RC23631 (W0506-079) June 17, 2005
Mathematics

IBM Research Report

Ranking Pilots in Aerobatic Flight Competitions

Andrew Davenport, Douglas Lovell
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Ranking Pilots in Aerobatic Flight Competitions

Andrew Davenport and Douglas Lovell
IBM T.J.Watson Research Center, 1101 Kitchawan Road, Yorktown Heights,
New York, 10598, USA
davenport@us.ibm.com, dclo@us.ibm.com

Abstract

We present initial results of a feasibility study into using rank aggregation methods from Social Choice theory to determine the outcome of aerobatic flying competitions. In such competitions, multiple judges grade individual pilots over a number of aerobatic maneuvers. The judges' grades are aggregated to determine a final, consensus ranking of the pilots in each competition. The current system in use, based on manipulating the mean of the judges' grades for each pilot, is complex, opaque and unpopular with many pilots and judges. We present results of applying two popular rank aggregation methods-- the Borda Count and the Kemeny Rule --to data from International Aerobatics Club sanctioned contests flown in 2004. The Kemeny Rule provides a promising, practical solution for determining winners of aerobatics competitions.

1 Introduction

Many times a year in the United States and around the world, pilots of high-performance small aircraft compete in the exacting sport of aerobatics. The contestants have invested hundreds of thousands of dollars and hundreds of hours of practice in dedicated efforts to guide their airplanes in precise figures through the air.

The process of judging competition aerobatics is similar to that found in figure skating. In both sports, contestants perform a series of precise figures graded by a panel of judges. In aerobatics, each judge gives a grade to each figure from zero to ten in half point increments. The judges follow comprehensive and detailed guidelines for grading the figures, beginning with a perfect figure grade of ten, then deducting for variations from an ideal performance.

The contest requires a system that converts the individual figure grades from all judges into a single ranking of the pilots. A contest will typically reward the top three to six pilots with trophies or placards commemorating their accomplishments.

The contest ranking system must have the confidence of the participants as a fair and accurate means for selecting the best pilots. Three key properties inspire that confidence: First, participants need an intuitive grasp of how the system operates. Second, the system must be robust in the presence of bias from individual judges. Third, it must be easy to audit.

Truchon [Truchon 1998] has investigated the application of rank aggregation techniques to the sport of Figure Skating. In this study we investigate the application of two rank aggregation techniques, the Borda Count and the Kemeny Rule, to the sport of aerobatics competitions. We use data for 343 flights from 32 contests flown in 2004 sanctioned by the International Aerobatics Club (IAC)[1]. The current system in use by the IAC has a number of drawbacks, which we describe in detail below. The goal of this study is to investigate and recommend alternative rank aggregation procedures for use by the IAC in future competitions.

2 Background of methods in use

The simplest system for ranking pilots computes the mean of the scores for each pilot from each of the judges. This system has a number of drawbacks: First, it is very easy for a single judge to manipulate the mean. One judge's high score can raise a pilot a place or two in the rankings.

A second problem with the mean is that it assumes all judges represent comparable samplings of the ideal score. It's true that they all scored the same event; but, in spite of the best efforts at training for use of consistent, objective criterion, judges may have idiosyncratic methods of scoring. One judge may never give a score higher than eight. Another judge will use the full range of scores, but hardly ever give a score lower than a six.

Bias in judging was apparent in world contests of the 1960's, when east-west rivalry was highly pronounced. A statistical method was designed for normalizing judges' scores and comparing them. The system replaced with the mean score any individual judge's score that was more than 1.2 standard deviations beyond the mean after normalization. This system, first used in 1978, discourages judges from intentionally raising the scores of their favorites. The international organization that oversees aerobatics has refined the method over the years to the method currently in use, called TBLP [TBLP 2005].

[1] The IAC sanctions contests in the United States and Canada.

For many years, TBLP has been the standard for scoring aerobatics contests worldwide. This year the organization that oversees contests in the United States, IAC, has elected to abandon TBLP in favor of the mean. It did so because of several perceived faults with TBLP. First, TBLP makes computations based on the mean and standard deviation. Many IAC regional contests have few pilots competing in a category, yielding too small a sample for meaningful computation of a standard deviation. Second, many pilots and judges in the United States object to the opaque manipulation TBLP applies to the scores.

3 Rank Aggregation Techniques

The *rank aggregation* problem is to combine many different rank orderings on the same set of candidates (often referred to as alternatives), in order to obtain a "consensus" ordering. Rank aggregation has been studied extensively in the context of Social Choice theory and arises in many settings, such as determining the outcome of sporting contests [Truchon 1998], multi-criteria and word association queries in databases [Dwork et al 2001], and fighting search engine spam [Dwork et al 2001]. In this study we considered two widely studied rank aggregation methods: the Borda Count [Borda 1781] and the Kemeny Rule [Kemeny 1959].

The Borda Count is a positional rank aggregation procedure that assigns a score to each candidate corresponding to the positions in which it appears within each judge's ranked list of candidates. The candidates are then sorted by their total score, and the winner is the candidate with the highest score. The Borda Count is used quite widely in ranking sports competitions: in the United States for example, it is used in baseball to determine the winner of the most valuable player award, as well as determining the winner of the national championship of American football.

An advantage of positional methods such as the Borda Count is that they are very easy to compute. They also satisfy properties of anonymity, neutrality, and consistency in the Social Choice literature [Young 1974]. However their simplicity makes them easy to manipulate, and it is known that no positional rank aggregation procedure can satisfy the Condorcet criterion. (The Condorcet criterion states that if every judge ranks x ahead of y, for all alternatives $y \neq x$, then x is ranked as the winner in the aggregated ranking.)

Majority ranking methods determine an outcome in terms of the majority ranking for each pair of alternatives: alternative x is ranked ahead of alternative y in the aggregated ranking if more judges prefer x to y. Unfortunately, even when all of the judges' ranking preferences are transitive (if judge A ranks x ahead of y and y ahead of z then they rank x ahead of z), the majority ranking of all of the judges may contain cycles[2] (the majority of judges prefers x to y, y to z and z to x). This is known as the Condorcet Paradox [Condorcet 1785]. As a result, simple majority rank aggregation methods, such as Condorcet methods, may fail to select any winners at all.

The Kemeny Rule has been proposed as a way of seeking a compromise ranking in the majority vote when there are cycles present in the majority voting relation. The Kemeny Rule is defined in terms of the Kendall-Tau distance [Diaconis 1988]. The Kendall-Tau distance defined over two ordered lists counts the number of adjacent pair-wise disagreements between these lists. That is the number of pair-wise adjacent transpositions needed to transform from one list into the other, sometimes referred to as the "bubble sort distance."

The Kemeny Rule determines a ranking σ applied to a set of judge's rankings $(\zeta 1, \zeta 1, \zeta n)$ that minimizes the total Kendall-Tau distance between σ and each of the rankings in $(\zeta 1, \zeta 1, \zeta n)$. (It minimizes the number of pair wise disagreements between the judge's rankings and the aggregated ranking.) The Kemeny ranking satisfies the Condorcet criterion, as well as properties of neutrality, consistency and local independence of irrelevant alternatives. The main drawback of the Kemeny Rule is that its computation is known to be NP-Hard [Cohen, et. al. 1999; Dwork et al. 2001].

4 Experimental Study

Theoretical properties of rank aggregation procedures have been widely studied in the field of Social Choice theory. Practical considerations concerning the use of different methods have received little attention. First, although computing the outcome of the Kemeny Rule is known to be NP-hard, in practice the size of the input for many sports competitions will be bounded. We rarely expect human judges to accurately rank more than a few tens of competition participants. Second, in sports competitions it may not be acceptable for a rank aggregation procedure to determine that two or more candidates are tied for a particular place: a clear winner in the competition may be required. Finally, for many NP-hard problems it is known that there may exist multiple optimal solutions. In the context of rank aggregation using the Kemeny Rule there may be multiple optimal Kemeny rankings which rank the candidates in significantly different ways.

We were fortunate to obtain from the IAC detailed data for thirty-two contests held in 2004 with 343 flight competitions. The data contains the TBLP computed scores for every pilot in every flight, as well as over 134,000 raw grades given every pilot by every judge for every flight. Most of the flights had five judges: eighteen flights had four judges and fifty flights had three judges. The median number of pilots in a flight was six. Some flights had up to twenty-five pilots.

We have converted the set of scores that each judge gave to the pilots in each flight competition into a set of rankings (one from each judge) of the pilots for each flight. To determine the Kemeny ranking, we used the branch and bound procedure described in [Davenport and Kalagnanam 2004].

[2] When there are more than two alternatives to be ranked.

4.1 CPU Time

The procedure for computing the Borda Count ranking scales linearly in the number of pilots to be ranked, and in practice has negligible running time for all problems. The run time of the branch and bound procedure for computing the Kemeny rankings scales exponentially in problem size. However, for 257 out of the 343 competitions we examined, the branch and bound procedure was able to find an optimal Kemeny ranking with no search at all.

Propagation rules and good lower bounds in the branch and bound procedure are sufficient to find many Kemeny rankings without search. In such cases, the CPU time required to find the Kemeny ranking is negligible (less than 0.1 seconds). The longest CPU time required to find a Kemeny ranking was 16 seconds for a problem with 25 pilots and significant disagreement between the judge's rankings. 97% of the computations completed in one tenth of a second or less[3]. The distribution of CPU times to find a Kemeny ranking is presented in Figure 1.

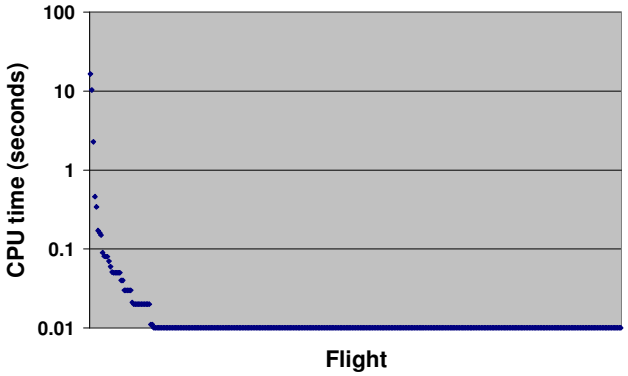

Figure 1: Distribution of the CPU time required to find the Kemeny ranking for all competitions ranked.

4.2 Kemeny Distance

Figure 2 presents the distribution of the normalized Kemeny distances for the competitions ranked by the Kemeny Rule. The normalized Kemeny distance *K'* scales the Kemeny distance to lie between 0 (where all judges agree with each other) and 1 (maximum disagreement between judges). It is calculated from the Kemeny distance *K* for a competition with *n* pilots and *j* judges in the following way:

$$K' = K / jn(n-1).$$

From this chart we can see that there was some disagreement between the judges for almost all the competitions ranked.

[3] The computations were made using an IBM T41 ThinkPad® computer.

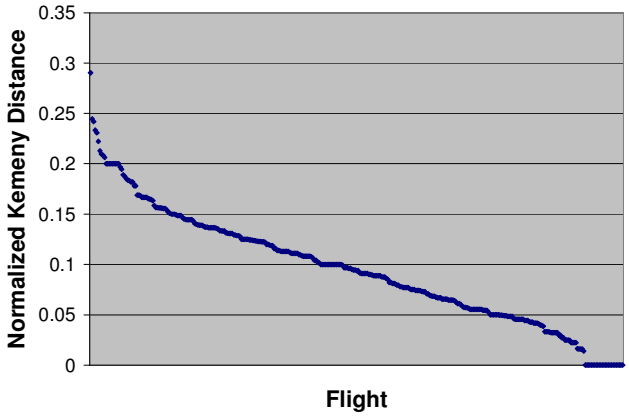

Figure 2: Distribution of normalized Kemeny distance for Kemeny rankings of all competitions.

4.3 Multiple solutions

We found multiple solutions with respect to the Kemeny ranking in 18 out of the 343 flight competitions we looked at. Figure 3 presents details regarding the distribution of these multiple solutions, for the competitions where multiple solutions were found.

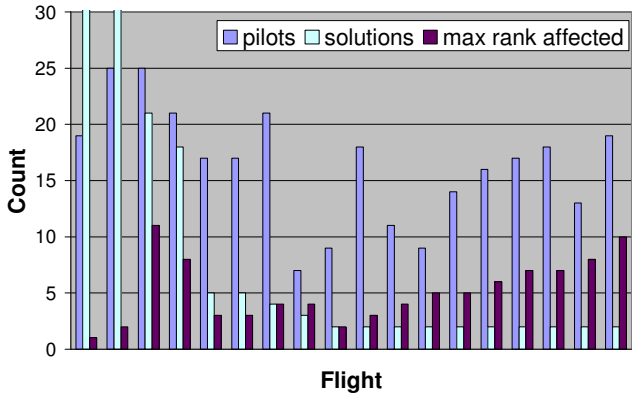

Figure 3: Multiple solutions for Kemeny rankings.

For each competition we present in the figure a histogram showing the number of pilots in the competition, the number of Kemeny rankings found and the highest rank at which the multiple solutions presented different candidates for that rank. For instance, for the third flight in the chart where there were 25 pilots we found 21 Kemeny ranking solutions. All solutions agreed on the ranking of the first ten pilots, but there were some disagreements on the rankings of the pilots ranked 11th and lower.

The chart has been scaled for clarity: for the first two (leftmost) flights presented in the chart, the number of solutions found was 99 and 101 respectively. We found disagreements among the multiple solutions for one first place ranking, four second place rankings, and five third place rankings.

4.4 Ties

Positional rank aggregation methods such as the Borda Count have the potential to declare ties for certain places. Ties can be problematic for sports competitions when they occur in the rankings for first, second or third places. The mean and TBLP methods currently in use by the IAC seldom produce ties. The judges' numerical grades for each pilot are usually different enough that these methods almost never produce the same aggregated score for two pilots.

We compared the number of ties for each place produced by the Borda Count ranking for each flight to the number of pilots ranked for each place by one or more rankings found by the Kemeny Rule. When the Kemeny Rule finds just one ranking, then there are no ties for any of the places. When there are multiple Kemeny rankings, there may be disagreement among these rankings over which pilots are ranked in certain places. Although this is not directly comparable to a tie in the Borda Count ranking, it provides a basis of comparison for the two methods in terms of the practical necessity that a rank aggregation method for sports competitions should produce as unambiguous a ranking as possible.

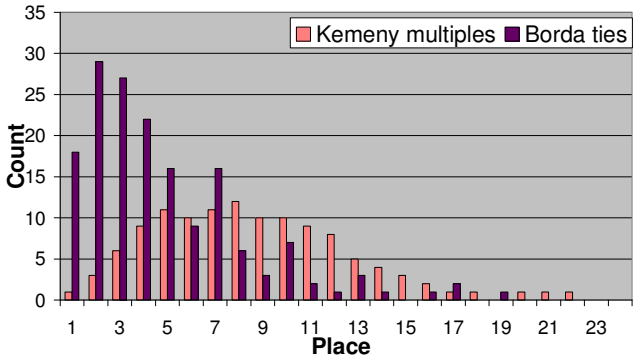

Figure 4: A comparison of the number of pilots ranked in each place by each of the rank aggregation methods.

Figure 4 presents a histogram showing for each place in each competition, the number of times the Kemeny Rule produced multiple candidates to be ranked in that place, and the number of times the Borda Count produced ties for that place. Overall the Borda Count produced ties more often than the Kemeny Rule, especially for the critical first three places. The Kemeny Rule produced only one tie for first place out of the 343 competitions, whereas the Borda Count found ties for first place in 18 of these competitions.

5 Three examples

We now present a more detailed discussion of three flights where we observed interesting behavior from the ranking methods.

5.1 Example 1

The first example ranks a 25 pilot competition. ([Flight 693] gives a URL for the data for this competition.) There was a significant amount of disagreement between the judges in the ranking of this competition. For instance, one pilot was ranked first by one judge, fourth by another, and no higher than tenth by the remaining three judges.

The computation of the Kemeny rankings required almost 16 CPU seconds (this was the longest of all of the Kemeny Rule computations). The Borda Count produced a tie for the first place ranking. The Kemeny Rule computation found 99 possible Kemeny rankings. There was agreement among all of these 99 rankings for the pilot to be ranked in first place; however, this first place pilot in the Kemeny ranking was not ranked first by any of the individual judges. There was disagreement concerning the second and third place rankings between the multiple Kemeny solutions.

5.2 Example 2

In the second example [Flight 574], which ranked 19 pilots, there was also a significant amount of disagreement between the judges. The Kemeny Rule computation found 101 distinct Kemeny orders. Different Kemeny orders ranked different pilots for first place: in total four pilots were ranked in first place among the multiple orders. Furthermore, none of the pilots which were ranked first place by the Kemeny Rule were ranked first by the Borda Count ranking. Only the last five places were the same in every Kemeny order.

5.3 Example 3

The first two examples we have given are fairly extreme cases where the Kemeny order produced many possible rankings. The more usual case for the data was that the judges were in much greater agreement concerning their rankings of the pilots. In these cases, applying the Kemeny Rule resulted in a unique ranking (for example, see [Flight 437].) The Kemeny Rule computation found a unique ranking in 18 of these cases where the Borda method tied a ranking for first place between two pilots. There was a clear Condorcet winner in all of these cases.

6 Discussion

6.1 Comparison with absolute scores

Rank aggregation methods give no indication regarding the absolute performance of each pilot. They do not tell a pilot how well the judges thought they were flying in an absolute sense, only how well they flew relative to other pilots. They also do not give any indication of how much better the first ranked pilot flew that the second ranked pilot.

We believe that the raw scores and mean values are useful for telling an individual pilot how well he or she flew their sequence. The raw scores show which figures graded best and which figures need the most improvement. The overall mean gives a rough metric of absolute performance quality.

6.2 Dealing with multiple solutions

The presence of multiple solutions for some flights raises the need for some method for producing a single aggregate ranking in those cases. Some methods for resolving multiple solutions have been studied in [Truchon 1998]. This is an area for further study, but we have briefly explored several alternatives:

1. Select among the Kemeny rankings one that minimizes the maximum broken majority vote.
2. Combine the solutions into a partial order placing each item at the best rank where it occurs in any of the solutions.

Use of any of these methods will require further theoretical justification. Some of them seem promising based on this data set.

7 Conclusions

We have investigated the feasibility of using two rank aggregation methods from Social Choice theory, the Borda Count and the Kemeny Rule, to determine the outcome of aerobatic flight competitions. One advantage of both methods studied, compared with the current system in use (TBLP), is that the candidates being ranked may easily understand and verify the outcomes of the competitions with respect to these rank aggregation methods.

The Borda Count, a positional rank aggregation scheme, is simple to implement; however we believe that it would not be accepted by the flight aerobatics community due to the frequency with which it produces ties. Furthermore, it does not address the very real concerns of manipulation.

The Kemeny Rule, a majority vote aggregation scheme, is intuitively clear. Also, due to the NP-hardness of determining its outcome, it is more resistant to manipulation by individual judges than the Borda Count. The NP-hardness of the Kemeny Rule computation has not so far resulted in run times problems for the size of competitions involved in this study.

One difficulty remaining with use of the Kemeny Rule is the presence of multiple solutions for a small number of competitions where there is significant disagreement among the judges. In further work we plan to investigate methods for resolving this issue.

The IAC are considering using the Kemeny Rule for ranking flight competitions in 2006.

References

[Borda. 1781] Borda. Mémoire sur les élections au scrutin. In *Histoire de l'Académie Royale des Sciences*, 1781.

[Cohen et. Al. 1999] Cohen, W.; Schapire, R.; and Singer, Y. Learning to order things. *Journal of Artificial Intelligence Research* 10:213-270, 1999.

[Condorcet. 1785] Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendue à la pluralité des voix. In *Paris: Imprimerie royale,* 1785. Also reproduced in Condorcet, Sur les élections et autres textes, edited by O. de Bernon, Fayard, 1986.

[Davenport and Kalagnanam 2004] Davenport, A. and Kalagnanam, J. A Computational Study of the Kemeny Rule for Preference Aggregation. *Proc. 19th National Conference on Artificial Intelligence (AAAI 2004).*

[Diaconis 1988] Diaconis, P. *Group representation in probability and statistics.* IMS Lecture Series 11, Institute of Mathematical Statistics, 1988.

[Dwork et. al. 2001] Dwork, C.; Kumar, R.; Naor, M.; and Sivakumar, D. Rank aggregation methods for the web. In *Proc. 10th WWW*, 613-622. 2001.

[Flight 437] Third flight of the Advanced category at Casa Grande, AZ. March, 2005.
`http://www.research.ibm.com/cr_aerobatics/flight437.html`

[Flight 574] Second flight of the Sportsman category at Paso Robles, CA. June, 2005.
`http://www.research.ibm.com/cr_aerobatics/flight574.html`

[Flight 693] Third flight of the Sportsman category at Delano, CA, September, 2005.
`http://www.research.ibm.com/cr_aerobatics/flight693.html`

[Kemeny 1959] Kemeny, J. Mathematics without numbers. *Daedalus* 88:571-591. 1959.

[TBLP, 2005] Regulations for the Conduct of International Aerobatic Events. *Fédération Aéronautique Internationale*. `http://www.fai.org/aerobatics/documents`

[Truchon 1998] Truchon, M. Figure skating and the theory of social choice. Technical Report Cahier 98-16, Centre de Recherche en Economie et Finance Appliquées, Université Laval, Canada. 1988.

[Young 1974] Young, H. An axiomatization of Borda's rule. *Journal of Economic Theory* 9:43-52. 1974.