A unified approach to understanding randomness in sport

Benjamin S. Baumer
Smith College
bbaumer@smith.edu

Michael Lopez
Skidmore College
mlopez1@skidmore.edu

Gregory J. Matthews
Loyola University Chicago
gmatthews1@luc.edu

September 19, 2016

Abstract

With easily accessible data, worldwide interest, and continuously recurring out-of-sample observations, sports provides an excellent framework for testing predictive accuracy and understanding randomness. This has led to the creation of many innovative statistical models and metrics that estimate individual or team talent, as well as measures of within-league parity. However, current approaches tend to be restricted to a single sport or limited by a reliance on won-loss outcomes—which are noisy, particularly in a short run of games. Building on Glickman and Stern (1998), we present a modified Bayesian state-space approach that uses game-level probability information provided by betting markets to estimate perceived team strengths. This model provides a unique advantage in that posterior draws allow for a uniform understanding across leagues with respect to differences in between-season, within-season, and game-to-game variability. In addition, topics such as competitive balance, talent dispersion, and the value of home advantage can more rigorously be explored and contrasted across sports. We implement our model using 10 seasons of competition from the National Football League, National Basketball Association, Major League Baseball, and the National Hockey League.

Keywords: sports analytics, randomness, Bayesian modeling, competitive balance, MCMC

Keywords: baseball, statistical modeling, simulation, R, reproducibility

1 Introduction

As interest in sports analytics has grown over the past decade, formal methods of estimation have increasingly been used by statisticians to make meaningful contributions to our understanding of competitive sports. One specific application in which statistics has played an important role is in the estimation of team quality. For example, is team i better than team j? And if so, how confident are we in making these types of claims?

Central to such an understanding of sporting outcomes is that if we know each team's relative strength, then, *a priori*, game outcomes—including wins and losses—can be viewed as unobserved realizations of random variables. As an easy example, a 75% probability of team i beating team j at time k entails that, in the hypothetical infinite number games between the two teams at time k, i wins three times as often as j.

While sporting leagues fail in practice to give us an infinite number of games, they do give us a recurring regular season in which each team plays the same number of games. In addition, at least in the four major North American sports leagues (i.e. Major League Baseball (MLB), the National Basketball Association (NBA), the National Football League (NFL), and the National Hockey League (NHL)), the top teams battle annually for a league championship in a shortened tournament known as the *postseason*. National interest, linked with an academic curiosity across disciplines, has fostered many approaches that provide a better understanding of team quality than examination of league standings alone allows. Such exercises do more than drive water-cooler conversation. Indeed, estimating team rankings has driven the development of advanced statistical models (Bradley and Terry, 1952; Glickman and Stern, 1998) and occasionally played a role in the decision of which teams are eligible for continued postseason play (CFP, 2014). Moreover, there are also financial ramifications to accurately understanding team quality, as outcome uncertainty (i.e., that each of the participating teams has a decent chance of winning) is positively correlated with game attendance (Knowles et al., 1992). In fact, it is in each league's best interest to promote some level of *parity*—in short, a narrower distribution of team quality—to maximize revenue (Crooker and Fenn, 2007).

Because of the differences in the way that games in each sport are structured, researchers who hope to contrast one league to another often focus on the one outcome common to all sports: won-loss ratio. Among other flaws, measuring team strength based only on wins and losses performs poorly in a small sample size, ignores the game's final score (which is known to be more predictive of future performance than won-loss ratio (Boulier and Stekler, 2003)), and is unduly impacted by, among other sources, fluctuations in league scheduling, injury to key players, and the general advantage of playing in a home arena.

Although more technical approaches for team strength estimation have also been developed, most of these have focused on a single sport. These approaches typically blend past game scores with game, team and player characteristics in a statistical model. Corresponding estimates of talent are often checked or calibrated by comparing out-of-sample estimated probabilities of wins and losses to observed outcomes. While excellent work has been done in this domain, such a process suffers from a few limitations. First, given the array of differences between sports, models built using one sport cannot generally be applied to another. Second, model-estimated probabilities rarely, if ever, outperform a simple benchmark: the implied probability from betting markets. In other words, even the most sophisticated statistical models are generally unable to predict games more accurately than simple models built on sports betting market data (Harville, 1980; Stern, 1991). Third, there is a notable absence of technical and interpretable cross-sport measurements to help us better understand league-level characteristics. That is, after accounting for league characteristics including schedule, home advantage, and season length, what are the inherent differences in the dispersion and evolution of team strength across sports? [ML]: 'effect of randomness' seemed strange: this was the old sentence - That is, are there inherent differences in the effect of randomness across sports? Or are the observed differences in sports entirely a function of talent dispersion, length of schedule, etc.?

This manuscript aims to fill these voids. Instead of estimating team strengths within a single sport and using those estimates to generate estimated win probabilities, we work backwards. First, we assume, and work to validate, an assumption that betting market probabilities provide unbiased and low-variance estimates of the true probabilities of wins and losses in each game. Second, using the logit transform of those probabilities, we propose a modified Bayesian state-space model that captures implied team strength and variability. An advantage of this model is that, apart from a few user-defined inputs, it can be applied uniformly across leagues. Finally, by looking at posterior estimates of within and between season variability, as well as the overall dispersion in team strength estimates, we present unique league-level contrasts which, to this point, have been difficult to capture. As examples, we find that [ML]: fill in blanks with major findings here. Additionally, we estimate [ML]: fill in blanks with secondary findings here. All together, our results better inform an understanding of the dispersion of both talent and randomness in sport. [GM]: This paper is awesome [BB]: :)

2 Studies of sport and team characteristics

The importance of quantifying team strength in sport extends across disciplines. This includes contrasting league-level characteristics in economics (Leeds and Von Allmen, 2004), estimating game-level probabilities in statistics (Glickman and Stern, 1998), and classifying future game winners in forecasting (Boulier and Stekler, 2003). We highlight and attempt to coalesce general approaches below.

2.1 Competitive Balance

Assessing the competitive balance of organizations is particularly important in economics and management (Leeds and Von Allmen, 2004). While competitive balance can purportedly measure several different quantities, in general, it refers to levels of equivalence between teams. This could be equivalence within one time frame (such as how similar was the distribution of talent within a season), between time frames (such as year-to-year variations in talent), or from the beginning of a time frame until the end (such as the likelihood of each team winning a championship at the start of a season).

The most widely accepted within-season competitive balance are the Noll-Scully (Noll, 1988; Scully, 1989) and Gini coefficient (Mizak et al., 2005), which each use won-loss ratio as a proxy for team strength.[1] Related, the Hirfindahl-Hirschman Index (Owen et al., 2007) and Competitive Balance Ratio (Humphreys, 2002) attempt to quantify the relative chances of success that teams have in each season and the season-to-season changes in team quality, respectively.

While a benefit to each of these metrics is that they allow for interpretable cross-sport comparisons, a reliance on winning percentages also yields unattractive properties (Owen, 2010; Owen and King, 2015). For example, Noll-Scully increases, on average, with the number of games played (Owen and King, 2015), hindering any comparisons of the NFL (16 games) with MLB (162). Additionally, each of the league's employ some form of an unbalanced schedule. Teams in each of the MLB, NFL, NBA, and NHL play intradivisional opponents more often than interdivisional ones, and intraconference opponents more often than interconference ones, meaning that one team's won-loss record may

[1] Noll-Scully is defined as the ratio of the observed standard deviation in team win totals to the idealized standard deviation, defined as that which would have been observed due to chance alone. It is argued that larger Noll-Scully values correspond to higher levels of imbalance in team strength. Somewhat related, the Gini coefficient compares the proportion of total wins cumulatively earned by the bottom proportion of teams, relative to idealized value. If fewer teams are earning a larger share of wins, that contributes to larger league-level imbalance.

not be comparable to another team's. Moreover, the NFL structures each season's schedule so that teams play interdivisional games against opponents that finished in the same spot in the standings. In expectation, this punishes teams that finish atop standings with tougher games in the following year, and potentially drives winning percentages towards 0.500. Unsurprisingly, unbalanced scheduling and interleague play has yielded unintended consequences with several common competitive balance metrics (Owen and King, 2015; Utt and Fort, 2002). As one final weakness, varying home advantages between sports, as shown in (Moskowitz and Wertheim, 2011), can also impact comparisons of relative team equality using wins and losses.

Although metrics for league-level comparisons have been continually debated, the importance of competitive balance in sports is more uniformly accepted, in large part due to the uncertainty of outcome hypothesis (UOH, (Knowles et al., 1992; Lee and Fort, 2008)). Under UOH, league success, as judged by attendance, engagement, and television revenue correlate positively with teams having equal chances. Outcome uncertainty is generally considered on a game-level basis, but can also extend to season-level success (i.e, teams having equivalent chances at making the postseason).

2.2 Approaches to estimating team strength

Competitive balance and outcome uncertainty can be considered proxies for understanding distributions of team talent. For example, when two teams of equal talent play a game without a home advantage, outcome uncertainty is maximized; e.g., the game is a coin flip. Such relative comparisons of team talent began in statistics with paired comparison models. More generally, paired comparison models are those which are designed to calibrate the equivalence of two entities. In the case of sports, the entities are teams or individuals.

The Bradley-Terry model (BTM, (Bradley and Terry, 1952)) is generally considered to be the first paired comparison model. Consider an experiment with t treatment levels, compared in pairs. The BTM assumes that there is some true ordering $\pi_1, \cdots, \pi_t$ with the constraints that $\pi_i \geq 0$ and $\sum \pi_i = 1$. When comparing treatment i to treatment j, the probability that treatment i is preferable to j (i.e a win in a sports setting) is computed as $\frac{\pi_i}{\pi_i + \pi_j}$.

Glickman and Stern (1998) and Glickman and Stern (2016)) built on the BTM by allowing team-strength estimates to vary over time using a state-space model in the NFL. Let y_{ij} be the outcome of the game between i and j, where i and j take on values between 1 and t, where t is the number of teams in the league (at the time, $t = 28$). Game outcomes are assumed to follow an approximately normal distribution. Let $\theta_{(s,k)i}$ be the strength of team i in season s during week k, and let α_i be the home advantage parameter for team i, for $i = 1 \ldots t$. As in Glickman and Stern (1998), the expected point differential between i and j in a game played at team i during week k in season s is as follows:

$$\theta_{(s,k)i} - \theta_{(s,k)j} + \alpha_i$$

The distribution of the outcomes of games during season s and week k is:

$$\mathbf{y_{(s,k)}} | \mathbf{\tilde{X}_{(s,k)}}, \tilde{\theta}_{\mathbf{(s,k)}}, \phi \sim \mathbf{N}(\mathbf{\tilde{X}_{(s,k)}} \tilde{\theta}_{\mathbf{(s,k)}}, \phi^{\mathbf{-1}} \mathbf{I_{n_{(s,k)}}}),$$

where $\mathbf{y_{(s,k)}}$ is a vector of score differentials, $\tilde{\theta}_{(s,k)}$ is a vector containing the team strengths and the home advantage parameters, ϕ is the precision, and $n_{(s,k)}$ is the number of games played in week k during season s. The matrix $\tilde{X}_{(s,k)}$ consists of $n_{(s,k)}$ rows and $2t$ columns. The first t columns contain the values 1, 0, and -1 where a 1/-1 in the i-th column indicates that team i was the home/away team for the game corresponding to that row and a 0 otherwise. The second set of t columns (i.e. t+1 to 2t) contains a 1 in the i-th column (i.e column t+i) if team i is playing at home and 0 otherwise.

The model of Glickman and Stern (1998) allows for team strength parameters to vary stochastically in two distinct ways: 1) from the last week of season s to the first week of season $s+1$, and 2) from week k of season s to week $k+1$ of season s.

The variation from the last week of one season to the first week of the next season is expressed as:

$$\theta_{(s+1,1)} | \gamma_{seas}, \theta_{\mathbf{(s,g_s)}}, \phi, \omega_{\mathbf{seas}} \sim \mathbf{N}(\gamma_{\mathbf{seas}} \mathbf{G} \theta_{\mathbf{(s,g_s)}}, (\phi \omega_{\mathbf{seas}})^{\mathbf{-1}} \mathbf{I_t})$$

where $\mathbf{G}$ is the matrix that transforms $\theta_{\mathbf{(s,g_s)}}$ to $\theta_{\mathbf{(s,g_s)}} - \bar{\theta}_{\mathbf{(s,g_s)}}$ and g_s is the number of weeks in season s.

Team strength parameters also vary from week to week as follows:

$$\theta_{(s,k+1)} | \gamma_{week}, \theta_{\mathbf{(s,k)}}, \phi, \omega_{\mathbf{week}} \sim \mathbf{N}(\gamma_{\mathbf{week}} \mathbf{G} \theta_{\mathbf{(s,k)}}, (\phi \omega_{\mathbf{week}})^{\mathbf{-1}} \mathbf{I_t})$$

An attractive property of the state-space model is that it prior and future season performances are incorporated into season-specific measurements of team quality. Perhaps as a result, Koopmeiners (2012) identified stronger fits when comparing state-space models to BTM's fit separately within each season. Additionally, state space models to not typically suffer from identifiability problems were a team to win or lose all of its games in a single season.

Alternative specifications to the original state-space model have been proposed. Knorr-Held (2000) considered time-dependent team abilities in soccer and basketball using a first-order random walk. Baker and McHale (2015) assumed a dynamic, non mean-reverting model of team strength to answer the question of 'best team ever.' An exponentially weighted moving average of team strength was suggested by (Cattelan et al., 2013), who found roughly similar performances when comparing weighted and unweighted versions. Most recently, (Manner, 2015) combined predictions from a state-space model (with an AR(1) assumption) to those from betting markets. These authors also found differences in the variability of team strength parameters, although those levels of variability appeared to operate independent of team strength (e.g. low variances were possible for both good and bad teams). Interestingly, errors were random and normally distributed (e.g., no streakiness).

As additional and related BTM resources, team-specific home advantages using a BTM were compared to a constant HFA model by (Tutz and Schauberger, 2015) in soccer and Koopmeiners (2012) in football. Matthews (2005) considered data transformations of NFL score outcomes to account for blowouts in BTM's, and Owen (2011) used dynamic Bayesian models in association football with evolution variance parameters. Koopmeiners (2012) explicitly modeled the variance parameters of team strength in football, finding little change over time.

In place of paired comparison models, alternative measures for estimating team strength have also been developed. (Massey, 1997) used maximum likelihood estimation to develop a rating system using American football outcomes, called Massey rankings. A more general summary of other rating systems for forecasting use is explored by Boulier and Stekler (2003). In addition, support vector machines and simulation models have been proposed in hockey (Demers, 2015; Buttrey, 2016), neural networks and Naive Bayes implemented in basketball (Loeffelholz et al., 2009; Miljković et al., 2010), linear models and probit regressions in football Harville (1980); Boulier and Stekler (2003), and two stage Bayesian models in baseball (Yang and Swartz, 2004). While this is no doubt an incomplete list, it speaks to the depth and variety of coverage that sports prediction models have generated.

One final predictive measure that has often been used for sake of comparison is that provided by betting markets. Before each contest, sports books, including those in Las Vegas and in overseas markets, provide a price for each team, more commonly known as the money line. For example, assume team i was -140 on the money line against team j, which was priced at +120. A bet of \$140 on team i would yield a \$100 profit with an i victory, while backing j for \$100 would produce a \$120 profit were j to win. A savvy bettor would back team i if there was a belief that i's actual win probability against j was greater than 58.3% (140/240). Alternatively, that bettor would back j if the expectation was that j would win more than 45.4% (100/220) of the time. Note that these two probabilities sum to more than 1; this accounts for the 'vigorish' taken in by markets which helps them, over the long run, remain profitable. However, it is straightforward to normalize money-line prices to sum to unity. In our example, dividing each probability by 1.038 (140/240 + 100/220) implies i has a 56.2% chance of defeating j.

In principal, money line prices account for all determinants of game outcomes, including team strength, location, and injuries. Across time and sporting leagues, researchers have identified that these prices are incredibly difficult to beat; i.e, that the betting markets are efficient. As an incomplete list, see (Harville, 1980), (Stern, 1991), Carlin (1996), Colquitt et al. (2001), Nichols (2012), Paul and Weinbach (2014). Interestingly, (Colquitt et al., 2001) suggested that the efficiency of college basketball markets was proportional to the amount of pre-game information available, which would suggest that markets in professional sports are as efficient as they come. Manner (2015) merged predictions from a state-space model with those from betting markets, finding that the combination of predictions only occasionally outperformed betting markets alone.

In addition to being used as standards by which to judge predictive accuracy, betting markets have been used to suggest that NBA teams 'tank' (Soebbing and Humphreys, 2013), that bettors place larger shares of bets on the home team (Paul and Weinbach, 2011), and that markets do not move lines to attract an even number of bets on each side (Paul and Weinbach, 2008; Humphreys et al., 2014). We are not aware of any published findings that have compared leagues using implied probabilities. In one somewhat related cross-sport project, Wolfson and Koopmeiners (2015) used BTM's and margin-of-victory BTM's to find a clear separation between two pairs of leagues, the NFL and the NBA versus the NHL and MLB, as far as predictive accuracy when testing out of sample.

Paragraph setting up our goals

3 Validation of betting market data

A summary of our data set is shown in Table 1.

sport	N	earliest	latest	num_teams	home_wp	N_bets	mean_home_p	N_results	coverage
mlb	24242	2005-04-03	2014-09-28	30	0.542	24242	0.549	24299.000	0.998
nba	12000	2004-11-02	2014-04-16	30	0.599	12000	0.620	12059.000	0.995
nfl	2549	2005-09-08	2014-12-28	32	0.570	2549	0.590	2560.000	0.996
nhl	10527	2005-10-06	2014-04-13	30	0.550	10527	0.566	10564.000	0.996

Table 1: Summary of cross-sport data. Note that we have near perfect coverage (betting odds for every game) across all four major sports during the 2005–2014 regular seasons.

As expected, the probabilities implied by the betting markets are unbiased and quite accurate. Were this not the case, there would be easy arbitrage opportunities.

4 Bayesian state-space model

Whereas in many previous articles, the focus is on one particular league or sport, here we examine all four of the major north American sports together. There fore, for each of the 4 sports considered here (NFL, MLB, NHL, NBA), the following model described below was fit for each sport and indexed by $q \in \{NFL, MLB, NHL, NBA\}$.

Now, let $p_{(q,s,k)ij}$ correspond to the probability that team i will beat team j from sports league q in season s during week k. Further, let $\mathbf{p}_{(q,s,k)}$ represent the vector of length $g_{(q,s,k)}$ (i.e. the number of games in week k of season s in league q) containing all the probabilities that team i will beat team j in league q in week k of season s. [GM]: These probabilities are assumed to be known and are derived based on the implied odds of the betting market given the money line odds for a game where team i plays team j from league q in week k of season s.

We then fit the model

$$logit(p_{(q,s,k)}) \sim N(\theta_{(\mathbf{q,s,k})}\mathbf{X_{q,s,k}} + \alpha_{\mathbf{q0}}\mathbf{J_{g_{q,s,k}}} + \alpha_{\mathbf{q}}\mathbf{Z_{q,s,k}}, \tau^{\mathbf{2}}_{\mathbf{q,game}}\mathbf{I_{g_{(q,s,k)}}}) \tag{1}$$

$\theta_{(\mathbf{q,s,k})}$ is a vector of length t_q, where t_q is the number of teams in league q, containing the team

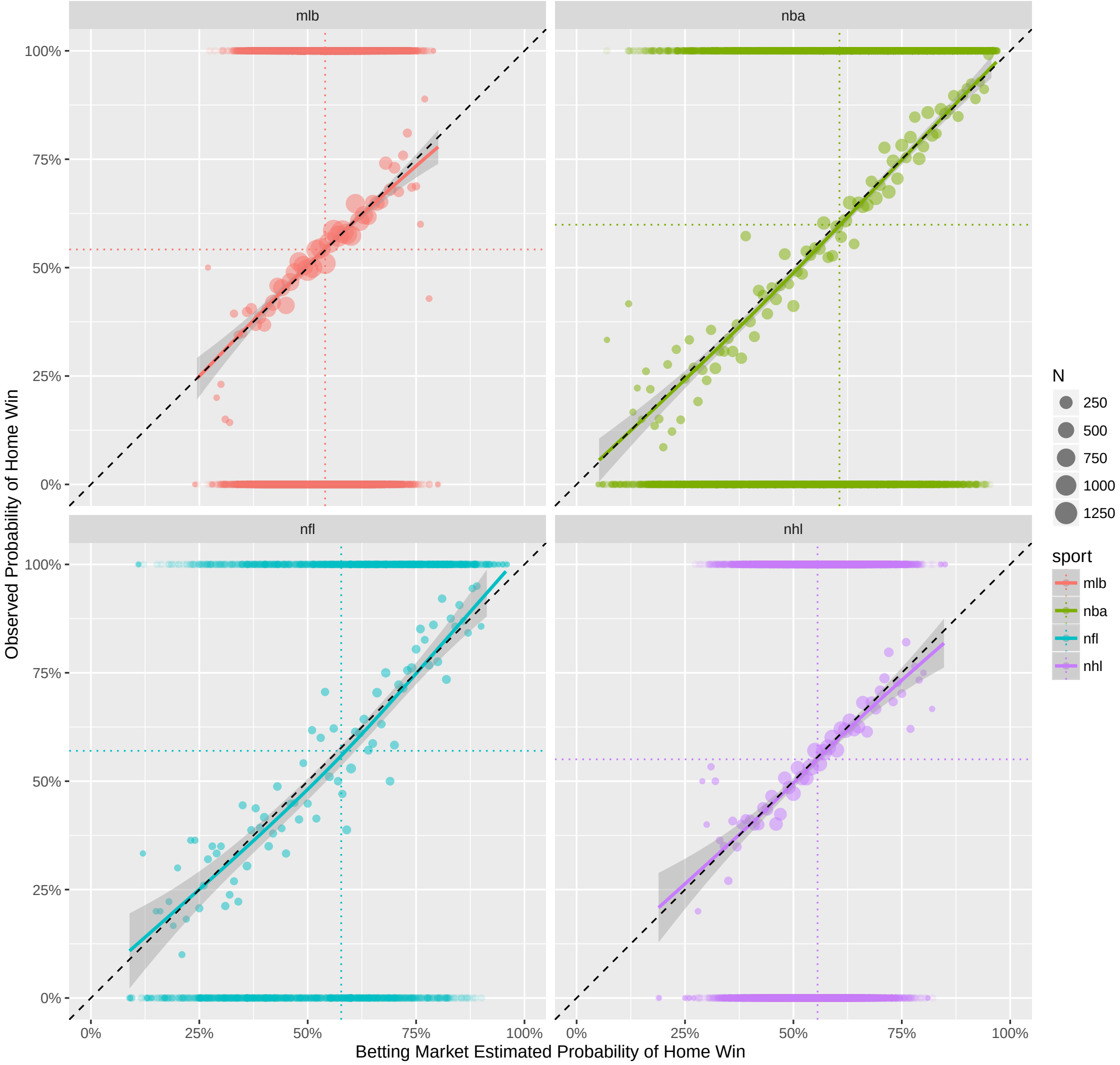

Figure 1: Accuracy of probabilities implied by betting markets. We note that across all four major sports, the observed winning percentage accord with those implied by the betting markets.

strength parameters in season s during week k, α_{q0} is the overall home field advantage parameter for sports league q, and $\alpha_{\mathbf{q}}$ is a vector of length [GM]: $t_q^{\star}$ containing [GM]: *arena* specific home field advantage parameters for league q that do not vary over time (i.e. HFA is assumed to be constant for a team over weeks and seasons). [GM]: (Note: $t_q^{\star}$, the number of home venues in league q, is greater than t_q, the number of teams in league q, due to franchise relocation.) $\mathbf{X_{q,s,k}}$ and $\mathbf{Z_{q,s,k}}$ both contain $g_{(q,s,k)}$, the number of games in league q during week k of season s, rows and t_q and $t_q^{\star}$ columns, respectively. The matrix $\mathbf{X_{q,s,k}}$ contains the values 1, 0, and -1 where for a given row (i.e. one game) the value of i-th column in that row is a 1/-1 if the i-th team played at home/away in the given game and 0 otherwise. $\mathbf{Z_{q,s,k}}$ is a matrix containing a 1 in the i-th column if the i-th team played the game corresponding to that row at home and 0 otherwise. [GM]: Finally, $\tau^2_{q,game} = \frac{1}{\sigma^2_{q,game}}$ is the precision, $\mathbf{J_{g_{q,s,k}}}$ is a column vector of length $g_{q,s,k}$ containing all 1's,and $\mathbf{I_{g_{(q,s,k)}}}$ is an identity matrix with dimension $g_{(q,s,k)}$ by $g_{(q,s,k)}$.

Similarly to (Glickman and Stern, 1998), we allow the strength parameters of the teams to vary from week to week and from season to season.

$$\theta_{(q,s+1,1)}|\gamma_{q,seas}, \theta_{\mathbf{q,s,g_{q,s,.}}}, \tau^{\mathbf{2}}_{\mathbf{q,seas}}, \sim \mathbf{N}(\gamma_{\mathbf{q,seas}}\theta_{(\mathbf{q,s,g_{q,s,.}})}, (\tau^{\mathbf{2}}_{\mathbf{q,seas}})\mathbf{I_{t_q}})$$

and

$$\theta_{(q,s,k+1)}|\gamma_{q,week}, \theta_{\mathbf{q,s,k}}, \tau^{\mathbf{2}}_{\mathbf{q,week}}, \sim \mathbf{N}(\gamma_{\mathbf{q,week}}\theta_{(\mathbf{q,s,k})}, (\tau^{\mathbf{2}}_{\mathbf{q,week}})\mathbf{I_{t_q}})$$

where $g_{q,s,.} = \sum_{k=1}^{k_q^{\star}} g_{q,s,k}$ where $k_q^{\star}$ is the number of weeks in a season in league q, $\gamma_{q,week}$ is the autoregressive parameter from week to week and $\gamma_{q,season}$ is the autoregressive parameter from season to season.

We depart from (Glickman and Stern, 1998) here in our specification of precision. Here, we estimate three separate parameters, $\tau^2_{q,game}$, $\tau^2_{q,season}$, and $\tau^2_{q,week}$. This allows these three parameters to be estimated separately.

For sport q, the team strength parameters for week $k = 1$ and season $s = 1$ have a prior of

$$\theta_{(q,1,1)i} \sim N(0, \tau^2_{q,season})$$

for i in $1, \cdots, t_q$.

The team specific home field advantage parameter has a similar prior, namely,

$$\alpha_{(q)i} \sim N(0, \tau^2_{q,\alpha})$$

for i in $1, \cdots ,$

.

Finally, we assume the following prior distributions:

$$\tau^2_{q,game} \sim \Gamma(0.0001, 0.0001)$$

$$\tau^2_{q,season} \sim \Gamma(0.0001, 0.0001)$$

$$\tau^2_{q,week} \sim \Gamma(0.0001, 0.0001)$$

$$\tau^2_{q,\alpha} \sim \Gamma(0.0001, 0.0001)$$

$$\alpha_q \sim N(0, 10000)$$

$$\gamma_{q,week} \sim Uniform(0, 2)$$

$$\gamma_{q,season} \sim Uniform(0, 2)$$

In addition, we propose a reduced version of Model (1) that assumes a constant sport-level home advantage for each time,

$$logit(p_{(q,s,k)}) \sim N(\theta_{\mathbf{(q,s,k)}}\mathbf{X_{q,s,k}} + \alpha_{\mathbf{q0}}\mathbf{J_{g_{q,s,k}}}, \tau^{\mathbf{2}}_{\mathbf{q,game}}\mathbf{I_{g_{(q,s,k)}}}) \tag{2}$$

[GM]: I thought we used JAGS? Posterior distributions of each parameter are estimated using Markov Chian Monte Carlo (MCMC) methods. We used Gibbs sampling via the R2Winbugs package in the R statistical software Sturtz et al. (2005) to obtain the posterior distributions, obtained separately within each q. Three chains, using 40,000 iterations after a burn-in of 2,000 draws, fit with a thin of 5 to reduce the autocorrelation within chains, yielded 8,000 posterior samples in each q. Visual inspection of trace plots with parallel chains are used to confirm convergence; comparisons of Models (1) and (2) are made using the Deviance Information Criterion (DIC, Spiegelhalter et al. (2002)), as implemented with a Bayesian state-space model by Koopmeiners (2012).

While we are unable to share our data, the data wrangling code, Gibbs Sampling code (via R2Winbugs), posterior draws, and the code used to obtain posterior estimates and figures are all posted to a Github repository, available at `https://github.com/bigfour/competitiveness`.

5 Results

5.1 Model Fitting

Paragraph about model checks being satisfied. Do we include trace plots?

Model fit comparisons differed by league. Table 2 shows the DIC for each fit in each league, along with the difference in DIC values and its standard error. In the NBA and NHL, fits with a team-specific HFA yielded significantly lower DIC's (lower is better), while in the NFL and MLB, DIC values were roughly similar between fits.

	DIC (unique HFA)	DIC (constant HFA)	Difference (SE)
MLB	-7580	-7569	-10.9 (7.7)
NBA	5219	5547	-327.9 (25.6)
NFL	-1427	-1427	0.8 (3.0)
NHL	-14912	-14739	-172.2 (19.0)

Table 2: Deviance information criterion (DIC) by sport and model, along with difference in DIC (standard errors of the difference shown in parenthesis))

Taken wholly, this suggests non-random differences in the home advantage between NBA and NHL arenas, while it cannot be ruled out that any differences in the home advantage between NFL and MLB stadia are due to chance.

5.2 Posterior estimates

A summary of of the posterior draws of several coefficients is shown in Table XX, which uses draws from Model (1).

Figure XX-XY show estimated team coefficients over time, approximated by using posterior mean draws from Model (1) for all weeks k and seasons s. Teams in Figures XX-XY are depicted using their two primary colors, scraped from `http://jim-nielsen.com/teamcolors/` via the 'teamcolors' package (`https://github.com/beanumber/teamcolors`) in R.

Overall, there tends to be larger amounts of variability in team strength parameters at any given point in time in both the NFL and NBA, with posterior team strength coefficients tending to vary between -1.5 and 1.5 (on the logit scale). Team-to-team level variability is substantially lower in the MLB (roughly between -0.5 and 0.5) and the NHL (-0.6 to 0.6). Among other findings in the four figures, the New England Patriots' 2007 season (NFL) stands out as the top individual performance of the last decade. In that season, New England finished the regular season 16-0 before eventually losing in the Super Bowl. Additionally, in the NBA, it is interesting to notice that the team strength estimates of the bad teams - for example, Cleveland in 2007-08 - lie further from 0 than the the estimates for the good teams. This possibly relates to the tendency for teams in this league to 'tank', which involves losing games as to increase the chances of improved positioning in the upcoming league draft (Soebbing and Humphreys, 2013).

In the NHL and only the NHL, there seems to be a peculiar convergence of team strength estimates towards 0 over time, perhaps implying that team talent is less variable in recent years. Alternatively, the league's point system, which changed before the 2005-06 season and encouraged teams to play more overtime games (**?**), could be responsible. If teams were purposefully playing overtime contests more often, it could lead to different perceptions in how betting markets view team strenths, as overtime sessions and the resulting shootouts are, by and large, coin flips (Lopez and Schuckers, 2016). As one final point of clarification, the periods with straight lines of team strength estimates in the 2013 season (NHL) and 2012 season (NBA) reflect time lost due to lockouts.

Contour plots using posterior draws of season-to-season $\sigma_{q,seas}$ and week-to-week $\sigma_{q,week}$ variability in team strength are shown in Figure WW using separate colors for each sport. There is the highest variability in team strength estimates in the NBA, followed in order by the NFL, NHL, and MLB. The importance of star players in the NBA is one potential driver of these findings (Berri and Schmidt, 2006); injuries or trades involving the league's best players perhaps have a larger influence in the NBA than in other leagues, causing immediate shifts in the estimates of team strength.

The joint posterior distribution of $\gamma_{q,seas}$ and $\gamma_{q,week}$ is likewise shown in Figure YY via contour plots for each sport. For each of the NHL, NBA, and NFL, the posterior estimates of $\gamma_{q,week}$ (and 95% credible intervals) do not cross 1, implying some auto-regressive nature to team caliber within each season. In the MLB, meanwhile, team strength estimates quite possibly follow a random walk (as in $\gamma_{q,week} = 1$).

On a season-to-season basis, team strengths in each of the leagues tend to revert towards the league average (0). Reversion towards the mean is largest in the NHL (posterior mean $\gamma_{q,seas} = 0.55$, implying 45% reversion), followed by MLB (38% reversion), NBA (36%), and the NFL (30%). However, there is enough uncertainty in the posterior distribution of $\gamma_{q,seas}$ that we are unable to say for certain if the differences in the season-to-season autocorrelation between MLB, the NBA, and the NBA are significant. In the NFL, the slight inverse association between the two autoregressive parameters, as shown in Figure YY, matches that obseved in Glickman and Stern (1998).

6 Results

6.1 Visualizing the home advantage

Informally, team-level home advantage coefficients similar in effect size, both in magnitude and with respect to team ranking, to those depicted by (Koopmeiners, 2012) in the NFL.

6.2 Uncertainty of postseason tournaments

One application of our work extends estimated team strengths to explore the each league's respective postseason The four league's all use a knockout-style postseason tournament, in which teams with stronger regular season performances are matched against lower-performing teams, with the winner

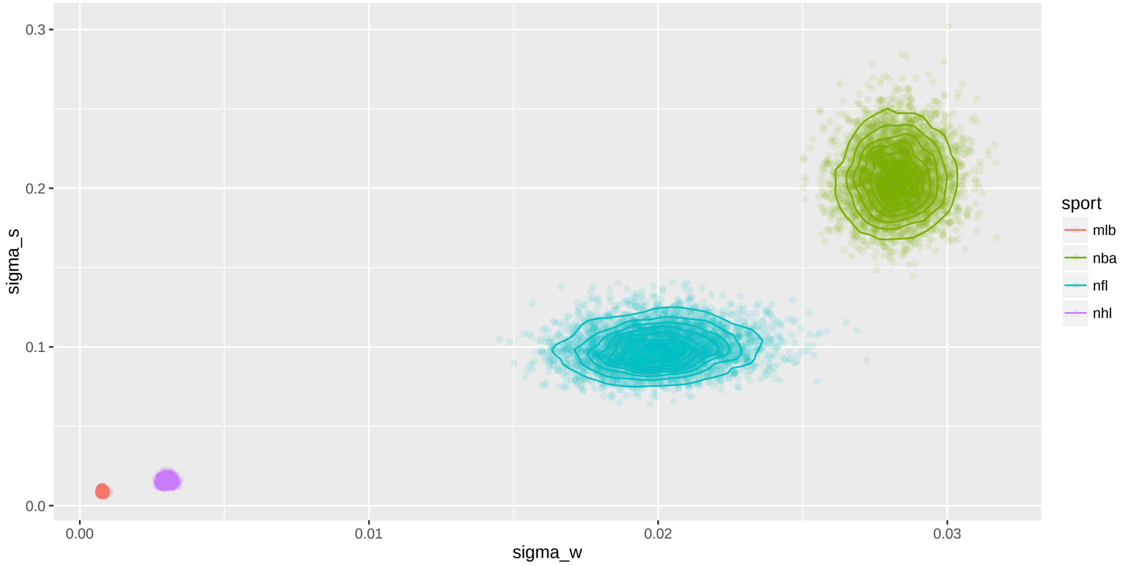

Figure 2: Contour plot of the season-to-season and week-to-week variability across all four major sports.

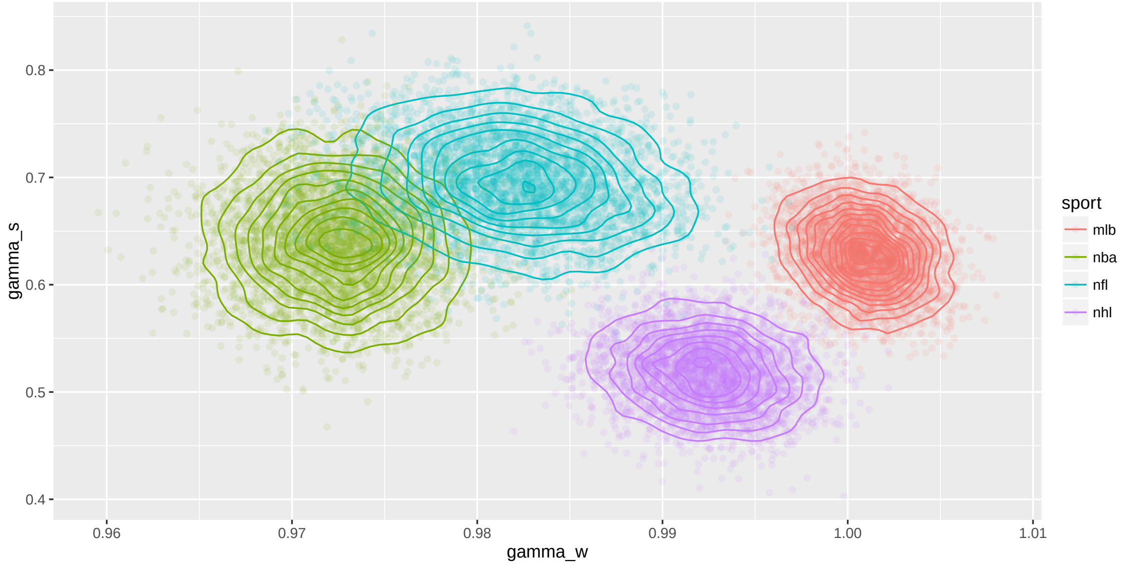

Figure 3: Contour plot of the season-to-season and week-to-week auto-regressive parameter across all four major sports.

advancing to future rounds. In the MLB, NHL, and NBA, match-ups use a series format, where the series winner is generally determined by which of the two participating teams wins the majority of games. Generally, series lengths are capped at seven games (the MLB, for example, has also used five games in earlier rounds), entailing that the series winner is the first team to win four games. Informally, these are known as 'best of 7' series. In these formats, one team is given a home advantage, which allows for one extra home game over the duration of the series. Traditionally, the team with the stronger regular season performance receives the extra home game, although league's allow for slight adaptations of this requirement. In contrast to the other three leagues, NFL postseason tournaments have consisted only of compilations of one game competitions; there is no 'best of n' series in its history.

In addition to slightly different rules regarding postseason eligibility, leagues also differ with respect to the number of teams eligible for postseason play. As of 2016, 10 (MLB), 12 (NFL), and 16 (NFL, NBA) teams make the postseason in each season, although for the duration of our data, the MLB used only 8 teams each year. Given these general guidelines, we assess the randomness of postseason formats by comparing the top 8 teams in each league.

Let $\widehat{\theta}_{q,s,k,i}$ be the posterior distribution of the estimated team strength of i at week k. Teams in league q at season s were ranked based on $\bar{\widehat{\theta}}_{q,s,i}$, where

$$\bar{\widehat{\theta}}_{q,s,i} = (\widehat{\theta}_{q,s,(k^*-4)} + \widehat{\theta}_{q,s,(k^*-3)} + \widehat{\theta}_{q,s,(k^*-2)} + \widehat{\theta}_{q,s,(k^*-1)} + \widehat{\theta}_{q,s,(k^*)})/5,$$

the average posterior team strength over the last five weeks of the regular season. These draws are meant to roughly reflect team quality entering the postseason; five weeks are used in place of using k^* alone to account for the possibility that top performing teams rest top players in the final few games, thereby lowering their perceived team quality.

Next, let $\bar{\widehat{\theta}}_{q,s,|r|}$ be the r^{th} ranked team strength in sport q in season s. For example, $\bar{\widehat{\theta}}_{MLB,2005,|1|} = 0.499$ represents the average posterior draw of team strength for 2005 New York Yankees, which ranks first in MLB for that season. To simulate a 'best of n' series between the first ranked team (team i_1, where $\bar{\widehat{\theta}}_{q,s,|i_1|} = \bar{\widehat{\theta}}_{q,s,|1|}$) and the second ranked team (i_2), the following procedure was used in each q and s.

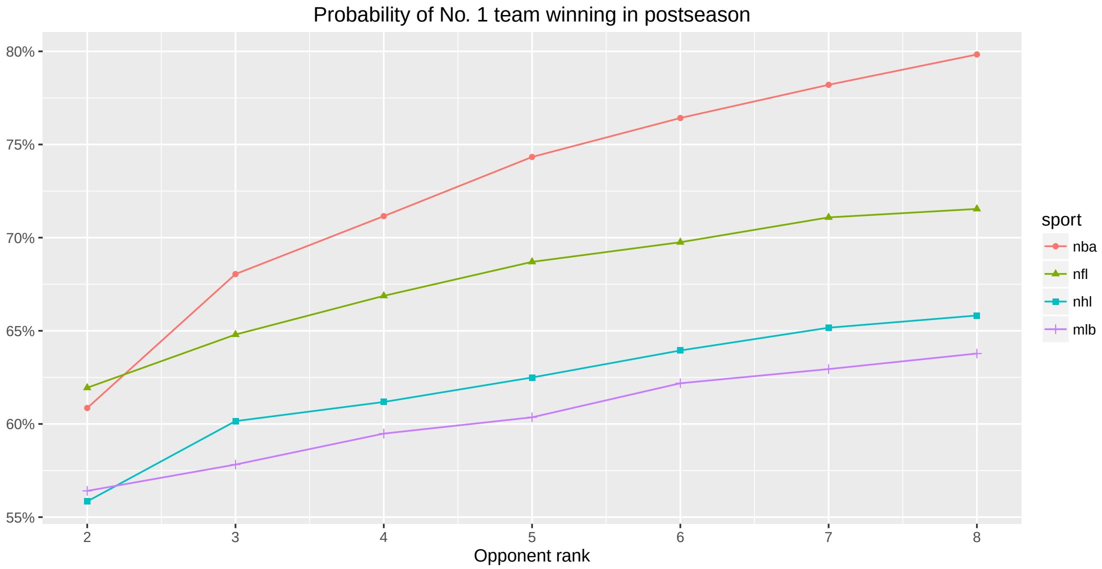

Figure 4: Probability top ranked team wins a postseason series by sport and opponent rank

1. Draw n samples from each the posterior distributions of $\widehat{\theta}_{q,s,i_1}$ and $\widehat{\theta}_{q,s,i_2}$, respectively, labeled as $\widetilde{\theta}_{q,s,i_1}$ amd $\widetilde{\theta}_{q,s,i_2}$, respectively.

2. Draw n samples from the posterior distribution of $\widehat{\alpha}_q$, labeled as $\widetilde{\alpha}_q$.

3. Estimate $\widetilde{logit}(p_{q,s,i_1,i_2}) = \widetilde{\theta}_{q,s,i_1} - \widetilde{\theta}_{q,s,i_2} + \widetilde{\alpha}_q * \mathbb{1}^*$, where $\mathbb{1}^*$ is an n-dimensional vector of alternating 1's and -1's to account for rotating home advantages.

4. Sample from $exp\{\widetilde{logit}(p_{i_1,i_2,q,s})\}/(1 + exp\{\widetilde{logit}(p_{i_1,i_2,q,s})\}$ to obtain 1's and 0's reflecting simulated postseason outcomes. If the sum of these simulations is greater than $n/2$, team i_1 wins the series. Otherwise, team i_2 wins the series.

5. Repeat 10,000 times for each q and s, and for teams ranked third (i_3) through eight (i_8).

Figure XX shows a plot of the estimated probabilities that the top-ranked team beats opponents ranked No. 2 - No. 8 in each sport, averaged over s.

In general, the likelihood that the better team wins is highest in the NBA, ranging from roughly 60% (No. 1 over No. 2) to 80% (No. 1 over No. 8). Most often, the top NFL team expects to win between 60% and 70% of its postseason match-ups, while those numbers are closer to between 55% and 65% in the MLB and NHL. Across seeds No. 3 - No. 8, the MLB shows the most randomness; even in an average match-up between its top team and its eighth best team, the top team wins no more than 65% of the time.

One likely reason that the NFL falls short of the NBA in this chart is that the NBA's series length gives the better team more chances to win. An alternative question explores the number of games that other sports must play in order to match the frequency with which the best team in the NBA wins. To ensure an 80% chance that the league's top team has defeated its eighth best team, the NFL (which would require a 'Best of 9'), NHL ('Best of 39'), and MLB ('Best of 51') must each play a substantially larger number of games than current league standards allow. Likewise, to ensure a 72% chance that the league's top team has defeated its fourth best team, simlar game thresholds must be reached (9 games the NFL, 41 in the NHL, and 55 in MLB).

7 Conclusion

Measurements of competitive balance: Competitive balance at a single point in time (independent of scheduling, compare to ???), Competitive balance between seasons (compare to Humphreys), competitiveness within a season (compare to)

Alternative idea: Fix point in time. Draw from posterior distribution of each team's beta at that time. Calculate standard deviation of those posterior draws. Repeat 10000 times to get a posterior distribution of the standard deviation of team strength (CBR at that time)

Competitive balance between seasons

Posterior distribution of gamma season

Competitive balance within a season

Posterior distribution of the gammaWeek -Function of overall differences in team strength (2% in NBA is not same as 2% in MLB) -Function of injuries and players resting, which aren't necessarily getting at team strength

7.1 Summary

7.2 Open Problems

value of unbalanced vs balanced schedule: fix team strengths, estimate results relative to current system.

Change to a day-by-day structure to account for MLB starting pitchers

Different model specification: use stochastic variance terms which may more aptly account for sudden changes in team strength. Related: out of sample testing with respect to model fits.

Simulate balanced versus unbalanced scheduling to understand how each league's schedule impacts winning percentages (and the resulting measurements of competitive balance in the econ literature)

8 Assorted notes

Points from (Glickman and Stern, 2016):

Restriction of ρ so that stochastic process on θ_{jt} is stationary. Considered to be a normal linear state space model and is also example of the Kalman filter (see paper for citations).

Alternative specification: stochastic volatility process (Kim 98 Jacquier 2002, Glickman (2001). Assumes stochastic process on the τ_t such that $\tau_{t+1} \sim N(\log\tau_t^2, \omega^2)$. 'This model accounts for sudden changes in team ability that are not well captured by normal state space model'. Unknown question: which assumption is preferred? Not addressed in Glickman (2001).

Fitting process: Rjags process, 30k iterations, burn in 10k, every 5th sampled value beyond burn-in. Estimated $\rho = 0.67$ for season to season reversion towards mean ability.

Interesting finding: adding in several seasons lessons the error on team strength within a single season. That is, the season prior and after linked to performance in a single season.

Neat visualization on shrinkage toward league average

Suggestion for varying HFA in hockey: West coast teams more consistently awarded rest advantages (suggested by Mike's hockey people).

Notes on DIC https://users.soe.ucsc.edu/~draper/rjags.pdf The problem of determining what is a noteworthy difference in DIC (or other penalized deviance) between two models is currently unsolved. Following the results of Ripley (1996) on the Akaike Information Criterion, Plummer (2008) argues that there is no absolute scale for comparison of two penalized deviance statistics, and proposes that the difference should be calibrated with respect to the sample standard deviation of the individual contributions from each observed stochastic node.

Gelman http://andrewgelman.com/2011/06/22/deviance_dic_ai/ One of my practical problems with DIC is that it seems to take a lot of simulations to calculate it precisely. Long after weve achieved good mixing of the chains and good inference for parameters of interest and were ready to go on, it turns out that DIC is still unstable. In the example in our book we ran for a zillion iterations to make sure the DIC was ok.

9 Appendix

(Humphreys, 2002) derives the 'Competitive Balance Ratio' to assess between season changes in standings. CBR is a function of within team variation (over time) and between team variation (within a season), measured using win percentage. It can account for varying season lengths, and potentially boasts positive links to attendance.

Lenten (2015) proposed measurements of competitive balance to account for conference and division-based unbalanced scheduling. Implementing on the NFL, it is suggested that teams with high win ratios have generally played easier schedules, and teams with low win ratios have tended to play more difficult schedules. 'The Seahawks effectively had an advantage of more than 8.6 wins over 10 seasons distributed to them via mere virtue of the schedule.'

Hirfindahl-Hirschman Indexes (HHI) measure of competitive balance looks at the concentration of championships or other related outcomes (such as finishing in first place in a division).

Noll-Scully most common metric of assessing competitive balance (Noll, 1988; Scully, 1989). N-S is the ratio of the observed standard deviation of number of wins and the idealized standard deviation, that which would have been observed under binomial distribution (Ex: ISD in NFL = 8/sqrt(16)).

N-S is influence by season length (fluctuation of observed SD, and influence of ISD). In general, N-S increasing in terms of number of games played.

Gini coefficient as one measure of within season inequality of league outcomes, comparing proportion of the total wins of the population which has been cumulatively earned by the bottom proportion of teams, relative to idealized value. Issues with Gini given unbalanced schedule, inter-league play (Utt and Fort, 2002).

Lee (2009) contrasts parity in the NFL before and after the 1993 collective bargaining agreement, finding interseasonal parity, as measured by looking at winning percentages, increased as a result. The author finds first-order autoregressive nature to winning percentages; with coefficients suggesting team win percentages will revert towards league average about 70%.

Fort and Lee (2007) identify that most break points in the four major sports, including expansion and team relocation, have corresponded with increased competitive balance. As a result, leagues tend to be more balanced than in the past.

(Owen, 2010; Owen and King, 2015) uses simulations to identify that RSD is sensitive to changes in season length when there are imbalances in team strength. Team strength parameters for the simulations are estimated using BTM. ASD (RSD without dividing by ISD) less sensitive to changes in season length, although it tends to overestimate imbalance in shorter seasons. Additionally, RSD has an upper bound that is a function of games played (Owen, 2010).

References

Baker, R. D. and McHale, I. G. (2015), "Time varying ratings in association football: the all-time greatest team is.." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178, 481–492.

Berri, D. J. and Schmidt, M. B. (2006), "On the road with the National Basketball Association's superstar externality," *Journal of Sports Economics*, 7, 347–358.

Boulier, B. L. and Stekler, H. O. (2003), "Predicting the outcomes of National Football League games," *International Journal of Forecasting*, 19, 257–270.

Bradley, R. A. and Terry, M. E. (1952), “Rank analysis of incomplete block designs: I. The method of paired comparisons,” *Biometrika*, 39, 324–345.

Buttrey, S. E. (2016), “Beating the market betting on NHL hockey games,” *Journal of Quantitative Analysis in Sports*, 12, 87–98.

Carlin, B. P. (1996), “Improved NCAA basketball tournament modeling via point spread and team strength information,” *The American Statistician*, 50, 39–43.

Cattelan, M., Varin, C., and Firth, D. (2013), “Dynamic Bradley–Terry modelling of sports tournaments,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62, 135–150.

CFP (2014), “Bowl Championship Series explained,” http://www.collegefootballpoll.com/bcs_explained.html, accessed May 19, 2016.

Colquitt, L. L., Godwin, N. H., and Caudill, S. B. (2001), “Testing efficiency across markets: Evidence from the NCAA basketball betting market,” *Journal of Business Finance & Accounting*, 28, 231–248.

Crooker, J. R. and Fenn, A. J. (2007), “Sports leagues and parity when league parity generates fan enthusiasm,” *Journal of Sports Economics*, 8, 139–164.

Demers, S. (2015), “Riding a probabilistic support vector machine to the Stanley Cup,” *Journal of Quantitative Analysis in Sports*, 11, 205–218.

Fort, R. and Lee, Y. H. (2007), “Structural change, competitive balance, and the rest of the major leagues,” *Economic Inquiry*, 45, 519–532.

Glickman, M. E. (2001), “Dynamic paired comparison models with stochastic variances,” *Journal of Applied Statistics*, 28, 673–689.

Glickman, M. E. and Stern, H. S. (1998), “A state-space model for National Football League scores,” *Journal of the American Statistical Association*, 93, 25–35.

— (2016), “Estimating team strength in the NFL,” .

Harville, D. (1980), “Predictions for National Football League games via linear-model methodology,” *Journal of the American Statistical Association*, 75, 516–524.

Humphreys, B. R. (2002), “Alternative measures of competitive balance in sports leagues,” *Journal of Sports Economics*, 3, 133–148.

Humphreys, B. R., Paul, R. J., and Weinbach, A. (2014), “Understanding Price Movements in Point-Spread Betting Markets: Evidence from NCAA Basketball,” *Eastern Economic Journal*, 40, 518–534.

Knorr-Held, L. (2000), “Dynamic rating of sports teams,” *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49, 261–276.

Knowles, G., Sherony, K., and Haupert, M. (1992), “The demand for Major League Baseball: A test of the uncertainty of outcome hypothesis,” *The American Economist*, 36, 72–80.

Koopmeiners, J. S. (2012), “A Comparison of the Autocorrelation and Variance of NFL Team Strengths Over Time using a Bayesian State-Space Model,” *Journal of Quantitative Analysis in Sports*, 8.

Lee, T. (2009), “Competitive balance in the national football league after the 1993 collective bargaining agreement,” *Journal of Sports Economics*.

Lee, Y. H. and Fort, R. (2008), “Attendance and the uncertainty-of-outcome hypothesis in baseball,” *Review of Industrial Organization*, 33, 281–295.

Leeds, M. and Von Allmen, P. (2004), “The economics of sports,” *The business of sports*, 361–366.

Lenten, L. J. (2015), “Measurement of competitive balance in conference and divisional tournament design,” *Journal of Sports Economics*, 16, 3–25.

Loeffelholz, B., Bednar, E., Bauer, K. W., et al. (2009), “Predicting NBA games using neural networks,” *Journal of Quantitative Analysis in Sports*, 5, 1–15.

Lopez, M. J. and Schuckers, M. (2016), “Predicting coin flips: using resampling and hierarchical models to help untangle the NHLs shoot-out,” *Journal of Sports Sciences*, 1–10.

Manner, H. (2015), “Modeling and forecasting the outcomes of NBA basketball games,” *Journal of Quantitative Analysis in Sports*.

Massey, K. (1997), “Statistical models applied to the rating of sports teams,” *Bluefield College*.

Matthews, G. J. (2005), “Improving paired comparison models for NFL point spreads by data transformation,” Ph.D. thesis, WORCESTER POLYTECHNIC INSTITUTE.

Miljković, D., Gajić, L., Kovačević, A., and Konjović, Z. (2010), “The use of data mining for basketball matches outcomes prediction,” in *IEEE 8th International Symposium on Intelligent Systems and Informatics*, IEEE, pp. 309–312.

Mizak, D., Stair, A., and Rossi, A. (2005), "Assessing alternative competitive balance measures for sports leagues: a theoretical examination of standard deviations, Gini coefficients, the index of dissimilarity," *Economics Bulletin*, 12, 1–11.

Moskowitz, T. and Wertheim, L. J. (2011), *Scorecasting: The hidden influences behind how sports are played and games are won*, Crown Archetype.

Nichols, M. W. (2012), "The impact of visiting team travel on game outcome and biases in NFL betting markets," *Journal of Sports Economics*, 1527002512440580.

Noll, R. G. (1988), "Professional basketball," *Studies in Industrial Economics Paper.*

Owen, A. (2011), "Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter," *IMA Journal of Management Mathematics*, 22, 99–113.

Owen, P. D. (2010), "Limitations of the relative standard deviation of win percentages for measuring competitive balance in sports leagues," *Economics Letters*, 109, 38–41.

Owen, P. D. and King, N. (2015), "Competitive balance measures in sports leagues: the effects of variation in season length," *Economic Inquiry*, 53, 731–744.

Owen, P. D., Ryan, M., and Weatherston, C. R. (2007), "Measuring competitive balance in professional team sports using the Herfindahl-Hirschman index," *Review of Industrial Organization*, 31, 289–302.

Paul, R. J. and Weinbach, A. P. (2008), "Price setting in the NBA gambling market: Tests of the Levitt model of sportsbook behavior," *International Journal of Sport Finance*, 3, 137.

— (2011), "NFL bettor biases and price setting: further tests of the Levitt hypothesis of sportsbook behaviour," *Applied Economics Letters*, 18, 193–197.

— (2014), "Market efficiency and behavioral biases in the wnba betting market," *International Journal of Financial Studies*, 2, 193–202.

Scully, G. W. (1989), *The business of major league baseball*, JSTOR.

Soebbing, B. P. and Humphreys, B. R. (2013), "Do gamblers think that teams tank? Evidence from the NBA," *Contemporary Economic Policy*, 31, 301–313.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 583–639.

Stern, H. (1991), "On the probability of winning a football game," *The American Statistician*, 45, 179–183.

Sturtz, S., Ligges, U., Gelman, A., et al. (2005), "R2WinBUGS: a package for running WinBUGS from R," *Journal of Statistical software*, 12, 1–16.

Tutz, G. and Schauberger, G. (2015), "Extended ordered paired comparison models with application to football data from German Bundesliga," *AStA Advances in Statistical Analysis*, 99, 209–227.

Utt, J. and Fort, R. (2002), "Pitfalls to measuring competitive balance with Gini coefficients," *Journal of Sports Economics*, 3, 367–373.

Wolfson, J. and Koopmeiners, J. S. (2015), "Who's good this year? Comparing the Information Content of Games in the Four Major US Sports," *arXiv preprint arXiv:1501.07179*.

Yang, T. Y. and Swartz, T. (2004), "A two-stage Bayesian model for predicting winners in major league baseball," *Journal of Data Science*, 2, 61–73.

Supplementary Materials for

"A unified approach to understanding randomness in sport"