XXI IMEKO World Congress “Measurement in Research and Industry”
August 30 − September 4, 2015, Prague, Czech Republic

A SOFTWARE TOOL FOR SIMULATION OF INTERLABORATORY COMPARISON DATA EVALUATION METHOD USING PREFERENCE AGGREGATION

Sergey Muravyov, *Irina Marinushkina*

Tomsk Polytechnic University, Department of Computer-aided Measurement Systems and Metrology
Tomsk, Russia, muravyov@tpu.ru

Abstract − It is presented an integrated software for experimental testing preference aggregation method for interlaboratory comparison data processing. 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.

Keywords: interlaboratory comparisons, reference value, largest consistency subset, preference aggregation, robust method

1. INTRODUCTION

Interlaboratory comparisons (IC) are now 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 are not participated in establishing final reference value.

There are different approaches to check consistency of laboratory measurement results and to find the reference value x_{ref}, see, for example [3-6]. Choice of 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 difference of laboratory measurement result and assigned by comparison provider, percent differences, percentiles, or ranks [8]. However, these methods usually imposes limitations on a feasible IC participating laboratories number. Moreover, statistical methods may evince low discriminating ability, that is ability to differ truly unreliable laboratories from laboratories providing results to be trusted to.

In [1], a method had been presented that is based on rather widely known so called Procedure A [4-5]. The procedure uses weighted mean value *y*:

$$y = \frac{\sum_{i=1}^{m} \frac{x_1}{u^2(x_i)}}{\sum_{i=1}^{m} \frac{1}{u^2(x_i)}}, \tag{1}$$

where x_i is nominal value estimate provided by *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:

$$u^2(y) = \left(\sum_{i=1}^{m} \frac{1}{u^2(x_i)} \right)^{-1}. \tag{2}$$

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 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 remainder in limits of claimed uncertainties. A result is deemed to be inconsistent if $|E_{\text{n}}| > 2$, where

$$E_{\text{n}} = \frac{x_i - y}{\sqrt{u^2(x_i) \pm u^2(y)}}, \; i = 1, \ldots, m. \tag{3}$$

The process of exclusion of one inconsistent result is repeated until a consistency of results by the criterion χ^2 is achieved. For obtained in this way LCS the reference value is determined by formula (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 with 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 which successful application 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 of reference value x_{ref} determination that is insensitive to outliers, i.e. results with the uncertainty considerably lower than those of other participants.

This paper is devoted to software implementation of comparison reference value determination method that is 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, constituent an initial preference profile, can serve as input data for determination of consensus ranking by Kemeny rule that allows to find the reference value of measurand and to assess an ability of participating laboratories to provide reliable measurement results. In Section 3 a specially developed software is discussed to carry out numerical experimental researches of IC methods including Procedure A, Nielsen algorithm and the proposed preference aggregation method.

2. IC DATA PROCESSING ON THE BASE OF PREFERENCE AGGREGATION

Define procedure of transformation of uncertainty intervals provided by laboratories into rankings. For this aim, designate an uncertainty interval gained by *i*-th laboratory through $u(x_i) = [u_{\text{l}}(x_i), u_{\text{u}}(x_i)]$.

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 $a_1 = \min\{u_{\text{l}}(x_i) \mid i = 1,...,m\}$ provided by laboratories. The finite value a_n of *A* is chosen to be equal to a largest upper bound of laboratories uncertainty intervals $a_n = \max\{u_{\text{u}}(x_i) \mid i = 1,...,m\}$.

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 Fig. 1.

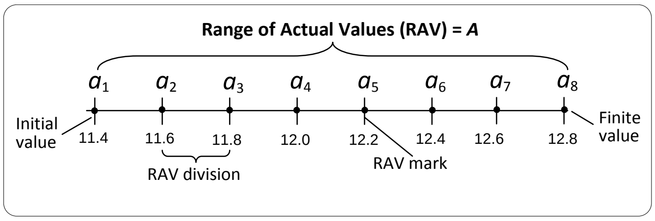


Fig. 1. An example of shaping a range of actual values *A*.

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 indicated by some laboratory measurement result is represented by a ranking of the measurand values where more preferable are one or more equivalent values which belong to uncertainty interval of the laboratory. 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 ~.

To aggregate *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 [12,14,15] it was shown that Kemeny median can be used in the capacity of consensus ranking. One of possible algorithms on the base of branch and bound technique is 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 measurand.

Subset of consistent results will consist of laboratories uncertainty intervals of which include the revealed reference value. In opposite case corresponding laboratories are excluded of forming the largest consistent subset.

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

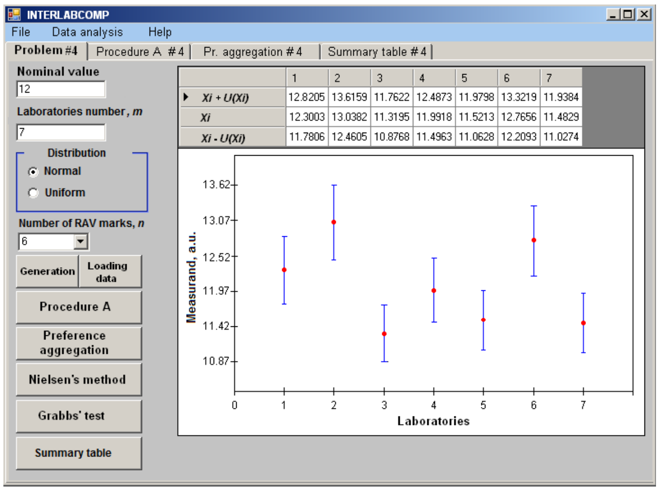
3. EXPERIMENTAL INVESTIGATIONS OF IC DATA PROCESSING METHODS

To investigate experimentally the proposed method for IC data processing on the base of preference aggregation there was developed special software called INTERLABCOM in the environment Microsoft Visual C#. The software has 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 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 Monte-Carlo method when conducting numerical computing experiments. There is a possibility of choice of uniform or normal distributions of generated measurement results. Uniformly distributed data of comparison x_i и $u(x_i)$ can be generated at a given value x_{nom} using standard library function `rand()`. Normally distributed data of comparison results are obtained of uniformly distributed data using well known Box–Muller transform [16].

When preparing for an experiment, in a special window, one can preset a nominal measurand value x_{nom}, number of participating laboratories *m*, and number of the measurand values *n*. By pushing button "Generation" 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. Graph of the initial generated IC data is indicated in a special window (Fig. 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), graph of comparison processed data and conclusion on consistency of each participating laboratory results.



Nominal value: 12

Laboratories number, *m*: 7

Distribution: ● Normal ○ Uniform

Number of RAV marks, *n*: 6

	1	2	3	4	5	6	7
Xi + U(Xi)	12.8205	13.6159	11.7622	12.4873	11.9798	13.3219	11.9384
Xi	12.3003	13.0382	11.3195	11.9918	11.5213	12.7656	11.4829
Xi - U(Xi)	11.7806	12.4605	10.8768	11.4963	11.0628	12.2093	11.0274

Fig. 2. One of the software user interface windows.

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

In order to demonstrate the developed software tool operation, some IC measurement data for 7 participating laboratories are shown in Fig. 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.

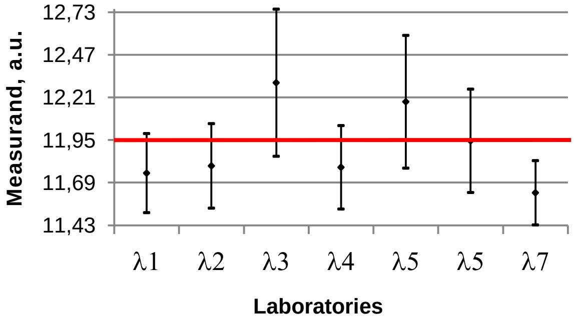


Fig. 3. Example of IC measurement results.

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:

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

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

In Table 1 and Table 2 the numerical experimental investigations results of PAM as compared with Procedure A and 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 simple calculation of the difference

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

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.

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 of 100 individual problems and organized in ascending order.

Fig. 4 show graph of deviations ξ obtained by the proposed PAM compared to Procedure A for uniform (U) and normal (N) distributions of comparison result. 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 for it under the uniform law are given here in order to demonstrate the non-robust method behavior compared to the robust ones over the same data. One can see in Fig. 4 that a particular kind of measured results probability distribution practically does not influence to 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 considerable increasing of ξ when passing from normally to uniformly distributed measurements.

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

Table 1. A fragment comparison generated data procession results for $x_{\text{ном}}$ = 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

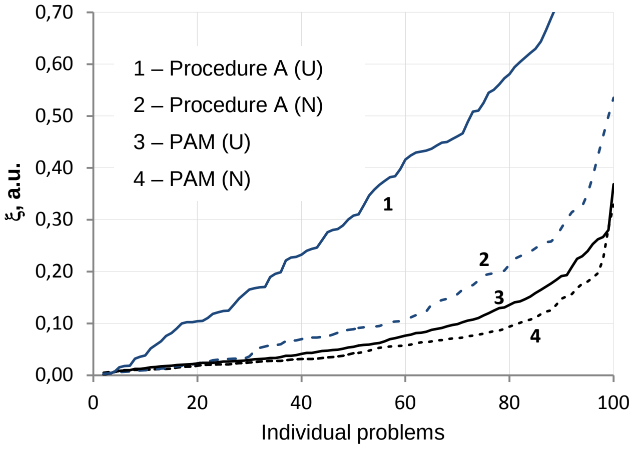


Fig. 4. Deviations ξ of obtained by PAM and Procedure A for uniform (U) and normal (N) distributions of comparison results.

4. CONCLUSION

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

Table 2. A fragment comparison generated data procession results for $x_{\text{ном}}$ = 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

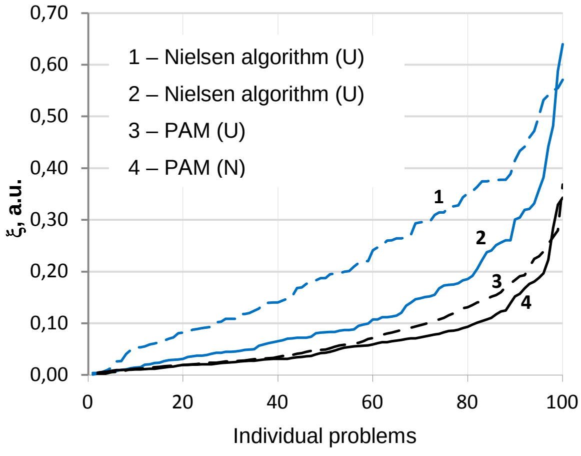


Fig. 5. Deviations ξ of obtained by PAM and Nielsen algorithm for uniform (U) and normal (N) distributions of comparison results

A software tool was considered that is intended for experimental researches of the proposed method and other methods of IC data processing. Numerical experiments carried out with its help have shown that the PAM is indeed a robust procedure that does not depend on measured results probability distribution. The PAM provides an estimates of a reference value closer to the nominal value than other robust method (Nielsen algorithm). The latter method has shown discrepancy between normally and uniformly distributed data is more than twice bigger against PAM.

ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Education and Science of Russian Federation, basic part of the state task, project No. 2078.

REFERENCES

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

[2] ISO/IEC 17043:2010 Conformity assessment – General requirements for proficiency testing, pp. 1–39.

[3] M.G. Cox, "The evaluation of key comparison data: determining the largest consistent subset", *Metrologia*, vol. 44, pp. 187–200, 2007.

[4] M.G. Cox, "The evaluation of key comparison data", *Metrologia*, vol. 39, pp. 589–595, 2002.

[5] Efremova N. Yu., Chunovkina A.G. Experience in evaluating the data of interlaboratory comparisons for calibration and verification laboratories, *Measurement Techniques*, vol. 50, nº. 6. pp. 584–592, 2007.

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

[7] I. Lira, A.G. Chunovkina, C. Elster, W. Woeger, "Analysis of key comparisons incorporating knowledge about bias", *IEEE Transactions on Instrumentation and Measurement*, vol. 61, nº. 8, pp. 2079–2084, 2012.

[8] ISO 13528:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons, pp. 1–66.

[9] H.S. Nielsen, "Determining consensus values in interlaboratory comparisons and proficiency testing", *NCSLI Newsletter*, vol. 44, nº. 2, pp. 12-15, 2004.

[10] L. Brunetti, L. Oberto, M. Sellone, P. Terzi, "Establishing reference value in high frequency power comparisons", *Measurement*, vol. 42, pp. 1318–1323, 2009.

[11] S.V. Muravyov, I.A. Marinushkina, "Largest consistent subsets in interlaboratory comparisons: preference aggregation approach", Joint International IMEKO TC1+TC7+TC13 Symposium pp. 287–290, Jena, Germany, Sept. 2011.

[12] S.V. Muravyov, "Ordinal measurement, preference aggregation and interlaboratory comparisons", *Measurement*, vol. 46, Issue 8, pp. 2927–2935, 2013.

[13] S.V. Muravyov, "Aggregation of preferences as a method of solving problems in metrology and measurement technique", *Measurement Techniques*, vol. 57, nº. 2, pp. 132–138, May 2014.

[14] S.V. Muravyov, Marinushkina I.A., "Intransitivity in multiple solutions of Kemeny Ranking Problem", *Journal of Physics: Conference Series*, vol. 459, issue 1, Atricle number 012006, pp. 1–6, 2013.

[15] S.V. Muravyov, "Dealing with chaotic results of Kemeny ranking determination", *Measurement*, vol. 51, pp. 328-334, 2014.

[16] G.E.P. Box, M.E. Muller "A note on the generation of random normal deviates" // *The Annals of Mathematical Statistics*, vol. 29, nº. 2, pp. 610–611, 1958.