ELSEVIER

European Journal of Operational Research 140 (2002) 142–147

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Decision Aiding

Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting

W.J. Hurley *, D.U. Lior

Department of Business Administration, Royal Military College of Canada, CP 17000, Succursale Forces, Kingston, Ont., Canada K7K 7B4

Received 15 October 1999; accepted 13 June 2001

Abstract

Analytic group decision techniques for selecting a subset of alternatives range between multicriteria decision analysis techniques such as multiattribute utility theory and the analytic hierarchy process to voting techniques where each member of the decision group submits a ranking of the alternatives, and these individual rankings are then aggregated into an overall ranking. The obvious advantage of voting is that it bypasses the rather intensive data generation requirements of multicriteria techniques. In this paper we compare the performance of trimmed mean rank-order aggregation procedures in the case where a subset of the individuals in the group charged with the decision vote strategically. We employ a Monte Carlo simulation experiment on a specific decision instance and find that trimmed mean aggregation compares favorably with other procedures. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Voting theory; Expert judgment

1. Introduction

An important area of research in the field of decision analysis is the way in which expert judgments are aggregated into a single judgment or decision. In this paper we study methods for aggregating expert rank-orders into an overall group rank-order. We study this problem in the case where there are two interesting characteristics associated with individual submissions. The first is measurement error: we assume that the alternatives under consideration have an objective true value and our experts can only observe an imperfect signal of this value. Hence our expert judges may arrive at different rank-orders due to measurement error. The second is strategic voting: some members of the group may give "their" alternative a better ranking than one they know is in the best interest of the organization.

There are many number of organizational settings where this kind of strategic voting is a problem. For instance, take the example of judging an international figure skating competition. Does the

* Corresponding author. Tel.: +1-613-541-6000x6468; fax: +1-613-541-6315.
E-mail address: hurley_w@rmc.ca (W.J. Hurley).

PII: S0377-2217(01)00226-0

scoring system mitigate the effects of those judges who have a preference, either conscious or subconscious, for their country's skater? Another example is the way in which the Canadian Army determines annually which of the projects valued at under $2.5 million will be undertaken. A board made up of officers from the various land arms is convened and charged with the task of rank-ordering all projects. There is certainly 'ownership' of projects at this table. An artillery officer will not see armored projects the same way an armored officer sees them.

There is an immense literature on the way expert opinion should be aggregated. A part of this literature considers aggregation based on a simple average and would include Armstrong (1985), Ashton (1986), Ashton and Ashton (1985), Einhorn and Hogarth (1975), Einhorn et al. (1977), Hastie (1986), Hogarth (1978), Zajonc (1962), and Zarnowitz (1984). Lansdowne (1996) compares a number of popular vote aggregation procedures. A good summary of the literature on the specific problem of aggregation of probabilities and probability distributions can be found in Ginest and Zidek (1986). Major contributions include Agnew (1985), Clemen and Winkler (1985), Lindley (1983), Winkler and Poses (1993), Winkler (1968), and Lipscomb et al. (1998). We should also note that there is a large literature on formal techniques to arrive at group decisions. Analytic group decision techniques for selecting a subset of alternatives range between multicriteria decision analysis techniques such as multiattribute utility theory and the analytic hierarchy process (Saaty, 1980) and voting techniques where individual rankings are aggregated directly into an overall ranking. One of the advantages of voting techniques is that they bypass the data requirements of multicriteria approaches.

Our work is also clearly related to the social choice literature on vote aggregation. This literature is characterized by seminal work on a number of difficulties in the construction of a social preference function and would include the famous theorems of Arrow and Gibbard–Satterthwaite. More recently, Austen-Smith and Banks (1996) have examined the Condorcet jury theorem. They examine the conventional wisdom that each voter will cast a "sincere" vote for his or her alternative of choice and find that such voting does not constitute Nash equilibrium behavior. In our simulation experiment, we take as given that a fraction of judges will vote strategically.

However, the work of this paper relies heavily on the work of Bassett and Persky (BP) (1994). BP study the way figure skating competitions are judged. They point out that the rank-order of the skaters is determined by median rank voting, a system that has some interesting properties in the presence of strategic voting. One is that it is 'resistant to manipulation by a minority subset of judges ...' (p. 1075). In view of the fact that other sports use different aggregation techniques (for instance, diving and gymnastics use a trimmed mean), it is reasonable to ask which of these trimmed mean vote aggregation techniques is preferred in the case where there is strategic voting. Yaniv (1997) also considers trimmed mean as an aggregation tool but does not consider strategic voting explicitly.

Hence, in this paper, we consider the performance of trimmed mean rank aggregation procedures in the presence of measurement error and strategic voting. We denote by $Trim(k)$ the trimmed mean rank-order which throws out the best k ranks and the worst k ranks for each alternative. Note that the average rank and median rank measures are special cases of a trimmed mean. Average rank, denoted $Trim(0)$, is the case where no observations are thrown out. At the other extreme, median rank throws out all but one observation. For instance, if there are seven rank observations (seven judges), $Trim(3)$ is the median rank-order.

To examine this issue, we study a specific instance using Monte Carlo simulation. Our general finding is that, in the presence of strategic voting, the median rank procedure is superior to other trimmed mean procedures including the average rank procedure. And in the absence of strategic voting, it does no worse than these other procedures. Hence the evidence of this paper suggests that median rank aggregation is a reasonable aggregation tool.

Finally, we point out that these results have an analog in the theory of robust estimation. For instance, consider a random sample from an unknown symmetric distribution and suppose it is necessary to estimate the median of the distribution.

If the underlying distribution were normal, then a maximum likelihood estimate of the median is the sample average. However, if the underlying distribution has fat tails (such as the Cauchy), a trimmed mean will produce a lower mean square error than the sample average. Hence, for some fat-tailed distributions, a trimmed mean is preferred to a sample average. [1]

2. Median rank aggregation and strategic voting

To give some idea of the way median rank voting handles strategic voting, consider the following example taken from Hurley (1998). Suppose that all-stars in a football conference are selected by the teams making up the conference. Each team ranks the players nominated at a position, and the best player is to be selected on the basis of median rank. Suppose the following rank-orders have been submitted for the quarterback position:

	Teams						
QBs	*A*	*B*	*C*	*D*	*E*	*F*	*G*
a	1	1	1	3	3	1	1
b	2	2	2	1	1	2	2
c	3	3	3	2	2	3	3

where *a* is team *A*'s quarterback, *b* is team *B*'s, and *c* is team *C*'s. Moreover, suppose that these rankings are each team's true unbiased rankings. The first step is to simply write the ranks for each quarterback, in order, from highest to lowest:

QBs				Median			
a	1	1	1	1	1	3	3
b	1	1	2	2	2	2	2
c	2	2	3	3	3	3	3

The median rank, then, is the rank of the middle observation, or, in this case, the fourth highest number in the list. Hence, QB *a* has a median rank of 1, QB *b* has a median rank of 2, and QB *c* has a median rank of 3. Therefore, the median rank-order is $a \succ b \succ c$. If it happened that two players had a median rank of 1, there are a number of tie-breaking mechanisms. Some of these are set out in BP.

Table 1
Results for the various aggregation rules

	Rank average	Truncated average rank	Median rank
a	1.857	1.8	1
b	1.571	1.6	2
c	2.571	2.6	3

To see how the median rank scheme handles strategic voting, suppose that team *B* feels that the rank order is $a \succ b \succ c$, but, in the interests of having their quarterback be the all-star, they submit the ranking $b \succ c \succ a$. The results for the median rank aggregation technique and two others (average rank and trimmed mean rank where a player's best and worst ranks are thrown out) are shown in Table 1. The median rank scheme returns the correct rank order, $a \succ b \succ c$, whereas the average-rank and truncated-average-rank return $b \succ a \succ c$. This is no more than the basic BP result that median rank aggregation is immune to manipulation by a minority of judges.

3. The simulation design

In the absence of strategic voting, the simulation works in this way. Suppose there are N objects and M judges. The objective, true values of the objects, $v_1, v_2, \ldots, v_N$, are selected randomly from the unit interval, [0,1]. The *unbiased assessment* of judge j for alternative i is then drawn from a normal distribution with mean v_i and variance σ_i^2. Once judge j obtains an unbiased assessment for all objects, he can rank-order the objects. We do this for each of the judges. These are then aggregated using $Trim(k)$ for various values of k. Effectively we are modelling measurement error on the part of the judges.

To measure how well each aggregation technique does, we measure the frequency with which each picks:

[1] See Degroot (1975).

1. *Unordered objective*: the best L alternatives *without* regard for rank-order;
2. *Ordered objective*: the best L alternatives in the proper rank-order.

For the *unordered objective*, we are only concerned that a vote aggregation technique pick the top L objects; it does not matter what order these top L objects end up in, only that they make the top L spots in the list. The *ordered objective* is more specific. Not only do the top L objects have to be picked, but they also have to appear in the correct order.

We introduce strategic voting as follows. We assume that each judge will vote strategically with probability p. Effectively p controls the "amount" of strategic voting: the expected number of judges voting strategically is Mp. If a judge is going to vote strategically, he will do so in an extreme way. Suppose the judge's unbiased assessment of three objects is $A \succ B \succ C$ but, for other reasons, he would prefer that object B is selected (he is an artillery officer and A is an armored project). Then we assume that he will submit the ranking $B \succ C \succ A$. That is he puts his object first, and the remaining objects in reverse order. He reasons that this gives his project the best chance of succeeding. We term this rank-order the *biased assessment*, and in the case at hand, the judge is said to have ownership of object B (the personally preferred choice). While this form of strategic voting is particularly perverse, we only introduce it in this way to give an extreme version. In reality strategic voting is likely to be more subtle, especially if the judge has a reputation to maintain.

In summary, the steps of the simulation are:

1. Generate a set of true values, $v_1, v_2, \ldots, v_N$, randomly from [0,1].
2. Generate an unbiased assessment for each judge.
3. For each judge, generate a biased assessment with probability p. (If a judge is biased, ownership of an alternative is assigned randomly.)
4. Aggregate the assessments using *Trim*(k) for various values of k, and for each, determine whether the unordered and ordered objectives are satisfied.

Before reporting our experimental results, it is important to clarify why judges vote strategically. As we have argued above, organizational decision-makers are sometimes of two minds on what the best course of action is. For instance, in the context of choosing military projects, an armored officer may be presented with objective, rational evidence that an artillery project has higher value than the armoured project he has put up. A part of him says that he ought to vote for the artillery project because it is best for the force structure as a whole. Another part says he ought to vote for the armoured project because it is "his" project. In addition to this motivation for strategic voting, each judge observes an imperfect, unbiased signal of each alternative's value. Non-strategic-voters are assumed to submit a rank-order based on these signals; strategic voters submit the perverse rank-order described above.

4. Results

For the first set of simulations, we fix the following parameters:

Number of objects	7
Number of judges	7
Standard deviation (σ_i)	0.10
Number of simulation iterations	100,000

What we vary is the probability, p, that judges vote strategically.

The frequencies for the *ordered objective* for various values of the strategic voting parameter, p are shown in Table 2 and for the *unordered objective*, in Table 3. For instance, consider the element for $p = 0$, $L = 2$, and *Trim*(0) in Table 2: 0.714. This means that in 71.4% of the 100,000 iterations, the average-rank aggregation procedure picked the top two elements in the correct order, when there was no strategic voting. Similarly, in Table 3, for $p = 0.20$, $L = 3$, and *Trim*(3), the element is 0.731. In this case, the median rank aggregation procedure is able to pick the top three elements without regard for order in 73% of the 100,000 iterations, when there was a 20% probability of strategic voting.

We first consider the ability of each aggregation method to find the best object $L = 1$. In the absence of strategic voting ($p = 0$) we would expect the average rank procedure *Trim*(0) to perform the best.

Table 2
Results for the ordered objective

L	1	2	3	4	5	6	7
$p = 0$							
Trim(0)	0.844	0.714	0.600	0.503	0.420	0.354	0.354
Trim(1)	0.843	0.710	0.596	0.498	0.414	0.348	0.348
Trim(2)	0.844	0.710	0.595	0.497	0.412	0.347	0.347
Trim(3)	0.844	0.710	0.595	0.497	0.413	0.347	0.347
$p = 0.20$							
Trim(0)	0.717	0.541	0.423	0.338	0.272	0.221	0.221
Trim(1)	0.746	0.572	0.449	0.363	0.295	0.245	0.245
Trim(2)	0.772	0.596	0.465	0.376	0.306	0.255	0.255
Trim(3)	0.781	0.603	0.470	0.380	0.309	0.257	0.257
$p = 0.40$							
Trim(0)	0.492	0.299	0.208	0.157	0.122	0.096	0.096
Trim(1)	0.527	0.329	0.232	0.179	0.141	0.115	0.115
Trim(2)	0.573	0.360	0.253	0.198	0.159	0.131	0.131
Trim(3)	0.612	0.392	0.273	0.213	0.171	0.141	0.141

Table 3
Results for the unordered objective

L	1	2	3	4	5	6	7
$p = 0$							
Trim(0)	0.844	0.837	0.836	0.838	0.835	0.844	1.000
Trim(1)	0.843	0.834	0.834	0.835	0.833	0.843	1.000
Trim(2)	0.844	0.833	0.832	0.833	0.831	0.842	1.000
Trim(3)	0.844	0.833	0.832	0.832	0.831	0.842	1.000
$p = 0.20$							
Trim(0)	0.717	0.708	0.726	0.749	0.785	0.807	1.000
Trim(1)	0.746	0.727	0.733	0.755	0.794	0.828	1.000
Trim(2)	0.772	0.736	0.731	0.761	0.798	0.834	1.000
Trim(3)	0.781	0.742	0.731	0.760	0.800	0.834	1.000
$p = 0.40$							
Trim(0)	0.492	0.449	0.475	0.527	0.656	0.774	1.000
Trim(1)	0.527	0.471	0.483	0.531	0.659	0.795	1.000
Trim(2)	0.573	0.488	0.482	0.542	0.666	0.810	1.000
Trim(3)	0.612	0.512	0.486	0.540	0.679	0.820	1.000

However it does not. In fact, it appears as though all of the aggregation procedures are statistically equivalent. When strategic voting is introduced ($p = 0.20$, 0.40), the trimmed mean procedures outperform the average rank procedure. Of these, the median rank procedure is best. To give some idea of the error in these numbers, a 95% confidence interval for median rank frequency of 0.781 (*Trim*(3), $p = 0.20$) is [0.778,0.784]. Note that when there is excessive strategic voting ($p = 0.40$), *Trim*(3) picks the best alternative 61.2% of the time and *Trim*(0) picks it 49.2%. In the presence of strategic voting, the median rank procedure is clearly better.

Now consider the case where the objective is to identify the best L objects without regard for order. For instance, a voting method might be used to prescreen a set of investment alternatives: the top L alternatives will be given further consideration. To assess the methods on this objective, consider the output in Table 3. Note that, in the absence of strategic voting (the panel with $p = 0$), the trimmed means do as well as average rank for all values of L. On the other hand, in the presence of strategic voting ($p = 0.20$, 0.40), the trimmed means outperform average rank, and of these median rank does the best.

We have done an extensive sensitivity analysis of this example, varying the number of judges, the number of alternatives, and the standard deviation σ_i. In each case the same result obtains: with only measurement error, median rank does as well as the other trimmed mean and average rank measures; with measurement error and strategic voting, median rank outperforms these other aggregation procedures.

5. Conclusions

We have presented results of a Monte Carlo simulation which suggests that, in the presence of strategic voting, median rank aggregation works better than other trimmed mean procedures and the average rank procedure. In view of the fact that all of these aggregation procedures are about the same when there is no strategic voting, the firm conclusion of this simulation example is that median rank voting is a robust method for aggregating individual rank orders.

However this example is in no way a proof that median rank would always be superior. Our results are based only on a simulation for a specific parameter set. Future research will explore a more general approach.

References

Agnew, C.E., 1985. Multiple probability assessments by dependent experts. Journal of the American Statistical Association 80, 343–347.

Armstrong, J.S., 1985. Long-range Forecasting: From Crystal Ball to Computer, second ed. Wiley, New York.

Ashton, R.H., 1986. Combining the judgment of experts: How many and which ones? Organizational Behavior and Human Decision Processes 38, 405–414.

Ashton, A.H., Ashton, R.H., 1985. Aggregating subjective forecasts: Some empirical results. Management Science 31 (12), 1499–1508.

Austen-Smith, D., Banks, J.S., 1996. Information aggregation, rationality, and the Condorcet jury theorem. American Political Science Review 90, 34–45.

Bassett, G.W., Persky, J., 1994. Rating skating. Journal of the American Statistical Association 89 (427), 1075–1079.

Clemen, R.T., Winkler, R.L., 1985. Limits for the precision and value of information from dependent sources. Operations Research 33, 427–442.

Degroot, M.H., 1975. Probability and Statistics. Addison-Wesley, London.

Einhorn, H.J., Hogarth, R.M., 1975. Unit weighting schemes for decision making. Organizational Behavior and Human Performance 13, 171–192.

Einhorn, H.J., Hogarth, R.M., Klempner, E., 1977. Quality of group judgment. Psychological Bulletin 84, 158–172.

Ginest, C., Zidek, J.V., 1986. Combining probability distributions: A critique and an annotated bibliography. Statistical Science 1, 114–148.

Hastie, R., 1986. Experimental evidence on group accuracy. In: Grofman, B., Owen, G. (Eds.), Decision Research. JAI Press, Greenwich, CT, pp. 129–157.

Hogarth, R.M., 1978. A note on aggregating opinions. Organizational Behavior and Human Performance 21, 40–46.

Hurley, W.J., 1998. An efficient, objective technique for selecting an all-star team. Interfaces 28, 51–57.

Lansdowne, Z.F., 1996. Ordinal ranking methods for multicriteria decision making. Naval Research Logistics 43 (5), 613–627.

Lindley, D.V., 1983. Reconciliation of probability distributions. Operations Research 31, 806–886.

Lipscomb, J., Parmigiani, G., Hasselblad, V., 1998. Combining expert judgment by hierarchical modeling: An application to physician staffing. Management Science 44, 149–161.

Saaty, T.L., 1980. The Analytic Hierarchy Process. McGraw-Hill, New York.

Winkler, R.L., 1968. The consensus of subjective probability distributions. Management Science 15, 61–75.

Winkler, R.L., Poses, R.M., 1993. Evaluating and combining physicians' probabilities of survival in intensive care units. Management Science 39, 1526–1543.

Yaniv, I., 1997. Weighting and trimming: Heuristics for aggregating judgments under uncertainty. Organizational Behavior and Human Decision Processes 69 (3), 237–249.

Zajonc, R.B., 1962. A note on group judgments and group size. Human Relations 15, 177–180.

Zarnowitz, V., 1984. The accuracy of individual and group forecasts from business and group forecasts. Journal of Forecasting 3, 11–26.