12th IMEKO TC1 & TC7 Joint Symposium on
Man Science & Measurement
September, 3 – 5, 2008, Annecy, France

MANAGEMENT OF SUBJECTIVE EVALUATIONS REPRESENTED BY ORDINAL SCALES

Giulio Barbato [1], *Stefano Farné* [1], *Gianfranco Genta* [1]

[1] Politecnico di Torino, Department of Production Systems and Business Economics,
Corso Duca degli Abruzzi 24, 10129 Torino, Italy, gianfranco.genta@polito.it

Abstract: Common statistical methods, like arithmetic average, are, in principle, not adequate for ordinal scale evaluations, but in practice are frequently used. Specific methods are necessary, more complicated and with limits due to coarse resolution. This paper indicates some ways for overcoming resolution problems, allowing to compare specific and common analysis results. Thereafter it was possible to consider chi-square and skewness as indicators of the possibility of safely using common methods on datasets produced by non-linear metrics.

Keywords: ordinal scale, OWA, Condorcet method.

1. INTRODUCTION

Subjective evaluations are quite common in many fields of science: social sciences require frequently subjective evaluations; politics involves elections [1-2], a very important example of subjective evaluations. Psychology defines many scales of characteristics that are not based on objective measurements. Not to forget the large amount of evaluations obtained by questionnaires and interviews and even scores given to students. Also in the field of engineering a number of decisions are taken with the help of subjective evaluations, for instance, to choose which “click” of an electromechanical switch sounds better, or which “slam” of a car door conveys a higher feeling of reliability.

Technical uses of subjective evaluations are so common and important to be widely present in the technical literature, arousing a deep debate on the limits of management of subjective evaluations. Every one can agree that subjective evaluations are frequently given by means of linguistic quantifiers: on-off condition, as *good* - *no good* or *reject* - *accept*, or ordinal scales, as *inadequate* - *sufficient* - *good* - *very good* - *excellent*. In the representational theory, the concept of *measurement*, originally defined by Ernest Nagel [3] as "the correlation of numbers with entities that are not numbers", was broadened by Stanley S. Stevens [4]. He defined types of evaluations to include nominal, ordinal, interval, and ratio levels. In practice, this scheme is used mainly in the social sciences but even there its use is controversial because it includes definitions that do not meet the more strict requirements of the classical theory and additive conjoint measurement. However, the classifications of *interval* and *ratio* level measurement are not controversial. In particular, ordinal scales includes variables that can be ordered but for which the difference between successive levels cannot be considered as equal and zero point cannot be defined. For example: preference ranks (Thurstone rating scale), Mohs hardness scale, hotel ratings, shirt sizes (S,M,L,XL), and grades for academic performance (A, B, C, ...). Also includes the Likert scale used in surveys (strongly agree, agree, undecided, disagree, strongly disagree). It is easy to notice that distances between each ordered category are not necessarily the same (a four star hotel isn't necessarily just "twice" as good as a two star hotel). There are different levels of measurement that involve different properties (relations and operations) of the numbers or symbols that constitute the measurements. Associated with each level of measurement is a set of permissible transformations. For ordinal scale, the most commonly discussed levels of measurement are as follows:

- things are assigned numbers such that the order of the numbers reflects an order relation defined on the attribute. Two things x and y with attribute values $a(x)$ and $a(y)$ are assigned numbers $m(x)$ and $m(y)$ such that if $m(x) > m(y)$, then $a(x) > a(y)$;
- permissible transformations are any monotone increasing transformation, although a transformation that is not strictly increasing loses information; in other words, if numbers are used, they are only relevant up to strictly monotonically increasing transformations (order isomorphism).

The point is that ordinal scales do not follow some hypotheses established as mathematical basis for common data analysis: the simple use of an arithmetic average requires, in fact, that the intervals between each step of evaluation be equal, that is, based on a linear metrics: as stated before, this cannot be verified for subjective quantifiers. Other properties, as the mutual exclusion of each of the different states, are required for applying common statistical distributions, as the binomial or hypergeometrical, or the regression method to identify a relationship between two variables (independent variables should have no uncertainty, in principle).

On the other side we shall observe that these limits of basic mathematical hypotheses are frequently neglected or “approximated” in a spirit of practical application: normal distribution is widely used, even if its mathematical basis requires an *infinite* number of random contributions to be present, which is, in practice, not possible. A common practice is to switch from linguistic quantifiers to numbers, just because thereafter it is possible to apply calculation methods to evaluate positions and dispersion indexes, like average and standard deviation, not allowed if one is not sure that the evaluation to be analyzed is, at least, an equal-interval scale. But this condition is not always true even for measurement results obtained by physical instruments, and often questionable for many technological quantities. A frequently adopted practical way for overcoming these problems consists in considering the frequency distribution, and to decide if normal distribution is acceptable or if it is better to switch to log-normal distribution. Summarizing the practical way consists in having some form of indicator showing whether the usual analysis methods can be applied, or if it is better to transform somehow the metrics to get results in a more suitable form.

Our aim is to define and evaluate some indicators to show whether operations like average or variance, commonly used in many different fields for data analysis, bring to valid results or not. So, in this paper, we would examine some indicators in order to evidence their capability to show on a given dataset whether methods specifically developed for ordinal scales shall be adopted or common methods can be applied for obtaining compatible results.

2. METHODS FOR ANALYSIS OF SUBJECTIVE EVALUATIONS

The absence of a linear metric in subjective evaluations requires, in principle, to use specific methods for analyzing results obtained by judgments and usually reported by literal quantifiers corresponding to poorly defined ordinal scales. Even the simple operation of defining an index of position from a group of evaluations could not be done by the arithmetic average, if the constant dimension of the successive intervals is not demonstrated. In this section two common statistical methods and two corresponding specific ordinal methods are treated, then enhancements of the two specific methods and a procedure for uncertainty evaluation are proposed, finally two case studies on comparison of common and enhanced specific methods are described.

2.1. Common methods

As common methods the arithmetic average and the Borda count for ranking are considered.

2.1.1. Arithmetic average

As well known, the arithmetic average of a list of numbers is the sum of all the members of the list divided by the number of items in the list.

2.1.2. Borda count

The Borda count is named after Jean-Charles de Borda [5]. The method, developed in the late 18th century at the French Academy of Sciences for electoral purposes, is still applied in several fields to process data in which each "voter" declares a preference order among different candidates [6]. This method transforms the ranking provided by each voter, given *n* candidates to choose from, into a numerical representation assigning *n* points to the candidate placed first, *n*-1 to the second and so on, down to one point for the candidate placed last. Therefore it is a procedure involving the concept of equal-interval scale. The overall ranking is obtained adding up the points given to each candidate [7].

2.2. Specific methods

As specific ordinal methods OWA (ordered weighted average) and Condorcet method are considered

2.2.1. OWA method

For ordinal scales OWA, a specific emulator of arithmetic average, was introduced in 1993 by Yager [8-9]. This operator is typically used with linguistic scales. It is defined as:

$$\text{OWA}=\underset{k=1,\ldots,n}{\text{Max}}[\text{Min}\{Q(k),b_k\}] \qquad (1)$$

where $Q(k)=S_{g(k)}$, $k=1, \ldots, n$, with:

- S an ordinal random variable whose values belong to the set $\{S_1,\ldots, S_t\}$, where S_i is the i-th level of the ordinal scale and t is the number of levels of the scale;
- $g(k)$=floor$\{1+[k((t-1)/n)]\}$;
- $Q(k)$ the average linguistic quantifier (the weights of the OWA operator);
- b_k the k-th element of the sample previously ordered in a decreasing order.

This OWA operator is said to be an emulator of arithmetic average since it operates, in an ordinal environment, in the same way as the arithmetic average works in a cardinal one. It can take value only in the set of levels of the ordinal scale, while a numerical codification of these levels could lead to some intermediate mean values. Figure 1 shows an example of graphical representation of the OWA calculation. The value of the OWA emulator of arithmetic average is given by the intersection of the "ascending stair" (OWA weights) and the "descending stair" (ordered sample elements) [10].

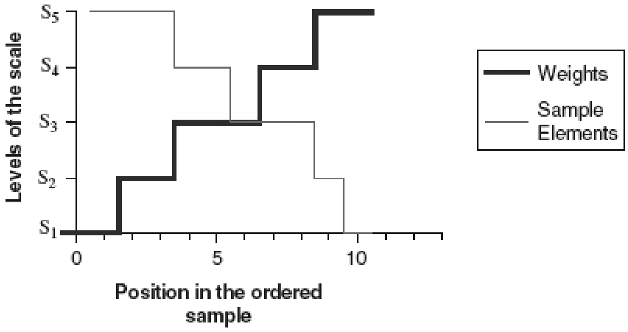

Fig. 1. Example of graphical representation of the OWA calculation. The value of the OWA emulator of arithmetic average is given by the intersection of the ''ascending stair'' (OWA weights) and the ''descending stair'' (ordered sample elements).

2.2.2. Condorcet method

The Condorcet method is named after Marie Jean Antoine Nicolas de Caritat, Marquis de Condorcet [11]. The method, developed in the late 18th century at the French Academy of Sciences as the Borda count, is still applied in several fields [6]. Votes are counted by pitting each candidate against all the others in a series of imaginary one-to-one contests. The winner of each pairing is considered the candidate preferred by the majority of voters. When all possible pairings of candidates have been made, if one candidate beats all the other candidates in these contests he is declared the "Condorcet winner". The latter check is usually done by creating the so-called "pairwise comparison matrix". A vicious circle occurs if there is no Condorcet winner; a further criterion must therefore be introduced to break the cycle (e.g. Schulze method [7-12]). In order to find the overall ranking, it is necessary to reiterate the algorithm taking away the winner of the previous iteration at each iteration.

2.3. Enhancement of specific methods

The two specific methods have a common problem: the low resolution that is evident when one tries to associate an uncertainty band to the results of the evaluation obtained. This section describes an enhancement of OWA and the Condorcet method, aimed to produce results with a better resolution. Only in this way it was possible to compare results of the two specific methods for ordinal scales to that obtained with the corresponding common statistical methods. The arithmetic average was compared with the results of enhanced OWA method and the Borda count with the enhanced Condorcet method.

2.3.1. Enhanced OWA method

As illustrated in figure 1, it is possible to obtain the OWA emulator of arithmetic average as the intersection of two stair functions. In order to get results with a better resolution, the authors propose to determine the OWA as the intersection of the two corresponding regression curve (second-order polynomials are chosen). Therefore, the enhanced OWA value is the ordinate of the intersection point.

2.3.2. Enhanced Condorcet method

The authors propose a significant change to the Condorcet method and introduce the "Enhanced Condorcet method", also described in [13]. The first part of the algorithm is the same until the creation of the pairwise comparison matrix, and then instead of verifying whether one candidate beats all the other candidates in one-to-one contests, the ranking was obtained by row sums of the pairwise comparison matrix. This is justified by the fact that each row represents the number of successes of each contender against his opponents. In this way, if each voter expresses a complete priority classification, the overall ranking obtained with the enhanced Condorcet method corresponds to that obtained with the Borda count. Because of various "draw" conditions, the obtained overall ranking has a lower weight than that of a voter giving a complete ranking. Therefore, a constant term, that is equal to the difference between the above mentioned weights divided by the number of candidates, was added to all the candidates.

2.4. Uncertainty evaluation

The comparison between specific and common statistical methods is possible adding the evaluation of a dispersion index. At present, there are no generally accepted procedures for uncertainty assessment in the context of subjective evaluations. The main standard for uncertainty calculation is ISO GUM [14], that defines "uncertainty" in clause 2.2.3 and the more specific term "expanded uncertainty" in clause 2.3.5, but does not consider subjective data. Moreover, the relationship between the evaluated variables and the derived response could be complicated. In order to overcome these limits, the authors propose the use of the statistical bootstrap method, introduced in 1979 by Efron [15-16]: a computer-based method to estimate the distribution of statistical estimators, e.g. the variable representing the subjective response. The method introduces the notion of bootstrap sample: *n* being the number of data of the original experimental sample, a bootstrap sample is obtained by *n* extractions with replacement of the data contained in the original sample. According to the theory proposed by Efron, when the number of bootstrap samples is sufficient, it is possible to estimate even complex population parameters.

So the dispersion index is derived, whenever possible, by calculating standard deviation, and, when not possible, by the application of the bootstrap method. The comparison was applied to some group of simulated data and to real experimental case studies.

2.5. Case studies on comparison of common and enhanced specific methods

Application to experimental data showed the advantages of increasing the resolution of OWA and Condorcet methods, but the evaluation of the uncertainty band pertaining to each method indicates that corresponding common methods often lead to a smaller uncertainty.

2.5.1. Borda count vs. enhanced Condorcet method

Figure 2 shows the comparison of Borda count, Condorcet method and enhanced Condorcet method with the relevant band of uncertainty, when applied to the ranking of ten acoustical attributes of a musical space under examination [17]. The comparison was made examining with the three methods a set of 44 votes given by expert musicians that were asked to assign to each acoustical attribute a number from 1 (the most important) to 10 (the least important). The mean scores were obtained averaging on the number of musicians. In order to present results in a comparable way, data were normalized in the 0÷1 interval. The uncertainty bands were evaluated using bootstrap in case of Condorcet and enhanced Condorcet methods, while for Borda count the usual standard deviation was adopted. Uncertainties are clearly much larger for Condorcet method, probably due to the coarse resolution, but even if enhanced Condorcet method produces reduced uncertainty bands, Borda count seems to be always better.

2.5.2. Arithmetic average vs. enhanced OWA method

Figure 3 shows the comparison between the simple arithmetic average and the enhanced OWA method, applied to the results of the subjective evaluation of the "click" of five electromechanical switches.

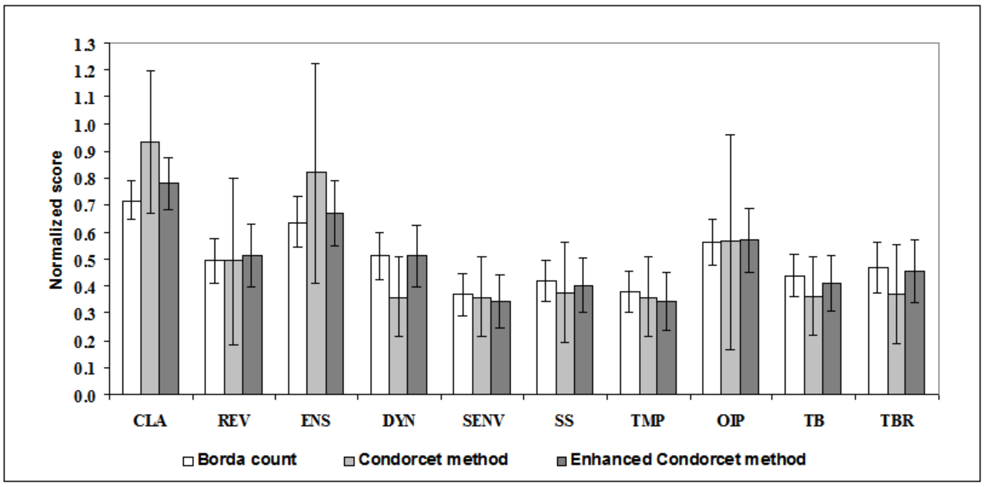

Fig. 2. Comparison of acoustical attributes' results obtained with Borda count, Condorcet method and enhanced Condorcet method, with their relevant uncertainty bands. It is evident that, while all results can be considered compatible, Condorcet method has the largest uncertainty, probably due to the coarse resolution; in fact enhanced Condorcet method, that has the merit of a better resolution, produces a smaller uncertainty. Nevertheless Borda count seems to be the best one.

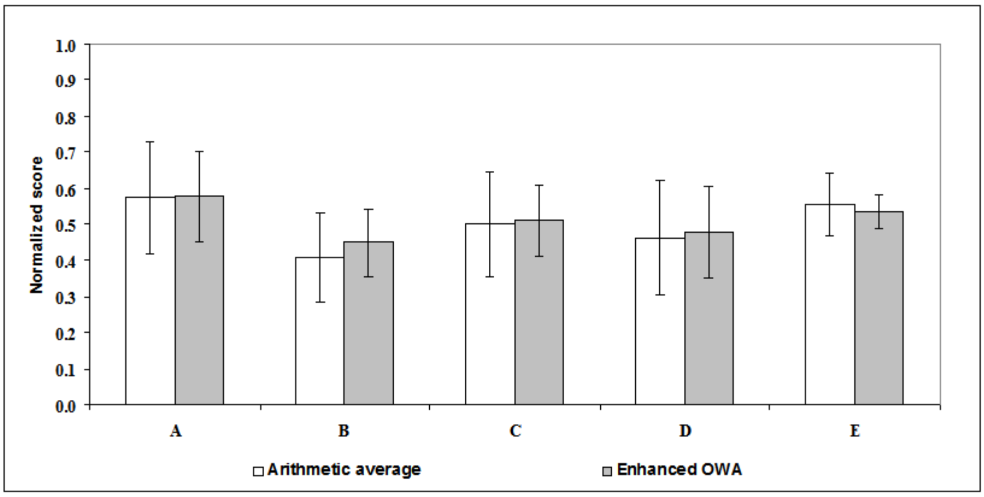

Fig. 3. Comparison of the results obtained with simple arithmetic average and enhanced OWA method, with their relevant uncertainty bands, applied to the subjective evaluation of the "click" of five electromechanical switches (A, B, C, D and E). It is evident a general good compatibility between the two methods, with comparable uncertainties. Standard OWA is not reported, having a constant value, due to its coarse resolution.

The comparison was made examining with the two methods a set of votes given by 27 judges, that were asked to assign a number from 1 (the most important) to 5 (the least important) to the switches. The mean scores were obtained averaging on the number of judges after the normalization of the data in the 0÷1 interval. The uncertainty bands were evaluated using bootstrap in case of enhanced OWA method, while for arithmetic average the usual standard deviation was adopted. The two compared methods show fully compatible results, with comparable uncertainties. Traditional OWA results are not showed, having a constant value, due to its coarse resolution.

The case studies presented allow to compare results and uncertainties obtained by methods of evaluation specific for ordinal scales (Condorcet and OWA), in their standard and enhanced versions, with that obtained with common methods (Borda and arithmetical average), showing a good compatibility justifying the wide use of common methods, because they, generally, are much simpler in practice and produce a lower uncertainty. Nevertheless common methods, in principle, cannot be applied when important distortion of the metric is present, possible with subjective evaluations but also with objective measurements in case of significant non-linearity. Therefore a research of possible indicators, showing the limits of application of common methods, was done comparing the results of arithmetic average and enhanced OWA for simulated data with given non-linearity distortion.

3. COMPARISON BETWEEN COMMON AND SPECIFIC METHODS ON NON-LINEAR DATA

Simulated data are produced starting from a normally distributed set modified by a non linear transformation to simulate the distorted evaluation of an observer. Note that this way is quite severe, as in practice even distortions of evaluation are subjective, therefore different among the various members of the jury, so that a sort of compensation is involved when the group of data are examined. Nevertheless one shall note that there are also situations showing a sort of systematic distortion produced by the matter analyzed. A very clear example is when the evaluation scale, even having a number of steps, has also an implicit important boundary of the type on-off between two of its steps: for instance the scale *inadequate* - *sufficient* - *good* - *very good* - *excellent* has between the steps inadequate and sufficient a marked boundary between the region of "refuse" and the region of "accept". In that condition frequently there is a significant gap of probability levels between the two steps, clearly higher than the probability differences between the other evaluation steps.

One hundred data sets of simulated data are produced, for each of six different levels of non-linearity, and qualified, evaluating them by chi-square and skewness. Results obtained with each data set are analyzed with arithmetic average and enhanced OWA, considering, also, the relevant uncertainties, so that it is possible to understand if the differences are significant or not. As the indicators adopted are statistical, the number of data produced for each set, from 50 to 1000, is also considered. Table 1 shows some mean results obtained for non-linearity ranging from 0.5%, a value typical for instrumentation, to 20%, that can represent the scale distortion of some technological measurement or of subjective evaluations. The chi-square column indicates the percentage of the relevant dataset having a χ^2 value inside of a 95% confidence interval. The "Sk." column shows skewness values. The "Com." column indicates the percentage of the relevant dataset showing compatibility between common and specific methods, evaluated by a normalized error [18] lower than 1.5 at a confidence level of 95%. It is possible to see that there is a generic, not well defined connection between compatibility and chi-square; while a defined relationship was extrapolated considering skewness *Sk*. The level of 95% of compatible results is taken as boundary between compatibility and non-compatibility regions. This boundary depends on the number *n* of data of the dataset and is shown in figure 4, together with its tendency defined by the logarithmic path $Sk = -0.12 \ln(n) + 1.0$ (dashed line). Therefore skewness can be considered to be a good indicator of the possibility of using common methods for analyzing subjective data.

Table 1. Analysis of simulated evaluations. For different data numbers and different levels of distortion (given as amount of non-linearity), the percentage of χ^2 test accepted at 95% confidence and the value of skewness are shown in relation to the percentage of compatibility (normalized error lower than 1.5 at 95% confidence) between arithmetic average and enhanced OWA.

Data number	50			100			200			300			500			1000		
Indicator Distort.	χ^2	Sk.	Com.	χ^2	Sk.	Com.	χ^2	Sk.	Com.	χ^2	Sk.	Com.	χ^2	Sk.	Com.	χ^2	Sk.	Com.
0.5%	0.90	0.02	1.00	0.94	0.04	1.00	0.96	0.03	1.00	0.96	0.02	1.00	0.86	0.03	1.00	0.36	0.02	1.00
1%	0.83	0.05	1.00	0.96	0.05	1.00	0.96	0.05	1.00	0.95	0.05	1.00	0.80	0.05	1.00	0.17	0.05	1.00
2%	0.89	0.10	1.00	0.97	0.10	1.00	0.95	0.11	1.00	0.95	0.09	1.00	0.81	0.10	1.00	0.15	0.09	0.99
5%	0.96	0.23	1.00	0.93	0.22	1.00	0.84	0.24	1.00	0.64	0.24	0.99	0.11	0.24	1.00	0.00	0.24	0.90
10%	0.95	0.41	1.00	0.91	0.38	0.97	0.31	0.42	0.94	0.02	0.43	0.81	0.00	0.43	0.47	0.00	0.43	0.04
20%	0.80	0.54	0.98	0.38	0.58	0.82	0.00	0.59	0.57	0.00	0.60	0.18	0.00	0.60	0.00	0.00	0.60	0.00

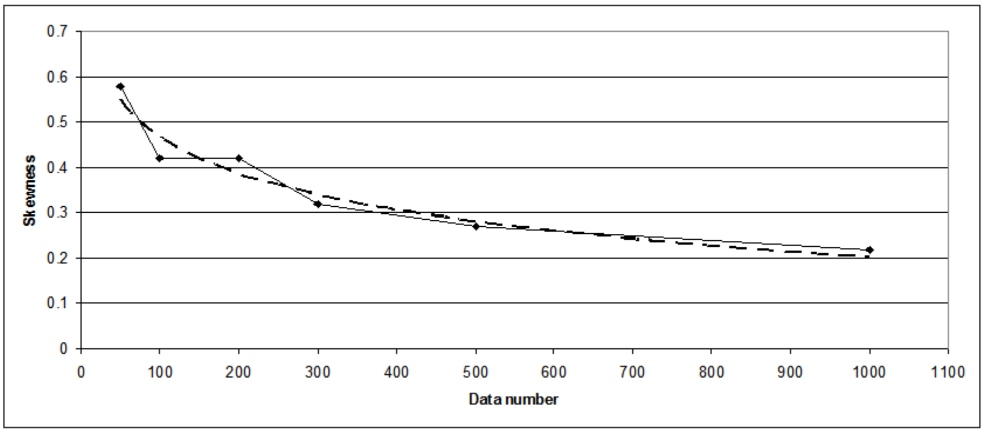

Fig. 4. Limit values of skewness representing the boundary between the lower region of compatibility (between arithmetic average and enhanced OWA) and the upper region of non-compatibility.

4. CONCLUSIONS

The advantages of common analysis methods as compared with methods developed specifically for ordinal scales are important: from a practical point of view common methods are much friendly, as generally well known and because the analysis may be performed using traditional procedures and widely available software. Nevertheless in principle one should not apply common analysis methods to results obtained in ordinal scales, because the mathematical basis of common methods is the condition of an equal-interval scale. On the side of specific methods, like Condorcet or OWA methods, the main drawback is the coarse resolution, that can be overcome using the developed relevant enhanced methods.

Advantages of common methods were showed to be significant, even more considering the often lower uncertainty of results, so that it is clear the usefulness of simple indicators apt to alert when the metric distortion is small enough to allow the use of common methods or when, on the contrary, specific ordinal scale methods are mandatory. Skewness was found to be a good indicator for dividing the two regions of application of common methods or specific ordinal scale methods. The boundary depends also from the number of data considered in the analysis. A logarithmic shape boundary line appears to be adequate.

REFERENCES

[1] K. Arrow, “Social Choice and Individual Values”, John Wiley & Sons, New York, 1951.

[2] D. G. Saari, F. Valognes, “Geometry, Voting and Paradoxes”, Mathematics Magazine, Vol. 71, No. 4, pp. 243-259, 1998.

[3] E. Nagel, “Measurement”, Erkenntnis, 2, pp. 313-33, 1932.

[4] S. S. Stevens, “On the Theory of Scales and Measurement”, Science, No. 103, pp. 677-680, 1946.

[5] J. C. Borda, “Mémoire sur les Élections au Scrutin”, Histoire de l'Académie Royale des Sciences, 1781.

[6] http://stv.sourceforge.net/votingmethods/other, visited May 2008.

[7] P. E. Johnson, “Voting Systems”, http://pj.freefaculty.org/Ukraine/PJ3_VotingSystemsEssay.pdf, 2005, visited May 2008.

[8] R. R. Yager, “Non-Numeric Multi-Criteria Multi-Person Decision Making”, Group Decision and Negotiation, Vol. 2, pp. 81–93, 1993.

[9] R. R Yager, D. P. Filev, “Essentials of Fuzzy Modelling and Control”, John Wiley & Sons, New York, 1994.

[10] F. Franceschini, M. Galetto, M. Varetto, “Qualitative Ordinal Scales: The Concept of Ordinal Range”, Quality Engineering, Vol. 1

[11] M. Condorcet, “Essai Sur l'Application de l'Analyse à la Probabilité des Décisions Rendues à la Pluralité des Voix”, 1785.

[12] M. Schulze, “A New Monotonic and Clone-Independent Single-Winner Election Method”, Voting Matters, No. 17, pp. 9-19, 2003.

[13] G. Genta, M. Giovannini, A. Astolfi, G. Barbato, “Investigation on Subjective Acoustical Attributes Preferred by Performers through Ranking Data Analyses”, submitted to Acta Acustica united with Acustica, 2008. No. 4, pp. 515–524, 2004.

[14] ISO Guide 98, “Guide to the Expression of Uncertainty in Measurement (GUM)”, International Organization for Standardization, Genève, 1995.

[15] B. Efron, “Bootstrap Methods: Another Look at the Jackknife”, The Annals of Statistics Vol. 7, No. 1, pp. 1–26, 1979.

[16] B. Efron, R. J. Tibshirani, “An Introduction to the Bootstrap”, Chapman & Hall, New York, 1993.

[17] G. Genta, M. Giovannini, A. Astolfi, G. Barbato, “The Subjective Investigation of Acoustic Perception of Musicians: a Proposed Method for Interpretation of Results”, Proceedings of the 19th International Congress on Acoustics, Madrid, September 2007.

[18] ISO/IEC Guide 43-1:1997 “Proficiency testing by interlaboratory comparisons -- Part 1: Development and operation of proficiency testing schemes”, 1997.