travelling standard has been sent to 12 laboratories. The comparison aim was to confirm the claimed uncertainties of laboratories accredited in the national system of accreditation in the field of microwave measurements.

Table 5 and Figure 10 show one of the series of comparison data of the power sensor calibration factor K measurements at a frequency of 1 GHz.

To process the comparison data the Nielsen algorithm (see Section 1) was used. According to the data analysis outcomes,

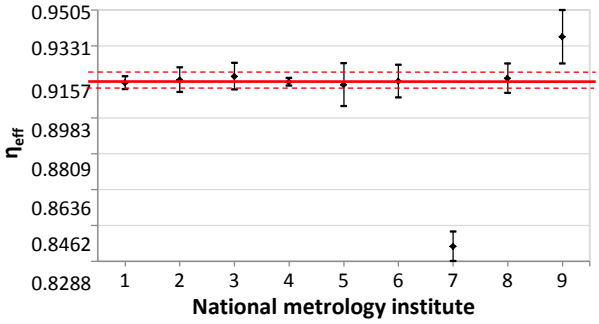

Figure 8. Uncertainty intervals of effective efficiency value provided by NMIs and reference value obtained by the PAM.

Table 5. Comparisons data on power sensor calibration factor K at 1 GHz.

Laboratories	x_i	$u(x_i)$
1	0.985	0.013
2	0.989	0.008
3	0.982	0.013
4	0.982	0.035
5	0.984	0.014
6	0.980	0.028
7	0.981	0.017
8	0.990	0.021
9	0.982	0.011
10	0.989	0.017
11	1.017	0.014
12	0.987	0.019

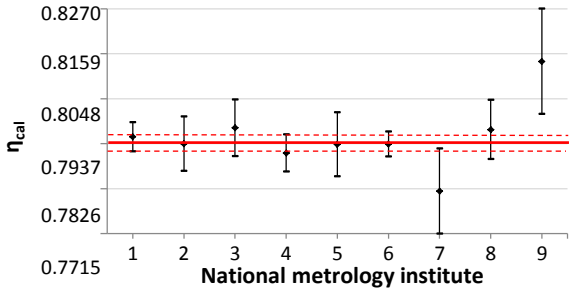

Figure 9. Uncertainty intervals of the calibration factor value provided by NMIs and reference value obtained by the PAM.

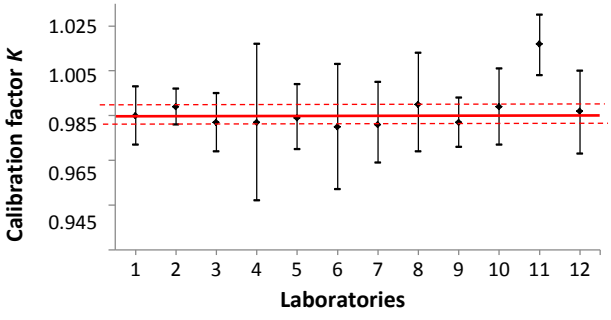

Figure 10. Uncertainty intervals of calibration factor K value provided by participating laboratories and corresponding reference value.

the LCS, formed as a result of Nielsen's algorithm processing, included the eleven laboratories. Laboratory 11 was excluded, because its result, in accordance with the algorithm conditions, was deemed to be unreliable. The reference value was obtained as $x_{ref} = 0.985$ in correspondence with the greatest number of laboratory "votes".

The data of Table 5 were processed using the PAM at $n = 5$. The RAV was divided into $n - 1 = 4$ equal divisions. The bounds of intervals corresponded to five values a of the measurand: $a_1 = 0.947$, $a_2 = 0.968$, $a_3 = 0.989$, $a_4 = 1.009$, $a_5 = 1.030$ (Figure 11).

The corresponding preference profile was as follows:

λ_1: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_2: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_3: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_4: $a_1 \sim a_2 \sim a_3 \sim a_4 \succ a_5$
λ_5: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_6: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5$
λ_7: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5$
λ_8: $a_3 \sim a_4 \succ a_1 \sim a_2 \sim a_5$
λ_9: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_{10}: $a_3 \succ a_1 \sim a_2 \sim a_4 \sim a_5$
λ_{11}: $a_4 \sim a_5 \succ a_1 \sim a_2 \sim a_3$
λ_{12}: $a_2 \sim a_3 \succ a_1 \sim a_4 \sim a_5$.

The final consensus ranking was $\beta_{fin} = \{\underline{a_3} \succ a_2 \succ a_4 \succ a_1 \sim a_5\}$. The value a_3 was chosen as the reference value $x_{ref} = 0.989$ with a corresponding uncertainty $u(x_{ref}) = 0.004$. The LCS formed by the PAM included 11 laboratories, just as in the project SIT.AF-01 (Figure 11).

5. CONCLUSION

A method, called preference aggregation method (PAM) has been described, aimed to process IC data. The PAM is based on transformation of uncertainty intervals provided by participating laboratories into rankings of measured quantity values. For a preference profile composed in this way, a consensus ranking is determined by the Kemeny rule that allows to find the reference value of a measurand. The operation of this method was demonstrated.

A software tool has been considered that is intended for experimental investigations of the proposed method and other methods of processing generated normally and uniformly distributed IC data. Numerical experiments, carried out with its help, have shown that the PAM is indeed a robust procedure that does not depend on the probability distribution of the measurement results.

It also follows from numerical experiments that the PAM provides an estimate of a reference value being closer to the nominal value than the other robust method (the Nielsen algorithm) with half the discrepancy between normally and uniformly distributed comparison data.

The PAM performance was experimentally verified on real comparison results. In all cases, the reference value and associated uncertainty, determined by the proposed method, were very close to the outcomes obtained by the comparison coordinators.

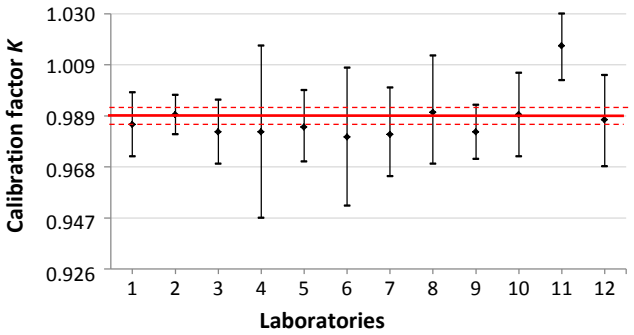

Figure 11. Uncertainty intervals provided by participating laboratories and corresponding reference value obtained by the PAM.

ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Education and Science of Russian Federation, basic part of the state task in 2014-2016, project 2078 and in 2017-2019, project 4.1763.GZB.2017. The authors would like to thank the anonymous referee for helpful comments.

REFERENCES

[1] CIPM MRA-D-05. Measurement comparisons in the CIPM MRA, Version 1.5, p. 28.

[2] ISO/IEC 17043 (2010) Conformity assessment – General requirements for proficiency testing. International Organization for Standardisation, Geneva, Switzerland.

[3] M.G. Cox, The evaluation of key comparison data: determining the largest consistent subset, Metrologia 44 (2007) pp. 187-200.

[4] M.G. Cox, The evaluation of key comparison data, Metrologia 39 (2002) pp. 589-595.

[5] N.Yu. Efremova, A.G. Chunovkina, Experience in evaluating the data of interlaboratory comparisons for calibration and verification laboratories, Meas. Tech. 50(6) (2007) pp. 584-592.

[6] C. Elster, B. Toman, Analysis of key comparisons data: critical assessment of elements of current practice with suggested improvements, Metrologia 50 (2013) pp. 549-555.

[7] I. Lira, A.G. Chunovkina, C. Elster, W. Woeger, Analysis of key comparisons incorporating knowledge about bias, IEEE Trans. Instrum. Meas. 61(8) (2012) pp. 2079-2084.

[8] ISO 13528 (2005) Statistical methods for use in proficiency testing by interlaboratory comparisons. International Organization for Standardisation, Geneva, Switzerland.

[9] H.S. Nielsen, Determining consensus values in interlaboratory comparisons and proficiency testing, NCSLI Newsletter 44(2) (2004) pp. 12-15.

[10] L. Brunetti, L. Oberto, M. Sellone, P. Terzi, Establishing reference value in high frequency power comparisons, Measurement 42 (2009), pp. 318-1323.

[11] S.V. Muravyov, I.A. Marinushkina, "Largest consistent subsets in interlaboratory comparisons: preference aggregation approach", Proc. of 14th Joint International IMEKO TC1, TC7, TC13 Symposium, Aug. 31-Sept. 2, 2011, Jena, Germany, pp. 69-73.

[12] S.V. Muravyov, Ordinal measurement, preference aggregation and interlaboratory comparisons, Measurement 46(8) (2013) pp. 2927-2935.

[13] S.V. Murav'ev, Aggregation of preferences as a method of solving problems in metrology and measurement technique, Meas. Tech. 57(2) (2014) pp. 132-138.

[14] S.V. Muravyov, I.A. Marinushkina, Processing of interlaboratory comparison data by preference aggregation method, Meas. Tech. 58(12) (2016) pp. 1285-1291.

[15] S.V. Muravyov, I.A. Marinushkina, Intransitivity in multiple solutions of Kemeny Ranking Problem, J. Phys. Conf. Ser. 459(1) (2013) 012006.

[16] S.V. Muravyov, Dealing with chaotic results of Kemeny ranking determination, Measurement 51 (2014) pp. 328-334.

[17] G.E.P. Box, M.E. Muller, A note on the generation of random normal deviates, Ann. Math. Stat. 29(2) (1958) pp. 610-611.

[18] R. Judaschke, Final report of the pilot laboratory CCEM Key Comparison CCEM.RF-K25.W RF power from 33 GHz to 50 GHz in waveguide, Physikalisch-Technische Bundesanstalt, Germany, 2014.