Anal Bioanal Chem (2013) 405:8363–8375
DOI 10.1007/s00216-013-7206-5

Comparison of Multianalyte Proficiency Test Results by Sum of Ranking Differences, Principal Component- and Hierarchical Cluster Analysis

Biljana Škrbić, Károly Héberger, Nataša Đurišić-Mladenović

Comparison of Multianalyte Proficiency Test Results by Sum of Ranking Differences, Principal Component- and Hierarchical Cluster Analysis

Biljana Škrbić[a], Károly Héberger[b,*], Nataša Đurišić-Mladenović[a]

[a] University of Novi Sad, Faculty of Technology, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia

[b] Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1025 Budapest, Pusztaszeri út 59-67, Hungary

* Corresponding author: Károly Héberger, E-mail: heberger.karoly@ttk.mta.hu

ABSTRACT

Sum of ranking differences (SRD) was applied for comparing multianalyte results obtained by several analytical methods used in one or in different laboratories, i.e. for ranking the overall performances of the methods (or laboratories) in simultaneous determination of the same set of analytes. The data sets for testing of the SRD applicability contained the results reported during one of proficiency tests (PTs) organized by EU Reference Laboratory for Polycyclic Aromatic Hydrocarbons (EU-RL-PAH). In this way, the SRD was also tested as a discriminant method alternative to existing average performance scores used to compare mutlianalyte PT results. SRD should be used along with the *z*-scores – the most commonly used PT performance statistics.

SRD was further developed to handle the same rankings (ties) among laboratories. Two benchmark concentration series were selected as reference: (i) the assigned PAH concentrations (determined precisely beforehand by the EU-RL-PAH), (ii) the averages of all individual PAH concentrations determined by each laboratory.

Ranking relative to the assigned values and also to the average (or median) values pointed the same laboratories with the most extreme results, as well as revealed groups of laboratories with similar overall performances. SRD reveals differences between methods or laboratories even if classical test(s) cannot. The ranking was validated using comparison of ranks by random numbers (a randomization test) and using seven folds cross-validation, which highlighted the similarities among the (methods used in) laboratories. Principal component analysis and hierarchical cluster analysis justified the findings based on SRD ranking/grouping. If the PAH-concentrations are row-scaled (i.e. *z*-scores are analyzed as input for ranking) SRD can still be used for checking the normality of errors. Moreover, cross-validation of SRD on *z*-scores groups the laboratories similarly. The SRD technique is general in nature, i.e. it can be applied to any experimental problem in which the multianalyte results obtained either by several analytical procedures, analysts, instruments, or laboratories need to be compared.

Keywords: ~~interlaboratory~~ multianalyte results, comparison, ~~polycyclic aromatic hydrocarbons,~~ ~~round-robin test~~, sum of ranking differences, principal component analysis, comparison of ranks by random numbers

1. Introduction

Due to public health concerns there has been a need in different domains (e.g. food safety, environmental protection) for development of analysis that can identify and measure the numerous contaminants belonging to the same or similar chemical groups in order to get as many as possible data in one analytical run for the risk assessment. For instance, there are several lists of contaminants belonging to different chemical classes (polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), organochlorine pesticides, etc.) required or advised to be monitored in food and/or environmental samples. In response to this, a number of methods have been developed and applied routinely for the control of contaminants levels. Those methods that can identify and measure a number of analytes concurrently are called “multianalyte (i.e. multiresidue) methods” [1]. Although these methods are in routine use, they are often quite complex and differ among themselves in terms of the sample preparation step, instrumental techniques available, applied working parameters, etc. Multianalyte methods require not only careful performance but also continuous monitoring to check the reliability of the measurements [1].
In order to verify the confidence in measurement results (or the competence of the laboratory either accredited or non-accredited), including such multianalyte results, there is a request for the laboratories to have quality control procedures for monitoring the performances of the analysis undertaken (ISO/IEC 17025). One of the means to monitor the laboratory performance is its participation in interlaboratory comparison programs. In an interlaboratory comparison ~~experiment~~, different laboratories determine some characteristic, e.g. concentration of the same analyte(-s) in one or various homogenous samples under documented conditions, assuming that the systematic errors of methods in different laboratories follow normal distribution [2-6]. For simplicity, we use the term interlaboratory comparison further on knowing that it is essentially analytical methods comparison, (c.f. Table 1). The typical purposes for interlaboratory comparisons include evaluation of the performance of laboratories for specific measurements, identification of problems in

laboratories and initiation of actions for improvement, establishment of the effectiveness and comparability of test or measurement methods, provision of additional confidence to laboratory customer, etc. [6]. In general there are two sub-types of interlaboratory studies: i) collaborative trials or method performance (used to check the performance, generally the precision) of a single analytical method, and ii) proficiency testing or laboratory performance studies (sometimes, the term “round robin test” is also used) [3].

The laboratories participating in proficiency tests receive test material from the proficiency testing provider; the material should be analyzed by measurement procedure of the choice, which is consistent with the routine procedure in the laboratory. In the specified time period, the results of the test material analysis should be reported to the proficiency testing provider, who further analyze the results by appropriate statistical methods, generating summary statistics and performance statistics in order to aid interpretation and to allow comparison with defined objectives. In fact, the purpose is to measure deviation from the assigned value – a value attributed to a particular property of a proficiency test material (e.g. concentration of analyte(-s)). Determination of the assigned values belongs to the responsibility of the proficiency testing providers. The assigned value is not disclosed to the participants until they have reported their results. Different statistical methods may be used for calculation of the performance statistics; generally simple numerical or graphical criteria, described in ISO 13528 [5] and ISO/IEC17043 [6] have been used to interpret the results reported by laboratories participating in a proficiency test. The majority of these performance statistics are generated from the results referring to the single analyte. If several analytes are subjects of the proficiency test, performance statistics are generally given for each analyte separately (i.e. the results for each analyte are analyzed separately). Additionally, in the case of results for several analytes in one proficiency test material (multianalyte results), the use of some graphical methods are recommended by ISO 13528 [5], describing the conditions and limitations of these approaches. Youden [2] also describes a protocol how to complete an interlaboratory examination, how to present data and what to do with the problems arisen (missing data, outliers,

and ranking laboratories). Youden suggests an approximate test to decide whether a laboratory "passed" the test in measuring a single analyte (i.e. if produces acceptable results or not). The test is based on *sum of ranks* and a special table with *critical values* could be used for comparison only in the case if the number of participating laboratories is 15 or less. The other limitation of Youden protocol is the number of objects (e.g. compounds content), which are also restricted and decision on the laboratory accuracy needs a more sophisticated evaluation.

One of the most commonly used performance statistics is the *z*-scores calculated by Equation (1):

$$z = \frac{(x - X)}{\hat{s}} \quad (1)$$

where x is the participant's result, X is the assigned value and $\hat{s}$ is the sample standard deviation for proficiency assessment, which can be calculated by applying one of five proposed approaches [5,6]. The standard deviation for proficiency testing is used to assess laboratory bias, i.e. deviation from the assigned value found in a proficiency test [5].

"Satisfactory" performance is indicated by an absolute value of *z*-score less or equal to 2. Absolute values of *z*-score between 2 and 3 suggest "questionable" performance, while results are considered "unsatisfactory" if absolute values of *z*-scores are above 3.0.

However, some authors highlighted that the *z*-score statistics can present pitfalls and have limitations, so they should be interpreted cautiously [7,8].

Organization of the interlaboratory comparisons (ILCs) ~~(~~PAHs~~)~~ in food is one of the core duties of the European Union Reference Laboratory for PAHs in food (EU-RL-PAH) hosted at the Institute for Reference Materials and Measurements (IRMM) of the European Commission's Joint Research Centre. PAHs are a group of about ten thousand compounds, a few of them occurring in considerable amounts in the environment and food, many being classified as probable or possible human carcinogens. Human beings are exposed to PAHs mostly by intake of food, which is also the reason why reliable analysis of PAHs in foodstuffs is of great importance. The activities of EU-RL-PAH refer to Commission Regulation (EC) 1881/2006 [9] as amended by Commission Regulation 835/2011 [10] setting maximum levels of selected PAHs in various types of food, and to

Commission Regulation 333/2007 [11] as amended by Commission Regulation 836/2011 [12] laying down sampling and analysis measures for the official control of the selected PAH levels in foodstuffs.

Till now, there have been nine rounds of ILCs organized by EU-RL-PAH for 15+1 EU priority PAHs (5-Methylchrysene-5MC, Benzo[a]anthracene-BAA, Benzo[a]pyrene-BAP, Benzo[b]fluoranthene-BBF, Benzo[c]fluorine-BCF, Benzo[ghi]perylene-BGP, Benzo[j]fluoranthene-BJF, Benzo[k]fluoranthene-BKF, Chrysene-CHR, Cyclopenta[cd]pyrene-CPP, Dibenzo[a,e]pyrene-DEP, Dibenzo[a,h]anthracene-DHA, Dibenzo[a,h]pyrene-DHP, Dibenzo[a,i]pyrene-DIP, Dibenzo[a,l]pyrene-DLP, and Indeno[1,2,3-cd]pyrene-ICP) in various matrices, e.g. olive oils, sausages, solvent solutions, etc. Reports of these ILCs are readily available on the official web site of IRMM:
http://irmm.jrc.ec.europa.eu/interlaboratory_comparisons/Pages/index.aspx.

These ILC studies aimed to evaluate trueness and precision of analytical results reported by the participating laboratories for compounds belonging to the group of 15+1 EU priority PAHs in different food matrices and to assess the influence of standard preparation and instrument calibration on the performance of individual laboratories. The ILCs organized by EU-RL-PAH till 2010 have been designed and evaluated along the guidelines given in well approved ISO/IEC Guide 43 [13], while the latest proficiency tests have been conducted in accordance with ISO/IEC 17043 [6]. Additionally, the IUPAC International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories has been also used in all proficiency tests of EU-RL-PAH [14].

The performance of the laboratories in determination of the target PAHs in selected food items during the proficiency tests organized by EU-RL-PAH has been evaluated by *z*-score (Eq.1), in which standard deviation for proficiency testing, $\hat{\sigma}$, for benzo[a]pyrene has been set to be equal to the maximum tolerated standard measurement uncertainty, U_f, as defined by Commission Regulation (EC) No 333/2007 [11] amended by Regulation (EC) 836/2011 [12]:

$$U_f = ((LOD/2)^2+(\alpha C)^2)^{0.5} \quad (2)$$

where LOD relates to the required limit of detection (which is 0.3 μg kg^{-1} [11,12]), α is a numeric factor depending on the concentration C (for C less or equal to 50 μg kg^{-1}, α is 0.2 [11,12]). For instance, the application of Eq. 2 with the assigned value of 3.0 μg kg^{-1} for benzo[a]pyrene and the required limit of detection of 0.3 μg kg^{-1} results in a U_f value of 0.62 μg kg^{-1} (i.e. 20.6% of the assigned value of 3.0 μg kg^{-1}). For all other PAHs in the group of 15+1 EU priority PAH compounds, standard deviation for proficiency testing was set to 22% of the assigned values of the compounds of interest, as suggested by Thompson [15].

In this way, *z*-scores obtained for each analyzed PAH was used to assess the performance of the laboratory (i.e. analytical method) taking into account PAH-compounds separately. Usually, bar-plots of the *z*-scores grouped for each participating laboratory, have been used for visualisation of the overall performance of the laboratories to analyze simultaneously all 16 PAHs. Such bar-plots reveal common features in the *z*-scores for a laboratory (for instance, if one laboratory achieved several high *z*-scores (higher than 2), a bar-plot would easily indicate a laboratory with poor performance for these analyzed PAH compounds) [5]. Besides bar-plots, ISO 13528 [5] and ISO/IEC 17043 [6] recomend use of other graphical methods in case of multianalyte proficiency testing results, which combine performance scores for all analytes. For example, histogram type plot of *z*-scores is a suitable method, when the number of measured characteristics is small. An individual participating laboratory is identified by the position of its scores, which are used to assess the lab performance. Nevertheless, these two documents discourage application of composite or averaged performance scores (e.g. average absolute *z*-score) because they can mask poor performance on one or more analytes, also suggesting that simply the number (or percentage) of results determined to be acceptable could be used in case of multianalyte proficiency tests.

There has also been an attempt to improve well established combined *z*-scores for evaluation of the overall laboratory performance in application of multianalyte method [8]. There is a definite scarcity of the works on introducing alternatives to the existing procedures for assessment of the laboratory performance in multianalyte determination. Thus, the aim of this work is to contribute to

those scarce alternatives and to test a simple ~~alternative~~ method based on sum of ranking differences (SRD) for comparative assessment of the overall performances of laboratories in multianalyte determinations.

SRD is simple, entirely general technique suitable to order methods, models, to find their similarities and the differences among them [16]. The SRD procedure is easy to apply and it provides a unique ranking. So far, this technique (SRD) has been used in different fields (e.g. for column selection in chromatography [16], for selection of the best polarity measure for small organic molecules [17], for sensory panel testing [18-20], for comparison and ranking of QSAR/QSPR models, including selection of metric for QSAR models [21-24], for PLS model comparison in near infra-red spectroscopy [25], for testing performance for Raman spectra resolution [26], for Hansen's solubility parameters [27], for comparison of biochemical assay (Elisa veratox) and liquid chromatography in determination of mycotoxin contents [28], for comparative evaluation of acidic dissociation constants [29]. There has not been any attempt to apply it for comparison of analytical results obtained in different laboratories, including also those from interlaboratory comparisons. Here we extend the SRD procedure to evaluate laboratories according to the overall performance taking into account multianalyte results simultaneously not just evaluating the quality in measuring one individual compound.

The data reported for 15+1 EU PAHs during the ILCs organized by EU-RL-PAH were taken for testing this new technique; one of the major reasons for using these data is their availability and abundance, providing the source for SRD validation on different data sets. In this way, the SRD was also tested as a discriminant method alternative to existing average performance scores used to compare mutlianalyte PT results. SRD should be used along with the *z*-scores, and it was compared with well-known chemometric techniques, too. Additionally, the ranking was validated by Comparison of Ranks of Random Numbers (CRNN procedure), which is a kind of permutation test [16,30] and by leave-many-out cross-validation (CV) [31].The ranking made by SRD was compared to the results of principal component analysis and hierarchical cluster analysis.

2. Experimental

2.1. Data sets

Data published in Report on the 5th ILC for determination of 15+1 EU priority PAHs in edible oil [32] were used to form the input matrices: 16 PAH-compounds (samples) were enumerated in the rows, whereas laboratories (analytical methods) were arranged in columns and were coded as L1, L2, …, L13. The edible oil sample was provided by the ILC organizer and it was an olive oil spiked with 15+1 EU priority PAHs. Of all laboratories ~~included~~ in the ILC, only those (13 laboratories) that reported the results for all 16 PAHs of interest, were included in this study, since the input matrices for SRD testing should be without empty cells, which is the case when results for some PAH compounds were not reported. The laboratories were free in the selection of the test method for sample preparation and PAH determination. The reported results, corresponding *z*-scores and methods used by participating laboratories, taken from the report of the 5thILC of EU-RL-PAH, are summarized in Table 1. The percentages of acceptable results (*z*-scores less or equal to 2) are also presented in Table 1 for each selected laboratory.

Table 1

Two data sets based on the experimental results [32] were formed for testing the applicability of SRD procedure:

- "OIL" set was formed of the PAHs contents in edible oil sample reported by each participating laboratory ("reported" results presented in Table 1); the set size was 16 rows (PAH-compounds) × 13 columns (laboratories or methods);
- "OIL+As" set was in fact the "OIL" set extended with the column containing the assigned values – analyte concentrations in ILC test material (spiked edible oil sample) determined beforehand by EU-RL-PAH (i.e. calculated from gravimetric preparation data); thus, its size was 16 × 14.

Furthermore, an additional data set, so-called "Z-SCORE" set (16 × 13), was created of the absolute values of *z*-scores calculated by the ILC organizer using Eq. (1) (presented in Table 1 [32]).

2.2. Sum of ranking differences

The key step in SRD procedure is the selection of the reference for ranking, when the true (ideal, benchmark) ranking is not known [16]. Often the ranking by average values can be accepted as "ideal", since the errors cancel each other. The maximum likelihood principle will ensure that the most probable ranking will be provided by the average. The methods that deviate from the average less are ranked ahead. The best ranking is not necessarily provided by the average values, as it can be a known sequence (here the assigned values), the maximum (if comparing best classification rates) or the minimum (in the case of error rates and residuals). For the sets created in this study the following references for ranking of the laboratories ~~values~~ were chosen:

a) the assigned values of 15+1 EU PAHs contents in edible oil sample (last column in Table 1) as a reference for ranking within "OIL" set,
b) the averages of the reported results (values in μg/kg presented in Table 1) and the assigned value for each compound (row averages) as a reference for "OIL+As" set, while
c) the minimums of the absolute values of *z*-scores for each compound (presented in Table 1; row minimums) for "Z-SCORE" set.

These selection were the logical choices in order to test SRD procedure: a) ranking of the reported values on the base of the known (assigned) values would indicate laboratories that obtained multianalyte results most similar to the assigned values; b) similar indication might be expected if the assigned values would be included into the input set and then using the "overall" averages of reported and assigned results (which, by the way, could be assumed to converge to the true values), leading to the simultaneous ranking of the assigned and reported values, and finally c) ranking of the laboratories according to their absolute *z*-score values in comparison to the minimal (absolute) *z*-scores (representing the minimum deviation from the assigned value) per each compound. The absolute values of the *z*-scores would allow a direct estimation of the performance of the laboratories, but calculation of z-scores realizes a row-standardization (c.f. Eq. (1)), i.e. differences needed for ordering are destroyed by row standardization. Hence, the absolute values of *z*-scores

order the laboratories randomly, and hence they are suitable to check whether the initial assumption (normality) is valid or not.

Each (individual) laboratory was ranked and compared to the above mentioned references in following way: the ascending reference values of PAH concentrations were ordered giving them consecutive numbers from 1 to 16 (this is so-called “reference (benchmark) ranking”). Then, ranking of data within each column (i.e. ranking of the results of each laboratory) was made (so-called “individual ranking”); the absolute values of the differences between the reference and the individual rankings for all compounds were calculated and summed for each laboratory. In this way, the sum of (absolute) ranking differences, SRD values, were calculated for each laboratory. The closer is the SRD value to zero (i.e. the closer is the sum of differences of individual ranking to the reference one) the better is the analytical method for simultaneous determination of all analytes. The proximity of SRD values shows that the methods used by the laboratories have similar (overall) performance in the multianalyte (PAHs) determination. Equal concentrations (so-called ties) to two digits received the same rank number during the ranking procedure.

2.3. Validation

Two types of validations have been carried out (i) comparison of ranks by random numbers (CRRN), which is in fact a randomization test [16,30], and (ii) leave-many-out (seven folds) CV followed the literature recommendation [31]. Namely, (i) CRRN procedure includes the determination of the theoretical distribution for ranking using solely random numbers and the distribution is compared to the actual rankings; (ii) during the seven folds CV (approximately) 1/7 of the objects were left out and the ranking was made on the remaining 6/7th number of objects just seven times. The different rankings provided uncertainties for the SRD values.

2.4. Exploratory statistics

In the exploratory phase box and whisker plots were used to graphically present numerical data like *z*-scores and cross-validated SRD values, while hierarchical cluster analysis (HCA) and principal component analysis (PCA) were applied on the above sets in order to observe the similarity and dissimilarity of laboratories (methods), to analyze quantitatively the relationships among the results of laboratories (i.e. their analytical efficiency) and to compare these results with the SRD ones. Mean centering and scaled to unit standard deviation were applied as data preprocessing step before principal component- and hierarchical cluster analysis. Standard procedures were applied (Statistica™, version 7.0, StatSoft Inc., Tulsa Oklahoma, USA).

3. Results and discussion

3.1. Exploratory statistics

PCA shows (Figure 1) the grouping of the laboratories within the "OIL+As" set (thus, grouping relative to–the assigned values similar grouping can be observed for absolute values of *z*-scores ("*Z*-SCORES" set). Figures 1a and 1b show the loading plots of two main PCs retained in both cases that accounted similar share of the total data variance (~70%). The L5 was by far the most outlying laboratory when the reported values were compared to the assigned (Figure1a); there were few more points (L2, L6, L7, L8, L9, L12) diverging from the central cluster comprising of the laboratories (L1, L3, L4, L10, L11, L13) closest to the assigned value. The score plot for the *z*-scores (Figure 1b) also pointed out L5 as an outlier and similarities among L1, L3, L4, L10, L11, and L13.

Figures 1a and 1b

The dendrogram of Figure 2 indicates clustering of the laboratories similarly to the PCA groupings. Laboratories L6 and L9 and particularly L5 reported the most dissimilar results to those reported by the other labs (Figures 2a and 2b) and also to the assigned values determined by the proficiency testing provider, EU-RL-PAH (Figure 2a).

Figures 2a and 2b

The results reported by laboratories L1, L3, L4, L10, L11 and L13 form a dense cluster (the assigned values also belong to this cluster on Figure 1a). The same pattern can be observed on the PCA plots (Figures 1a and 1b).

Z-scores of these six laboratories were all below 2, while the rest of laboratories had at least one *z*-score (its absolute value) higher than 2, indicating questionable ($2 < |z| \leq 3$) or unsatisfactory ($|z| > 3$) performances for one (or more) particular PAH compound(s). Box and whisker plots of the absolute values of *z*-scores of the laboratories are given in Figure 3. The outlying laboratory L5 could be easily seen in Figure 3a; after its exclusion (Figure 3b) the laboratories might be ordered according to the median absolute values of *z*-scores as follows (median absolute values of *z*-scores are given in parentheses): L4 (0.25) ~ L11 (0.265) < L3 (0.355) ~ L10 (0.37) < L1 (0.39) < L13 (0.455) < L9 (0.675) ~ L12 (0.685) < L2 (0.715) <L8 (0.885) < L7 (1.03) < L6 (2.41). Apart from L5, the highest standard deviations (SD) of the absolute values of *z*-scores were observed for L6 and L9 (SD for both laboratories SD = 1.26), while for others, the SDs were in the range from 0.25 (for L13) to 0.78 (for L8).

Figures 3a and 3b

3.2. Sum of ranking differences

The SRDs calculated for “OIL” and “OIL+As” data sets can be seen in Figure ~~4~~Table 2. Similarities (i.e. groupings) of laboratories can also be observed, as well as their dissimilarities from the ordering point of view, i.e. SRD can also be considered as a dissimilarity measure (the higher its value, the more dissimilar to the reference value) [16,30]. Thus, the best ranked laboratories according to the lowest SRD values in “OIL” and “OIL+As” sets appeared to be L2 and L3 (~~Table 2~~Figure 4); they showed the best overall performance in simultaneous determination of 15+1 EU PAHs.

Figure 4

It is interesting to note that both laboratories differed from others by using the sample preparation method based on size-exclusion chromatography (gel permeation chromatography) followed by high performance liquid chromatography with fluorescence detection (Table 1). It could also be seen that proximity of the SRD values indicates similar performances in analyzing 15+1 EU priority PAHs among majority of the laboratories (ten laboratories in "OIL" set had SRD between 3-11, while in "OIL+As", SRDs of eleven laboratories ranged from 8 to 14). Three laboratories (L9, L6 and L5) had distinguishable higher SRDs (~~Table 2~~Figure 4) as a consequence of significantly lower performances in analyzing 15+1 EU PAHs. The L5 was the worst ranked laboratory in "OIL" and "OIL+As" sets; it should be noted that only this lab used method for determination of PAHs based on liquid-liquid/solid-phase extraction followed by gas chromatography coupled to mass spectrometry. The best two laboratories (L2, L3) are somewhat better than the assigned values (L10 is equivalent) if accepting the ~~mean~~ average value as reference for ranking within the "OIL+As" set. The ranking of laboratories in these two sets, other than those ranked as "the best" and "the worst", was slightly different.

Even though L2 had one *z*-score (its absolute value) slightly higher than 2 ($z = 2.02$) it was ranked exactly on the same way as L3, indicating that SRD procedure might conceal one result very close to the questionable performance, but it clearly depicts the laboratories with the poorest performances (outlier).

In order to check the influence of the outlier on the ranking in "OIL+As" set, the SRD procedure was also applied on the set without L5 (so-called "OIL+As-OUT" set) and the resulting SRDs (calculated on the base of the averages used for the reference ranking) are also presented in ~~Table 2~~Figure 4. The rationale behind this lies in fact that the average values selected for the reference ranking in "OIL+As" set were directly affected by the all input values (including the outlier), contrary to the reference chosen for "OIL" set (i.e. the assigned values cannot be influenced by the presence of outlier). Removing the outlier ("OIL+As-OUT" set, using averages as the reference) caused slightly less SRDs for L8 and L10 (Figure 4). An alternative would be the selection of

median (or other robust measure) instead of the averages triangles in Figure 4). Interestingly L13 is ranked first (slightly better than the assigned values), which exhibits the smallest range on Figure 3b. Other patterns are mostly similar to the remaining rankings of-Figure 4. From the comparison of SRD rankings Figure 4 it could be concluded that median is the best choice. Figure 4 also contains the normalized sum of squared *z*-scores (SZ2norm, a Euclidean distance) suggested as the most optimal overall performance indicator by Medina-Pastor et al. [8]. All indicators in Figure 4 were placed on the same scale between 0 and 100. As a non-robust measure, SZ2norm is sensitive to the outlying observation most. Almost all variability in the data (>94%) is carried by the L5 outlier. Any variants of SRD ranking are robust and allowed observing differences in other laboratories as well (on the expense of the outlying L5).

The SRDs for laboratories were scaled between 0 and 100 (Figures 5a and 5b) in order to be comparable among each other [16]. It could easily be seen that the location of the scaled values for majority of laboratories was far from the SRDs of random numbers in the case of "OIL" and "OIL+As" sets (Figures 5a and 5b, respectively), showing that their ranking was far from being random. The L5 was the worst ranked laboratory in "OIL" and "OIL+As" sets; its SRD value in both sets was close to the first icosaile (5%).

Figures 5a and 5b

The SRDs calculated for the "Z-SCORE" set were quite different than those obtained for the reported values (i.e. for "OIL", "OIL+As", and "OIL+As-OUT" sets), as expected, because row-standardization eliminates the differences needed for ordering. However, the overlapping with normal distribution for the *z*-scores can easily be seen on Figure 6. All SRD values for the "Z-SCORES" set overlapped with random distribution, except for L3 (Figure 6), which was also located very close to the first icosaile, indicating that ordering of labs based on the absolute *z*-scores for all compounds is not better than the random ordering (ordering of random numbers). In order to check this observation, the SRD with CRNN procedure was also used on absolute values of *z*-scores calculated for 24 laboratories participating in the 7thILC on PAHs in edible oil [33] and for 14

laboratories during the 4thILC on PAHs in fish [34], and, again, SRDs overlapped with the random distribution (data not shown). These observations can be considered as a proof that the errors of labs (i.e. the deviation of their results from the assigned values, not the individual PAH concentrations) expressed as *z*-scores follow a normal distribution.

Figures 6a and 6b

To reveal uncertainties for SRD, cross-validation (seven-fold CV [30,31]) has been carried out. Box and whisker plots clearly exhibit the difference between classical (statistical) and present (SRD) approach (Figures 7a and 7b, respectively). Figure 7a allows observing one outlying laboratory (L5) nothing else, whereas seven-fold CV of SRD values allow us to group the laboratories similarly to Figures 2a, 2b and 5a.

Figures 7a and 7b

Figure 7b shows the same pattern as Figure 5a with subtle, negligible differences suggesting that cross-validation does not change the ranking of laboratories just helps in grouping them.

Comparing the results of PCA, HCA, SRD and CV-SRD, shows the very same (or almost the same) clustering pattern. Moreover, CV-SRD reveals the uncertainties in the ranking and clustering. Sign test or Wilcoxon's matched pair test is suitable to decide about the significance of CV-SRD grouping.

4. Conclusions

Sum of Ranking Differences methodology (SRD) is a simple technique general in nature that can be ~~used as~~ applied to any experimental problem in which the multianalyte results obtained either by several analytical procedures, analysts, instruments, or laboratories need to be compared. Besides the *z*-scores, the most commonly used PT performance statistics that assess the results of each analyte separately, SRD could be regarded as an alternative way for ranking of measurement methods and laboratories involved in interlaboratory comparison tests according to their multianalyte results. SRD provides similar groupings as classical techniques (principal component

and hierarchical cluster analysis) and it is more influential than the (normalized) sum of the squared *z*-scores.

The overall bias covering simultaneously the results on the whole group of targeted analytes is taken into account (the bias follows normal distribution). SRD takes the disadvantages of the earlier evaluation methods out (e.g. the discrepancies in ranking for individual compounds).

SRD proved to be a useful tool in choosing the analytical methods or the laboratories with the best overall performances in multianalyte determinations. An unambiguous selection of the laboratory(ies) or analytical methods could be made that produce results the most similar to the assigned values, if comparison of the overall (multianalyte) performances of laboratories participating in PT programs is made. SRD could point out the method(s) that produce(s) the best results with respect to the overall averages (or medians), if the comparison of several multianalyte methods should be taken. Similarly, the laboratories with the most extreme results could be easily pointed out in any of the above two cases. Additionally, grouping of laboratories with similar overall (multianalyte) performances can be obtained in similar manner by multivariate techniques such as principal component analysis and hierarchical cluster analysis.

Acknowledgement

The results presented here are the part of project no. 172050 "Development and application of the advanced chromatographic and spectrometric methods in the analysis of xenobiotics and their degradation pathways in biotic and abiotic matrices", coordinated by Prof. B. Škrbić and supported by the Ministry of Education and Science of the Republic of Serbia, as well of the bilateral project of the Hungarian-Serbian Intergovernmental S&T Cooperation Programfor 2010–2011"Comparison of various analytical and chemometric methods" funded by the Ministry of Education and Science of the Republic of Serbia and National Innovation Office of Hungary.

References:

[1] W. Horwitz, T. Jacksin, S.J. Chirtel, J. AOAC Int. 84 (2001) 919-935.

[2] W. J. Youden, Statistical Manual of the Association of Official Analytical Chemists: Statistical Techniques for Collaborative Test, AOAC International, Gaithersburg, 1975.

[3] E. Hund, D. Luc Massart, J. Smeyers-Verbeke, Anal. Chim. Acta 423 (2000) 145–165.

[4] M. Thompson, R. Wood, J. AOAC Int. 76 (1993) 926–940.

[5] ISO 13528 – Statistical methods for use in proficiency testing by interlaboratory comparisons,

[6] ISO/IEC 17043 – Conformity assessment – General requirements for proficiency testing, 2010.

[7] M. Ricci, O. Bercaru, R. Morabito, C. Brunori, I. Ipolyi, C. Pellegrino, A. Sahuquillo, F. Ulberth, TrAC - Trends Anal. Chem. 26 (2007) 818-827.

[8] P. Medina-Pastor, M. Mezcua, C. Rodriguez-Torreblanca, A.R. Fernandez-Alba, Anal. Bioanal. Chem. 397 (2010) 3061-3070.

[9] COMMISSION REGULATION (EC) No 1881/2006 of 19 December 2006 setting maximum levels for certain contaminants in foodstuffs, Official Journal of the European Union 364, 20.12.2006.

[10] COMMISSION REGULATION (EU) No 835/2011 of 19 August 2011 amending Regulation (EC) No 1881/2006 as regards maximum levels for polycyclic aromatic hydrocarbons in foodstuffs, Official Journal of the European Union L 215, 20.8.2011.

[11] COMMISSION REGULATION (EC) No 333/2007 of 28 March 2007 laying down the methods of sampling and analysis for hte official control of the levels of lead, cadmium, mercury, inorganic tin, 3–MCPD and benzo(a)pyrene in foodstuffs. Official Journal of the European Union L 88, 29.3.2007.

[12] COMMISSION REGULATION (EU) No 836/2011 of 19 August 2011 amending Regulation (EC) No 333/2007 laying down the methods of sampling and analysis for the official control of the levels of lead, cadmium, mercury, inorganic tin, 3-MCPD and benzo(a)pyrene in foodstuffs, Official Journal of the European Union L 215, 20.8.2011

[13] ISO/IEC Guide 43-1:1997 (E), Proficiency testing by interlaboratory comparisons: Part 1: Development and operation of proficiency testing schemes

[14] M. Thompson, S.L.R. Ellison, R. Wood, Pure Appl. Chem. 78 (2006) 145–196.

[15] M. Thompson, Analyst 125 (2000) 385–386.

[16] K. Héberger, TrAC - Trends Anal Chem, 29(2010)101–109.

[17] K. Héberger, I.G. Zenkevich, J. Chromatogr. A, 1217 (2010) 2895–2902.

[18] K. Kollár-Hunek, J. Heszberger, Z. Kókai, M. Láng-Lázi, E. Papp, J. Chemometr. 22 (2008) 218–226.

[19] L. Sipos, Z.Kovács,D. Szöllösi,Z. Kókai, I.Dalmádi; A. Fekete, J. Chemometr. 25 (2011) 275–286.

[20] V. Losó, A. Tóth, A. Gere, J. Heszberger, G. Székely, Z. Kókai, L. Sipos, Acta Alimentaria 41 (2012) 109–119 .

[21] Z. Garkani-Nejad,M. Ahmadvand, Chromatographia 73 (2011) 733–742.

[22] X. Liu, Y. Ren, P.Zhou, Z. Shang, J. Mol.Struct. 995 (2011)163–172.

[23] K. Héberger, B. Škrbić, Anal. Chim. Acta 716 (2012) 92–100.

[24] K. Roy, I. Mitra, P. K. Ojha, S. Kar, R. N. Das, H. Kabir, Chemometr. Intell. Lab. Syst. 118 (2012) 200–210.

[25] A. A. Gowen, G. Downey, C. Esquerre, C. P. O'Donnell, J. Chemometr. 25 (2011) 375–381.

[26] B. Vajna, A. Farkas, H. Pataki, Z. Zsigmond, T. Igricz, G. Marosi, Anal. Chim. Acta, 712 (2012) 45–55.

[27] K. Bielicka-Daszkiewicz, A. Voelkel, K. Héberger, M. Pietrzyńska, J. Chromatogr. A 1217 (2010) 5564-5570.

[28] E. K. Tangni, J.–C. Motte, A. Callebaut, A. Chandelier, M. De Schrijver, L. Marnix, L. Pussemier, Mycotoxin Res. 27 (2011)105–113.

[29] G. T. Balogh, A. Tarcsay, G. M. Keserű, J. Pharm. Biomed. Anal., 67-68 (2012) 63–70.

[30] K. Héberger, K. Kollár-Hunek, J. Chemometr. 25 (2011) 151–158.

[31] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, New York, 2009.

[32] D. Lerda, L. Hollósi, P. Lopez, S. Szilágyi, T. Wenzl, Report on the 5th interlaboratory comparison test organized by the European Union Reference Laboratory for Polycyclic Aromatic Hydrocarbons-15+1 EU priority PAHs in edible oil and acetonitrile, JRC Scientific and Technical Reports, European Commission, Joint Research Centre, Institute for Reference Materials and Measurements, 2010.

[33] D. Lerda, P. L. Sanchez, S. Szilágyi, P. Verlinde, T. Wenzl, Report on the 7th inter-laboratory comparison test organised by the European Union Reference Laboratory for Polycyclic Aromatic Hydrocarbons “15+1 EU priority PAHs in spiked olive oil and solvent solution”,JRC Scientific and Technical Reports, European Commission, Joint Research Centre, Institute for Reference Materials and Measurements,2011.

[34] D. Lerda, L. Hollósi, P. Lopez, S. Szilágyi, T. Wenzl, Report on the 4th interlaboratory comparison test organised by the Community Reference Laboratory for Polycyclic Aromatic Hydrocarbons “15+1 EU Priority PAHs in fish and acetonitrile”,JRC Scientific and Technical Reports, European Commission, Joint Research Centre, Institute for Reference Materials and Measurements, 2010.

.

514 Table 1 Summary of data (reported results in µg/kg, *z*-scores and analytical methods) selected from the report of the 5thILC on PAHs in edible oil [32]

515 used for checking the applicability of SRDs.

	L1		L2		L3		L4		L5		L6		L7		L8		L9		L10		L11		L12		L13		As.
	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg	*z*	µg/kg
5MC	1.20	0.35	1.61	2.02	1.11	0	1.01	-0.43	1.28	0.67	1.72	2.47	0.87	-1	1.60	1.98	1.50	1.57	1.30	0.75	1.13	0.06	1.30	0.75	0.98	-0.55	1.1
BAA	2.40	0.04	2.75	0.7	2.05	-0.63	2.48	0.19	2.81	0.82	1.57	-1.55	2.13	-0.48	2.60	0.42	3.80	2.71	2.30	-0.15	2.72	0.65	2.70	0.61	2.13	-0.48	2.4
BAP	2.90	-0.13	3.43	0.73	3.28	0.48	3.06	0.13	4.22	2.01	2.61	-0.61	2.40	-0.95	3.30	0.52	2.90	-0.13	2.80	-0.3	3.13	0.24	3.20	0.35	2.72	-0.43	3.0
BBF	5.20	-0.2	6.04	0.5	5.44	0	5.52	0.06	4.95	-0.41	4.64	-0.67	4.20	-1.04	7.00	1.3	6.20	0.63	5.40	-0.04	5.66	0.18	4.70	-0.62	5.16	-0.24	5.4
BCL	2.20	1.01	1.90	0.25	1.89	0.23	1.91	0.28	2.89	2.75	1.99	0.48	1.40	-1.01	2.00	0.5	1.80	0	1.30	-1.26	1.83	0.07	2.10	0.76	1.60	-0.51	1.8
BGP	6.10	-0.04	6.44	0.21	6.71	0.41	6.44	0.21	8.96	2.07	6.52	0.27	5.80	-0.26	6.80	0.48	5.50	-0.48	5.80	-0.26	6.58	0.31	6.30	0.11	5.97	-0.14	6.2
BJF	1.40	-0.07	1.78	1.15	1.78	1.15	1.49	0.22	11.73	32.99	1.40	-0.1	1.47	0.16	0.50	-2.95	1.90	1.53	1.70	0.89	1.01	-1.31	1.80	1.21	1.53	0.35	1.4
BKF	8.20	-0.03	10.08	1.01	8.65	0.22	8.23	-0.01	1.66	-3.63	8.53	0.16	6.00	-1.24	9.90	0.91	8.30	0.03	8.80	0.31	8.61	0.2	9.20	0.53	7.96	-0.16	8.2
CHR	3.70	0.45	4.12	1.02	3.60	0.31	3.58	0.29	4.21	1.14	4.17	1.08	3.33	-0.05	4.30	1.26	6.60	4.36	3.30	-0.09	3.87	0.68	4.40	1.39	3.28	-0.12	3.4
CPP	8.60	0.55	8.98	0.78	8.07	0.24	8.28	0.36	2.84	-2.86	11.00	1.98	5.33	-1.39	8.70	0.61	13.20	3.28	6.20	-0.87	7.17	-0.29	6.80	-0.51	6.20	-0.87	7.7
DEP	0.80	-0.97	1.02	0.02	1.03	0.06	0.78	-1.06	1.83	3.64	0.49	-2.35	1.00	-0.07	0.90	-0.52	0.80	-0.97	0.80	-0.97	0.78	-1.06	1.20	0.82	0.85	-0.75	1.0
DHA	4.90	1.33	5.17	1.65	5.08	1.54	4.83	1.24	4.98	1.42	8.00	5.05	4.00	0.25	5.70	2.29	4.00	0.25	4.40	0.73	5.05	1.51	4.80	1.21	4.52	0.87	3.8
DHP	2.10	-0.67	2.83	0.66	2.78	0.57	2.23	-0.44	4.60	3.92	2.20	-0.5	1.87	-1.1	3.00	0.98	2.80	0.61	2.10	-0.68	1.95	-0.96	4.20	3.18	2.11	-0.66	2.5
DIP	9.10	-0.3	10.69	0.44	10.60	0.4	9.31	-0.21	33.59	11.11	11.92	1.01	6.67	-1.44	10.30	0.26	11.20	0.68	10.30	0.26	9.41	-0.16	11.80	0.96	9.30	-0.21	9.8
DLP	1.60	0.43	1.77	0.95	1.51	0.15	1.13	-1.03	15.81	44.57	1.80	1.05	1.20	-0.82	1.70	0.74	1.70	0.74	1.60	0.43	1.41	-0.16	1.60	0.43	1.37	-0.29	1.5
ICP	3.40	-0.45	4.27	0.6	4.34	0.68	3.82	0.05	4.61	1.01	1.97	-2.17	3.53	-0.29	3.50	-0.33	3.40	-0.45	3.80	0.03	3.81	0.04	4.10	0.39	3.35	-0.51	3.8
% of acceptable results (|*z*|≤2)	100		94		100		100		31		75		100		88		81		100		100		94		100		
Method used for PAHs analysis	SAP + LLE + GC-MS		SEC + LC-FLU		SEC + LC-FLU		SPE + GC-MS/MS		LLE + SPE + GC-MS		SAP + LLE + LC-FLU		LLE + SEC + GC-MS		SEC + LC-MS		SAP + SPE + GC-MS		SAP + SPE + GC-MS		SAP + LLE + GC-MS		SPE + SPE + GC-MS/MS		LLE + SEC + GC-MS		

516 SAP: saponification; LLE: liquid-liquid extraction; SEC: size exclusion chroatography; SPE: solid phase extraction; GC-MS: gas chromatography with mass spectrometry; LC-FLU:
517 liquid chromatography with fluorescence detection; GC-MS/MS: gas chromatography with tandem mass spectrometry; LC-MS: liquid chromatography with mass spectrometry

Figure captions

Figure 1

PCA score plots PC1 *vs*. PC2 for the sets consisted of (a) the reported and assigned values ("OIL+As"), and (b) the absolute values of *z*-scores ("Z-SCORE") for the laboratories (methods) (L1, L2, ..., L13)

Figure 2

The dendrogram of the laboratories according to a) the reported and assigned values ("OIL+As" data set), and b) the absolute values of *z*-scores ("Z-SCORE" set)

Figure 3

Box and whisker plots of the absolute *z*-scores calculated for the laboratories (methods) (L1, L2, …, L13) a) all 13 laboratories included in the "Z-SCORE" formed in this study, b) after excluding L5 as an outlier.

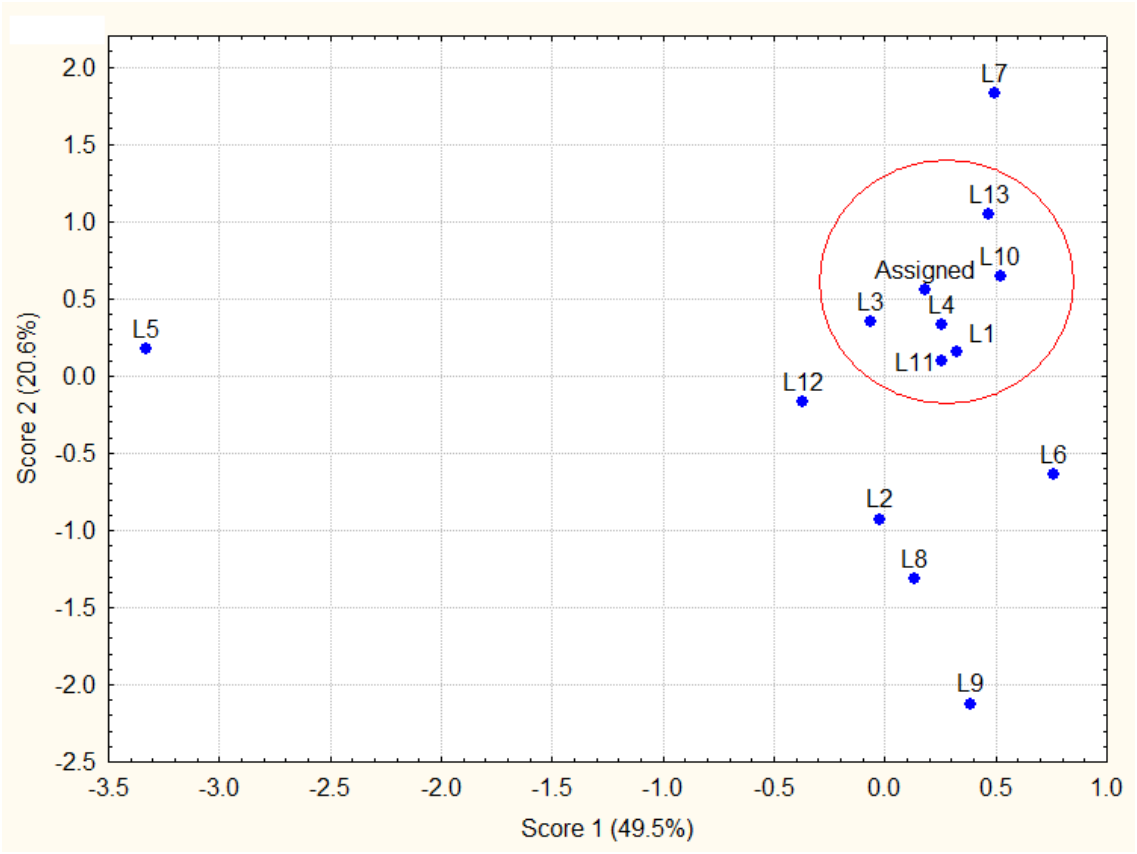
Figure 4

Line plots for SRD rankings: "OIL" set, reference: assigned value (full circles, blue); "OIL+As" set, reference: averages (full boxes, red); "OIL+As-OUT" set, reference: averages (full rhombuses, green); "OIL+As", reference: medians (full triangles, pink); normalized sum of squared *z*-scores, SZ2norm (black full circles, dotted line)

Figure 5

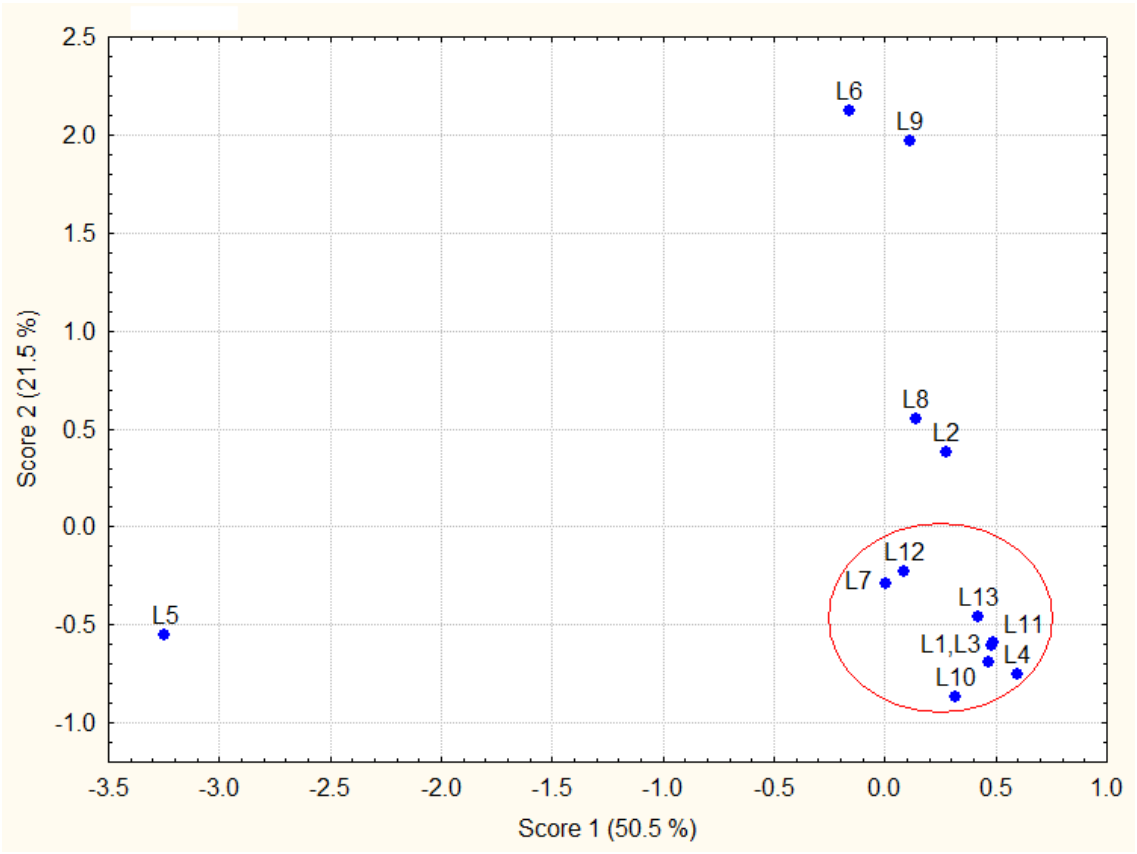
SRD ranking with CRNN validation of 13 laboratories for a) "OIL" set, b) "OIL+As" set. The Y left-hand side-axis and X-axis are SRD values scaled between 0 and 100. The Y right-hand side-axis represents relative frequencies of the theoretical distribution for ranking random numbers.

543 Statistical characteristics of this distribution (CRRN procedure) are defined by the first icosaile
544 (5%), XX1; the first quartile, Q1; median, Med; the last quartile, Q3; the last icosaile (95%), XX19.
545
546 Figure 6
547 SRD ranking with CRNN validation of 13 laboratories according to the absolute values of *z*-scores
548 ("*Z*-SCORE" set) calculated according the contents of EU 15+1 PAHs reported during the 5thILC
549 on PAHs organized by IRMM, Geel, Belgium [32]. The Y left-hand side-axis and X-axis are SRD
550 values scaled between 0 and 100. The Y right-hand side-axis represents relative frequencies of the
551 theoretical distribution for ranking random numbers. Statistical characteristics of this distribution
552 (CRRN procedure) are defined by the first icosaile (5%), XX1; the first quartile, Q1; median, Med;
553 the last quartile, Q3; the last icosaile (95%), XX19.
554
555 Figure 7
556 Box and whisker plot of the original PAH concentrations (a); box and whisker plot of sum of
557 ranking difference values obtained from a seven segments cross-validation (b).
558



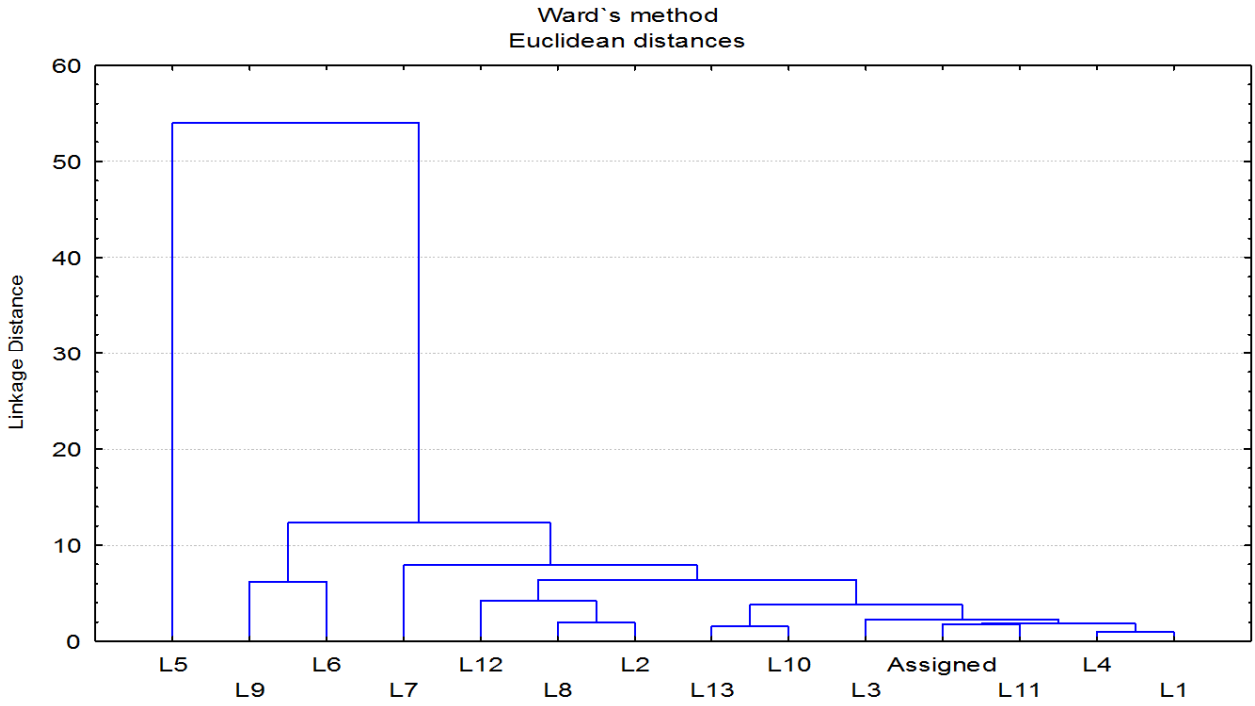
559

560 Figure 1a



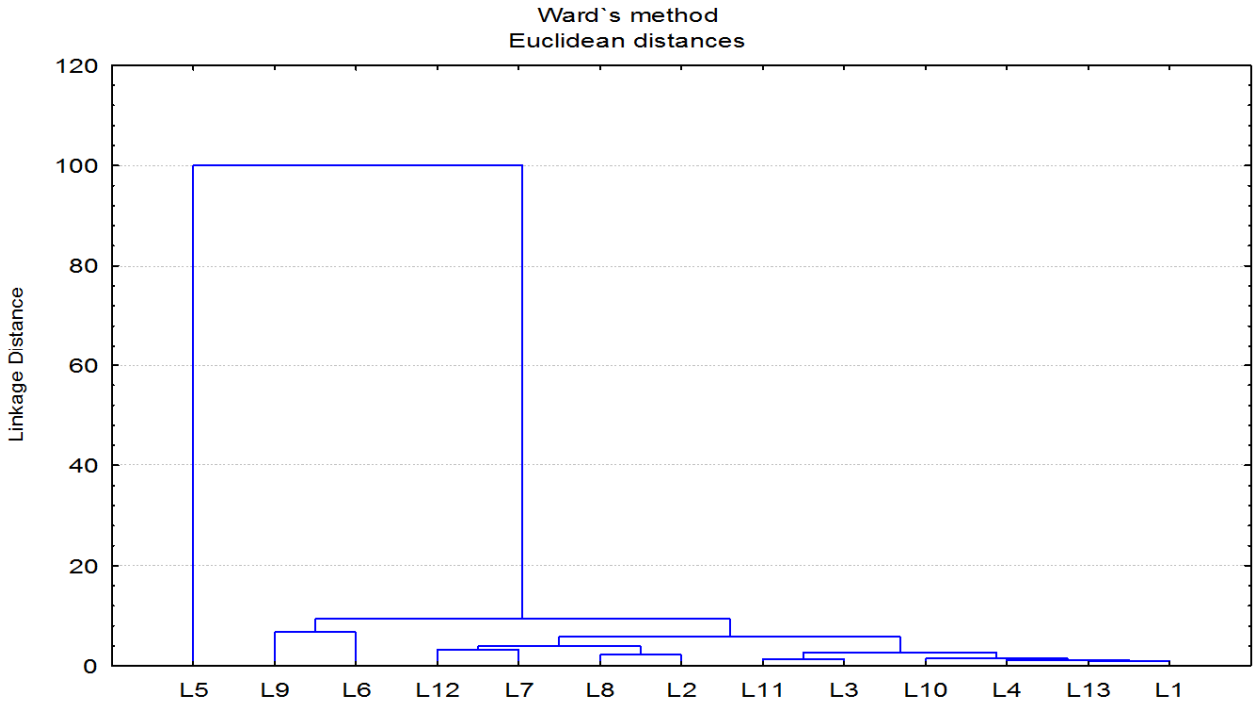
561

562 Figure 1b



563

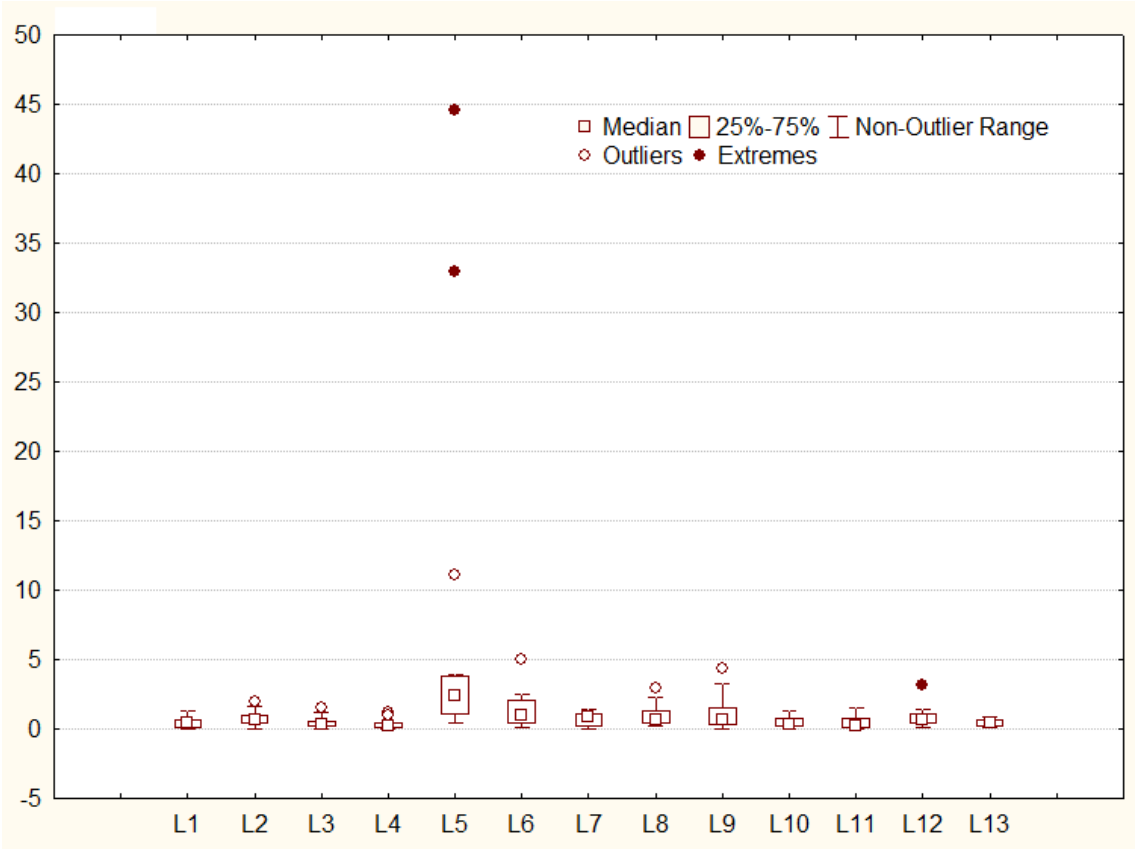
564 Figure 2a



565

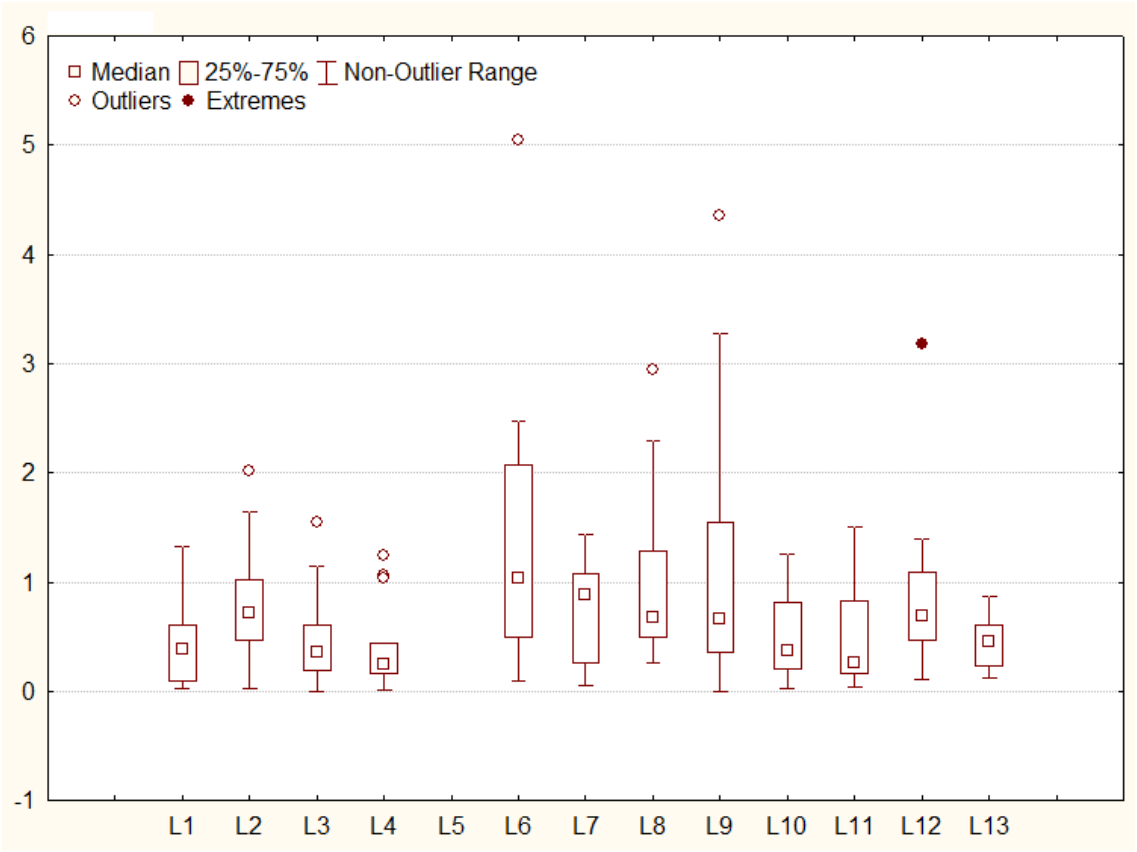
566 Figure 2b

567



568

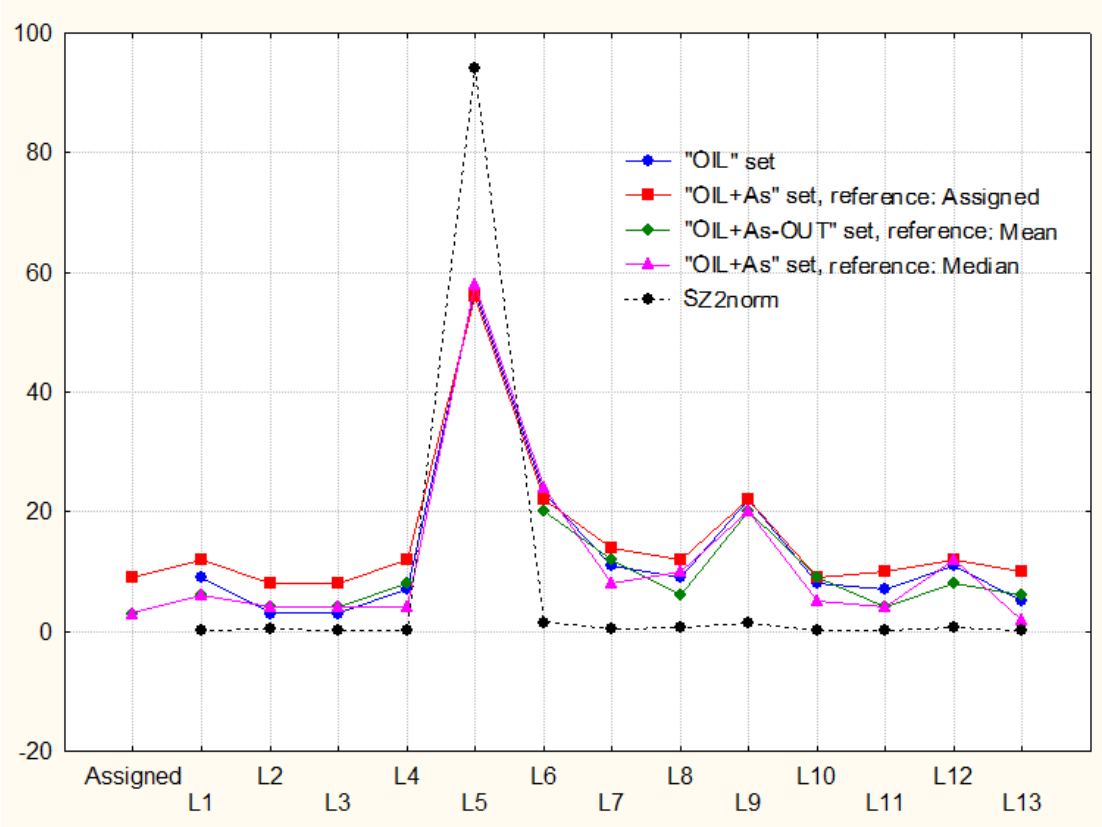
569 Figure 3a



570

571 Figure 3b

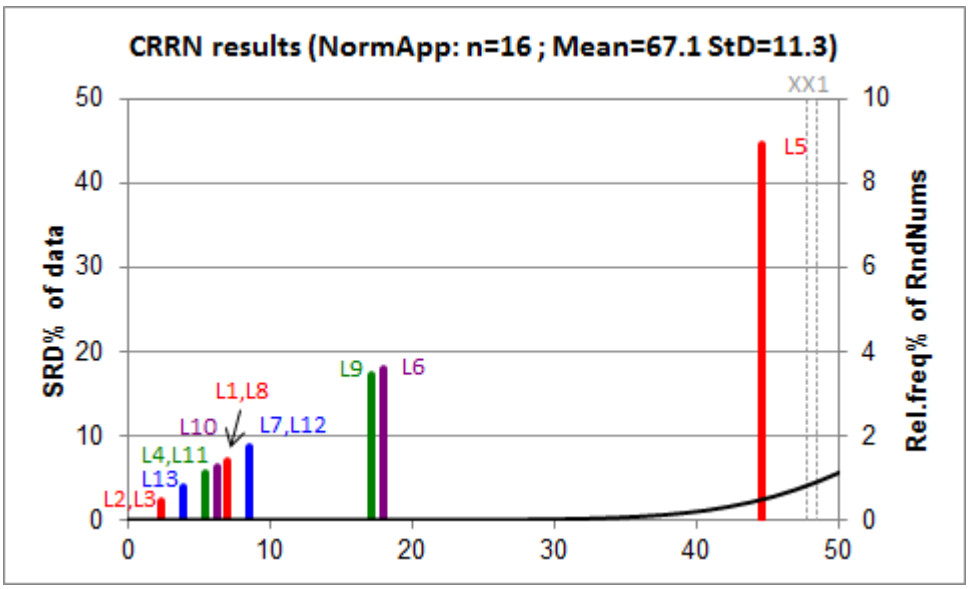
572



573

574 Figure 4

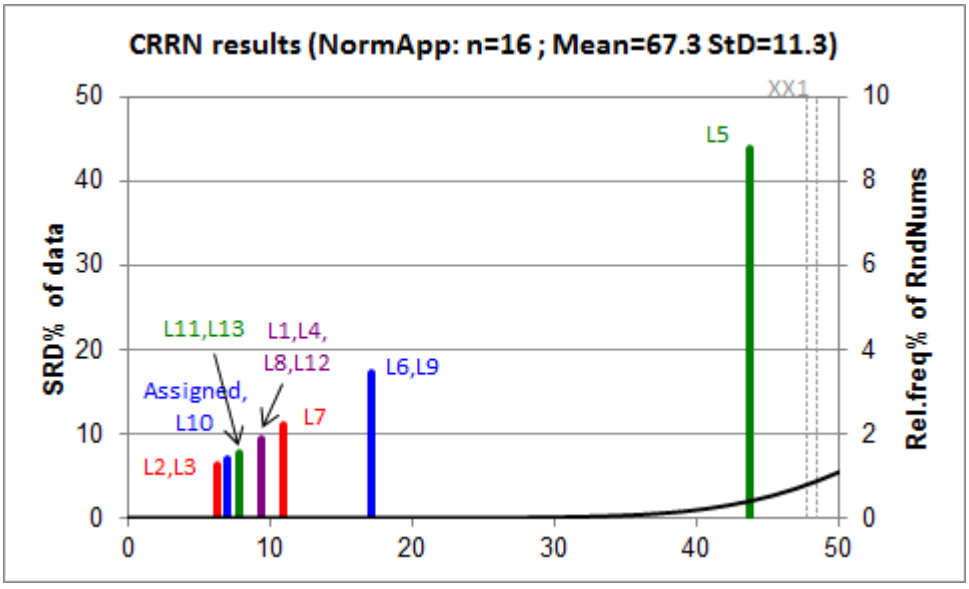
575



576

577 Figure 5a

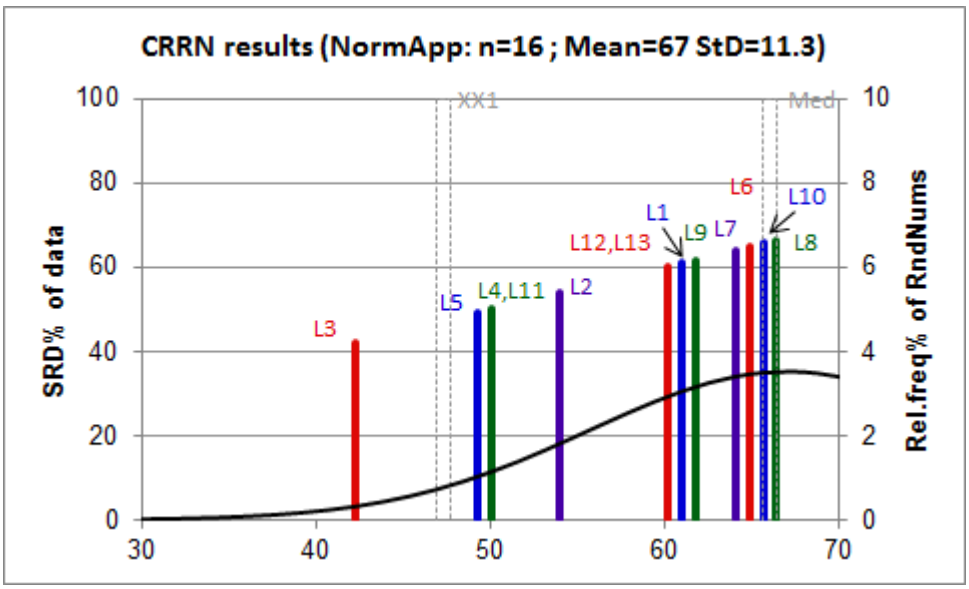
578



579

580 Figure 5b

581



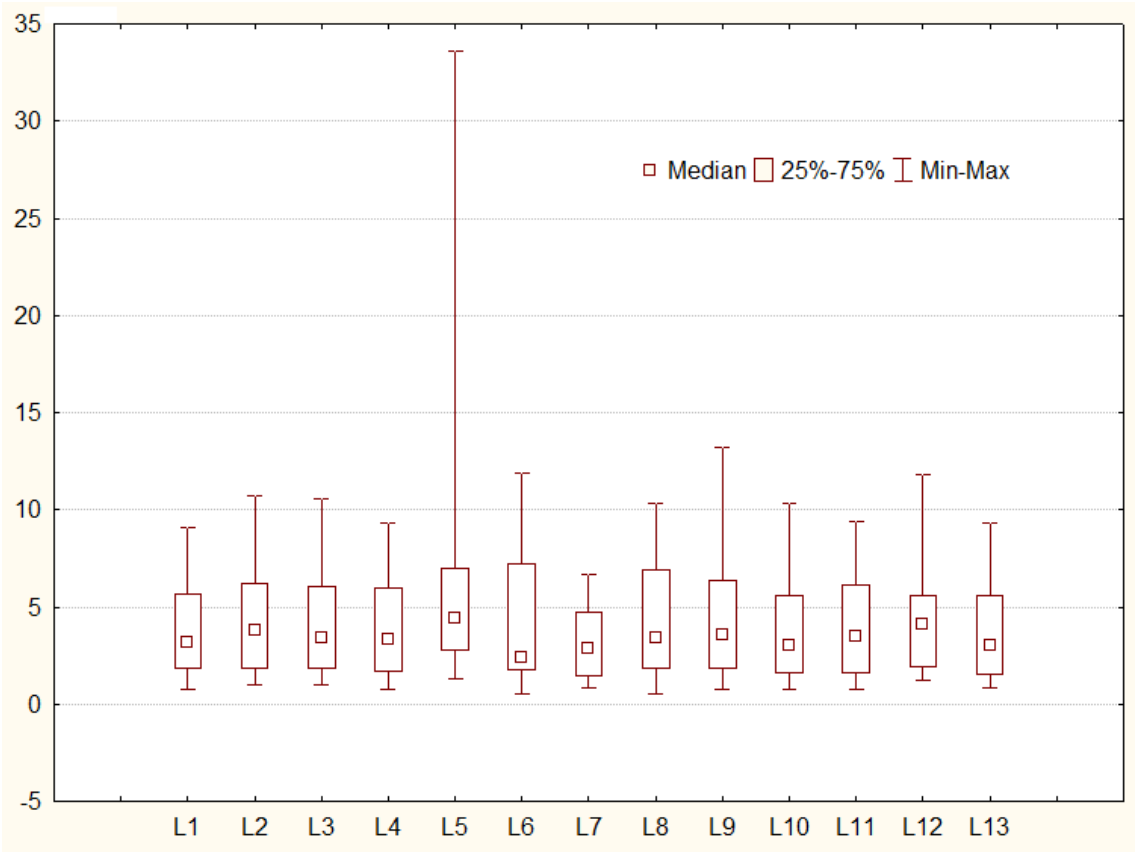
582

583 Figure 6

584

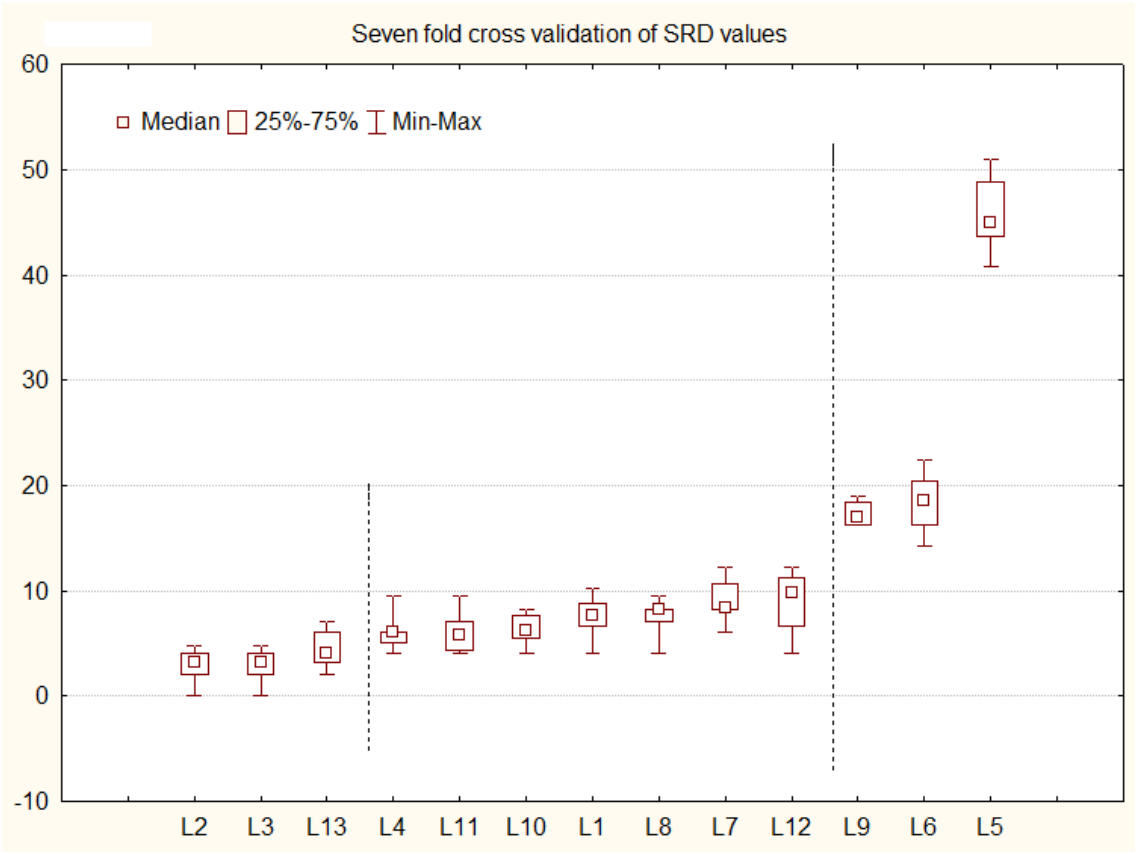
585

586



587

588 Figure 7a



589

590 Figure 7b