2008

Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series

Christopher M. Rump
Bowling Green State University, cmrump@bgsu.edu

Journal of Quantitative Analysis in Sports

Volume 4, *Issue* 1 2008 *Article* 2

Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series

Christopher M. Rump*

*Bowling Green State University, cmrump@bgsu.edu

Data Clustering for Fitting Parameters of a Markov Chain Model of Multi-Game Playoff Series

Christopher M. Rump

Abstract

We propose a Markov chain model of a best-of-7 game playoff series that involves game-to-game dependence on the current status of the series. To create a relatively parsimonious model, we seek to group transition probabilities of the Markov chain into clusters of similar game-winning frequency. To do so, we formulate a binary optimization problem to minimize several measures of cluster dissimilarity. We apply these techniques on Major League Baseball (MLB) data and test the goodness of fit to historical playoff outcomes. These state-dependent Markov models improve significantly on probability models based solely on home-away game dependence. It turns out that a better two-parameter model ignores where the games are played and instead focuses simply on, for each possible series status, whether or not the team with home-field advantage in the series has been the historical favorite - the more likely winner - in the next game of the series.

KEYWORDS: baseball, statistics, probability, set partitioning

1 Introduction

This article examines best-of-7 game playoff series in which two teams play a series of games that ends as soon as one of the teams has earned 4 wins. Following Mosteller's (1952) seminal analysis of the World Series of baseball, most analysis of playoff series in sport employ a simple binomial model (truncated as in a negative binomial distribution as soon as a team has 4 wins) in which the favored team has a probability p of winning each game of the series. Noticing a preponderance of full 7-game World Series, Groeneveld and Meeden (1975) extended Mosteller's work to include the fairly high conditional probability that the trailing team after five games — whether favored or not — wins the next game, forcing a seventh game.

Bassett and Hurley (1998) extended the simple binomial model to a two-parameter model incorporating home and away winning probabilities p_H and p_A, respectively, for the team favored with home advantage, i.e., the team who plays host to four of the seven games in the series. Each game played at home is an independent Bernoulli trial with probability p_H for a win by the favored team. Likewise, each game played away is an independent Bernoulli trial with probability p_A for a win by the favored team.

Compared to the binomial model, this home-away model provides a very good fit for best-of-7 game playoffs in the National Basketball Association (NBA) (Rump, 2006b). In the National Hockey League (NHL), however, neither the binomial model nor the home-away model has any explanatory power. Slight modifications to the two-parameter home-away model into a state-dependent Markov chain that captures a few of the series-ending "surrender" effects makes all the difference (Rump, 2006a).

The application of a Markov chain model to sports competition is not novel. Earlier work included single Bernoulli-parameter models of best-of-7 game playoff series (Brunner, 1987), tennis matches (Kemeny and Snell, 1960; Sadovskiĭ and Sadovskiĭ, 1993), as well as a home-away model of a squash match (Broadie and Joneja, 1993), where in this case "home" and "away" indicate who is serving. However, these Markov chain implementations are a bit of "window dressing" for what are essentially simple random walks with homogeneous transition probabilities (Stewart, 1989).

Whereas identification of state dependencies in multi-game playoff series was done on an ad hoc basis in Groeneveld and Meeden (1975) and Rump (2006a), this paper proposes an efficient optimization routine for partitioning the state space into clusters of similar game-winning transition probabilities. We apply these techniques to Major League Baseball (MLB) playoff data.

2 Partitioning the Markov Chain State Space

Following Rump (2006a), we model game-to-game dependence in a playoff series via a Markov chain. In a best-of-$(2m-1)$ game series whereby the series ends when a team first wins m games, we denote the status of the series as (α-β) for $\alpha, \beta = 0, \ldots, m$ (excluding the case $\alpha = \beta = m$). Here α and β represent the number of games currently won by the favored team and their opponent, respectively. The transient states that indicate a series still in progress are the subset $\mathcal{S} = \{(\alpha\text{-}\beta) : \alpha, \beta = 0, \ldots, m-1\}$. The number of such transient states is $M = m^2$; the number of series-ending absorbing states (where either $\alpha = m$ or $\beta = m$) is $2m$. For a best-of-7 game series, for example, a team must win $m = 4$ games, yielding $M = 16$ transient and 8 series-ending absorbing states.

To simplify the notation, we index these transient states with a scalar parameter s, $s = 1, \ldots, M$. Let n_s be the number of games played in state s and w_s be the number of those games won by the favored team. A maximum likelihood estimate (MLE) of the probability that the favored team wins in state s is thus given by the relative frequency of wins, $p_s = w_s/n_s$ (Bhat and Miller, 2002).

Our goal is to partition the states into sequential clusters or strings (Vinod, 1969) of states sharing similar game-winning frequency. To facilitate this process, we assume these states are sorted in decreasing order of p_s, so that $j > i$ implies $p_j \leq p_i$. We desire to limit the number of clusters used, K, in order to avoid over-fitting the Markov model with too many parameters. For a given number of clusters, $K \leq M$, there are $\binom{M-1}{K-1}$ possible partitions of the M states. Since there are 8 multinomial absorbing state outcomes of our best-of-7 game playoff series, we will impose an upper limit of $K \leq 6$ parameter estimates and thereby retain at least one degree of freedom in our statistical fitting process. Hence, we intend to explore the set of all $\sum_{k=0}^{5} \binom{15}{k} = 4944$ possible partitions containing 6 or fewer clusters of the $M = 16$ states.

Letting c_{ij} denote a measure of dissimilarity within a cluster consisting of consecutive states i through $j \geq i$, Rao (1971) provided a dynamic programming formulation of the clustering problem. We instead create an integer linear programming version of the problem by defining binary decision variables x_{ij} that indicate whether ($= 1$) or not ($= 0$) a cluster consisting of consecutive states i through j is formed, $i = 1, \ldots, M$, $j = i, \ldots, M$. This entails a total of only $M(M+1)/2 = 137$ decision variables. For a desired number of clusters, K, the set partitioning problem can then be formulated as a binary integer programming (BIP) problem:

$$
\begin{aligned}
\min_{\boldsymbol{x}} \quad & \sum_{i=1}^{M}\sum_{j=i}^{M} c_{ij}x_{ij} \\
\text{subject to} \quad & \sum_{i=1}^{M}\sum_{j=i}^{M} x_{ij} = K \\
& \sum_{i=1}^{s}\sum_{j=s}^{M} x_{ij} = 1 \qquad\qquad s = 1,\ldots,M \\
& x_{ij} \in \{0,1\} \qquad\qquad i = 1,\ldots,M,\ j = i,\ldots,M.
\end{aligned}
$$

The objective minimizes the dissimilarity within clusters for within-cluster homogeneity. The first constraint sums all $M(M+1)/2$ decision variables to ensure that exactly K clusters are selected. The main set of structural constraints that follows ensures that each state s is assigned to exactly one selected cluster.

The first constraint creates an obstacle to total unimodularity of the constraint coefficient matrix. If this constraint were not needed, the problem could be solved as a linear program by relaxing the binary integrality constraints to simple non-negativity conditions $x_{ij} \geq 0,\ i = 1,\ldots,M,\ j = i,\ldots,M$ (Nemhauser and Wolsey, 1988).

Joseph and Bryson (1997) point out that this continuous relaxation still can be used to identify so-called *w-inefficient* partitions, i.e., a cluster count K that does not offer an attractive rate (per cluster) of improvement in the chosen clustering criterion. They show that a cluster size is w-efficient if and only if the relaxation yields an integral solution. Thus, an optimal solution to the relaxation that is non-integer is w-inefficient, creating a so-called "fuzzy" partitioning consisting of partial cluster assignments (Gan et al., 2007).

To identify possible clustering criteria, we first define

$$p_{ij} = \frac{\sum_{s=i}^{j} n_s p_s}{\sum_{s=i}^{j} n_s} = \frac{\sum_{s=i}^{j} w_s}{\sum_{s=i}^{j} n_s} = \frac{w_{ij}}{n_{ij}}$$

as the winning frequency for the aggregated group of states i through j. As a measure of disparity for this cluster of states, Rao (1971) proposed using the within group

- *sum of squares* (SS), $c_{ij} = \sum_{s=i}^{j} n_s(p_s - p_{ij})^2$.

A related measure is the within group

- *sum of absolutes* (SA), $c_{ij} = \sum_{s=i}^{j} n_s\,|p_s - p_{ij}|$.

Other proposed metrics (Hansen and Jaumard, 1997) include the within group

- *clique* or sum of dissimilarities, $c_{ij} = \sum_{r=i}^{j} \sum_{s=i}^{j} n_s \, |p_s - p_r|$,
- *star* or minisum dissimilarity, $c_{ij} = \min_{r=i}^{j} \sum_{s=i}^{j} n_s \, |p_s - p_r|$,
- *radius* or minimax dissimilarity, $c_{ij} = \min_{r=i}^{j} \max_{s=i}^{j} n_s \, |p_s - p_r|$,
- *diameter* or maximax dissimilarity, $c_{ij} = \max_{r=i}^{j} \max_{s=i}^{j} n_s \, |p_s - p_r|$.

For completeness, we also investigate another criterion, which we call the

- *sun* or maxisum dissimilarity, $c_{ij} = \max_{r=i}^{j} \sum_{s=i}^{j} n_s \, |p_s - p_r|$.

Whereas the star criterion sums the radii eminating from a central node of a cluster, the sun criterion sums the rays eminating from a node on the periphery. We intend to use all these measures to find and compare the performance of various clustering alternatives for MLB best-of-7 game playoff data.

3 MLB Playoff Data

World Series competition dates back over a century. The first fall classic was played in 1903, pitting the champions of the two major baseball leagues, the National League (NL) and the American League (AL). The series has been played every year since, except for 1904 (due to the reluctance of John McGraw, manager of the NL champion New York Giants) and 1994 (due to a labor strike).

For the first twenty World Series (through 1923), MLB organizers tinkered with a variety of home-away playoff formats of various lengths (either seven or nine games). They eventually settled on a best-of-7 game playoff series, played under a so-called 2-3 format (HHAAAHH) in which the first two and (if necessary) last two games are played on the favored team's home field. The three intermediate games are played away from the favored team's home on their opponent's field.

Although MLB was first to establish a best-of-7 game championship series, it has been slow to do so in earlier rounds of the playoffs where a best-of-5 game series is standard. In 1985, a switch from best-of-5 game series to best-of-7 game series was adopted for the League Championship Series. This semi-final round identifies the NL and AL champions who will meet in the World Series final round. In contrast, the NBA and NHL employ the longer best-of-7 game

series for all four rounds of their playoffs. This limits the comparative amount of MLB playoff data available.

There have been 83 World Series played between 1924 and 2007. In addition, since 1985 there have been 22 years in which two League Championship Series were played. This yields a total of 127 best-of-7 game playoff series in the history of MLB. We collected game-by-game data on these playoff series from MLB historical records (Major League Baseball, 1996). We supplemented these records with the results of more recent play (through year 2007) as documented at the league web site: `mlb.com`.

There have been 736 games played in these 127 series, for an average series length of just over 5.795 games. This is slightly less than the maximum expected length of 5.8125 games between two evenly matched teams under a binomial model as shown by Groeneveld and Meeden (1975), Brunner (1987), Woodside (1989), Nahin (2000) and Ross (2004).

Of these 736 games, the favored team won 378, for a winning percentage of about 51.4%. Thus, a straight-forward MLE of the Bernoulli success probability p is $\hat{p} = 378/736 \approx 0.514$. For the home-away model of Bassett and Hurley (1998), we compute similar estimates based on whether the favored team was at home or away. These estimates are $\hat{p}_H = 213/377 \approx 0.565$ and $\hat{p}_A = 165/359 \approx 0.460$.

Table 1 further breaks down these game outcomes based on the status of the series when the game was played. The top row shows that, leading (3-0), the favored team wins game 4 — on the road no less — in a large majority (about 78.6%) of those cases. Thus, the trailing team seems to "surrender" the series, resulting in a fair amount of series "sweeps" by the favored team.

Psychological effects also seem to be at work in game 6, as the team trailing in the series three games to two faces a "do-or-die" situation. When trailing 2 games to 3 the favored team is particularly strong, winning nearly 64.3% of the time, which is the second-largest winning percentage (state 2). In contrast, when leading the series 3 games to 2 the favored team is not nearly as effective, winning only 45.7% of the time (state 11).

These strong "back-to-the-wall" performances on part of the trailing team in game 6 lead to many more 7-game series outcomes than would be otherwise expected, an anomaly first noticed by Groeneveld and Meeden (1975) and Simon (1977) in World Series play. This "back-to-the-wall" effort is in sharp contrast to the opposite behavior in the NHL playoffs, where the favored team, trailing 2-3 to start game 6, surrenders the series more often than not (Rump, 2006a). Groeneveld and Meeden (1975) create a so-called "do-or-die" model

State, s	Status	Games, n_s	Wins, w_s	Percent, p_s
1	(3-0)	14	11	0.786
2	**(2-3)**	42	27	0.643
3	**(0-0)**	127	75	0.591
4	**(0-1)**	52	30	0.577
5	**(3-3)**	46	26	0.565
6	(1-3)	26	14	0.538
7	**(1-0)**	75	39	0.520
8	(2-1)	54	28	0.519
9	(3-1)	31	16	0.516
10	(1-2)	46	22	0.478
11	**(3-2)**	35	16	0.457
12	(1-1)	66	29	0.439
13	(2-2)	48	20	0.417
14	(0-2)	22	9	0.409
15	(2-0)	39	14	0.359
16	(0-3)	13	2	0.154
Total		736	378	0.514

Table 1: MLB Game-to-Game Transition Data (home games in **bold**).

that incorporates an extra "surrender" parameter

$$\lambda = \Pr\{N = 6 | N > 5\} = 1 - \Pr\{N = 7 | N > 5\},$$

where the random variable N indicates the number of games played in the series. This parameter reflects the conditional probability, given the series is not over after 5 games, that the trailing team will lose the sixth game and thus the series, making a seventh game unnecessary.

Of the 77 such game 6 instances in the data shown in Table 1 — 42 from a (2-3) and 35 from a (3-2) playoff status — there were 46 series that went to a decisive seventh game knotted (3-3). Thus, we estimate the "surrender" parameter as $\tilde{\lambda} = (77 - 46)/77 \approx 0.403$. For the "do-or-die" model, $\tilde{\lambda}$ serves as an estimate for the game-winning probability when the series status is (3-2) and for the complement when the series status is (2-3). Thus, we estimate the game-winning probability when the series status is (2-3) as $1 - \tilde{\lambda} \approx 0.597$.

Removing the 27+16 = 43 wins by the favored team in these 77 games from the totals of Table 1, we revise the Bernoulli probability estimate for all other states in the "do-or-die" model as $\tilde{p} = 335/659 \approx 0.508$.

4 Clustering MLB Playoff Data

The data in Table 1 was clustered by solving the set-partitioning problem of Section 2 under each of the clustering criteria outlined therein. The optimization model was implemented in a computer spreadsheet that contains a table of size $M(M+1)/2$ by M, with one row for each of the clustering decision variables x_{ij} and a column for each state s, $s = 1, \ldots, M$. The table contains binary data indicating whether (1) or not (0) state s is contained in the cluster x_{ij}, i.e., whether or not $i \leq s \leq j$. For each state s, a sum-product of its column of indicators with a column of binary decision variables is constrained to equal 1, ensuring that each state is assigned to exactly one cluster. We also constrain the column of binary decision variables to sum to K. Each of the decision variables has pre-defined within-cluster dissimilarity measures, so a sum-product of a particular choice of clustering metric column with the column of decision variables forms the objective function.

Table 2 reports the optimal value of each clustering metric for each choice of cluster total K, $K = 1, \ldots, 6$. Most of the optimal partitions were found by solving the LP relaxation. Those w-inefficient partitions — limited to the radius and diameter metrics — for which the relaxation failed to yield a binary solution were found instead by solving the full BIP.[1] The inefficiency (italicized in Table 2) is apparent when examining the rate of decrease in the criterion metric as a function of K. For example, under the diameter metric, the second-order decrease at $K = 3$ is $(29.410 - 24.786) - (24.786 - 14.939) = -5.222 < 0$, indicating that the drop in the diameter metric from $K = 2$ to $K = 3$ clusters was not as much as the drop from $K = 3$ to $K = 4$ clusters. Thus, a choice of $K = 3$ clusters is relatively inefficient in terms of minimizing within-cluster diameters.

Since there is only one single clustering ($K = 1$) of the data, it is obviously optimal for any and all clustering metrics. For a choice of $K = 2$ to 6 clusters, Tables 3 through 7 respectively list the non-dominated partitions, i.e., optimal partitions found under each individual clustering metric. For each row of a table, a comparison is made between that partition's performance on each clustering metric and the performance of the optimal partition (as reported in

[1] Even the full BIP usually can be solved in about 1 sec. using commercial optimization software such as Frontline Systems Premium Solver v7.1 in Excel as done here.

	Clustering Metric						
Clusters, K	SS	SA	Clique	Star	Sun	Radius	Diam
1	6.681	52.310	1445.065	51.440	264.769	7.000	55.462
2	2.673	29.912	444.786	29.284	90.182	*6.709*	29.410
3	1.732	21.046	217.671	18.109	51.130	*6.000*	*24.786*
4	0.953	16.795	129.063	14.316	33.654	*5.091*	14.939
5	0.453	12.834	62.179	11.387	24.763	4.167	*12.077*
6	0.280	9.506	37.970	8.721	16.156	*3.934*	9.077

Table 2: Best Partition Performance on MLB Data (w-inefficiencies in *italics*).

Table 2) on that metric. Reported is the percentage difference in performance relative to the optimal partition. Of course, this so-called relative error is 0 for the optimal partition itself, which have been highlighted in bold. On the other hand, a relative error of 1, for example, reflects a partition that has twice as much (100% more) disparity on that clustering metric.

For $K = 2$ desired clusters (see Table 3), partition 9/7 (a grouping of the first 9 states with a shared game-winning frequency $p_{1,9} = 0.570$ and the last 7 states with game-winning frequency $p_{10,16} = 0.416$) optimizes all but the sun and diameter metrics, which have respective values of 108.83 and 43.632 that are nearly 20.7% and 48.4% larger than the optimal values of 90.182 and 29.41 (reported in Table 2) achieved by partitions 14/2 and 15/1, respectively.

	Clustering Metric								
Partition	SS	SA	Cliq	Star	Sun	Rad	Diam	Sum	Max
9 /7	**0**	**0**	**0**	**0**	0.207	**0**	0.484	**0.690**	**0.484**
14/2	0.612	0.506	1.042	0.471	**0**	0.383	0.115	3.129	1.042
15/1	0.859	0.561	1.493	0.594	0.291	0.043	**0**	3.841	1.493

Table 3: Relative Error for $K = 2$ Clusters of MLB Data.

The partition 9/7 is also the minisum and minimax solution across all clustering metrics in that it provides (highlighted in bold) the minimal sum (or average) of the relative errors as well as the smallest maximum error. For $K = 3$ (see Table 4), the minimax partition does not actually optimize any one of the individual objectives. It is, however, Pareto optimal (non-dominated) in

that no other partition simultaneously scores at least as well on all clustering objectives. A similar situation arises for $K = 4$ and $K = 6$ desired clusters and for the minisum solution for $K = 4$.

	Clustering Metric								
Partition	SS	SA	Cliq	Star	Sun	Rad	Diam	Sum	Max
9 /6/1	**0**	0.234	0.621	0.412	0.101	0.037	0.221	**1.627**	0.621
6 /3/7	0.180	**0**	0.282	0.069	0.748	0.027	0.772	2.079	0.772
5 /6/5	0.034	0.116	**0**	0.040	0.441	0.352	0.951	1.934	0.951
5 /5/6	0.048	0.036	0.031	**0**	0.552	0.299	0.887	1.853	0.887
6 /8/2	0.051	0.312	0.269	0.385	**0**	0.422	0.658	2.097	0.658
11/4/1	0.296	0.446	1.258	0.657	0.226	**0**	0.214	3.096	1.258
14/1/1	1.251	0.950	3.123	1.231	0.607	0.103	**0**	7.266	3.123
9 /5/2	0.104	0.265	0.514	0.386	0.056	0.308	0.451	2.085	**0.514**

Table 4: Relative Error for $K = 3$ Clusters of MLB Data.

	Clustering Metric								
Partition	SS	SA	Cliq	Star	Sun	Rad	Diam	Sum	Max
5 /5/5/1	**0**	0.023	0.175	0.026	0.079	0.272	1.224	1.799	1.224
1 /4/4/7	0.529	**0**	0.483	0.101	1.597	0.178	0.799	3.687	1.597
4 /5/5/2	0.254	0.181	**0**	0.064	0.069	0.684	1.635	2.886	1.635
5 /4/5/2	0.255	0.063	0.076	**0**	0.028	0.506	1.500	2.428	1.500
2 /7/5/2	0.213	0.411	0.137	0.452	**0**	0.934	0.783	2.930	0.934
11/3/1/1	1.209	0.664	2.680	0.939	0.660	**0**	0.793	6.944	2.680
1 /8/6/1	0.109	0.374	0.755	0.573	0.614	0.223	**0**	2.648	0.755
5 /4/6/1	0.066	0.024	0.256	0.033	0.096	0.188	1.119	**1.782**	1.119
1 /4/8/3	0.436	0.395	0.492	0.547	0.560	0.570	0.499	3.499	**0.570**

Table 5: Relative Error for $K = 4$ Clusters of MLB Data.

Since each measure of cluster dissimilarity almost always suggests a different solution, in Section 5 we focus on the minisum solutions that minimize the average percentage deviation across all metrics. One could also consider the minimax partitions for those cases ($K = 3$, 4 and 6) where this solution differed from the identified minisum solution. In general, however, we found that the minisum solution slightly outperforms its minimax counterpart (omitted here for brevity) in a goodness-of-fit test to the actual series outcomes.

	Clustering Metric								
Partition	SS	SA	Cliq	Star	Sun	Rad	Diam	Sum	Max
1 /4/5/ 5 /1	**0**	**0**	0.434	0.050	0.385	0.425	0.249	**1.544**	**0.434**
2 /3/5/ 4 /2	0.710	0.160	**0**	0.008	0.052	1.080	0.850	2.859	1.080
2 /3/4/ 5 /2	0.723	0.116	0.004	**0**	**0**	0.944	0.804	2.591	0.944
11/1/2/ 1 /1	3.601	1.062	6.555	1.306	1.175	**0**	1.082	14.780	6.555
1 /1/1/12/1	3.695	1.288	6.518	1.507	1.793	0.231	**0**	15.032	6.518

Table 6: Relative Error for $K = 5$ Clusters of MLB Data.

	Clustering Metric								
Partition	SS	SA	Cliq	Star	Sun	Rad	Diam	Sum	Max
1/4/5/ 4 /1/1	**0**	0.059	0.722	0.033	0.600	0.278	0.297	1.990	0.722
1/4/4/ 5 /1/1	0.022	**0**	0.728	0.023	0.521	0.135	0.235	**1.664**	0.728
2/3/4/ 3 /3/1	0.304	0.150	**0**	0.034	**0**	0.721	0.756	1.965	0.756
2/3/4/ 5 /1/1	0.322	0.087	0.363	**0**	0.038	0.381	0.519	1.709	0.519
6/3/1/ 4 /1/1	1.846	0.263	2.601	0.229	0.420	**0**	1.983	7.343	2.601
1/1/1/11/1/1	4.131	1.493	7.963	1.569	1.726	0.287	**0**	17.170	7.963
1/4/4/ 3 /3/1	0.004	0.063	0.365	0.056	0.484	0.475	0.472	1.920	**0.484**

Table 7: Relative Error for $K = 6$ Clusters of MLB Data.

5 Goodness of Fit

The game-winning percentages suggested by each minisum partition are found in Table 8. Notice that, save for a small abberation for $K = 5$, the minisum partitions are nested (or hierarchical), i.e., a subsequent partition of $K + 1$ clusters is formed by splitting one of the K clusters in two.

We now compare how well these minisum partitions do at predicting the outcomes of the 127 playoff series. Table 9 presents the Pearson (χ^2) goodness of fit of these partition assignments to the actual MLB series outcomes (α-β), where either α or β is equal to 4 (cf., e.g., Sec. 14.2 of Devore (2004)). Here again, α and β represent the number of games won by the favored team and their opponent, respectively, in a series that lasted $\alpha + \beta$ games.

The first row of Table 9 indicates the actual observed frequencies of such outcomes to compare against the expected number as predicted by the various models. These predictions are found by feeding the game-to-game transition probabilities of Table 8 as input into a Markov chain model of a best-of-7 game

			Assigned Game-Winning Percentages					
State	Status	Actual	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
1	(3-0)	0.786	0.514	0.570	0.570	0.601	0.786	0.786
2	**(2-3)**	0.643	0.514	0.570	0.570	0.601	0.592	0.592
3	**(0-0)**	0.591	0.514	0.570	0.570	0.601	0.592	0.592
4	**(0-1)**	0.577	0.514	0.570	0.570	0.601	0.592	0.592
5	**(3-3)**	0.565	0.514	0.570	0.570	0.601	0.592	0.592
6	(1-3)	0.538	0.514	0.570	0.570	0.522	0.513	0.522
7	**(1-0)**	0.520	0.514	0.570	0.570	0.522	0.513	0.522
8	(2-1)	0.519	0.514	0.570	0.570	0.522	0.513	0.522
9	(3-1)	0.516	0.514	0.570	0.570	0.522	0.513	0.522
10	(1-2)	0.478	0.514	0.416	0.430	0.430	0.513	0.442
11	**(3-2)**	0.457	0.514	0.416	0.430	0.430	0.419	0.442
12	(1-1)	0.439	0.514	0.416	0.430	0.430	0.419	0.442
13	(2-2)	0.417	0.514	0.416	0.430	0.430	0.419	0.442
14	(0-2)	0.409	0.514	0.416	0.430	0.430	0.419	0.442
15	(2-0)	0.359	0.514	0.416	0.430	0.430	0.419	0.359
16	(0-3)	0.154	0.514	0.416	0.154	0.154	0.154	0.154

Table 8: MLB Transition Probabilities Assigned by Minisum Partitions.

playoff (Kemeny and Snell, 1960; Brunner, 1987). The computed absorption probabilities for the 8 series-ending states are then multiplied by the total number of series played (127) to give an expectation of the number of such outcomes. Also computed are the predictions made by both the home-away model (Bassett and Hurley, 1998) and the "do-or-die" model (Groeneveld and Meeden, 1975) using the parameter estimates obtained in Section 3.

The models are listed in Table 9 (and plotted in Figure 1) in decreasing order of *raw* fit, ignoring the number of parameters needed to obtain that fit. It is interesting to note that the binomial model ($K = 1$) provides a better fit than the home-away model, even though the latter model includes two fitted parameters rather than one. This lack of fit is at least partially explained by fact that, unlike in the NBA and NHL where the home-field advantage is granted to the team with a stronger record, MLB World Series have historically alternated the home-field advantage from year to year between the two leagues.

	Series Outcome								
Model	(4-0)	(4-1)	(4-2)	(4-3)	(3-4)	(2-4)	(1-4)	(0-4)	p-value
Actual	11	16	16	26	20	15	12	11	
6-Minisum	11.05	16.42	15.81	25.98	17.92	16.54	13.28	9.99	0.429
5-Minisum	12.69	15.09	14.65	26.57	18.33	16.97	12.30	10.40	0.657
4-Minisum	10.29	17.56	15.19	26.74	17.72	16.11	13.64	9.74	0.806
3-Minisum	10.08	20.65	14.32	23.64	17.87	17.01	12.09	11.35	0.724
2-Minisum	9.77	20.42	13.48	24.51	18.52	18.23	14.05	8.01	0.582
Do/Die	8.48	16.68	16.24	24.09	23.30	15.70	15.09	7.42	0.580
1-Binomial	8.84	17.19	20.91	20.34	19.26	18.75	14.60	7.11	0.348
Home/Away	8.56	15.32	22.62	22.53	17.35	17.30	16.30	7.02	0.199

Table 9: Goodness of Fit to 127 MLB Best-of-7 Game Playoff Outcomes by Minisum Partitions.

In the past few years MLB began awarding this advantage to the league that wins the mid-season All-Star game.

These differences are accentuated once we remove the advantage of additional parameters by factoring in the degrees of statistical freedom, the number of freely determined cells (in this case $8 - 1 = 7$) less the number of parameters estimated. The goodness-of-fit calculation that incorporates these degrees of freedom appears in the last column of Table 9 as the so-called p-value or significance probability. For each model, this p-value represents the probability that we would observe outcome numbers that deviate at least as much as the actual observed figures given that the model perfectly characterizes the randomness of the series outcome. Therefore, small p-values near 0 indicate a poor model (that we reject as a good fit for the data), whereas large p-values near 1 indicate a good model.

As might be expected, the "do-or-die" model of Groeneveld and Meeden (1975) provides the closest fit on 6-game outcomes (4-2) and (2-4), in which the trailing team "surrenders" the series less frequently than predicted by the binomial or home-away models. The minisum partitions all over-correct here, under-estimating the number of (4-2) series outcomes due to the inclusion of a (3-2) series status in a cluster containing lower game-winning frequency states.

The partition 9/7 involving $K = 2$ clusters and the two-parameter "do-or-die" model both provide vast improvements on the relatively poor-fitting home-away model, which also requires two parameters. The partition 9/7 does an admirable job at predicting (4-0) sweeps by the favored team, but seriously

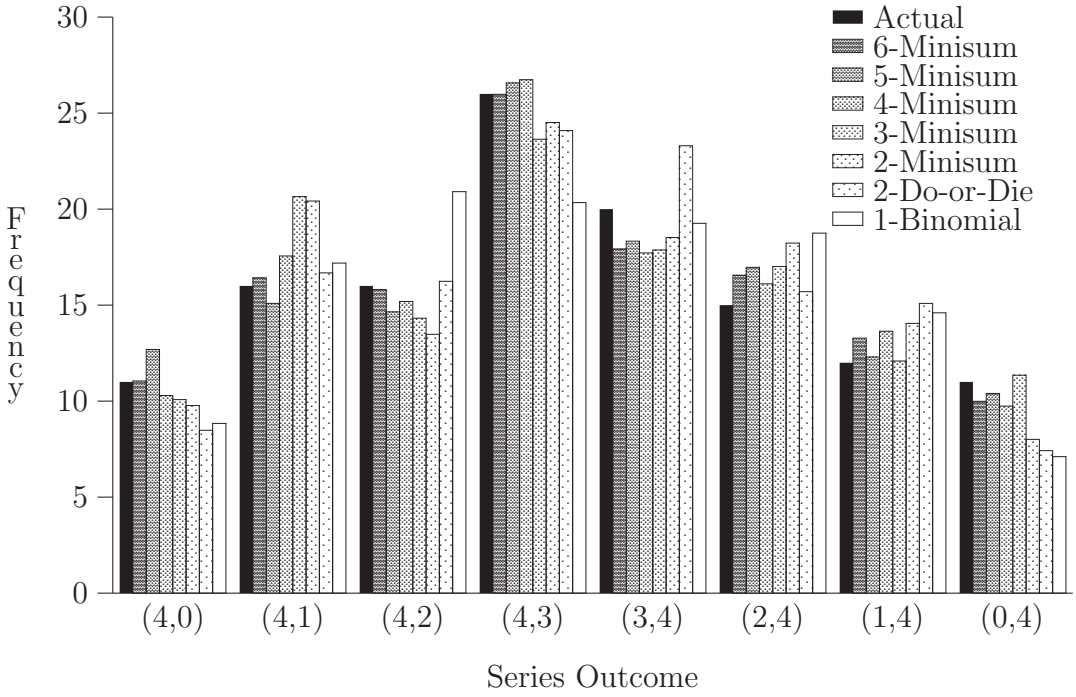

Figure 1: Predicted Outcomes of 127 MLB Best-of-7 Game Playoff Series.

over-estimates the number of (4-1) series outcomes due to the inclusion of a (3-1) status as the last member of the first cluster of higher game-winning frequency states. The weakness in the "do-or-die" model lies with its serious over-prediction of (3-4) series outcomes due to the relatively small probability $\tilde{p}$ afforded the favored team in game 7.

Overall, the 9/7 partition and the "do-or-die" model provide nearly identical fit. The slight edge afforded by the 9/7 partition may be deemed more significant when one considers its simplicity. We need only segregate the data into two categories: those games for which the favored team historically holds an edge (wins more than half the time) and those games for which they do not. As seen in Table 8, the favored team holds an edge in the first 9 of the 16 possible states (i) in all games played at home, except when leading the series 3-2, (ii) in games played on the road when leading the series 3-0, 3-1, 2-1, and (iii) even when trailing on the road 1-3, when the series is on the line.

An even better statistical fit can be obtained by $K = 3$ and $K = 4$ partitions of the data, where a singleton cluster is formed for a (0-3) series status. This helps the 9/6/1 minisum partition of $K = 3$ clusters provide the best fit

to (0-4) and (1-4) series outcomes. However, this partition suffers the same malady as the 9/7 partition on (4-1) series outcomes. This problem is remedied by the 5/4/6/1 minisum partition of $K = 4$ clusters, as it separates a (3-1) status from the first cluster of higher game-winning frequency states.

As one would expect, adding more partitions of the data increases the raw fit, but not the p-value. The minisum solution with $K = 6$ clusters is only a little better in terms of statistical fit than binomial model with $K = 1$. Indeed, a model using $K = 16$ parameter estimates from Table 1 — one for each transient state — would yield a perfect "curve fitting" to the 8 absorbing outcomes. However, we value a more parsimonious model with a limited number of uncertain parameters.

6 Conclusions

We found that a home-away model (Bassett and Hurley, 1998) of best-of-7 game playoffs does not provide a very good fit to historical playoff outcomes in Major League Baseball (MLB). Even a simple (truncated) binomial model (e.g., Mosteller (1952)) with a single game-winning probability parameter performs much better.

To improve on both these models, we formulated a Markov probability model that incorporates game-to-game dependence on the current status of the series. To create relatively parsimonious models, we proposed grouping transition probabilities of the Markov chain into clusters of similar game-winning frequency by solving a binary optimization problem. Most instances of these problems were efficiently solved as continuous linear programs, revealing so-called w-efficient solutions (Joseph and Bryson, 1997).

To capture the dissimilarity within a cluster of states, we computed both the sum of squared deviations and the sum of absolute deviations from the cluster mean, as well as complete pair-wise clique, star, radius, and diameter measures (Hansen and Jaumard, 1997) and a similar but new sun measure.

Each measure of cluster dissimilarity almost always suggested a different solution. Therefore, we chose to focus on minimax and minisum solutions that minimize the worst and average percentage deviation of any one metric, respectively. The minisum and minimax solutions are non-dominated or Pareto optimal, meaning that no other solution simultaneously scores at least as well on all clustering objectives.

The minisum/minimax solution involving only $K = 2$ clusters simply partitions the states of the series into two categories: those games for which the team favored with the home-field advantage in the series has an edge (wins

more frequently than not) and those games for which their opponent holds the edge. This model fit slightly better than a 2-parameter model that incorporates a strong "come-back" performance for a team trailing in the series after 5 games (Groeneveld and Meeden, 1975).

Better statistical fits were obtained with $K = 3$ to 5 partitions of the data, with the minisum partition involving $K = 4$ clusters the best among them. Adding more partitions of the data increases the raw fit at the expense of lost parsimony.

It is not clear how robust these models will be in light of future playoff series, an avenue for future exploration. Further work could also explore fitting probability models to best-of-5 game playoff series, which MLB still uses for the first-round (divisional) playoffs. These shorter series may be a dying breed, however, as they are now extinct in both the NBA and the NHL.

References

Bassett, G. W. and Hurley, W. J. (1998), "The Effects of Alternative HOME-AWAY Sequences in a Best-of-Seven Playoff Series," *The American Statistician*, 52, 51–53.

Bhat, U. N. and Miller, G. K. (2002), *Elements of Applied Stochastic Processes*, Hoboken, NJ: John Wiley & Sons, 3rd ed.

Broadie, M. and Joneja, D. (1993), "An Application of Markov Chain Analysis to the Game of Squash," *Decision Sciences*, 24, 1023–1035.

Brunner, J. (1987), "Absorbing Markov Chains and the Number of Games in a World Series," *The UMAP Journal*, 8, 99–108.

Devore, J. L. (2004), *Probability and Statistics for Engineering and the Sciences*, Thomson Brooks/Cole, 6th ed.

Gan, G., Ma, C., and Wu, J. (2007), *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, Philadelphia: SIAM.

Groeneveld, R. A. and Meeden, G. (1975), "Seven Game Series in Sports," *Mathematics Magazine*, 48, 187–192.

Hansen, P. and Jaumard, B. (1997), "Cluster Analysis and Mathematical Programming," *Mathematical Programming*, 79, 191–215.

Joseph, A. and Bryson, N. (1997), "W-Efficient Partitions and the Solution of the Sequential Clustering Problem," *Annals of Operations Research: Nontraditional Approaches to Statistical Classification and Regression*, 74, 305–319.

Kemeny, J. G. and Snell, J. L. (1960), *Finite Markov Chains*, The University Series in Undergraduate Mathematics, Princeton: D. Van Nostrand Co.

Major League Baseball (1996), *The Baseball Encyclopedia*, New York: Macmillan Publishing Co., Inc., 10th ed.

— (2007), mlb.com.

Mosteller, F. (1952), "The World Series Competition," *Journal of The American Statistical Association*, 47, 355–380.

Nahin, P. J. (2000), *Duelling Idiots and other Probability Puzzlers*, Princeton, NJ: Princeton University Press.

Nemhauser, G. L. and Wolsey, L. A. (1988), *Integer and Combinatorial Optimization*, New York: John Wiley and Sons.

Rao, M. R. (1971), "Cluster Analysis and Mathematical Programming," *Journal of The American Statistical Association*, 66, 622–626.

Ross, K. (2004), *Mathematician at the Ballpark: Odds and Probabilities for Baseball Fans*, New York: Pearson Education, Inc.

Rump, C. M. (2006a), "Andrei Markov in the Stanley Cup Playoffs," *Chance Magazine*, 19, 37–42.

— (2006b), "The Effects of Home-Away Sequencing on the Length of Best-of-Seven Game Playoff Series," *Journal of Quantitative Analysis in Sports*, 2, Article 5, www.bepress.com/jqas/vol2/iss1/5.

Sadovskiĭ, L. E. and Sadovskiĭ, A. L. (1993), *Mathematics and Sports*, vol. 3 of *Mathematical World*, Providence, RI: American Mathematical Society.

Simon, W. (1977), "Back-to-the-Wall Effect: 1976 Perspective," in *Optimal Strategies in Sports*, eds. Ladany, S. P. and Machol, R. E., North-Holland, Amsterdam, vol. 5 of *Studies in Management Science and Systems*.

Stewart, I. (1989), *Game, Set, & Math: Enigmas and Conundrums*, Cambridge, MA: Basil Blackwell.

Vinod, H. D. (1969), "Integer Programming and the Theory of Grouping," *Journal of The American Statistical Association*, 64, 506–519.

Woodside, W. (1989), "Winning Streaks, Shutouts, and the Length of the World Series," *The UMAP Journal*, 10, 99–113.