ELSEVIER

European Journal of Operational Research 130 (2001) 576–587

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

Comparison of three multicriteria methods to predict known outcomes

David L. Olson *

Department of Information and Operations Management, Texas A&M University, College Station, TX 77843-4217, USA

Received 1 March 1999; accepted 2 September 1999

Abstract

Major approaches to selection decisions include multiattribute utility theory and outranking methods. One of the most frustrating aspects of research in the relative performance of these methods is that data where the final outcome is known is not available. In the US, a great deal of effort has been devoted to statistically recording detailed performance characteristics of major league professional baseball. Every year there has been two to four seasonal competitions, with known outcome in terms of the proportion of contests won. Successful teams often have diverse characteristics, emphasizing different characteristics. SMART, PROMETHEE, and a centroid method were applied to baseball data over the period 1901–1991. Baseball has undergone a series of changes in style over that period, and different physical and administrative characteristics. Therefore the data was divided into decades, with the first five years used as a training set, and the last five years used for data collection. Regression was used to develop the input for preference selection in each method. Single-attribute utilities for criteria performance were generated from the first five years of data from each set. Relative accuracy of multicriteria methods was compared over 114 competitive seasons for both selecting the winning team, as well as for rank-ordering all teams. All the methods have value in supporting human decision making. PROMETHEE II using Gaussian preference functions and SMART were found to be the most accurate. The centroid method and PROMETHEE II using ordinal data were found to involve little sacrifice in predictive accuracy.

Keywords: Multiple criteria analysis; Decision theory

1. Introduction

One of the difficulties in comparing the many multicriteria methods available for analysis is that there is rarely any way to check the accuracy of the methods. This study applies three multicriteria methods (SMART, PROMETHEE, and a centroid method) to a multicriteria environment where measured statistics and known outcomes are present. At the same time, it must be understood that these methods have different relative advantages in dealing with other types of data, including ordinal data, and cases where not all data is objective. The intent is not to establish dominance of any one

* Tel.: +1-409-845-2254; fax: +1-409-845-5653.
E-mail address: dolson@tamvm1.tamu.edu (D.L. Olson).

0377-2217/01/$ - see front matter
PII: S0377-2217(99)00416-6

method, but rather to compare relative accuracy in a case of known outcome.

The data used concerns annual competitions in American major league professional baseball. This sport has been closely measured for the entire 20th century, and offers many data points over a number of criteria. Here we use five criteria that are often used to describe the abilities of baseball teams. This data is used to measure the relative accuracy of multicriteria methods, in this case in explaining the outcome of each competition. For each decade, the first five years are used to develop model parameters. The resulting model is then tested on the last half of that decade. The reason for building a model each decade is that baseball style changes quite dramatically over the years. While the models could be useful in a predictive sense to gambling, an industry that is large but neither open nor socially uplifting, that is by no means the intent. The comparative study is intended to demonstrate how the three methods work, and to focus on relative features of each method.

2. Multiattribute methods

This study will use SMART, a simple linear model reflecting multiattribute utility theory and an outranking method (PROMETHEE II using cardinal criteria). Two other methods that rely on less complete input data are tested for comparative purposes. One is PROMETHEE II with data analyzed using an ordinal criterion. The other is the centroid method is a version of SMART using swing weighting, using only ordinal input for relative weight estimation.

2.1. Smart

Multiple attribute utility theory (von Neumann and Morgenstern, 1947) can be implemented by the linear model

$$\text{value}_j = \sum_{i=1}^{k} w_i s_{ij},$$

where for each alternative j, value is measured as the weighted sum of measures s_{ij} for this alternative on each of the i criteria, weighted by the relative importance w_i which reflects both criterion importance and measurement scale. Regression output can be applied directly in this model.

To demonstrate with the 1906 season, using the National League (consisting of eight teams), the value function for each team is obtained by using the regression model over the data period 1901–1905 for weights w_i, and applying this to the scores s_{ij} for each of the eight teams. Data for the eight teams on each criterion are presented in Table 1.

The model ranking is quite close to the actual outcome, although there are two switches (second and third place, and fifth and sixth place). Possible explanations include that there might be other criteria that are not included in the model, as well as intangible factors such as better management, better team cooperation, or better performance in clutch situations. This model reflects MAUT data, where weights consider both scale of data and criterion importance.

Table 1

1906 NL	Percentage	Place	Hitting	Power	Speed	Fielding	Pitching	Value
Intercept	−3.4938	*Weights:*	*3.4058*	*0.0018*	*0.0003*	*3.4845*	*−0.0984*	
Chicago	0.763	1	0.262	20	283	0.969	1.75	0.7363 = 1
New York	0.632	2	0.255	15	288	0.963	2.49	0.6117 = 3
Pittsburgh	0.608	3	0.261	12	162	0.964	2.21	0.6141 = 2
Philadelphia	0.464	4	0.241	12	180	0.956	2.58	0.4880 = 4
Brooklyn	0.434	5	0.236	25	175	0.955	3.13	0.4345 = 6
Cincinnati	0.424	6	0.238	16	170	0.959	2.69	0.4809 = 5
St. Louis	0.347	7	0.235	10	110	0.957	3.04	0.3979 = 7
Boston	0.325	8	0.226	16	93	0.947	3.14	0.3272 = 8

The Simple MultAttribute Rating Technique (SMART_ – Edwards, 1977; von Winterfeldt and Edwards, 1986; Edwards and Barron, 1994) is a simplified version of MAUT, where scores are standardized to a 0–1 scale (with 0 representing the worst expected performance on a given criterion, and 1 representing the best expected performance). Weights in this case would not reflect scale, which was eliminated by the standardization. Standardized scores are given in the Table 2, and the regression model run on these standardized scores versus the winning percentage (not standardized). This yielded a different regression formula, applied to the standardized scores as shown.

The regression model output yields the weights given in Table 2, with precisely the same fit and *t*-scores (except for the intercept). The transformed values yielded exactly the same value scores. This implies that if scores are standardized to a 0–1 scale (or any other scale) consistent across all variables, the resulting regression coefficients will change to reflect this change of scale, and the final dependent variable values obtained will be identical. While MAUT allows greater flexibility in preference trade-off functions, if a linear MAUT model is used, transformation as applied in SMART yields precisely the same result. In MAUT, weights reflect both scale and importance. In SMART, scales are transformed to a common basis, so weights reflect importance. Note that this study does not reflect the psychological aspects of relative weight, where if all of the alternatives are very close in performance on a particular criterion, there is a tendency to place less emphasis on that criterion.

2.2. Outranking methods

The outranking method (Roy (1971, 1978)) PROMETHEE (Brans and Vincke, 1985) utilizes a function reflecting the degree of advantage of one alternative over another, along with the degree of disadvantage that same alternative has with respect to the other alternative it is compared against. Data are input in a spreadsheet form as shown in Table 3.

PROMETHEE requires that weights for each criterion be entered, and that criteria types be selected. The weights are meant to be rough indications of relative importance. Here we have used the proportions obtained from the standardized regression used in the SMART analysis (from which scale of measure was removed from weights, leaving only relative importance), and normalized them so that they add to 1.0. In PROMETHEE, when the Type 1 criterion is used, only relative advantage matters. When Type 6 (Gaussian distribution based on the standard deviation, where small differences have little importance, but importance increases following the normal distribution as differences increase) is used, differences play a major role in establishing outranking relationships. There are six options allowing the user to express meaningful differences by minimum gaps between observations. The simplest criterion type is I, which calculates the outgoing flow by identifying the proportion of the weights of criteria where the base alternative has advantage over the other alternatives, and incoming flow as the proportion of weights of criteria where the base alternative has a disadvantage relative to other alternatives. Total flow is outgoing flow minus

Table 2

1906 NL	Percentage	Place	Hitting	Power	Speed	Fielding	Pitching	Value
Intercept	0.1107	*Weights:*	*0.2384*	*0.0755*	*0.0660*	*0.1464*	*0.2332*	
Chicago	0.763	1	0.5429	0.3488	0.9579	1.0238	1.1013	0.7363 = 1
New York	0.632	2	0.4429	0.2326	0.9842	0.8810	0.7890	0.6117 = 3
Pittsburgh	0.608	3	0.5286	0.1628	0.3211	0.9048	0.9072	0.6141 = 2
Philadelphia	0.464	4	0.2429	0.1628	0.4158	0.7143	0.7511	0.4880 = 4
Brooklyn	0.434	5	0.1714	0.4651	0.3895	0.6905	0.5190	0.4345 = 6
Cincinnati	0.424	6	0.2000	0.2558	0.3632	0.7857	0.7046	0.4809 = 5
St. Louis	0.347	7	0.1571	0.1163	0.0474	0.7381	0.5570	0.3979 = 7
Boston	0.325	8	0.0286	0.2558	−0.0421	0.5000	0.5148	0.3272 = 8

Table 3

		C1	C2	C3	C4	C5
	Criterion	Hitting	Power	Speed	Fielding	Pitching
	Min/Max	Max	Max	Max	Max	Min
	Type	1	1	1	1	1
	Weight	0.31	0.10	0.09	0.19	0.31
A1	Chicago	0.262	20	283	0.969	1.75
A2	New York	0.255	15	288	0.963	2.49
A3	Pittsburgh	0.261	12	162	0.964	2.21
A4	Philadelph	0.241	12	180	0.956	2.58
A5	Brooklyn	0.236	25	175	0.955	3.13
A6	Cincinnati	0.238	16	170	0.959	2.69
A7	St. Louis	0.235	10	110	0.957	3.04
A8	Boston	0.226	16	93	0.947	3.14

incoming flow. Type VI involves more complex calculations, reflecting the degree of relative advantage. To demonstrate, we use Type I calculations (see Table 4).

To demonstrate calculation of preference indices, Chicago is better than New York in hitting, power, fielding and pitching, with respective weights of 0.31, 0.10, 0.19 and 0.31, totaling an outgoing flow of 0.91 for Chicago over New York. New York is better in stolen bases, with a weight of 0.09, giving a preference index of 0.09 for New York over Chicago.

Leaving preference flows are simply the average of outgoing flows for each alternative. The sum of the Chicago row in Table 4 is 6.81, versus 7 other alternatives. The average, $6.81/7 = 0.97286$, which is the Leaving Flow for Chicago. The Entering Flow for Chicago is the average of the column for that alternative, or $0.19/7 = 0.027144$. The Net Preference Flow for Chicago is the Leaving Flow minus the Entering Flow, or $0.97286 - 0.02714 = 0.94571$ (see Table 5).

2.2.1. PROMETHEE I

The PROMETHEE I method is designed to provide a partial ranking. Partial rankings focus on the best choice, not on a complete ranking. Pairs of alternatives are categorized by preference, indifference or incomparability. Two rankings are obtained. The positive outranking flows are used as the basis for the first ranking. Preference requires outflow of the base alternative to be greater than the outflow of the other alternative. The alternatives are indifferent in value if the outflows are equal. Preference in the second ranking requires inflow of the base alternative to be strictly less than inflow of the other alternative. The alternatives are indifferent in value if inflows are equal.

The PROMETHEE I partial ranking is the intersection of these two rankings. For alternative *a*

Table 4

PI preference indices

	Actions	A1	A2	A3	A4	A5	A6	A7	A8
A1	Chicago	0.00	0.91	1.00	1.00	0.90	1.00	1.00	1.00
A2	New York	0.09	0.00	0.19	0.81	0.90	0.71	0.81	0.90
A3	Pittsburgh	0.00	0.62	0.00	0.62	0.81	0.62	0.81	0.90
A4	Philadelph	0.00	0.00	0.09	0.00	0.90	0.71	0.81	0.90
A5	Brooklyn	0.10	0.10	0.19	0.10	0.00	0.19	0.50	0.81
A6	Cincinnati	0.00	0.10	0.19	0.10	0.81	0.00	0.81	0.90
A7	St. Louis	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.90
A8	Boston	0.00	0.10	0.10	0.10	0.00	0.00	0.10	0.00

Table 5

Preference flows			
Actions	Leaving	Entering	Net
Chicago	0.97286	0.02714	0.94571
New York	0.63000	0.26143	0.36857
Pittsburgh	0.62571	0.25143	0.37429
Philadelph	0.48714	0.39000	0.09714
Brooklyn	0.28429	0.68857	−0.40429
Cincinnati	0.41571	0.46143	−0.04571
St. Louis	0.20000	0.69143	−0.49143
Boston	0.05714	0.90143	−0.84429

to outrank alternative b, a must have a better positive flow than b, or a must have a smaller negative flow than alternative b, while it is at least equal on the other flow. The alternatives would be indifferent if both positive and negative outflows were equal. Otherwise, the alternatives are considered incomparable.

In the above example, Chicago has a better leaving flow (outflow, or positive flow) than any other alternative, and a lower entering flow (inflow, or negative flow) than any other alternative, and therefore Chicago outranks all of the other seven alternatives.

New York has a better leaving flow than any other remaining alternative, but Pittsburgh has a better entering flow than New York. Therefore, these two alternatives are incomparable. However, both outrank all of the other five alternatives.

Philadelphia outranks all of the other five alternatives. Cincinnati, which actually finished sixth, has both a better leaving and entering flow than does Brooklyn, which actually finished fifth. Therefore, Cincinnati outranks Brooklyn. Otherwise, the rankings are in order of finish.

The PROMETHEE I partial order for these eight alternatives is therefore:

1 Chicago	1 actual
2 New York & Pittsburgh	2 & 3 actual
4 Philadelphia	4 actual
5 Cincinnati	6 actual
6 Brooklyn	5 actual
7 St. Louis	7 actual
8 Boston	8 actual

PROMETHEE I partial order based on the Type VI criterion are only slightly different:

1 Chicago	1 actual
2 New York & Pittsburgh	2 & 3 actual
4 Philadelphia & Cincinnati	4 & 6 actual
6 Brooklyn	5 actual
7 St. Louis	7 actual
8 Boston	8 actual

Because the partial order can include a number of incomparable pairs of alternatives, it will not guarantee a complete ranking of alternatives. Therefore, it is evaluated on the basis of first place (and may have multiple alternatives ranked as either indifferent or incomparable in first place). In this case PROMETHEE I analysis identified the correct first-place team with both Type I and Type VI criteria.

2.2.2. PROMETHEE II

Net flows are the basis for the PROMETHEE II ranking. In this case, the ranking is:

1 Chicago	1 actual
2 Pittsburgh	3
3 New York	2
4 Philadelphia	4
5 Cincinnati	6
6 Brooklyn	5
7 St. Louis	7
8 Boston	8

This ranking (the same in this case for both Type I and Type VI criteria) is exactly the same as the ranking obtained emulating the MAUT method, with only two rank reversals among the eight positions.

2.3. The centroid method

The centroid method uses the same overall model, using ordinal input information. The core idea is to minimize the maximum error by finding the weights in the center of the region bounded by decision-maker ordinal ranking of factors. Olson and Dorai (1992) and Barron and Barrett (1996) applied the idea of the centroid to preference input based entirely on ordinal preference ranking of

criteria (within which single-attribute utility ranges could be considered). While continuous weight estimation methods, such as multiattribute utility theory models or analytic hierarchy models, would be expected to be more accurate estimators if preference input were accurate, the centroid approach is based on more reliable input, and is less subject to the errors introduced by inaccurate weight assessment. Flores et al. (1992) found that the centroid approach was useful when there were four or more criteria being considered, when criteria were close in relative importance, and when time available for analysis was short.

The only difference between the centroid method as implemented here and SMART is that the weights for the five measures are calculated differently. The idea behind the centroid approach is to identify the weights with the minimum–maximum error from the extreme points implied by the order of relative importance of the criteria. Given that scale has been eliminated by the normalization of measures in the SMART method above, the weights here would reflect relative importance of the criteria. For the 1901–1905 era, the relative weights identified from regression on the normalized data were:

Hitting	0.314	Rank 1
Home runs	0.099	Rank 4
Stolen bases	0.087	Rank 5
Fielding average	0.193	Rank 3
Earned run average	0.307	Rank 2

The centroid is calculated by taking the mean of the extreme sets of weights. In this case, these extremes are as shown in Table 6.

These weight estimates are then applied exactly as they were in SMART (see Table 7).

These values differ from those obtained by SMART weights, of course, but in this case yield the same rank order for all eight teams. Input data on relative weights with the centroid approach requires only ordinal input, expected to be more robust than continuous data obtained from SMART. Here, we have precise measures, so the centroid approach is expected to be less accurate than SMART in predicting rank order. Edwards and Barron (1994) reported very little loss in accuracy in moving from SMART to the centroid approach (SMARTER), however. We will compare the centroid results with SMART results to compare relative accuracy loss due to the less precise input.

Table 6

	Hitting	Home runs	Stolen bases	Fielding	Earned runs
1	1	0	0	0	0
2	1/2	0	0	0	1/2
3	1/3	0	0	1/3	1/3
4	1/4	1/4	0	1/4	1/4
5	1/5	1/5	1/5	1/5	1/5
Average	0.4567	0.09	0.04	0.1567	0.2567

Table 7

1906 NL	Percentage	Place	Hitting	Power	Speed	Fielding	Pitching	Value
		Weights:	*0.4567*	*0.09*	*0.04*	*0.1567*	*0.2567*	
Chicago	0.763	1	0.5429	0.3488	0.9579	1.0238	1.1013	0.7607 = 1
New York	0.632	2	0.4429	0.2326	0.9842	0.8810	0.7890	0.6031 = 3
Pittsburgh	0.608	3	0.5286	0.1628	0.3211	0.9048	0.9072	0.6435 = 2
Philadelphia	0.464	4	0.2429	0.1628	0.4158	0.7143	0.7511	0.4469 = 4
Brooklyn	0.434	5	0.1714	0.4651	0.3895	0.6905	0.5190	0.3771 = 6
Cincinnati	0.424	6	0.2000	0.2558	0.3632	0.7857	0.7046	0.4328 = 5
St. Louis	0.347	7	0.1571	0.1163	0.0474	0.7381	0.5570	0.3427 = 7
Boston	0.325	8	0.0286	0.2558	−0.0421	0.5000	0.5148	0.2449 = 8

3. Data

Each year there is a professional baseball season. The length of this season has increased from 140 scheduled games per team per year in 1901 to 154 games per year for most of the century, and currently is up to 162 games per year. Data was taken from Thorn and Palmer (1995). There were some shortened seasons: in 1918 due to World War I, and in 1981 due to a professional baseball players' strike. The 1918 season was played in one short season, and was included in the analysis. The 1981 season was split into two halves, and thus the team with the best season record did not necessarily win. (In fact, St. Louis and Cincinnati had the best overall records in their divisions, but won neither of the split seasons, so were not first-place finishers.) For this reason, the 1981 season was not included. There were eight teams per league until the 1960s, when expansion led to 10 teams per league, and on to four divisions of six or seven teams each. Currently this has been expanded to six divisions. Despite these changes, for the most part major league baseball has been quite stable.

There is detailed statistical data available for each team (indeed for each player) each season on many measures. The essence of baseball is that it involves offensive, defensive, and pitching team performance, and the better measures for a team, the better expected winning performance. Offensive performance is primarily represented by batting average (base hits per chances), but is supplemented by a power measure (number of home runs) and a measure of speed (number of stolen bases). Fielding is measured by the number of successful fielding events divided by the total number of chances. Pitching effectiveness is measured by the number of earned runs given up per nine innings pitched (earned run average, or ERA). While there are many other specific measures available that capture other details of team performance, these five measures are commonly used and capture the majority of relative team differences. In this study, these five measures are used as criteria that differentiate between the teams that competed against each other each season.

While professional baseball has been quite stable, there have been changes in the relative emphasis. In the first two decades of the 20th century, there were notably fewer home runs and many more stolen bases than was the case thereafter. For the next four decades, the emphasis was on power rather than speed. This has been followed by a reemergence of higher stolen base values. Therefore, the data has been treated by decade. Following the spirit of expert system methodology, the first five years of each decade have been used as the basis for a regression model to measure the relationship between each of the five independent variables and team winning percentage (the training phase). This regression model was then used to determine preference input for the multiattribute methods applied to the last five years of each decade (the predictive phase). One adjustment to this process was used. Because the 1981 season was much shorter than the others, it was held out, and the regression model was developed over the period 1982–1986, and this result was applied to the 1987–1991 seasons.

4. Data features

The methodology used provides a difficult test for the multiattribute models. Team performance is expected to be correlated to relative performance on the five criteria measured, but also is expected to be a function of team cooperation and management. There are a number of intangible features that might elude measurement.

Correlations of the data are appended. The correlation of earned run average (era, or runs given up by pitching per nine innings) with winning percentage is stronger than for any other criterion except hitting in the decade 1941–1950. The correlation between winning percentage and pitching declined with time after its peak in the decade 1951–1960. Hitting and fielding have always had strong correlation with winning percentage, with hitting having the strongest correlation of the two. Power (represented by home runs, or hr) had a fairly strong correlation from 1921 to 1980, but dropped over the period 1982–1991. Stolen bases (sb) have had low correlation with winning percentage except in the period 1951–1960.

Among criteria, the relationship between power (hr) and hitting has been quite high, with the exception of the period 1982–1991. In three of the decades (the 1930s, 1950s and 1970s) power has had stronger relationship with winning percentage than with hitting. Fielding had a strong correlation with pitching (era) in most decades, but had a stronger relationship with winning except in the initial decade. Therefore, while there is some overlap in criteria, they reflect different aspects of team performance.

The regression models reflected some changes over time with respect to coefficients for specific criteria. Hitting's β coefficient was up to 3.95 in the period 1941–1945, and dropped below three for three of the four decades from the 1950s to 1980s. Home runs and stolen bases had very low β coefficients, in great part due to the difference in measurement scale from hitting and fielding, which were measured in proportions. The β coefficient for era was smaller due to its measure being well over one hundred times greater than that of hitting.

As far as criteria significance, all were significant at the 0.95 confidence level except for stolen bases (variable sb) in the 1920s and 1950s, and fielding in the 1940s. The other three measures were always significant beyond the 0.95 level of confidence.

The fit of these models is relatively strong, with the r-square measure being around 0.8. The r-square measure was stronger until the 1970s. The decline in model fit may well have been due to expansion of the number of teams, which were split into more competitive groupings of slightly fewer teams than had been the case over the period 1901–1960.

5. Relative accuracy

The methods considered were compared over the entire data set available, ranging from 1901 to 1991 in nine decades. This involved a total of 114 cases, with case size ranging from six alternatives to ten alternatives. PROMETHEE I generates a partial order, with potential ties for each position. If the actual first place team was among those teams ranked first, the method was given credit for 1/(the number of teams in the first rank). The fit for selection of the team to finish first is resulted in Table 8.

These figures show almost no difference between the multiattribute methods shown. None of the methods reported here are terribly reliable, however, as all are roughly capable of predicting 70% of the winning teams.

Another measure of accuracy is obtained by calculating the sum of absolute differences between the actual ranking and the model ranking. This measure reflects the ability of the method to rank order, not just select the preferred alternative.

Table 8

Seasons	Teams	SMART	PROMI (VI)	PROMII (VI)	PROMII(I)	Centroid
1906–1910	8	7/10	6.83/10	7/10	7/10	7/10
1916–1920	8	7/10	8/10	9/10	5/10	8/10
1926–1930	8	4/10	4.33/10	4/10	4/10	5/10
1936–1940	8	7/10	7/10	7/10	7/10	7/10
1946–1950	8	7/10	7.5/10	7/10	8/10	7/10
1956–1960	8	7/10	7.33/10	7/10	6/10	6/10
1966–1968	10	6/6	4/6	5/6	4/6	4/6
1969–1970	6	6/8	6.17/8	6/8	7/8	6/8
1976–1980	6	9/12	8.83/12	10/12	11/12	10/12
1976–1980	7	6/8	3.5/8	5/8	5/8	6/8
1987–1991	6	7/10	7/10	7/10	8/10	6/10
1987–1991	7	6/10	6.33/10	6/10	6/10	6/10
Total		79/114	77.165/114	80/114	78/114	78/114
Ratio		0.693	0.677	0.702	0.684	0.684

Table 9

Seasons	Teams	SMART	PROMII (VI)	PROMII (I)	Centroid
1906–1910	8	53/10	52/10	51/10	57/10
1916–1920	8	65/10	60/10	70/10	61/10
1926–1930	8	46/10	52/10	64/10	54/10
1936–1940	8	50/10	54/10	65/10	53/10
1946–1950	8	60/10	58/10	64/10	60/10
1956–1960	8	59/10	57/10	71/10	65/10
1966–1968	10	74/6	78/6	78/6	73/6
1969–1970	6	29/8	27/8	34/8	29/8
1976–1980	6	37/14	30/14	35/14	31/14
1976–1980	7	17/6	24/6	27/6	31/6
1987–1991	6	44/10	42/10	44/10	63/10
1987–1991	7	61/10	67/10	76/10	73/10
Total		595	601	679	650

PROMETHEE I yields a partial order, which is highly problematic in estimating rank-order. Results for the other methods are presented in Table 9.

This data indicates an advantage in ranking accuracy on the part of SMART and PROMETHEE II using Gaussian data over the methods using ordinal input.

Table 10 sorts out accuracy using Spearman's rho statistic. The difference in ranks for each season was obtained, with each of the differences squared and summed. For example, the results for the 1906 National League season for all methods demonstrated are shown in Table 10.

Olds's (1939) limits at the 0.05 level were used to evaluate significance. These limits are a function of the number of items ranked. By these limits, a significant relationship is measured if the sum of squared differences no more than the following:

Six items	limit = 4
Seven items	limit = 12
Eight items	limit = 22
Ten items	limit = 56

Table 10

Actual rank	MAUT	Diff^2
1 Chicago	1	0
2 New York	3	1
3 Pittsburgh	2	1
4 Philadelphia	4	0
5 Brooklyn	6	1
6 Cincinnati	5	1
7 St. Louis	7	0
8 Boston	8	0
Sum diff^2		4

The number of cases of significant accuracy at these levels are given in Table 11.

These measures indicate that as the number of items ranked increased, the accuracy of the methods increased for the most part. The most accurate by this measure was PROMETHEE II using criterion type VI (Gaussian), which was significant in 110 out of 114 cases. SMART was significant in 95 cases, the centroid method 92, and PROMETHEE II using criterion Type I in 89 cases. The data here clearly indicates that it is much more difficult to be significant with fewer numbers of teams. PROMETHEE II using criterion Type VI was significant in all cases except for four involving only six teams.

Table 11

n	SMART	PROMII (VI)	PROMII (I)	Centroid
6	16 of 30	26 of 30	15 of 30	17 of 30
7	15 of 18	18 of 18	12 of 18	11 of 18
8	58 of 60	60 of 60	56 of 60	59 of 60
10	6 of 6	6 of 6	6 of 6	5 of 6

6. Conclusions

This study was intended to demonstrate how these three methods of preference modeling work on data with precise measures and known outcomes. The multiple attribute utility method and SMART using transformed measure scales (standardized to have the worst measure equal 0 and the best measure equal 1) were shown to be equivalent. Earlier phases of this study seemed to indicate that the PROMETHEE method was less precise than SMART, but final results clearly indicate that the two methods were very similar in accuracy. PROMETHEE models are more accurate (as expected) is more complete data is used. The criterion Type 6 uses the standard deviation of the data, and gives notably more accurate results than the ordinal input used with criterion Type 1. Using ordinal weight input with the centroid method, however, did not result in a great deal of lost accuracy relative to the SMART results.

Table 12
Correlations of winning percentage and five criteria by decade[a]

	pct	Hitting	hr	sb	Field	pct	Hitting	hr	sb	Field
	1901–1910					1951–1960				
pct	1.0000					1.0000				
Hitting	0.5373	1.0000				0.5952	1.0000			
hr	**0.2561**	**0.5309**	1.0000			0.4901	**0.3634**	1.0000		
sb	0.4855	0.2648	0.0020	1.0000		0.3194	0.1905	0.0476	1.0000	
Field	0.4547	−0.2442	−0.2107	0.1330	1.0000	0.5135	0.2515	0.1893	0.2588	1.0000
Era	−0.6032	0.1896	0.2395	**−0.3027**	**−0.7015**	−0.7616	−0.2181	−0.1744	**−0.3209**	**−0.4423**
	1911–1920					1961–1970				
pct	1.0000					1.0000				
Hitting	0.4338	1.0000				0.4852	1.0000			
hr	**0.1998**	**0.3434**	1.0000			0.3588	**0.4013**	1.0000		
sb	**0.2168**	0.0649	−0.1421	1.0000		**0.1278**	0.0483	**−0.3431**	1.0000	
Field	0.4971	0.1132	0.0721	**−0.3570**	1.0000	0.4370	0.0423	0.0862	0.0264	1.0000
Era	−0.6327	0.2427	0.1662	−0.1169	**−0.4389**	−0.5957	0.1664	0.2342	−0.2110	**−0.3630**
	1921–1930					1971–1980				
pct	1.0000					1.0000				
Hitting	0.6104	1.0000				0.5266	1.0000			
hr	0.3589	**0.4463**	1.0000			0.4428	**0.4074**	1.0000		
sb	**0.2833**	0.1082	−0.2040	1.0000		**0.2569**	0.2056	−0.1204	1.0000	
Field	0.3774	0.0855	0.0771	0.1159	1.0000	0.4532	0.1646	0.1994	0.0258	1.0000
Era	−0.6385	−0.0012	0.1628	**−0.3372**	−0.2744	−0.5585	0.1707	0.1235	−0.1092	−0.2941
	1931–1940					1982–1991				
pct	1.0000					1.0000				
Hitting	0.5466	1.0000				0.4332	1.0000			
hr	0.4875	**0.3931**	1.0000			**0.2144**	0.2895	1.0000		
sb	**0.2026**	0.2699	0.0927	1.0000		**0.1389**	−0.1357	−0.1067	1.0000	
Field	0.5797	0.1626	0.2135	−0.0613	1.0000	0.3683	0.1873	0.1460	−0.1414	1.0000
Era	−0.6473	0.0357	−0.0179	0.0894	**−0.4329**	−0.5204	0.1922	**0.4155**	−0.2018	−0.0618
	1941–1950									
pct	1.0000									
Hitting	0.6274	1.0000								
hr	0.3226	**0.4435**	1.0000							
sb	**0.0777**	−0.0649	**−0.3225**	1.0000						
Field	0.4742	**0.3228**	**0.4003**	−0.2636	1.0000					
Era	−0.6167	0.0091	0.2387	−0.2081	−0.1529					

[a] Bold characters indicate correlation with winning percentage of less than 0.3, and correlation between criteria greater than 0.3.

A last, but very important point, is that relative accuracy when accurate and complete data is available is only one factor in selecting a multiattribute model. Another very important aspect is the ability to collect accurate preference and alternative performance data. Sometimes, only general qualitative data is available. (There are also cases where precise values might be obtained, but they are not accurate.) In such cases, methods such as ZAPROS (Larichev and Moshkovich, 1991) are appropriate, and within the outranking methods, ELECTRE IV and PROMETHEE with criterion Type 1 have value. The outranking methods also have been proposed as means of incorporating decision-maker personal views into the model.

Appendix A

See Tables 12 and 13.

Table 13
Regression models by decade[a]

	Coefficients	Standard Error	*t* Stat	*P*-value	Coefficients	Standard Error	*t* Stat	*P*-value
1901–1905		R Square	0.8653		1951–1955	R Square	0.8661	
Intercept	−3.4938	0.7700	−4.5373	0.0000	−3.7090	1.6715	−2.2190	0.0296
Hitting	3.4058	0.3321	10.2562	0.0000	2.7463	0.5339	5.1436	0.0000
hr	0.0018	0.0006	3.1687	0.0022	0.0008	0.0001	5.6069	0.0000
sb	0.0003	0.0001	3.0428	0.0032	0.0001	0.0002	0.2621	**0.7939**
Field	3.4845	0.7655	4.5521	0.0000	3.8912	1.7267	2.2535	0.0272
Era	−0.0984	0.0112	−8.7483	0.0000	−0.1007	0.0088	−11.4308	0.0000
1911–1915		R Square	0.8089		1961–1965	R Square	0.8553	
Intercept	−3.0085	0.9894	−3.0407	0.0033	−2.5996	1.0520	−2.4711	0.0153
Hitting	3.1180	0.4028	7.7417	0.0000	3.0897	0.3090	9.9975	0.0000
hr	0.0010	0.0004	2.3790	0.0199	0.0007	0.0001	6.2184	0.0000
sb	0.0003	0.0001	2.3921	0.0193	0.0003	0.0001	2.0890	0.0395
Field	3.0988	1.0185	3.0424	0.0032	2.6844	1.0718	2.5047	0.0140
Era	−0.1139	0.0121	−9.4383	0.0000	−0.1116	0.0076	−14.7599	0.0000
1921–1925		R Square	0.8081		1971–1975	R Square	0.7913	
Intercept	−3.2951	1.0336	−3.1879	0.0021	−3.2974	0.9924	−3.3226	0.0012
Hitting	3.0708	0.3638	8.4414	0.0000	2.5892	0.2569	10.0771	0.0000
hr	0.0005	0.0002	2.9949	0.0037	0.0008	0.0001	6.2389	0.0000
Sb	0.0001	0.0002	0.9221	**0.3595**	0.0003	0.0001	3.7225	0.0003
Field	3.3793	1.0517	3.2131	0.0019	3.4224	0.9999	3.4226	0.0009
Era	−0.0983	0.0103	−9.5662	0.0000	−0.0891	0.0076	−11.6929	0.0000
1931–1935		R Square	0.8613		1982–1986	R Square	0.7815	
Intercept	−4.7854	1.2001	−3.9874	0.0002	−3.6375	1.0562	−3.4440	0.0008
Hitting	3.7867	0.4751	7.9697	0.0000	2.6553	0.2775	9.5684	0.0000
hr	0.0007	0.0002	4.6050	0.0000	0.0009	0.0001	9.4884	0.0000
Sb	0.0005	0.0002	2.7703	0.0071	0.0003	0.0001	4.8733	0.0000
Field	4.6357	1.2360	3.7506	0.0003	3.7769	1.0787	3.5012	0.0006
Era	−0.0832	0.0085	−9.8020	0.0000	−0.1058	0.0072	−14.7652	0.0000
1941–1945		R Square	0.8831					
Intercept	−1.6901	1.2335	−1.3701	0.1748				
Hitting	3.9510	0.3701	10.6769	0.0000				
hr	0.0007	0.0002	4.1884	0.0001				
Sb	0.0004	0.0002	2.2437	0.0278				
Field	1.5727	1.2474	1.2608	**0.2113**				
Era	−0.1201	0.0088	−13.6898	0.0000				

[a] Bold entries indicate probabilities of insignificance greater than 0.05.

References

Brans, J.P., Vincke, P., 1985. A preference-ranking organization method: The PROMETHEE method. Management Science 31, 647–656.

Edwards, W., 1977. How to use multiattribute utility measurement for social decisionmaking. IEEE Transactions on Systems, Man, and Cybernetics SMC 7 (5), 326–340.

Edwards, W., Barron, F.H., 1994. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. Organizational Behavior and Human Decision Processes 60, 306–325.

Flores, B.E., Olson, D.L., Wolfe, C., 1992. Judgmental adjustment of forecasts: A comparison of process and weighting schemes. International Journal of Forecasting 7, 421–433.

Larichev, O.I., Moshkovich, H.M., 1991. ZAPROS: A method and system for ordering multiattribute alternatives on the base of a decision-maker's preferences. All-Union Research Institute for Systems Studies, Moscow.

Olds, E.G., 1939. Distribution of sums of squares of rank differences for small samples. Annals of Mathematical Statistics, 9.

Olson, D.L., Dorai, V.K., 1992. Implementation of the centroid method of Solymosi and Dombi. European Journal of Operational Research 60 (1), 1–20.

Roy, B., 1971. Problems and methods with multiple objective functions. Mathematical Programming 1 (2), 239–266.

Roy, B., 1978. ELECTRE III: Un algorithme de classement fonde sur une representation floue des preferences en presence de criteres multiple. Cahiers du Centre Etudes Recheerche Operationelle 20, 3–24.

Thorn, J., Palmer, P. (Eds.), 1995. Total Baseball, fourth ed., Viking.

Von Neumann, J., Morgenstern, O., 1947. Theory of Games and Economic Behavior, second ed., Princeton University Press, Princeton, NJ.

Von Winterfeldt, D., Edwards, W., 1986. Decision Analysis and Behavioral Research. Cambridge University Press, New York.