Submitted to *Manufacturing & Service Operations Management*
manuscript (MSOM-22-558)

How to Conclude a Suspended Sports League?

Ali Hassanzadeh[a], Mojtaba Hosseini[b], John G. Turner[c]
[a]Alliance Manchester Business School, The University of Manchester, UK; [b] Tippie College of Business, University of Iowa, Iowa, USA; [c]The Paul Merage School of Business, University of California, Irvine, California, USA
ali.h@manchester.ac.uk, mojtaba-hosseini@uiowa.edu, john.turner@uci.edu

Problem definition: Professional sports leagues may be suspended due to various reasons such as the recent COVID-19 pandemic. A critical question the league must address when re-opening is how to appropriately select a subset of the remaining games to conclude the season in a shortened time frame. **Academic/practical relevance:** Despite the rich literature on scheduling an entire season starting from a blank slate, concluding an existing season is quite different. Our approach attempts to achieve team rankings similar to that which would have resulted had the season been played out in full. **Methodology:** We propose a data-driven model which exploits predictive and prescriptive analytics to produce a schedule for the remainder of the season comprised of a subset of originally-scheduled games. Our model introduces novel rankings-based objectives within a stochastic optimization model, whose parameters are first estimated using a predictive model. We introduce a deterministic equivalent reformulation along with a tailored Frank-Wolfe algorithm to efficiently solve our problem, as well as a robust counterpart based on min-max regret. **Results:** We present simulation-based numerical experiments from previous National Basketball Association (NBA) seasons 2004–2019, and show that our models are computationally efficient, outperform a greedy benchmark that approximates a non-rankings-based scheduling policy, and produce interpretable results. **Managerial implications:** Our data-driven decision-making framework may be used to produce a shortened season with 25-50% fewer games while still producing an end-of-season ranking similar to that of the full season, had it been played.

Key words: COVID-19 pandemic; sports scheduling; rankings; concordance; predictive analytics; stochastic optimization; Frank-Wolfe algorithm; min-max regret

1. Introduction

The global COVID-19 pandemic that began in December 2019 quickly led to unprecedented quarantines, lockdowns, and travel restrictions. Worldwide, professional sports leagues ceased activity, with the National Basketball Association (NBA) being the first major league in the US to suspend games, pausing the season as of March 11, 2020 (NBA 2020a). Fans eagerly followed media specu-

lation on the possible actions the NBA would take: whether the 2019–20 regular season would be resumed, and how it would be concluded. The main directions the NBA could have taken were:

1. **Cancel Everything.** Cancel the remaining regular-season games and the playoffs. Determine the champion using information already available (e.g., either the top-ranked team as of the suspension is crowned champion, or votes from players, coaches, and the public are solicited in a manner similar to how individual player awards are selected).
2. **Skip to Playoffs.** Cancel the remaining regular-season games. The top-ranked teams as of the suspension date qualify for the playoffs, which begin immediately.
3. **Full Season.** Resume the regular season and play all 259 scheduled games, followed by the playoffs as usual. The season will end late in the year.
4. **Shortened Season.** Select a subset of the remaining 259 regular-season games to be played, followed by the playoffs.

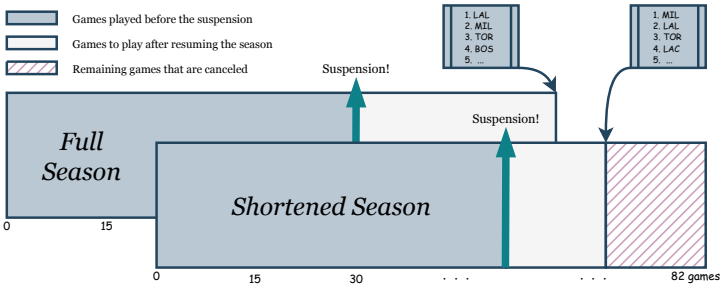


Figure 1 **Two strategies to conclude the league: full season vs. shortened season after resuming the league**

Options 1 and 2 would be unfair for several reasons. Indeed, teams may differ in the number of games played by the suspension date, along with the relative difficulties of the opponents they played. Options 3 and 4, which we compare extensively throughout the paper, are illustrated in Figure 1. Note that the rankings of the teams (by number of wins over the played games) depends on the specific set of games that are played, as well as the outcomes of those games.

Our focus in this paper is to propose a method which chooses a subset of games to conclude a shortened season that remains an asymmetrical round-robin tournament while producing end-of-season rankings that are as close as possible to the rankings that would result had the full season been played (i.e., no games canceled). There are, of course, many considerations that come into play when constructing a sports schedule. While our approach is less detailed than constructing a full timetable that incorporates not only the set of games to play but also their sequence (and corresponding travel schedule), our model is sufficiently general to allow for logical constraints on the subset of games chosen, which may be driven by specific practical considerations.

At a high level, our model selects which games to include in the remainder of the season. To make these decisions, we develop several model components. First, we use a predictive model to estimate

the likelihood of each game outcome for all games in the season that have not yet been played. Then, we generate one or more scenarios, and for each scenario we produce a "target ranking" which ranks the relative performance of the teams if all games in the remainder of the season were played. We then use a prescriptive model to select a subset of games so that, in expectation, the ranking we get from our shortened season is as close to the target ranking as possible.

Note that our methodology may be applied broadly to any sports league requiring schedule adjustments due to a suspension (i.e., pause in game play, typically for multiple games over an extended period). Our approach is most applicable when the league combines a round-robin regular season with an elimination postseason ("playoffs"), which is the case for most major sports leagues in North America, including the NBA, National Football League (NFL), National Hockey League (NHL), and Major League Baseball (MLB). Table 1 lists the number of suspensions in the NBA, NFL, NHL, and MLB from 1946–2021, along with the number of suspensions that led to shortened seasons. In summary, suspensions are an occasional occurrence of significant consequence, and our methodology can be applied to resuming a league's regular season from any such suspension.

League	NBA	NFL	NHL	MLB
Number of Suspensions	6	8	4	9
Suspensions with Shortened Seasons	4	2	2	3

Table 1 The number of disruptions to regular season games.

Our paper is organized as follows. We review related works in §2, and mathematically define our problem in §3. In §4 we introduce our predictive and prescriptive models (binary classifiers and stochastic optimization models, respectively), and in §5 we develop solution techniques (one based on the Frank-Wolfe algorithm) for solving our prescriptive models efficiently. Our best prescriptive model extends our base model using the concept of min-max regret, and produces shortened seasons that perform well while being robust to (predictive) model misspecification and overfitting. Next, in §6 we use Monte Carlo simulation to evaluate our models, and show that not only do the shortened seasons produced by our best model have rankings that are close to the counterfactual end-of-season ranking, but our rankings are in high agreement with respect to the teams that make the playoffs (95.65%), the teams that get home court advantage in the playoffs (92.28%), and the teams that receive double-digit lottery odds in the rookie draft (91.36%). We also provide a model extension that ensures each team's strength-of-schedule is not materially impacted by our choice of shortened season. Finally, we conclude our paper in §7.

2. Literature Review

A distinct feature of our study is the two-phase analytics approach that combines predictive and prescriptive models. We review existing literature in both streams: predicting single-game outcomes and end-of-season rankings, and scheduling sports leagues using algorithms and optimization.

Predictive models for single-game outcomes. There is an extensive literature on predicting the outcome of a single sports game using historical data. On the one hand, there is an inherent difficulty in making such predictions, as the outcome of a game depends both on luck and skill (Aoki et al. 2017), and there is a limit to how much one can disentangle team/player skills from randomness (Martin et al. 2016). On the other hand, with ever-growing access to sports data and advancements in the fields of data analysis and machine learning, there has been a growing interest in predicting the outcomes of sporting events, both among researchers and for-profit organizations (e.g., FiveThirtyEight 2022). Existing models analyze the outcomes of sporting events using (i) Bayesian inference and rule-based reasoning (Miljković et al. 2010), (ii) Markov chain modeling (Kvam and Sokol 2006, Brown and Sokol 2010), (iii) machine learning (Magel and Melnykov 2014, Prasetio et al. 2016), or (iv) wisdom of crowds (Halberstadt and Levine 1999). A key differentiator of our approach stems from the fact that we do not only predict the outcome of the next game to be played using all historical data available prior to that game. Instead, we predict the outcomes of all post-suspension games using only data from the pre-suspension period.

Predictive models for end-of-season rankings. A few researchers have developed models to directly predict end-of-season team rankings given historical data of game play up to a certain (e.g., suspension) date. To the best of our knowledge, the first such paper is Van Haaren and Davis (2015), who study the final league rankings in European football leagues both before the start of the season and during the course of the season. A body of research also focuses on determining the true ranking of individuals or teams in competitive sports based on network-based ranking systems (Motegi and Masuda 2012, Bozóki et al. 2016). More recently, inspired by the COVID-19 suspension in the European football leagues, researchers Van Eetvelde et al. (2021) and Csató (2021a) studied the suspended season problem (also known as "incomplete round robin league") with the aim of predicting the final team ranking without additional games being played (i.e., if put into practice, this model would declare the league champion directly without resuming the season or playing any additional games). These studies use descriptive and predictive models based on historical data from all games played prior to the suspension, to obtain a measure of strength for each team in the league and a measure of toughness of schedule. Using these two measures, the authors produce a projected final ranking and evaluate the accuracy of their proposed ranking using Monte Carlo simulation. A key differentiator of our approach is our substantial prescriptive component. These papers do not predict outcomes of individual games and instead directly predict end-of-season

rankings, which could be used to declare league standings without playing any additional games. In contrast, our predictive model produces predictions for each game that we then feed into our prescriptive model to determine the subset of games to play in our prescribed shortened season.

Prescriptive models for scheduling sports games. Within the sports scheduling literature, there are articles that focus on a primary objective (e.g., broadcast TV or travel logistics, fairness considerations), and others that propose multi-objective models. Among those that use optimization to schedule basketball games, Bean and Birge (1980) consider travel costs and player fatigue as the main goals, Weiss (1986) studies schedule bias between the regular season and post-season, while Westphal (2014) focuses on venue availability and broadcasting considerations. To propose a schedule for NCAA basketball games, Nemhauser and Trick (1998) and Henz (2001) apply integer programming and constraint programming, respectively. Other papers develop tailored algorithms, often based on graph theory, for scheduling basketball games, see Briskorn and Drexl (2009). We suggest the survey paper Rasmussen and Trick (2008) for an overview of round-robin scheduling. For a comprehensive list of articles in the broader scope of analytical methodologies applied to sports, including optimization and probabilistic modeling, see Fry and Ohlmann (2012a,b). Mixed-integer programs have been used to schedule games in different leagues: Fleurent and Ferland (1993) in hockey, Goossens and Spieksma (2009) in soccer, Jiaqi Xu et al. (2019) in baseball, and Cocchi et al. (2018) in volleyball.

In contrast to the existing literature that schedules an entire season from a blank slate, the problem we consider compresses the remainder of an already-started season by selecting to play a subset of previously-scheduled games. To the best of our knowledge, this problem has not been previously studied. A significant novelty in our approach is our objective function, which attempts to achieve rankings similar to those that would have resulted had the season been played in full; this motivates us to introduce several model components which are novel in the context of sports scheduling, including ranking-based objectives, related stochastic optimization models, and finally predictive models for estimating the parameters of our stochastic optimization models.

3. Problem

In this section, we first describe an NBA season and explain why a team's end-of-season ranking is important. We then frame the conclusion of the season as a problem of selecting a subset of the remaining games, which leads us to introduce several ranking similarity metrics.

3.1. Background

The NBA is composed of 30 teams, divided into 2 conferences of 3 divisions with 5 teams each; for details, see Figure 2. In a regular season spanning approximately 180 days from October through April, each team plays 82 games according to the following formula: four games against the other

four division opponents ($4 \times 4 = 16$ games), four games against six (out-of-division) conference opponents ($4 \times 6 = 24$ games), three games against the remaining four conference teams ($3 \times 4 = 12$ games), and finally two games against teams in the opposing conference ($2 \times 15 = 30$ games). A five-year rotation determines which out-of-division conference teams are played only three times. After 5 seasons, each team will have played 20 games against each in-division opponent, 18 games against each out-of-division opponent, and 10 games against each team from the opposing conference.

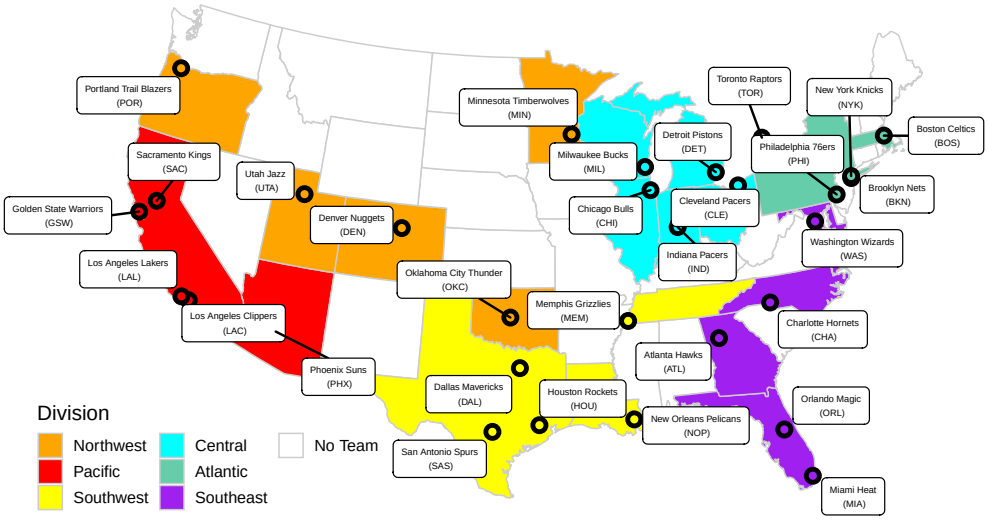


Figure 2 **The Eastern Conference is comprised of the Central, Atlantic, and Southeast divisions, while the Western Conference consists of the Northwest, Pacific, and Southwest divisions.**

In our paper, we assume the regular-season standings directly determine the teams that advance to the playoffs. This follows the practice prior to the 2020–21 season, in which the 8 top-ranked teams in each conference (16 in the league) advance to the playoffs. However, starting with the 2020–21 season, only the top 6 teams in each conference automatically advance to the playoffs, with teams ranked 7 through 10 competing in a *play-in tournament* to determine who is seeded 7^{th} and 8^{th} and who is eliminated; for details, see NBA.com (2023). Note that the new format doesn't materially change our models or testing methodology, and since our problem instances are all from pre-COVID seasons, we exclude the play-in tournament from our analysis.

Once the standings are finalized, the playoffs begin. In each conference, the i^{th}-ranked team is initially matched to the $(9-i)^{th}$-ranked team, for $i \in \{1 \dots 8\}$. All playoff matchups are best-of-seven series, i.e., a team needs to win four out of seven games against the same opponent to win the matchup and progress to the next round (the loser of the matchup is eliminated). All matchups occur within-conference until the final matchup, which pits the winning team of the Eastern conference against the winning team of the Western conference. It is also important to note that the higher-ranked team in each matchup is awarded home court advantage; this means

it hosts games 1, 2, 5, and 7, while the lower-ranked team hosts games 3, 4, and 6 (with games 5–7 played only if needed). Note that due to how teams are matched, the 4 top-ranked teams in each conference are given home court advantage in the first round of the playoffs.

There is a strong connection between a team's regular-season ranking and its playoff performance. Examining the history of 76 completed NBA seasons from 1946–2021, we find that (i) over 75% of playoff series were won by the team with home court advantage, (ii) in only 5 seasons did an 8^{th}-ranked team win a playoff series against a 1^{st}-ranked team, and strikingly (iii) of the 76 NBA champions, 73 were ranked among the top 3 teams in the league at the end of the regular season.

To the extent that end-of-season rankings give teams preferential treatment in the playoffs which boost a team's chances of winning a championship, it is in the league's best interest to ensure that the ranking is fair, i.e., reflects to the greatest extent possible which teams are truly the best. Fairness can be in question when the season ends early. This is because when only a shortened season is played, it is possible for some teams to be matched with relatively easy-to-beat teams while others are matched with harder-to-beat teams, and this may result in a ranking that is quite different than one which would have resulted had the season been played in full. (We assume the full season's ranking is fair, since the league constructs the full season schedule in a balanced and equitable manner, and in general the public accepts the ranking at the end of the full season).

Finally, the order in which teams pick rookie players in the annual NBA draft is also tied to the final ranking, with the lowest-ranked teams given a higher chance of drafting the best rookie player. Therefore, the ranking of teams outside of the top 8 in each conference is also important. The quality of the players in the draft varies from season to season, but some first-pick rookie players have included generational talents such as *LeBron James*, *Magic Johnson*, and *Hakeem Olajuwon*, who have had huge impacts in leading their respective teams to win multiple championships.

3.2. Problem Description

At the time of the 2019–20 COVID-19 suspension, 971 games in the 1230-game season were played leaving 259 games remaining; see Figure 3 for the ranking at the time of suspension. Given a target number of games each team should play in the full season, we wish to select a subset of the remaining 259 games that satisfies these targets. Typically, each of the 30 teams plays 41 home and 41 away games for a total of 82 games in the season. Shortening the season involves reducing the target from 82 games/team to a lower number (e.g., 70), with half the games at home and half away. Since the results of the 259 remaining games are uncertain, both the ranking produced by playing the full 82 game/team season and the ranking produced by playing a shortened (e.g., 70 game/team) season are uncertain. Our problem is to select a subset of games that minimizes the expected dissimilarity between the ranking of the full season and the ranking of the shortened season. Before we introduce our models, we introduce several metrics that may be used for measuring the similarity of rankings.

Western Conference						Eastern Conference					
League	West		Wins	Losses	Win%	League	East		Wins	Losses	Win%
2	1	Los Angeles Lakers	49	14	0.778	**1**	1	Milwaukee Bucks	53	12	0.815
4	2	Los Angeles Clippers	44	20	0.688	**3**	2	Toronto Raptors	46	18	0.719
6	3	Denver Nuggets	43	22	0.662	**5**	3	Boston Celtics	43	21	0.672
7	4	Utah Jazz	41	23	0.641	**8**	4	Miami Heat	41	24	0.631
9	5	Oklahoma City Thunder	40	24	0.625	**11**	5	Indiana Pacers	39	26	0.600
10	6	Houston Rockets	40	24	0.625	**12**	6	Philadelphia 76ers	39	26	0.600
13	7	Dallas Mavericks	40	27	0.597	**15**	7	Brooklyn Nets	30	34	0.469
14	8	Memphis Grizzlies	32	33	0.492	**16**	8	Orlando Magic	30	35	0.462
17	9	Portland Trail Blazers	29	37	0.439	**22**	9	Washington Wizards	24	40	0.375
18	10	New Orleans Pelicans	28	36	0.438	**23**	10	Charlotte Hornets	23	42	0.354
19	11	Sacramento Kings	28	36	0.438	**24**	11	Chicago Bulls	22	43	0.338
20	12	San Antonio Spurs	27	36	0.429	**25**	12	New York Knicks	21	45	0.318
21	13	Phoenix Suns	26	39	0.400	**26**	13	Detroit Pistons	20	46	0.303
28	14	Minnesota Timberwolves	19	45	0.297	**27**	14	Atlanta Hawks	20	47	0.299
30	15	Golden State Warriors	15	50	0.231	**29**	15	Cleveland Cavaliers	19	46	0.292

Figure 3 NBA ranking at the time of suspension on March 11, 2020.

3.3. Measures of Similarity/Dissimilarity between Rankings

We represent a ranking of n teams as a vector, with components 1 through n permuted in order from the highest to the lowest percentage of games won during the regular season. Throughout the paper, we follow the convention that $\hat{r}$ represents a ranking resulting from playing all games in the full season, while r represents a ranking resulting from playing a specific subset of the remaining games (i.e., the shortened season case). Furthermore, when we wish to distinguish between multiple rankings in the shortened season case, we use a superscript. For example, $r^{(1)}$ and $r^{(2)}$ represent two distinct rankings resulting from concluding a shortened season with two different sets of games.

Two widely used measures of similarity/dissimilarity between rankings are *Concordance*, and *Euclidean distance*. We now define these metrics, using the following small example.

Example 1: Assume there are only four teams in the league: `LAL`, `MIL`, `LAC`, and `BOS`. Table 2 contains the full-season ranking $\hat{r}$, and two alternative rankings $r^{(1)}$ and $r^{(2)}$.

Teams	Ranking ($\hat{r}$)	Ranking ($r^{(1)}$)	Ranking ($r^{(2)}$)
`LAL`	1	1	4
`BOS`	2	4	1
`MIL`	3	2	3
`LAC`	4	3	2

Table 2 Example 1: Three different rankings.

3.3.1. Concordance. Concordance is a metric used to measure the ordinal association between two measured quantities, each with n elements. Intuitively, concordance is high when observations in two variables have similar ranks, and it is low when observations have dissimilar (opposite) ranks. Concordance, as a metric, is inspired by Kendall's rank correlation coefficient, or simply Kendall's τ, introduced in Kendall (1938). For a given pair of team rankings $(r, \hat{r})$, we call a pair of teams (i, j) *concordant* if either i is above j in both rankings (i.e., $\hat{r}_i > \hat{r}_j$ and $r_i > r_j$) or i is below j in both rankings (i.e., $\hat{r}_i < \hat{r}_j$ and $r_i < r_j$). Conversely, we say a pair of teams (i, j) is

discordant if their relative positions in the two rankings do not agree (i.e., either $\hat{r}_i > \hat{r}_j$ and $r_i < r_j$, or alternatively $\hat{r}_i < \hat{r}_j$ and $r_i > r_j$). If $\hat{r}_i = \hat{r}_j$ or $r_i = r_j$, the pair of teams is neither concordant nor discordant. Following our example, when we compare rankings $\hat{r}$ and $r^{(1)}$ from Table 2, the pair (`LAL`, `BOS`) is concordant, since in both rankings `LAL` stands higher than `BOS`. The pair (`BOS`, `MIL`) however is discordant, since `BOS` has the higher rank in $\hat{r}$, while `MIL` is higher in $r^{(1)}$.

Using the number of concordant pairs given two rankings, concordance (τ_C) is defined as

$$\tau_C = \text{number of concordant pairs.} \tag{1}$$

Note that τ_C is a number between 0 and $\binom{n}{2}$. Continuing our example, concordance is $\tau_C(r^{(1)}, \hat{r}) = 4$ between $(r^{(1)}, \hat{r})$ and is $\tau_C(r^{(2)}, \hat{r}) = 2$ between $(r^{(2)}, \hat{r})$.

3.3.2. Euclidean distance. The (squared) Euclidean distance between two rank vectors is another metric used to measure dissimilarity between two alternative rankings, defined as

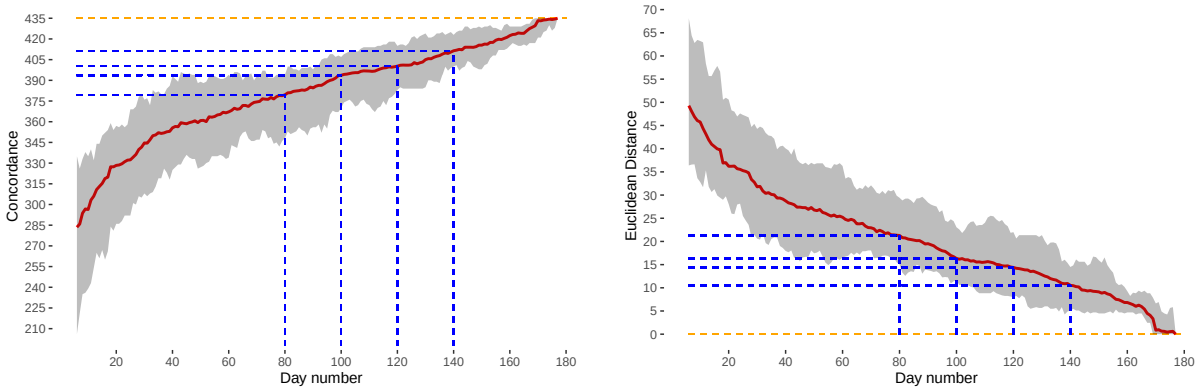
$$d_{\mathrm{E}}(r, \hat{r}) = \|r - \hat{r}\|^2 = \sum\nolimits_{i \in T} (r_i - \hat{r}_i)^2, \tag{2}$$

where T denotes the set of all the teams in the league. It can be easily verified that $d_{\mathrm{E}}(\hat{r}, r^{(1)}) = 6$ and $d_{\mathrm{E}}(\hat{r}, r^{(2)}) = 14$, which is consistent with our conclusion based on concordance τ_C that $r^{(1)}$, with shorter distance to $\hat{r}$, is the most similar to $\hat{r}$. We remark that Euclidean distance is equivalent to *Spearman's Rank Correlation Coefficient* (Spearman 1904) through an affine transformation with negative coefficient. We also note the following relationship between concordance and Euclidean distance of two rankings, which implies that highly concordant rankings have low Euclidean distance, and vice versa. In our experiments, the lower bound on $d_{\mathrm{E}}(r, \hat{r})$ in particular, proves to be very tight. Proof of this proposition uses the Durbin-Stuart inequality (Durbin and Stuart 1951) and Diaconis-Graham inequality (Diaconis and Graham 1977) and together with other proofs are provided in Appendix B.

Proposition 1. *For arbitrary rankings r and $\hat{r}$, the following relationship holds:*

$$\frac{4}{3n}\left(\frac{n(n-1)}{2} - \tau_C(r,\hat{r})\right)\left(\frac{n(n+1)}{2} - \tau_C(r,\hat{r})\right) \le d_E(r,\hat{r}) \le 2\left(\frac{n(n-1)}{2} - \tau_C(r,\hat{r})\right)^2. \tag{3}$$

3.3.3. Measuring ranking similarity/dissimilarity across time. It is interesting to plot how our metrics change over the course of an NBA season, as we compare daily team rankings (r) to the end-of-season ranking ($\hat{r}$). Note that with 30 teams in the NBA, the maximum concordance is $\binom{30}{2} = 435$. Figure 4(a) plots concordance over time, while Figure 4(b) plots Euclidean distance. The red line is the mean over 14 NBA seasons from 2004–2018, while the grey band illustrates the range of values over these 14 seasons. As we approach the end of the season, concordance approaches its maximum of 435 while Euclidean distance converges to its minimum of 0.



(a) Concordance ($\tau_{\mathrm{C}}(r, \hat{r})$) (b) Euclidean distance ($\sqrt{d_{\mathrm{E}}(r, \hat{r})}$)

Figure 4 **Measuring ranking similarity/dissimilarity across time; grey area shows the range across 14 NBA seasons.**

4. Models

Our two-phase modeling approach (c.f., Figure 5) falls under the "predict-then-optimize" paradigm (Mišić and Perakis 2020). To choose the best subset of games for our shortened season, we need estimates of the outcomes of all remaining games. Hence, in our first ("predictive") phase, we use historical data from all regular-season games played before the suspension to predict which teams win the remaining games. More specifically, we train binary classification models that incorporate game-related features (e.g., win percentage, point differential, and home-away indicator, among others). In our second ("prescriptive") phase, we determine which games to include and which games to cancel in the shortened season. We minimize the expected dissimilarity between the shortened season's ranking and the full season's ranking, where this expectation is taken over multiple possible scenarios that reflect the random chance each team has for winning each game. Specifically, we treat game outcomes as Bernoulli random variables whose parameters are estimated in the first ("predictive") phase, and formulate our prescriptive models as stochastic optimization problems.

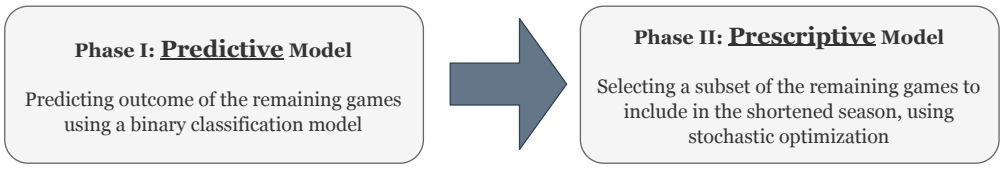


Figure 5 **The two main phases of our methodology.**

4.1. Predictive Models

We now present our models for predicting the outcomes of games postponed due to the suspension. Since the response variable (i.e., whether the home team wins or loses) is binary, we model our

prediction problem as a *binary classification task*. More specifically, given a set of games $\mathcal{G}$ from the pre-suspension period, each data point $g \in \mathcal{G}$ used for training our model consists of a vector of D features $\vec{x}_g = (x_{g,1}, \ldots, x_{g,D})$ and its binary class label y_g, where $y_g = 1$ if the home team wins and $y_g = 0$ otherwise. Our goal is to learn a discrimination rule $p : \mathbb{R}^D \to [0,1]$ representing the probability that a data point with feature vector $\vec{x}$ belongs to class 1 ("home team wins").

We numerically evaluate 8 popular binary classification models from the literature. These include Support Vector Machine (SVM; Cortes and Vapnik 1995), random forest (RF; Breiman 2001), bootstrap aggregating (Bagging; Breiman 1996), eXtreme Gradient Boosting (XGBoost; Chen and Guestrin 2016), Extreme Learning Machines (ELM; Huang et al. 2006), logistic regression (Logit), Gaussian Naïve Bayes (NB), and Multilayer Perceptron (MLP, also known as artificial neural network); see Hastie et al. (2009) for descriptions of the last three. The explanatory variables in our dataset and specific features we use in our models are documented in §6.1 and Appendix E.

Most binary classifiers initially estimate the class probability function $p(\vec{x})$. For each data point $\vec{x}$, they then apply a threshold (e.g., 0.5) to assign labels to observations: label 1 is assigned if $p(\vec{x}) > 0.5$, and 0 otherwise. One notable exception is traditional SVM, which directly assigns each data point to one of the two classes; here, to get probability estimates we follow common practice and employ a sigmoid calibration function (Platt et al. 1999, Niculescu-Mizil and Caruana 2005).

We perform model selection by evaluating our binary classification models using a performance metric, and choosing the best one. Two applicable categories of performance metric are (1) those that compare predicted class (which is binary) with actual outcomes (also binary), and (2) those that compare predicted probability with actual outcomes. The most natural criterion in the first category is *misclassification loss* (the complement of *accuracy*, or the rate of correctly classified data points). But, since we are primarily interested in class membership probabilities rather than zero-one labels, we adopt a performance metric from the second category. Specifically, we use *proper scoring rules*, which measure the quality of predicted probabilities (Gneiting and Raftery 2007).

For observations drawn from the distribution F, a scoring rule is called *proper* if its expectation is maximized when the forecaster issues the probabilistic forecast F. Moreover, if F is the unique maximizer, the scoring rule is *strictly proper*. Examples of strictly proper scoring rules include the logarithmic, quadratic, and spherical scoring rules. Brier score and LogLoss, which measure the distance between estimated and true outcomes, are the complements of the quadratic and logarithmic scoring rules, respectively. For details, see Bickel (2007) and Johnstone et al. (2011).

According to Bickel (2007), the logarithmic scoring rule favors *sharper* probability values (i.e., those that are closer to zero or one). As suggested by Johnstone et al. (2011), sharper probability values are preferred when the end use of estimated probabilities is a stochastic optimization problem, as we have in our second phase. This is because they reduce variance by focusing on the

most likely scenarios. Hence, we use the logarithmic scoring rule as our model selection criterion to select the best-performing of our binary classifiers. In line with machine learning terminology, we henceforth use LogLoss defined below, which is the negative of the logarithmic scoring rule:

$$\text{LogLoss} = -\frac{1}{|\mathcal{G}|}\sum\nolimits_{g\in\mathcal{G}}\Big(y_g\ \log(p_g) + (1-y_g)\ \log(1-p_g)\Big), \tag{4}$$

where $\{p_g\}_{g\in\mathcal{G}}$ is the vector of predicted probabilities and $\{y_g\}_{g\in\mathcal{G}}$ is the vector of true outcomes. We compare the predictive performance of candidate models according to LogLoss in §6.2.

4.2. Prescriptive Models

In a league with n teams, let T denote the set of teams. Assume that at the time of suspension, a set G of regular-season games remain to be played, and each team $i \in T$ has won a total of y_i^0 games before the suspension. We represent each game $g \in G$ with a tuple $g = (i, j, k)$, where $i(g) \in T$ and $j(g) \in T$ denote the host and guest teams, respectively, and $k(g)$ indexes the k^{th} match between these two teams (recall two teams may play each other more than once). We also define $G_i^h \subset G$ and $G_i^a \subset G$ as the set of remaining home and away games for team i, respectively.

We model the outcome of game g using the Bernoulli random variable W_g, which is one if the host team $i(g)$ wins, and zero if the guest team $j(g)$ wins. For each game g, we estimate the parameter $p_g = P(W_g = 1)$ using historical data as discussed in §4.1. Formally, we denote the set of all possible outcomes of all games in the remainder of the full season by Ξ, and use $\xi \in \Xi$ to index a specific realization of all games' outcomes. When helpful, we explicitly write $W_g(\xi)$ to indicate W_g's dependence on ξ. For a given outcome $\xi \in \Xi$, the win percentage of team i (total wins divided by games played) after playing all remaining games (i.e., at the end of the full regular season) is

$$\hat{y}_i(\xi) = \frac{1}{\hat{m}}\left(y_i^0 + \sum\nolimits_{g\in G_i^h} W_g(\xi) + \sum\nolimits_{g\in G_i^a}(1 - W_g(\xi))\right), \tag{5}$$

where $\hat{m}$ is the number of games played by each team in the full season (e.g., 82 for the NBA). We will continue to use the caret ($\hat{}$) to denote quantities that correspond to the full regular season.

For each game $g \in G$, we define a binary decision variable x_g that we set to one if game g is included in the shortened season, and zero otherwise. Note that these x-decisions are made before knowing the realization of ξ. We define X as the set of feasible solutions, i.e., restrictions placed on the x-variables to express tactical/fairness considerations such as each team having the same number of home/away games in the season, as well as binary requirements on the x-variables, i.e.,

$$X = \begin{Bmatrix} \sum_{g\in G_i^h} x_g = m_i^h, & \forall i \in T \\ \sum_{g\in G_i^a} x_g = m_i^a, & \forall i \in T \\ x_g \in \{0,1\}, & \forall g \in G \end{Bmatrix}, \tag{6}$$

where m_i^h and m_i^a denote the target number of home and away games for team i, respectively. For instance, if team i has played 33 home and 31 away games so far before the suspension, and we

decide to conclude the season with a total of 72 games for each team, then this team must play an additional $m_i^h = \frac{72}{2} - 33 = 3$ home and $m_i^a = \frac{72}{2} - 31 = 5$ away games. Another alternative would be to combine the constraints on the number of home/away games for each team into a single constraint of the form $\sum_{g \in G_i^h \cup G_i^a} x_g = m_i^h + m_i^a$ that sets a target for the total number of games to play without specific home/away sub-targets.

For a given shortened season $\boldsymbol{x} \in X$ and realization $\xi \in \Xi$, we denote the win percentage of team i at the end of the shortened season as $y_i(\boldsymbol{x}, \xi)$, where

$$y_i(\boldsymbol{x}, \xi) = \frac{1}{m} \left(y_i^0 + \sum\nolimits_{g \in G_i^h} W_g(\xi) x_g + \sum\nolimits_{g \in G_i^a} (1 - W_g(\xi)) x_g \right), \tag{7}$$

and m is the target total number of games for each team in the shortened season (e.g., 72).

Let $d(y(\boldsymbol{x}, \xi), \hat{y}(\xi))$ be a measure of dissimilarity between the vectors $y(\boldsymbol{x}, \xi)$ and $\hat{y}(\xi)$, i.e., a measure that compares the win percentage of each team at the end of the full season with the win percentage of each team at the end of the shortened season, for a specific shortened season $\boldsymbol{x}$ and outcome ξ. Note that there is a one-to-one correspondence between $\hat{y}(\xi)$ and the team rankings at the end of the full season, and between $y(\boldsymbol{x}, \xi)$ and team rankings at the end of the shortened season. Therefore, $d(y(\boldsymbol{x}, \xi), \hat{y}(\xi))$ can also be viewed as a dissimilarity measure between these rankings, and our goal is to find a shortened season $\boldsymbol{x}$ that minimizes the expected value of this dissimilarity. That is, we are interested in solving stochastic optimization problems of the form

$$\min_{\boldsymbol{x} \in X} \quad \mathbb{E}_\xi [d(y(\boldsymbol{x}, \xi), \hat{y}(\xi))], \tag{8}$$

for different choices of the dissimilarity measure d. We now introduce two such formulations.

4.2.1. Maximizing concordance. For a given outcome ξ, let $\hat{r}(\xi)$ and $r(\boldsymbol{x}, \xi)$ denote the ranking vector we get when the full season is played, and respectively, when the shortened season $\boldsymbol{x}$ is played. We solve the following stochastic optimization problem to maximize the expected similarity between these two rankings using the average concordance metric:

$$\max_{\boldsymbol{x} \in X} \quad \mathbb{E}_\xi \left[\tau_C \left(r(\boldsymbol{x}, \xi), \hat{r}(\xi) \right) \right]. \tag{9}$$

While this formulation is compact, its objective function is highly nonlinear; consequently, we linearize it as follows. First, we define a parameter $\hat{z}_{ij}(\xi)$ which takes value one if team i is above team j in the full-season ranking $\hat{r}(\xi)$, and zero otherwise. Similarly, we introduce a binary variable $z_{ij}(\boldsymbol{x}, \xi)$ which takes value one if team i is above team j in the shortened-season ranking $r(\boldsymbol{x}, \xi)$, and zero otherwise. Since $z_{ij}(\boldsymbol{x}, \xi) + z_{ji}(\boldsymbol{x}, \xi) = 1$, we introduce only the z_{ij}-variables where $i < j$ and use $1 - z_{ij}(\boldsymbol{x}, \xi)$ in place of $z_{ji}(\boldsymbol{x}, \xi)$ whenever it is needed. As well, we introduce continuous variables $y_i(\boldsymbol{x}, \xi)$, $i \in T$, to keep track of the win percentage of team i in the shortened season $\boldsymbol{x}$

under realization ξ. Finally, since it is clear that the solution to this optimization problem encodes a single x-vector, we henceforth suppress the x-argument for the y- and z-variables, and restate problem (9) as the following stochastic Mixed Integer Linear Program (MILP):

$$[\text{PC}] \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad \mathbb{E}_\xi \left[\sum_{i\in T} \sum_{j\in T:j>i} (z_{ij}(\xi)\hat{z}_{ij}(\xi) + (1-z_{ij}(\xi))(1-\hat{z}_{ij}(\xi))) \right] \tag{10}$$

$$\text{s.t.} \quad y_i(\xi) = \frac{1}{m}\left(y_i^0 + \sum_{g\in G_i^h} W_g(\xi)x_g + \sum_{g\in G_i^a} (1-W_g(\xi))x_g \right) \qquad \forall i\in T, \forall \xi\in\Xi \tag{11}$$

$$z_{ij}(\xi) - 1 \le y_i(\xi) - y_j(\xi) \le z_{ij}(\xi) \qquad \forall i,j\in T: i<j, \forall \xi\in\Xi \tag{12}$$

$$z_{ij}(\xi) \in \{0,1\} \qquad \forall i,j\in T: i<j, \forall \xi\in\Xi \tag{13}$$

$$\boldsymbol{x}\in X. \tag{14}$$

The objective function (10) counts the expected number of concordant pairs. Constraint (11) computes the win percentage for each team under each realization as defined by equation (7), and constraints (12) establish the relationship between win percentages and relative positions of teams.

4.2.2. Minimizing Euclidean distance between win percentages. We now consider the Euclidean distance between win percentages at the end of the shortened season $y(\boldsymbol{x},\xi)$ and win percentages at the end of the full season $\hat{y}(\xi)$. To minimize this dissimilarity measure, we solve the following stochastic mixed integer quadratic program:

$$[\text{PW}] \min_{\boldsymbol{x},\boldsymbol{y}} \quad \mathbb{E}_\xi \left[\sum\nolimits_{i\in T} (y_i(\xi) - \hat{y}_i(\xi))^2 \right] \tag{15}$$

$$\text{s.t.} \quad y_i(\xi) = \frac{1}{m}\left(y_i^0 + \sum_{g\in G_i^h} W_g(\xi)x_g + \sum_{g\in G_i^a} (1-W_g(\xi))x_g \right) \qquad \forall i\in T, \forall \xi\in\Xi \tag{16}$$

$$\boldsymbol{x}\in X. \tag{17}$$

While PW does not directly measure Euclidean distance between rankings, which is more closely tied to league outcomes than win percentages (see §3.3.2), it has several computational advantages. First, PW does not require binary variables $z_{ij}(\xi)$ and associated linking constraints, making it a lighter formulation than PC. Moreover, as we shall show in §5.1, we may derive a closed-form expression for the expected value in the objective function (15), which results in a much simpler deterministic equivalent problem, despite the objective being quadratic rather than linear.

PROPOSITION 2. *Let L be the least common multiple of m and $\hat{m}$. There exists a constant $D \le \frac{n(n^2-1)}{3}L^2$ such that $d_E(r(\boldsymbol{x},\xi),\hat{r}(\xi)) \le D\sum_{i\in T}(y_i(\boldsymbol{x},\xi)-\hat{y}_i(\xi))^2 \quad \forall \boldsymbol{x}\in X, \forall\xi\in\Xi$.*

This proposition formally shows that PW effectively minimizes the expected Euclidean distance between rankings. Obviously, the opposite does not necessarily hold (i.e., we can have identical rankings but different win percentages).

5. Solution Methodology

There are several practical considerations that we address here. Given that both PC and PW contain $2^{|G|}$ realizations of ξ, each with their own sets of second-stage decision variables and constraints, we begin by devising more tractable reformulations of these stochastic optimization models. In §5.1, we introduce an exact deterministic reformulation of PW, termed PW-DQIP. Although PW-DQIP forms the foundation for our subsequent models, we also introduce two benchmark solution methods for approximately solving PC. As defined in Appendix A, PC-MVP and PC-SAA, respectively, replace the random variables of PC with (i) their means and (ii) a small finite number of samples. Next, in §5.2, we present a fast tailored algorithm for approximately solving PW-DQIP based on the Frank-Wolfe method, named PW-FW. Finally, to reduce the impacts of (predictive) model misspecification and overfitting, in §5.3 we develop a robust optimization reformulation of PW-DQIP which uses as input (i) an ensemble of candidate predictions as well as (ii) optimal values computed from multiple runs of PW-FW.

5.1. Equivalent Deterministic Reformulation

We now show how the stochastic problem PW from §4.2.2 can be solved using an equivalent deterministic problem. We will use the notation $\mathbb{V}$ to refer to the variance of a random variable.

THEOREM 1. *The stochastic model PW can be solved using the following equivalent deterministic linearly-constrained convex quadratic mixed-integer optimization problem:*

$$\text{[PW-DQIP]} \min_{\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{v}} \quad \sum\nolimits_{i\in T}\left((\mu_i-\hat{\mu}_i)^2+v_i\left(1-\frac{2m}{\hat{m}}\right)+\hat{v}_i\right) \tag{18}$$

$$s.t. \quad \mu_i=\frac{1}{m}\left(y_i^0+\sum\nolimits_{g\in G_i^h}p_g x_g+\sum\nolimits_{g\in G_i^a}(1-p_g)x_g\right) \qquad \forall i\in T \tag{19}$$

$$v_i=\frac{1}{m^2}\sum\nolimits_{g\in G_i^h\cup G_i^a}p_g(1-p_g)x_g \qquad \forall i\in T \tag{20}$$

$$\boldsymbol{x}\in X, \tag{21}$$

where the decision variables, in addition to $\boldsymbol{x}=\{x_g, g\in G\}$, include μ_i and v_i which encode the mean and variance, respectively, of the win percentage of team i in the shortened season. Moreover, the following parameters represent the mean and variance, respectively, of the win percentage of team i in the full season:

$$\hat{\mu}_i=\mathbb{E}_\xi\left[\hat{y}_i(\xi)\right]=\frac{1}{\hat{m}}\left(y_i^0+\sum\nolimits_{g\in G_i^h}p_g+\sum\nolimits_{g\in G_i^a}(1-p_g)\right) \tag{22}$$

$$\hat{v}_i=\mathbb{V}_\xi\left[\hat{y}_i(\xi)\right]=\frac{1}{\hat{m}^2}\sum\nolimits_{g\in G_i^h\cup G_i^a}p_g(1-p_g). \tag{23}$$

We note that PW-DQIP can be solved using an off-the-shelf Mixed Integer Quadratic Programming (MIQP) solver such as Gurobi, but doing so is less efficient for large problem instances. In the following, we present a fast tailored algorithm that leverages the combinatorial properties of this problem.

5.2. Frank-Wolfe Algorithm

For the purpose of developing an efficient way to solve PW-DQIP, we now turn our attention to the combinatorial structure of PW-DQIP's feasible region.

PROPOSITION 3. *The coefficient matrix of the set of feasible schedules X is totally unimodular.*

Proposition 3 is based on the fact that incidence matrix of bipartite multigraphs are totally unimodular (Yannakakis 1985). As a result, and given that the right-hand-side values in X (i.e., $\{m_i^h\}$ and $\{m_j^a\}$) are integral, optimizing a linear function over the continuous relaxation of X (denoted $\bar{X}$) using the Simplex method yields an integral optimal solution. This property of PW-DQIP lends itself well to the Frank-Wolfe (FW) method (Frank and Wolfe 1956, Jaggi 2013). FW is an algorithm for solving non-linear convex optimization problems of the form $\min_{\boldsymbol{x}\in\bar{X}} f(\boldsymbol{x})$, where f is a smooth convex function and $\bar{X}$ is a compact convex set. At each iteration t, FW replaces f with its linear approximation at an incumbent point $\boldsymbol{x}^{(t)} \in \bar{X}$, to produce an "atomic" solution

$$\hat{\boldsymbol{x}}^{(t)} = \arg\min_{\boldsymbol{x}\in\bar{X}} \nabla f(\boldsymbol{x}^{(t)})^\top \boldsymbol{x}, \tag{24}$$

and then performs a line search between $\boldsymbol{x}^{(t)}$ and $\hat{\boldsymbol{x}}^{(t)}$ to produce the next iterate $\boldsymbol{x}^{(t+1)} \in \bar{X}$.

Algorithm 1 in Appendix D presents our implementation of the FW algorithm for solving the continuous relaxation of PW-DQIP. Our FW implementation is particularly efficient, since f is a convex quadratic function, and so (i) its gradient is easily computed, and (ii) the line search step admits closed-form optimal solution via the first-order optimality conditions. Moreover, given that the atomic solution $\hat{\boldsymbol{x}}^{(t)}$ produced by solving the transportation problem (24) is a feasible integer solution, it provides an upper bound on the optimal value of PW-DQIP. As the FW algorithm iterates, $\boldsymbol{x}^{(t)}$ converges to the optimal fractional solution and the upper bound $\hat{\boldsymbol{x}}^{(t)}$ becomes progressively tighter. Upon terminating at a finite iteration t, we use the best integer-feasible point $\hat{\boldsymbol{x}}^{(t)}$ found thus far, over iterations $1 \dots t$, as a near-optimal integer solution to PW-DQIP. We also note that we may produce a sub-optimality bound for the solution of FW as described in Appendix D. Henceforth, we will use PW-FW to refer to producing a PW solution using FW.

5.3. Robust Optimization Reformulation

In the data science literature, there are many examples of successful ensemble models, which combine the results of multiple predictive models to produce more robust results (c.f., Sagi and Rokach 2018). With this in mind, we derive a Min-Max Regret (MMR) formulation that combines the predicted probabilities from several of our predictive models, allowing our prescriptive model's results to be more robust to misspecification and model overfitting.

Given a set L of candidate predictions indexed by l, we define the following parameters: (i) $p_g^{(l)}$ is the probability that the home team wins game g under candidate l; (ii) $\theta^{(l)}$ is the optimal value

(or, alternatively, a lower bound on the optimal value) of PW-DQIP when solved using candidate l; and (iii) $\hat{\mu}_i^{(l)}$ and $\hat{v}_i^{(l)}$ correspond to the parameters defined by (22) and (23), respectively, when $p_g^{(l)}$ is used in place of p_g. In addition to x_g, the MMR counterpart of our PW model has decision variables $\mu_i^{(l)}$ and $v_i^{(l)}$, which are the candidate-specific versions of μ_i and v_i from (19) and (20), respectively. Finally, θ is a decision variable that captures the maximum regret over using all candidate win-probability vectors in the PW objective. Minimizing this maximum regret yields the following convex mixed-integer quadratically-constrained program, which can be solved using a commercial solver such as Gurobi:

$$
\begin{aligned}
[\text{PW-MMR}] \quad \min_{\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{v},\theta} \quad & \theta && && (25)\\
\text{s.t.} \quad & \theta \geq \sum\nolimits_{i\in T} \left(\left(\mu_i^{(l)} - \hat{\mu}_i^{(l)}\right)^2 + v_i^{(l)} \left(1 - \frac{2m}{\hat{m}}\right) + \hat{v}_i^{(l)} \right) - \theta^{(l)} && \forall l \in L && (26)\\
& \mu_i^{(l)} = \frac{1}{m}\left(y_i^0 + \sum\nolimits_{g\in G_i^h} p_g^{(l)} x_g + \sum\nolimits_{g\in G_i^a} (1-p_g^{(l)}) x_g \right) && \forall i \in T,\, l \in L && (27)\\
& v_i^{(l)} = \frac{1}{m^2} \sum\nolimits_{g\in G_i^h \cup G_i^a} p_g^{(l)}(1-p_g^{(l)}) x_g && \forall i \in T,\, l \in L && (28)\\
& \boldsymbol{x} \in X. && && (29)
\end{aligned}
$$

6. Computational Experiments

In this section, we (1) evaluate our predictive models and choose the best one(s) to use in our prescriptive phase ("model selection"), and (2) solve all variants of our prescriptive models using the home-team win-probabilities estimated during our predictive phase, and evaluate the quality of the shortened seasons produced by our prescriptive models. All models (predictive and prescriptive) were coded in `Python 3.8.8`. For our predictive models, we used the `scikit-learn` package (Pedregosa et al. 2011), and for solving our mixed integer programs in our prescriptive models we used `Gurobi 9.5.0` with all solver settings left at their default values. All experiments were conducted on a computer with a 2.6 GHz Intel Core i7 CPU and 16 GB of memory.

6.1. Dataset Description

We use historical data from 14 NBA seasons (2004–2010, 2012–2018), which are the years with the same regular season structure as today; i.e., 30 teams, each playing 82 games with schedules constructed in the manner described in §3.1. We omit the shortened seasons 2010–11 and 2020–21, and do not consider data prior to 2004 as back then the NBA had fewer teams. We used the box score datasets publicly available on the NBA's official website (NBA 2020b) which contains detailed information for each game, team and player. From this, we created 56 datasets that in turn consider, for each of the 14 seasons, 4 alternative hypothetical suspension dates (i.e., days 80, 100, 120, 140 of the season). Note that each regular season spans between 170–180 days.

In our experiments, we fixed the number of games per team in the shortened season so approximately half the post-suspension games are scheduled and the other half cancelled (alternative targets can easily be achieved by adjusting the parameters m_i^h and m_i^a, but we felt that varying the suspension day in our experiments already provides sufficient sensitivity analysis). Table 3 provides some summary statistics of our instances. For each suspension day, we list the number of Games per Team (GT) we wish to have in the shortened season, the average number of Games Played (GP) per team prior to suspension, the average number of Games we will Schedule (GS) per team post-suspension, and the number of Games we will Cancel (GC) per team post-suspension. Note that (i) $GP+GS+GC=82$ since a full season has 82 games per team; and (ii) $GP+GS=GT$.

Suspension	GT	GP	GS	GC
Day 80	62	38.4	23.6	20
Day 100	66	48.6	17.4	16
Day 120	70	56.4	13.6	12
Day 140	74	66.0	8.0	8

Table 3 Summary statistics of our problem instances, averaged over all teams and all 14 NBA seasons.

6.2. Predictive Model Results

In this subsection, we describe the explanatory features used in our predictive models, the preprocessing we performed on our datasets, and the cross-validation we used to assess our predictive models. Finally, we compare the predictive performance of the 8 binary classifiers described in §4.1.

6.2.1. Experimental setup. We use the basic and advanced statistics published in the box score datasets (NBA 2020b) to curate four groups of explanatory features, including i) overall team performance, ii) basic team-level statistics, iii) advanced team-level statistics, and iv) player-level statistics. When taken together, this results in 128 features, the details of which we provide in Appendix E. In our independent exploratory analysis, we have found that the features that tend to be the most important are, ranked in order from most to least important: i) overall team performance measured primarily by win percentage, ii) style of play metrics such as pace and number of possessions created, iii) shooting efficiency, iv) offensive fire power measured by metrics such as effective field goal percentage, true shooting percentage, points in the paint, and offensive rating in general, v) rebounding, and finally vi) assists and ball movement. We first normalize all features to the $[0,1]$ range, and then use *Principal Component Analysis (PCA)* to eliminate multicollinearity and prevent overfitting. We retain only the first 25 principal components with highest eigenvalue, which explain more than 90% of the total variance in our training data.

We use 5-fold cross-validation to tune our predictive models' hyper-parameters. For this, we partition our data into *training, validation,* and *test* datasets. Pre-suspension games are divided

into *training* and *validation* sets through 5 folds at random, with the validation dataset having the same size (30% of the of pre-suspension dataset) across all folds. Post-suspension games comprise the *test* dataset. On each fold, we fit each classifier using a training dataset and evaluate its performance using the corresponding validation dataset. Then, we measure the performance of each predictive model using the average LogLoss across the five folds. We reserve the test datasets for evaluating the combined performance of our predictive and prescriptive models.

As is well-known in machine learning, not all classification models produce unbiased class probabilities. Therefore, we employ Platt scaling (Platt et al. 1999), which involves transforming the home team's win probabilities using a sigmoid calibration function to recover unbiased win probability estimates. For this, we use the Python package `scikit-learn`, which uses multiple randomly-generated held-out samples from the training dataset to compute the calibrated probabilities.

6.2.2. Predictive performance. We evaluate the performance of our predictive models using LogLoss as defined in (4), which is a strictly proper scoring rule (see §4.1 for details). We also report the *accuracy* and the *Area Under the receiver operating characteristic Curve* (AUC), which are not proper scoring rules but are widely used for evaluating the performance of binary classifiers. Accuracy measures the proportion of correctly predicted class labels. AUC summarizes the trade-off between true positive rate and false positive rate when the threshold to predict class labels changes (see e.g., Fawcett 2006). Larger AUC values generally correspond to better predictive performance.

The LogLoss metric plays a crucial role in three distinct elements of our predictive phase. First, it serves as the internal loss function for all but three of our binary classifiers, with notable exceptions being SVM, Naïve Bayes, and Random Forest, which utilize hinge loss, Gini impurity, and no specific loss function, respectively; see (Shen 2005) for additional details. Second, LogLoss guides our hyper-parameter optimization process across all eight predictive models. Lastly, we use it as our model selection criterion to identify the most effective of our eight candidate models.

Figure 6 depicts, for 4 choices of suspension day and 9 classifiers, the distributions of accuracy, AUC, and LogLoss across 14 NBA seasons. In addition to the 8 classifiers described in §4.1, we also define "Best" as a composite classifier that, for each season, employs the lowest-LogLoss classifier among the 8 classifiers tested. Each boxplot's underlying distribution corresponds to 14 values (one per season), where each season's performance is taken as an average of the model's performance over the five validation datasets (cross-validation folds). Box heights correspond to Inter-Quartile Ranges (IQR), with medians marked by dark horizontal lines inside each box. Whiskers mark the most extreme-valued data points within 1.5IQR units above and below the boundary of the box, and any points lying outside this range are marked as outliers.

As seen in Figure 6, the Logit, ELM, SVM and MLP models generally have the lowest LogLoss values, which indicates their superiority in predicting well-calibrated and unbiased probabilities.

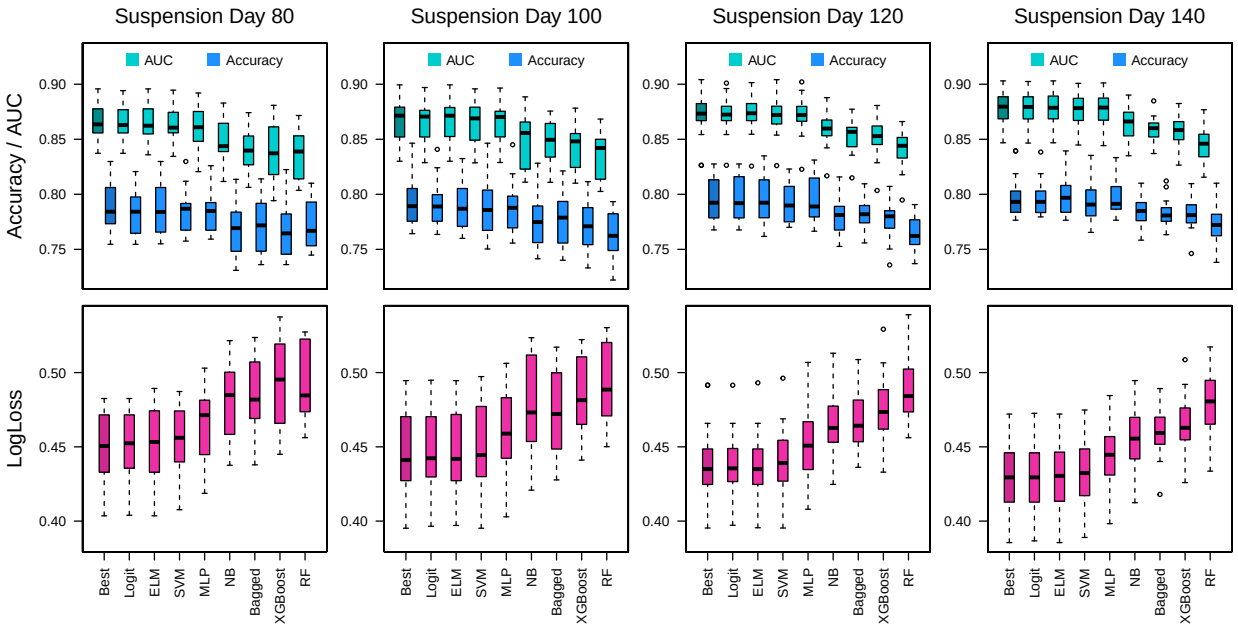


Figure 6 **Accuracy, AUC, and LogLoss for 9 different classifiers. The distributions are measured using the validation datasets across 14 NBA seasons.**

The same four models also outperform the other models according to accuracy and AUC, with Logit being the better model overall. It is worth noting that our best predictive models achieve an accuracy over 75%, which is in line with the accuracy of models that predict the outcome of the next game to be played (see, for example Kvam and Sokol (2006), Brown and Sokol (2010)).

We conjecture that two factors drive the performance of our predictive models: the level of league competition and how early we suspend the season. In a more competitive season, it becomes more difficult to predict game outcomes, since teams are more evenly-matched. Second, as we increase the suspension day from 80 to 140, the pre-suspension dataset grows and as a result, our predictive models generally perform better in all three metrics depicted in Figure 6, both in terms of the average value as well as the variation across seasons (i.e., length of the boxplot).

In general, since different predictive models perform better in different seasons, we use our composite "Best" classifier to produce the home-team win-probabilities for the post-suspension games used by our prescriptive models. That is, for each season and suspension day, we select the lowest-LogLoss of the 8 classifiers as measured by 5-fold cross-validation on our validation dataset, and then train the selected classifiers on all pre-suspension games to produce home-team win-probabilities for all post-suspension games.

6.3. Prescriptive Model Results

We now evaluate our prescriptive models and provide practical insights that apply to their use.

6.3.1. Experimental setup. We measure the runtimes and solution quality of all prescriptive models discussed in §4.2 (i.e., PW-DQIP, PW-FW, and PW-MMR), as well as the two introduced

Model	Objective	Solution Method
PW-DQIP	Min. distance of win percentages	Deterministic equivalent Quadratic Program
PW-FW	Min. distance of win percentages	Produce near-optimal solution for PW-DQIP using Frank-Wolfe
PW-MMR	Min. distance of win percentages	A Robust reformulation of the PW-DQIP model
PC-MVP	Max. concordance of rankings	Replace random variable W_g with its expected value p_g for each game $g \in G$
PC-SAA	Max. concordance of rankings	Replace distribution Ξ with sample $\mathcal{S}$, and expected value with sample average

Table 4 Summary of prescriptive models.

in Appendix A (i.e., PC-MVP and PC-SAA); results are summarized in Table 5. As described in detail in Appendix A.4, the trade-off between solution quality and runtime of SAA is balanced at 50 scenarios. Thus, we use 50 scenarios for our PC-SAA model. As well, to improve the computational efficiency of our PC model, we introduce variable fixing techniques (see Appendix A.3) to deduce and fix many z-variables at their optimal values. As detailed in Table 8 in Appendix A.3, our method eliminates 75% of the z-variables in PC-MVP and 60% in PC-SAA, which is significant, as it translates to closing the optimality gap at a faster rate (e.g., for suspension day 80, after 3,600 seconds of running PC-SAA the gap is 2.02% with variable fixing, and 6.93% without).

We implement two benchmarks. First, we are interested to know how well our models perform relative to an approach that does not explicitly optimize the end-of-season ranking when selecting the games in the shortened season. For this, we implement a greedy heuristic (henceforth known as "Greedy") that selects games according to their original scheduled dates, with earlier games assigned first until the target number of games for each team are met. Second, we are also interested to know how much better we can do by playing our optimally-chosen shortened season relative to not playing any more games after the suspension (the "Status Quo" ranking).

In line with previous studies in sports analytics (e.g., Van Eetvelde et al. 2021, Chater et al. 2021, Csató 2021b, Sziklai et al. 2022), we evaluate the expected performance of our models using Monte Carlo simulation. More specifically, we draw the outcomes of each post-suspension game from a Bernoulli distribution with home-team win-probability estimated by our "Best" predictive model. To measure the expected performance, we generate a sample of 10,000 game outcomes from these Bernoulli distributions. Each realization yields two rankings, one at the end of the shortened season, and the other at the end of the full season with all games played. We then compare these rankings using our primary performance metric, the number of concordant pairs between rankings.

Note that we use estimated home-team win-probabilities in both our prescriptive models and our simulation. To ensure robustness of our testing methodology, we use some *unseen* data points to estimate the home-team win-probabilities for our simulation. To this end, we hold out 20% of the pre-suspension data at random when we estimate the parameters p_g, $g \in G$, used as home-team win-probabilities in our prescriptive models, and use the entire pre-suspension dataset to estimate the home-team win-probabilities used by our simulation.

Finally, as an additional robustness check, we also ran simulations with probabilities estimated using multiple evaluator models with different functional forms. For each season and suspension day, we compared PW-FW to Greedy using all combinations of prescriptive model parameters p_g, $g \in G$, and simulation Bernoulli probabilities estimated by our 8 classifiers from §4.1 (for a total of 64 combinations). As illustrated in Appendix F.1, all PW-FW solutions outperformed the Greedy solutions by a considerable margin.

6.3.2. Prescriptive performance. Table 5 summarizes the results of running all four prescriptive models on the instances described in Table 3, averaged over 14 seasons. The "Time" column reports runtimes in seconds, while "Conc." is concordance (τ_C) measured by our simulation. PC-MVP solved all instances to optimality, while the optimality gap of PW-DQIP was never more than 10^{-3}. For PC-SAA and PW-MMR, no instance converged to optimality by the 3600-second time limit. We report the percentage optimality gap for PC-SAA, while for PW-MMR we provide the absolute optimality gap ($UB - LB$). The latter metric is more appropriate for PW-MMR, since its lower bound is always zero. For PW-FW, we report the sub-optimality gap defined in Appendix D, and observe that this gap is typically very low (less than 2% on average).

Comparing our prescriptive models, we observe the following. PC-MVP is fastest, but solution quality (concordance) is lowest. PC-SAA produces higher-quality solutions than PC-MVP, but is slow. All PW-based models generate high-quality solutions. While PW-DQIP typically produces slightly higher-quality solutions than PW-FW, our Frank-Wolfe algorithm is over **a thousand** times faster than the integer program of PW-DQIP. With regard to PW-MMR, it is interesting that although its solutions are more robust, overall solution quality is comparable to the other PW methods. While runtimes of PW-MMR are slower as we solve a mixed-integer quadratically-constrained program, it is worth pointing out that it may be possible to extend our Frank-Wolfe method to PW-MMR (see our comment at the end of Appendix D; we leave the details to future work). Finally, we also observe that as the season is suspended later, there are fewer games to choose from, and all models perform better as the solution space shrinks.

Sus. Day	PC-MVP		PC-SAA			PW-DQIP		PW-FW			PW-MMR		
(GT)	Time	Conc.	Time	Gap	Conc.	Time	Conc.	Time	Sub. Gap	Conc.	Time	Abs. Gap	Conc.
80 (62)	0.08	413.35	3600	4.98%	414.24	3600	415.96	0.39	1.19%	415.86	3600	0.00037	415.84
100 (66)	0.09	414.60	3600	1.66%	415.80	3600	417.94	0.24	1.32%	417.83	3600	0.0002	417.87
120 (70)	0.10	417.43	3600	1.49%	418.55	934.11	419.96	0.12	1.48%	419.88	3600	0.00015	419.92
140 (74)	0.06	420.64	3600	1.03%	421.45	82.33	422.24	0.05	1.66%	422.20	3600	0.00012	422.26

Table 5 Performance of the prescriptive models, averaged over 14 seasons

As different seasons unfold in different ways, it is interesting to compare our prescriptive models on a per-season basis. Figure 7 plots concordance as measured by our simulation for each of the

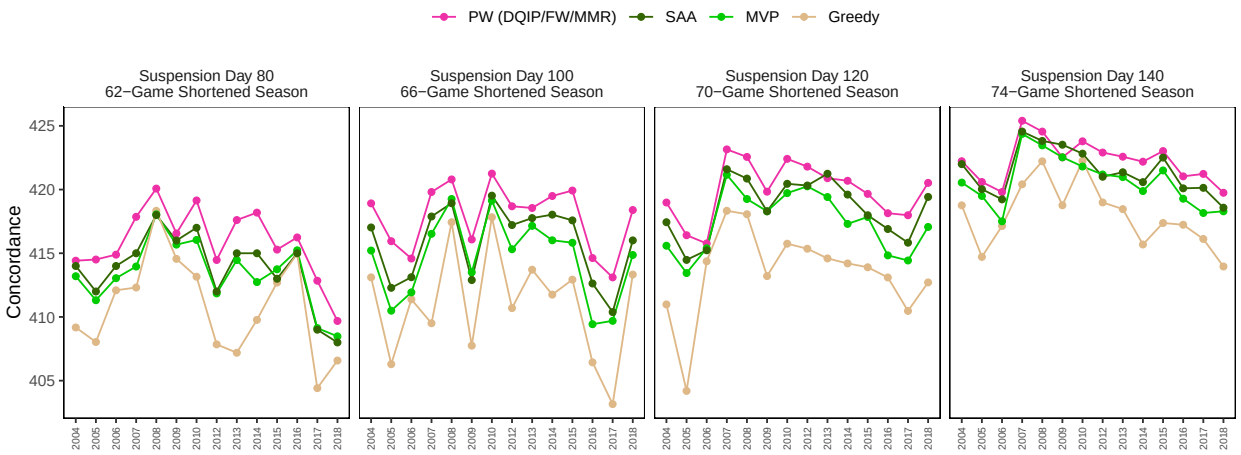


Figure 7 Simulation results for prescriptive models (concordance) across 14 NBA seasons.

14 NBA seasons and 4 suspension days. The three PW models (i.e., PW-DQIP, PW-FW, and PW-MMR) have very similar solution quality and thus we combine their plots. From Figure 7, we see that all five proposed models outperform the benchmark greedy algorithm on all instances (i.e., for all seasons and suspension days). Typically, PW dominates PC-SAA and PC-MVP as well. We remark that the results in Figure 7 and subsequent simulation-based results correspond to mean values estimated using 10,000 simulation replications. The standard errors of these estimates are negligible (i.e., within 0.1% of the means), and therefore we omit plotting confidence intervals for these figures.

We provide a few comments to explain the differences in performance across these models, which all approximate our stochastic optimization model PC in distinct ways. First, we note that although both PC-MVP and PC-SAA maximize concordance directly, PC-MVP uses only the means of the random variables while PC-SAA samples a small number of scenarios. Because PC-SAA more faithfully represents PC, it is not surprising that it typically outperforms PC-MVP. On the other hand, it is interesting to observe that the PW models, which take a different approach to approximating PC, outperform PC-SAA. Instead of sampling from the distributions of the random variables, the PW models approximate the ranking-based concordance objective with a win-percentage-based objective which allows us to formulate the problem as a deterministic equivalent. PW outperforms in practice because PC-SAA either suffers from (i) too few SAA scenarios causing a loose approximation of PC, or (ii) too many SAA scenarios resulting in combinatorial explosion and poor convergence (see Appendix A.4). Because PW is represented as a deterministic equivalent, it sidesteps this difficulty as it does not require sampling from the distributions.

Next, we illustrate the incremental value of playing a shortened season, relative to the "Status Quo" case of halting the season at the suspension date. Figure 8 compares our top-performing prescriptive model (i.e., PW) with "Status Quo". First, we note that the concordance of "Status Quo"

differs substantially across instances. Indeed, sometimes the ranking on the suspension date is close to the end-of-season ranking (e.g., 2009-10 with suspension day 80 and 2008-09 with suspension day 100), while in other cases the "Status Quo" ranking is far from the end-of-season ranking (e.g., 2013-14 with suspension day 80 and 2017-18 with suspension days 80 and 100). Second, we observe that playing additional games as chosen by PW significantly improves concordance, regardless of season and suspension date. And finally, as noted previously, as the season is suspended later, all models' rankings converge to the end-of-season ranking, leading to both higher concordance and lower variability in outcomes. Finally, it is important to note that our methodology is very different

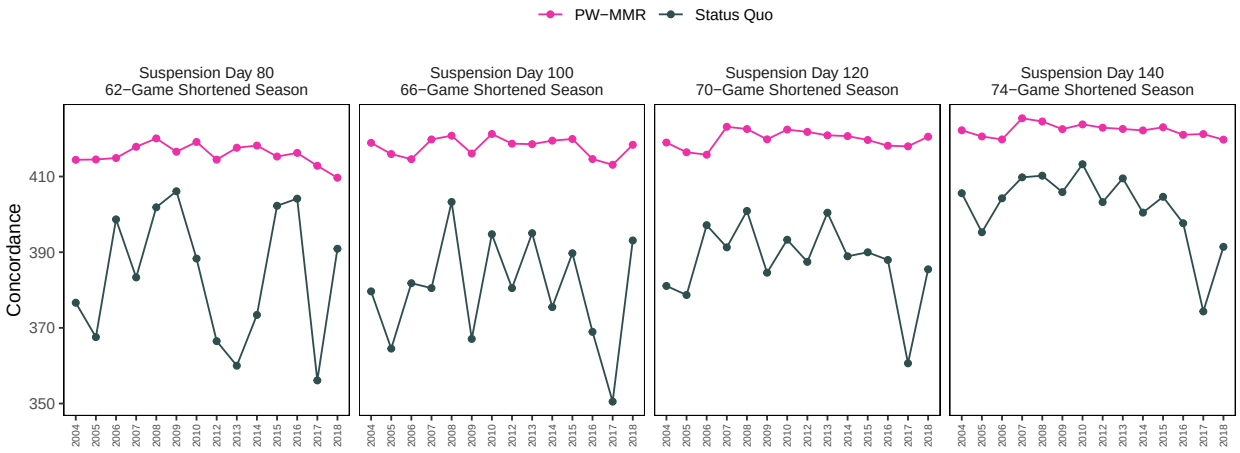


Figure 8 **Comparing the best-performing prescriptive model with the suspension day concordance (Status Quo) across 14 NBA seasons.**

from one that takes the ranking at the suspension date and tries to sustain this ranking through the end of the shortened season (had this been the case, the rankings from "Status Quo" and PW would be similar, and their concordance with the end-of-season ranking as plotted in Figure 8 would be much closer). Indeed, although our predictive model uses pre-suspension data to estimate our prescriptive model's parameters, a team with a high win rate pre-suspension will not necessarily have a high win rate post-suspension. The nature of the shortened season and its impact on rankings is complex, owing to the relative difficulty of the schedule before and after suspension (i.e., precisely when a team faces easy or hard-to-beat competitors). Because our prescriptive model aims to produce a shortened season with ranking similar to that of the full season, if a team had an easy schedule pre-suspension it would generally have a comparatively difficult schedule post-suspension, and our prescriptive model will naturally attempt to maintain this difficulty gradient in the shortened season post-suspension. Moreover, it is also interesting to note that the games that PW chooses also tend to be the more "competitive" games with uncertain outcomes, which are precisely the games that spectators enjoy watching (see Appendix C for details).

Appendix F.4 provides an illustration of a shortened season produced by our two-phase approach, which is the result of applying our methodology to the suspended 2019-20 NBA regular season.

6.3.3. Practical implications. We now measure the performance of our shortened seasons using several metrics directly tied to a team's ranking. In doing so, we verify that our PW-MMR model produces solutions that not only have high concordance, but are also practically appealing.

As discussed in §3.1, at the end of the regular season, within each conference the top 8 teams make the playoffs and the top 4 receive home court advantage. Moreover, the 5 bottom-ranked teams are given the highest (i.e., double-digit) lottery odds in next year's draft for rookie players. Thus, we define 3 categories of metrics: (i) "Playoff teams" (the top 8 teams in each conference), (ii) "Teams with home court advantage" (the top 4 teams in each conference), and (iii) "Teams with double-digit lottery odds" (the bottom 5 teams overall). Using the same 10,000 simulation replications described in §6.3.1, for each of these 3 metric categories we calculate the proportion of teams in the shortened season that agree with the full season. By *agree*, we mean that we check to see how many teams fall into that category in both the shortened season and the full season.

For example, imagine that in one particular realization the 5 bottom-ranked teams in the full-season are {`Warriors, Cavaliers, Timberwolves, Hawks, Pistons`}, while the 5 bottom-ranked teams in the shortened-season produced by PW-MMR are {`Warriors, Cavaliers, Timberwolves, Hawks, Knicks`}, and finally the 5 bottom-ranked teams in the shortened-season produced by Greedy are {`Warriors, Cavaliers, Timberwolves, Bulls, Knicks`}. Here, the PW-MMR model selected $\frac{4}{5} = 80\%$ of the bottom-5 teams correctly while for the Greedy ranking only $\frac{3}{5} = 60\%$ of the bottom-5 teams match those of the full-season. For this realization, we say models PW-MMR and Greedy have 80% and 60% *agreement* with the full season, respectively.

Table 6 compares the mean agreement percentage of our PW-MMR model with that of Greedy and Status Quo; agreement percentages have been averaged over all 4 suspension dates, 14 seasons, and 10,000 Monte Carlo replications. We can see that if we do not play a shortened season but instead select playoff teams based on the ranking as of the suspension date (Status Quo), then on average 88.96% of teams that would have made the playoffs if the full season was played do in fact advance to the playoffs; i.e., 88.96% of playoff teams are chosen correctly. Instead, if we play a shortened season but it is a Greedy rather than optimal one, then on average 94.27% of the teams that advance to the playoffs are chosen correctly. Finally, if we use PW-MMR to construct a shortened season, on average 95.65% of the teams advancing to the playoffs would also have been in the playoffs had the full season been played. The other two metrics also indicate significant improvements from using PW-MMR over Greedy and Status Quo. For further details, see Appendix F.2, which includes boxplots that show the distributions of the three agreement metrics over the 14 NBA seasons, for each of the 4 suspension days.

Agreement criteria	PW-MMR	Greedy	Status Quo
Playoff teams	**95.65**	94.27	88.96
Teams with home court advantage	**92.28**	89.83	79.10
Teams with double-digit lottery odds	**91.36**	89.24	78.10

Table 6 Mean agreement percentages on simulation for shortened seasons computed using three models.

To further validate the practical performance of our models, we also measured their performance using actual post-suspension game outcomes (i.e., we ran a backtest). In this experiment, for each season and suspension day, instead of 10,000 simulation replications we have only a single sample path. While this is a good robustness check, the backtest is susceptible to producing noisy outcomes. We find that, in terms of concordance between the shortened seasons and the full seasons, PW-MMR outperforms Greedy 65% of the time. Furthermore, when PW-MMR outperforms Greedy, it does so by a larger margin than when Greedy outperforms PW-MMR. Specifically, whenever PW-MMR beats Greedy, its concordance is higher by an average of 3.68 concordant pairs. This is more than twice the number of concordant pairs by which Greedy outperforms PW-MMR (1.57). Finally, Table 7 compares the backtest results for the PW-MMR, Greedy, and Status Quo models using the agreement criteria discussed earlier.

Agreement criteria - Backtest	PW-MMR	Greedy	Status Quo
Playoff teams	**95.64**	94.87	90.62
Teams with home court advantage	90.62	**91.07**	82.59
Teams with double-digit lottery odds	**87.14**	83.21	78.57

Table 7 Mean agreement percentages on backtest for shortened seasons computed using three models.

6.3.4. Strength of schedule extension. A practically appealing extension to our prescriptive model incorporates constraints that additionally ensure a *Strength of Schedule (SoS)* for each team that is not materially reduced from its full-season measure. There are several mathematical definitions of SoS (see NBAstuffer 2023), but essentially SoS is used by each team to quantify the difficulty of its remaining schedule of games. Pundits on television use SoS for arguments such as, "Although the Lakers are currently higher-ranked than the Clippers, the Clippers have a higher SoS and so are likely to make up some of this slack and could come out ahead by the end of the season." League managers may wish to assure each team that their SoS is not materially impacted by the shortened season being selected; this motivates the prescriptive model in this subsection.

As far as we know, there is limited related work on incorporating SoS in a prescriptive model, and the few articles that use such a measure take a dynamic scheduling approach where the schedule is updated according to certain criteria including SoS; see Bouzarth et al. (2020). As our approach is based on mathematical programming, we choose the *Opponent's Win percentage (OW)* as our SoS metric, as it is both simple to understand and linear in our decision variables.

Following our notation from §4.2, let $\bar{y}_i^0 = \frac{y_i^0}{m_i^0}$ denote the win percentage of team i at the time of suspension, where m_i^0 is the number of games played by team i pre-suspension. We define team i's strength of schedule in the remainder of the shortened season and full season, respectively, as:

$$OW_i = \frac{1}{m - m_i^0}\left(\sum_{g\in G_i^h} x_g \bar{y}_{j(g)}^0 + \sum_{g\in G_i^a} x_g \bar{y}_{i(g)}^0\right) \qquad \forall i \in T \tag{30}$$

$$\widehat{OW}_i = \frac{1}{\hat{m} - m_i^0}\left(\sum_{g\in G_i^h} \bar{y}_{j(g)}^0 + \sum_{g\in G_i^a} \bar{y}_{i(g)}^0\right) \qquad \forall i \in T. \tag{31}$$

The above expressions estimate the average win percentage of all opponents of team i in the remainder of the shortened season and full season, respectively. With slight abuse of notation, $i(g)$ refers to the home team in game g, and should not be confused with the focal team i. A larger value for OW_i or $\widehat{OW}_i$ implies a more difficult remainder of the season for team i, as this means future opponents are harder to beat.

Using the two quantities defined in (30) and (31), we modify our PW-DQIP formulation, (18)–(21), by adding the constraint (32) below, which ensures that for each team i, the strength of schedule (i.e., average opponents' win percentage) in the remainder of the shortened season is not materially higher than in the remainder of the full season, i.e., within ϵ of the full-season SoS. We refer to ϵ in percentage terms for ease of communication and to emphasize its relative scale. An extension to our PW-DQIP model incorporating strength-of-schedule constraints is as follows:

$$\begin{aligned} \text{[PW-SoS]} \quad \min_{\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{v}} \quad & \textstyle\sum_{i\in T}\left((\mu_i - \hat{\mu}_i)^2 + v_i\left(1 - \frac{2m}{\hat{m}}\right) + \hat{v}_i\right) \\ \text{s.t.} \quad & \frac{OW_i - \widehat{OW}_i}{\widehat{OW}_i} \le \epsilon \qquad \forall i \in T \qquad (32) \\ & (19)\text{–}(21), (30). \end{aligned}$$

Note that in the above, OW_i is an auxiliary decision variable as it depends on the choice of shortened season, while $\widehat{OW}_i$ is a constant since the full season is fixed.

To assess the strength of schedule in our solutions and given we are only interested in cases where the strength of schedule in the shortened season is larger than that of full season, we introduce the metric "Strength of Schedule Discrepancy (SSD)" which takes the *positive difference* between strength of schedule in the shortened and full seasons, defined as:

$$\text{SSD} = \frac{1}{n}\sum_{i\in T} \max\left\{\frac{OW_i - \widehat{OW}_i}{\widehat{OW}_i}, 0\right\}. \tag{33}$$

Figure 9 compares models PW (without SoS constraints), Greedy and PW-SoS with 4 choices for ϵ ranging from 2% to 10%. The left panel plots concordance as measured in our Monte Carlo

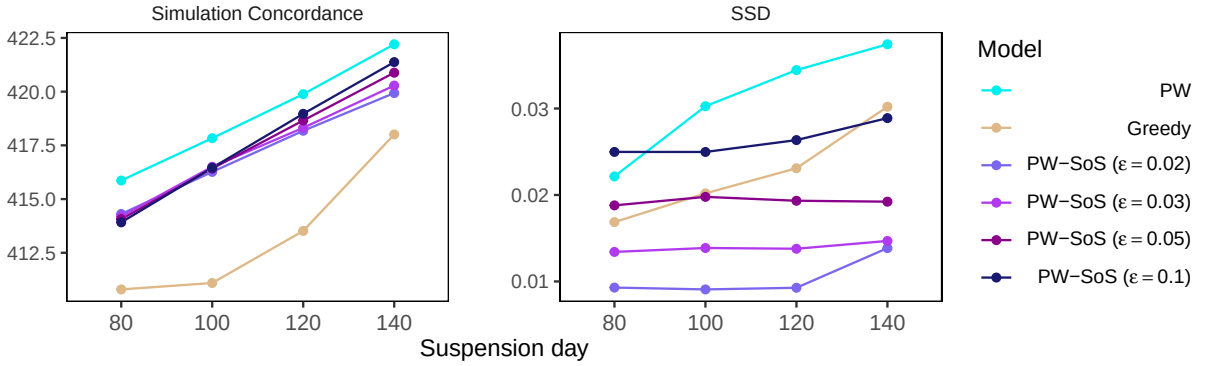


Figure 9 **Comparison of PW-SoS (with 4 different values for ϵ) to PW and Greedy based on simulation concordance (left panel), and SSD (right panel).**

simulation over 10,000 replications. Here, we see that the addition of SoS constraints sacrifices some amount of concordance, but in general PW-SoS still performs significantly better than Greedy. In the right panel, we investigate the sensitivity to the ϵ parameter, and show that for ϵ below 3% we can produce solutions using PW-SoS that have better (lower) SSD than both PW and Greedy. Each dot in Figure 9 represents an average value across 14 NBA seasons. Appendix F.3 includes the corresponding plots for individual NBA seasons.

7. Conclusions and Future Research

Professional sports leagues may be suspended due to various reasons, requiring the league to select which games to play in a shortened season. In this paper, we proposed a two-phase analytics approach for this problem. In phase one, we predicted game outcomes using a composite binary classifier, with a particular functional form for each season and for each suspension day, chosen based on LogLoss values. In phase two, we used stochastic optimization techniques to prescribe a data-driven decision which maximizes the expected similarity between the ranking at the end of the shortened season and the full season had it been played in full.

To solve one of our stochastic optimization problems (PW), we proposed three solution methodologies: i) a deterministic equivalent reformulation (i.e., PW-DQIP), ii) a Frank-Wolfe decomposition algorithm (i.e., PW-FW) which significantly reduces the running time of PW-DQIP while maintaining similar solution quality, and iii) a robust reformulation of PW-DQIP designed to handle misspecification in the input data (i.e., PW-MMR). For our second model (PC), we proposed approximation schemes (MVP and SAA), as well as variable fixing techniques. Our PC-SAA model approximates the distributions in PC but has an exact objective, while our PW models use the exact distribution but approximate PC's objective.

We evaluated our models' solutions using Monte Carlo simulation. Our computational experiments show PW outperforms PC-SAA even for reasonably-large 50-scenario instances. This suggests that PW (and specifically the PW-MMR variant) is the recommended prescriptive model

to maximize concordance. Finally, we verified that the higher-concordance solutions provided by PW-MMR outperform our benchmarks, leading to a higher agreement between shortened season and counterfactual full season in terms of the number of teams that (a) make the playoffs, (b) receive home court advantage, and (c) receive double-digit rookie draft lottery odds. We ran a backtest as a robustness check, and also provided a model extension (PW-SoS) that ensures each team's strength-of-schedule is not materially impacted by our choice of shortened season.

We envision several directions for future research. First, one potential improvement could come from considering more sophisticated loss functions for the predictive models that are custom-tailored to the specific needs of the downstream prescriptive models. Second, apart from concordance, other considerations (e.g., generated revenue, travel cost and distances, broadcasting restrictions, venue availability) may also be relevant, suggesting an alternative multi-criteria decision-making approach. Third, when faced with multiple stoppages during a single season, selecting the optimal subset of games between any two stoppages can be modeled as a dynamic stochastic optimization problem with a learning component. Finally, large-scale stochastic optimization techniques (e.g., progressive hedging) may be designed to tackle SAA with a larger sample.

Acknowledgments

The authors thank the department editors who handled this manuscript, Hau Lee and Mylvyn Sim, along with the anonymous associate editors and referees whose insightful comments significantly improved the manuscript. We also extend our thanks to those who offered valuable feedback at INFORMS 2020, INFORMS 2021, EURO 2022, EUROYoung Workshop 2022, and the Southern California OR/OM Day 2023, as well as during seminars at University of California Irvine (Merage), Sports Training and Research in DAta ScIence Methods for ANalytics and INjury Prevention Group (S-TRAINING), University of Manchester (AMBS), Lancaster University, University of Nottingham, University of Iowa, University of Melbourne (OPTIMA), Aarhus University, IÉSEG School of Management, University of Zurich, and Emory University.

References

Aoki RY, et al. (2017) Luck is hard to beat: The difficulty of sports prediction. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1367–1376.

Bean JC, Birge JR (1980) Reducing travelling costs and player fatigue in the national basketball association. *Interfaces* 10(3):98–102.

Bickel JE (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis* 4(2):49–65.

Bouzarth EL, Cromer AW, Fravel WJ, Grannan BC, Hutson KR (2020) Dynamically scheduling nfl games to reduce strength of schedule variability. *Journal of Sports Analytics* 6(4):281–293.

Bozóki S, Csató L, Temesi J (2016) An application of incomplete pairwise comparison matrices for ranking top tennis players. *European Journal of Operational Research* 248(1):211–218.

Breiman L (1996) Bagging predictors. *Machine Learning* 24(2):123–140.

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

Briskorn D, Drexl A (2009) A branch-and-price algorithm for scheduling sport leagues. *Journal of the Operational Research Society* 60(1):84–93.

Brown M, Sokol J (2010) An improved LRMC method for NCAA basketball prediction. *Journal of Quantitative Analysis in Sports* 6(3).

Chater M, et al. (2021) Fixing match-fixing: Optimal schedules to promote competitiveness. *European Journal of Operational Research* 294(2):673–683.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cocchi G, et al. (2018) Scheduling the Italian national volleyball tournament. *Interfaces* 48(3):271–284.

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3):273–297.

Csató L (2021a) Coronavirus and sports leagues: obtaining a fair ranking when the season cannot resume. *IMA Journal of Management Mathematics* 32(4):547–560.

Csató L (2021b) A simulation comparison of tournament designs for the world men's handball championships. *International Transactions in Operational Research* 28(5):2377–2401.

Diaconis P, Graham RL (1977) Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39(2):262–268.

Durbin J, Stuart A (1951) Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society: Series B (Methodological)* 13(2):303–309.

Fawcett T (2006) An introduction to ROC analysis. *Pattern Recognition Letters* 27(8):861–874.

FiveThirtyEight (2022) NBA Game Predictions. Retrieved on January 20, 2022. `https://projects.fivethirtyeight.com/2022-nba-predictions/games/`.

Fleurent C, Ferland JA (1993) Allocating games for the NHL using integer programming. *Operations Research* 41(4):649–654.

Frank M, Wolfe P (1956) An algorithm for quadratic programming. *Naval Res. Logis. Quart.* 3(1-2):95–110.

Fry MJ, Ohlmann JW (2012a) Introduction to the special issue on analytics in sports, part I: General sports applications. *Interfaces* 42(2):105–108.

Fry MJ, Ohlmann JW (2012b) Introduction to the special issue on analytics in sports, part II: Sports scheduling applications. *Interfaces* 42(3):229–231.

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Goossens D, Spieksma F (2009) Scheduling the Belgian soccer league. *Interfaces* 39(2):109–118.

Halberstadt JB, Levine GM (1999) Effects of reasons analysis on the accuracy of predicting basketball games 1. *Journal of Applied Social Psychology* 29(3):517–530.

Hastie T, et al. (2009) *The elements of statistical learning: data mining, inference, and prediction*, volume 2 (Springer).

Henz M (2001) Scheduling a major college basketball conference—revisited. *Operations Research* 49(1):163–168.

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70(1-3):489–501.

Jaggi M (2013) Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *International Conference on Machine Learning*, 427–435 (PMLR).

Jiaqi Xu J, et al. (2019) Designing and evaluating dynamic pricing policies for major league baseball tickets. *Manufacturing & Service Operations Management* 21(1):121–138.

Johnstone DJ, et al. (2011) Tailored scoring rules for probabilities. *Decision Analysis* 8(4):256–268.

Kendall MG (1938) A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Kvam P, Sokol JS (2006) A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NrL)* 53(8):788–803.

Magel R, Melnykov Y (2014) Examining influential factors and predicting outcomes in European soccer games. *International Journal of Sports Science* 4(3):91–96.

Martin T, et al. (2016) Exploring limits to prediction in complex social systems. *Proceedings of the 25th International Conference on World Wide Web*, 683–694.

Miljković D, et al. (2010) The use of data mining for basketball matches outcomes prediction. *IEEE 8th International Symposium on Intelligent Systems and Informatics*, 309–312 (IEEE).

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. *Manufacturing & Service Operations Management* 22(1):158–169.

Motegi S, Masuda N (2012) A network-based dynamical ranking system for competitive sports. *Scientific Reports* 2(1):1–7.

NBA (2020a) NBA Articles. Retrieved on May 31, 2020. `nba.com/article/2020/03/11/coronavirus-pandemic-causes-nba-suspend-season`.

NBA (2020b) NBA Statistics Homepage. Retrieved on May 31, 2020. `stats.nba.com`.

NBAcom (2023) NBA play-in tournament. Retrieved on December 21, 2023. `https://www.nba.com/news/nba-play-in-tournament`.

NBAstuffer (2023) Strength of schedule (sos) Retrieved on October 1, 2023. `https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/`.

Nemhauser GL, Trick MA (1998) Scheduling a major college basketball conference. *Operations Research* 46(1):1–8.

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning*, 625–632.

Pedregosa F, et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Platt J, et al. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 10(3):61–74.

Prasetio D, et al. (2016) Predicting football match results with logistic regression. *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 1–5 (IEEE).

Rasmussen RV, Trick MA (2008) Round robin scheduling–a survey. *European Journal of Operational Research* 188(3):617–636.

Sagi O, Rokach L (2018) Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4):e1249.

Shen Y (2005) *Loss functions for binary classification and class probability estimation* (University of Pennsylvania).

Spearman C (1904) The proof and measurement of association between two things. *The American Journal of Psychology* 15(1):72–101.

Sziklai BR, Biró P, Csató L (2022) The efficacy of tournament designs. *Computers & Operations Research* 144:105821.

Van Eetvelde H, et al. (2021) The probabilistic final standing calculator: a fair stochastic tool to handle abruptly stopped football seasons. *AStA Advances in Statistical Analysis* 1–19.

Van Haaren J, Davis J (2015) Predicting the final league tables of domestic football leagues. *Proceedings of the 5th International Conference on Mathematics in Sport*, 202–207.

Weiss HJ (1986) The bias of schedules and playoff systems in professional sports. *Management Science* 32(6):696–713.

Westphal S (2014) Scheduling the German basketball league. *Interfaces* 44(5):498–508.

Yannakakis M (1985) On a class of totally unimodular matrices. *Mathematics of Operations Research* 10(2):280–304.

Appendix A: Solution Methods for the Stochastic Model PC

In this section, we introduce two solution methodologies based on model PC, introduced in §4.2.1. Before elaborating on the solution methodologies, we remark that our formulation PC may be viewed as a stochastic program with recourse, where the x-variables are first-stage variables (for which there is only one choice to be made) and the y- and z-variables are second-stage "recourse" variables (for which there is one such variable for each possible outcome ξ). Note, however, that in our application there is no true recourse. Rather, y and z are auxiliary variables whose purpose is to linearize the objective function. In the next section, as the full stochastic optimization problems are too large to solve directly, we introduce two methods which approximately solve PC.

A.1. Mean Value Approximation

Replacing all random parameters in a stochastic optimization problem by their expected values yields a deterministic problem known as the Mean Value Problem (MVP). In our case, we may produce MVP for PC by replacing the random variables W_g which represent the outcome of each game g with their means $p_g = \mathbb{E}[W_g]$. The y-variables are then interpreted as expected values over all outcomes $\xi \in \Xi$, given the shortened season $\boldsymbol{x}$, and z variables capture relative positions of teams according to y. The MVP corresponding to PC is:

$$\text{[PC-MVP]} \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad \sum_{i\in T}\sum_{j\in T:j>i}(z_{ij}\hat{z}_{ij}+(1-z_{ij})(1-\hat{z}_{ij})) \tag{34}$$

$$\text{s.t.} \quad y_i = \frac{1}{m}\left(y_i^0 + \sum_{g\in G_i^h} p_g x_g + \sum_{g\in G_i^a}(1-p_g)x_g\right) \qquad \forall i\in T \tag{35}$$

$$z_{ij} \geq y_i - y_j \geq z_{ij} - 1 \qquad \forall i,j\in T: i<j \tag{36}$$

$$z_{ij} \in \{0,1\} \qquad \forall i,j\in T: i<j \tag{37}$$

$$\boldsymbol{x} \in X. \tag{38}$$

A.2. Sample Average Approximation

Sample Average Approximation (SAA) is a Monte Carlo simulation-based technique for approximating stochastic optimization problems (Kleywegt et al. 2002), Let $\mathcal{S} = \{\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(|\mathcal{S}|)}\}$ be an independently and identically distributed random sample of ξ. SAA reduces the size of the problem by approximating the expected value in the objective function with the sample average function. We use the superscript s to reference the second-stage variables and random parameters under scenario $s \in \mathcal{S}$. For instance, under scenario s, $W_g^{(s)}$ refers to the outcome of game g, $\hat{y}_i^{(s)}$ refers to the win percentage of team i at the end of the full season, and $y_i^{(s)}$ refers to the decision variable for the win percentage of team i at the end of the shortened season. We construct the SAA counterpart of the stochastic program PC by replacing the full set of outcomes Ξ with the sample set $\mathcal{S}$.

$$\text{[PC-SAA]} \max_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}} \quad \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\sum_{i\in T}\sum_{j\in T:j>i}\left(z_{ij}^{(s)}\hat{z}_{ij}^{(s)}+(1-z_{ij}^{(s)})(1-\hat{z}_{ij}^{(s)})\right) \tag{39}$$

$$\text{s.t.} \quad y_i^{(s)} = \frac{1}{m}\left(y_i^0 + \sum_{g\in G_i^h} W_g^{(s)} x_g + \sum_{g\in G_i^a}(1-W_g^{(s)})x_g\right) \qquad \forall i\in T, \forall s\in\mathcal{S} \tag{40}$$

$$z_{ij}^{(s)} \geq y_i^{(s)} - y_j^{(s)} \geq z_{ij}^{(s)} - 1 \qquad \forall i,j \in T : i < j, \forall s \in \mathcal{S} \tag{41}$$

$$z_{ij}^{(s)} \in \{0,1\} \qquad \forall i,j \in T : i < j, \forall s \in \mathcal{S} \tag{42}$$

$$\boldsymbol{x} \in X. \tag{43}$$

As the sample size increases, the optimal solution and the optimal value of the SAA problems converge to their 'true' stochastic counterparts with probability one (Kleywegt et al. 2002).

A.3. Variable Fixing and Preprocessing

We may improve the computational efficiency of both the SAA and MVP counterparts of PC by fixing certain variables at their optimal values and eliminating redundant constraints, as described by the following proposition.

PROPOSITION 4. *Let $\tilde{\xi}$ be an arbitrary realization or expected value of ξ. For each team i, sort G_i^h and G_i^a in non-decreasing order of $W(\tilde{\xi})$. Let U_i^h and L_i^h be the summation of $W_g(\tilde{\xi})$ values corresponding to the first and last m_i^h games in G_i^h, respectively. Similarly, let U_i^a and L_i^a be the summation of $W_g(\tilde{\xi})$ values corresponding to the first and last m_i^a games in G_i^a, respectively. Define $y_i^U = \frac{1}{m}(y_i^0 + U_i^h + m_i^a - L_i^a)$ and $y_i^L = \frac{1}{m}(y_i^0 + L_i^h + m_i^a - U_i^a)$ to be the optimistic and pessimistic win percentages of team i under $\tilde{\xi}$, respectively. For each pair of teams (i,j):*

(i) If $y_i^L - y_j^U > 0$, then $z_{ij}(\tilde{\xi}) = 1$, and the corresponding linking constraints are redundant.

(ii) If $y_i^U - y_j^L < 0$, then $z_{ij}(\tilde{\xi}) = 0$, and the corresponding linking constraints are redundant.

Proof. Using the definition of the z-variables, the statements follow from $z_{ij}(\tilde{\xi}) \geq y_i(\tilde{\xi}) - y_j(\tilde{\xi}) \geq y_i^L - y_j^U > 0$, and $0 > y_i^U - y_j^L \geq y_i(\tilde{\xi}) - y_j(\tilde{\xi}) \geq z_{ij}(\tilde{\xi}) - 1$, respectively. □

Table 8 summarizes the results of applying our variable fixing technique, introduced in Proposition 4, in PC variants. The high percentages under the columns "Percentage" in Table 8 highlight the effectiveness of the variable fixing technique in eliminating a large proportion of the z-variables across different scenarios in both MVP and SAA. More importantly, the technique is able to eliminate between 8,000–17,000 binary variables in the SAA problems with 50 scenarios. We also observe that as the suspension day increases (i.e., the season is suspended later), more pairs of teams become impossible to switch ranking positions, given the limited number of remaining games. For instance, when the season is suspended on day 140, PC-SAA has control over only 20% of these binary variables in a 74-game shortened season, with the remaining 80% of the variables fixed (i.e., eliminated).

A.4. Tuning the Sample Size for SAA

Here we analyze the impact of the number of scenarios on our PC-SAA model. Figure 10 presents the performance of PC-SAA across six choices of sample size $|\mathcal{S}| \in \{5, 10, 25, 50, 100, 200\}$. Each boxplot corresponds to 14 values for 14 NBA seasons, assuming a suspension day 100 and 66 as the target number of games in the shortened season. The concordance values are obtained after evaluating the solution $\boldsymbol{x}$ proposed by each model on 1,000 randomly-generated scenarios. The panel on the right presents the optimality gaps of the SAA problems after reaching a time limit of 3600 seconds. As the number of scenarios increases, one should expect to obtain a closer approximation of the true stochastic problem via SAA. However, a larger sample

Sus. Day	MVP		SAA	
(GT)	Percentage	Variables	Percentage	Variables
80 (62)	62.9%	273.7	38.6%	8,393.0
100 (66)	70.1%	304.9	53.9%	11,722.8
120 (70)	79.6%	346.4	65.5%	14,240.7
140 (74)	88.1%	383.2	80.1%	17,423.0
Average	75.2%	327.1	59.5%	12,944.9

Table 8 **Number of z-variables eliminated by our variable fixing technique, reported for different suspension days (80, 100, 120, 140) and target number of games/team (62,66,70,74). Results are averaged over 14 seasons.**

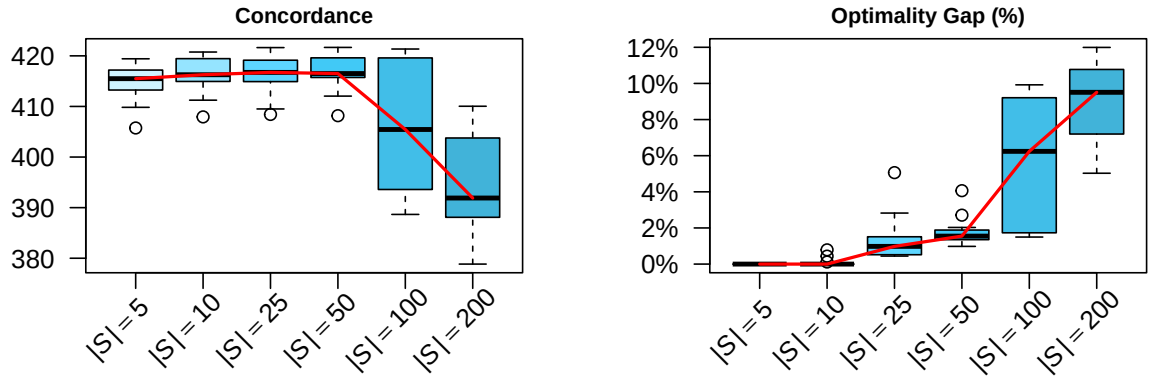


Figure 10 **Performance of the SAA algorithm across different choices of sample size.**

amounts to solving a more challenging SAA problem. As depicted in Figure 10, initially as the sample size increases, the quality of the solution improves in the simulation phase. However, after surpassing 50 scenarios, the SAA becomes computationally intractable, leading to larger optimality gaps and a degradation in the quality of the solution. Hence, the trade-off between the quality of the solution and the runtime is balanced at 50 scenarios. Thus, we select 50 scenarios for the SAA counterpart of our PC model in our main experiments.

Appendix B: Proof of Theorems and Propositions

Proof of Proposition 1. The first inequality in (3) is a direct result of the Durbin-Stuart inequality (Durbin and Stuart 1951, Theorem 2). To prove the second inequality, let us define the Manhattan distance between rankings r and $\hat{r}$ as $d_{\mathrm{M}}(r,\hat{r}) = \sum_{i\in T} |d_i|$, where $d_i = r_i - \hat{r}_i$. We first show the following:

$$d_{\mathrm{E}}(r,\hat{r}) \le \frac{1}{2} d^2_{\mathrm{M}}(r,\hat{r}). \tag{44}$$

Observe that $\sum_{i\in T} d_i = \sum_{i\in T} r_i - \sum_{i\in T} \hat{r}_i = 0$. Using the triangle inequality, this implies that for each $i \in T$:

$$|d_i| = |-\sum\nolimits_{j\neq i} d_j| \le \sum\nolimits_{j\neq i} |d_j| = d_{\mathrm{M}}(r,\hat{r}) - |d_i| \Rightarrow d_{\mathrm{M}}(r,\hat{r}) \ge 2|d_i|.$$

By multiplying the two sides of the last inequality above by $|d_i|$ and summing across all teams $i \in T$, we establish (44) by expanding $d^2_{\mathrm{M}}(r,\hat{r})$ as follows:

$$d^2_{\mathrm{M}}(r,\hat{r}) = d_{\mathrm{M}}(r,\hat{r}) \sum\nolimits_{i\in T} |d_i| \ge 2 \sum\nolimits_{i\in T} |d_i|^2 = 2 \sum\nolimits_{i\in T} d_i^2 = 2 d_{\mathrm{E}}(r,\hat{r}).$$

Finally, by the Diaconis-Graham inequality (Diaconis and Graham 1977, Theorem 2), we obtain

$$d_{\mathrm{M}}(r,\hat{r}) \leq 2\left(\frac{n(n-1)}{2} - \tau_{\mathrm{C}}(r,\hat{r})\right). \tag{45}$$

Inequalities (44) and (45) establish the second inequality in (3) and complete the proof. □

LEMMA 1. *Maximum quadratic Euclidean distance between any two permutations of* $\{1,\dots,n\}$ *is* $\frac{n}{3}(n^2-1)$.

Proof. Let P denote the set of all permutations of $N \coloneqq \{1,\dots,n\}$. Our goal is to find permutations $p \in P$ and $q \in P$ which maximize $\sum_{i\in N}(p_i - q_i)^2$. Note that, without loss of generality, we can fix $p = (1,2,\dots,n)$, and restate the problem as

$$\max_{q\in P} \sum\nolimits_{i\in N}(q_i - i)^2. \tag{46}$$

We first note that setting $q = (n, n-1, \dots, 1)$ (i.e., the exact reverse of p) yields $\sum_{i\in N}(q_i - i)^2 = \sum_{i\in N}\left((n+1-i)-i\right)^2 = \sum_{i\in N}(2i-n-1)^2 = 4\sum_{i\in N} i^2 - 4(n+1)\sum_{i\in N} i + n(n+1)^2 = \frac{n}{3}(n^2-1)$.

Next we use LP duality to show that this lower bound is tight. Note that we may state (46) as the following assignment problem

$$\begin{aligned}
\max_{\boldsymbol{x}} \quad & \sum\nolimits_{i\in N}\sum\nolimits_{j\in N} x_{i,j}(i-j)^2 & \\
\text{s.t.} \quad & \sum\nolimits_{j\in N} x_{i,j} = 1 & \forall i \in N \\
& \sum\nolimits_{i\in N} x_{i,j} = 1 & \forall j \in N \\
& x_{i,j} \geq 0 & \forall i,j \in N,
\end{aligned}$$

which can be stated in the dual form as

$$\begin{aligned}
\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum\nolimits_{i\in N}\alpha_i + \sum\nolimits_{j\in N}\beta_j & \\
\text{s.t.} \quad & \alpha_i + \beta_j \geq (i-j)^2 & \forall i,j \in N.
\end{aligned}$$

It is not difficult to verify that setting

$$\alpha_i = \beta_i = \begin{cases} \frac{1}{2}(n-i)^2 & \text{if } i \leq \frac{n}{2} \\ \frac{1}{2}(i-1)^2 & \text{if } i > \frac{n}{2} \end{cases} \qquad \forall i \in N$$

satisfies $\alpha_i + \beta_j = \alpha_i + \alpha_j \geq (i-j)^2$ for each i and j, and yields $\sum_{i\in N}\alpha_i + \sum_{j\in N}\beta_j = 2\sum_{i\in N}\alpha_i = \frac{n}{3}(n^2-1)$. Hence, by strong duality, the optimal value for (46) is $\frac{n}{3}(n^2-1)$. □

Proof of Proposition 2. The statement holds when win percentages are identical, which results in 0 on both sides. Now, assuming that win percentages are not identical, there exists team j such that $y_j(\boldsymbol{x},\xi) \neq \hat{y}_j(\xi)$. Note that $\hat{y}_j(\xi) \in \{0, \frac{1}{\hat{m}}, \frac{2}{\hat{m}}, \dots, 1\} \subseteq \{0, \frac{1}{L}, \frac{2}{L}, \dots, 1\}$, since $\hat{y}_j(\xi)$ is the number of wins for team j in the full season divided by $\hat{m}$. Similarly, $y_j(\boldsymbol{x},\xi) \in \{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\} \subseteq \{0, \frac{1}{L}, \frac{2}{L}, \dots, 1\}$. Therefore, the closest

that $y_j(\boldsymbol{x},\xi)$ and $\hat{y}_j(\xi)$ can get and still be different is $\frac{1}{L}$; thus for non-identical win percentage vectors we have

$$\frac{1}{L^2} \leq \sum\nolimits_{i\in T} \left(y_i(\boldsymbol{x},\xi) - \hat{y}_i(\xi)\right)^2. \tag{47}$$

On the other hand, by Lemma 1 we have

$$d_{\mathrm{E}}(r(\boldsymbol{x},\xi),\hat{r}(\xi)) \leq \frac{n}{3}(n^2-1). \tag{48}$$

Multiplying both sides of (47) and (48) and rearranging the resulting inequality yields

$$d_{\mathrm{E}}(r(\boldsymbol{x},\xi),\hat{r}(\xi)) \leq \frac{n}{3}(n^2-1)L^2 \sum\nolimits_{i\in T} \left(y_i(\boldsymbol{x},\xi) - \hat{y}_i(\xi)\right)^2,$$

which shows existence of constant $D \leq \frac{n(n^2-1)}{3}L^2$. □

Proof of Theorem 1. Using identity $\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \mathbb{V}[X]$, with $\mathbb{V}(\cdot)$ denoting variance, we obtain

$$\mathbb{E}_\xi\left[\sum\nolimits_{i\in T}\left(y_i(\xi)-\hat{y}_i(\xi)\right)^2\right] = \sum_{i\in T}\mathbb{E}_\xi\left[\left(y_i(\xi)-\hat{y}_i(\xi)\right)^2\right] = \sum_{i\in T}\mathbb{E}_\xi\left[y_i(\xi)-\hat{y}_i(\xi)\right]^2 + \sum_{i\in T}\mathbb{V}_\xi\left[y_i(\xi)-\hat{y}_i(\xi)\right].$$

Clearly, $\mathbb{E}_\xi\left[y_i(\xi)-\hat{y}_i(\xi)\right] = \mu_i - \hat{\mu}_i$. Moreover, given the definition of $y_i(\xi)$ and $\hat{y}_i(\xi)$, we have

$$\begin{aligned} y_i(\xi)-\hat{y}_i(\xi) =& (\frac{1}{m}-\frac{1}{\hat{m}})y_i^0 + \sum\nolimits_{g\in G_i^h} W_g(\xi)\left(\frac{1}{m}x_g - \frac{1}{\hat{m}}\right) + \sum\nolimits_{g\in G_i^a}(1-W_g(\xi))\left(\frac{1}{m}x_g-\frac{1}{\hat{m}}\right) \\ \Rightarrow \mathbb{V}_\xi\left[y_i(\xi)-\hat{y}_i(\xi)\right] =& \sum\nolimits_{g\in G_i^h\cup G_i^a} p_g(1-p_g)\left(\frac{1}{m}x_g-\frac{1}{\hat{m}}\right)^2, \end{aligned} \tag{49}$$

where we have used $\mathbb{V}_\xi\left[W_g(\xi)\right] = \mathbb{V}_\xi\left[1-W_g(\xi)\right] = p_g(1-p_g)$. Given that $x_g \in \{0,1\}$, we have

$$\left(\frac{1}{m}x_g - \frac{1}{\hat{m}}\right)^2 = \frac{1}{m^2}x_g^2 - \frac{2}{m\hat{m}}x_g + \frac{1}{\hat{m}^2} = \frac{x_g}{m^2}\left(1-\frac{2m}{\hat{m}}\right) + \frac{1}{\hat{m}^2}, \tag{50}$$

where we have used $x_g^2 = x_g$. Replacing (50) into (49) yields

$$\begin{aligned} \mathbb{V}_\xi\left[y_i(\xi)-\hat{y}_i(\xi)\right] &= \sum\nolimits_{g\in G_i^h\cup G_i^a} p_g(1-p_g)\left(\frac{x_g}{m^2}\left(1-\frac{2m}{\hat{m}}\right)+\frac{1}{\hat{m}^2}\right) \\ &= \left(1-\frac{2m}{\hat{m}}\right)\frac{1}{m^2}\sum_{g\in G_i^h\cup G_i^a} p_g(1-p_g)x_g + \frac{1}{\hat{m}^2}\sum_{g\in G_i^h\cup G_i^a} p_g(1-p_g) = \left(1-\frac{2m}{\hat{m}}\right)v_i + \hat{v}_i, \end{aligned}$$

which completes the proof. □

Proof of Proposition 3. As illustrated in Figure 11, the set X corresponds to a bipartite multigraph with $2n$ nodes (n home teams and n away teams) and each game $g \in G$ corresponds to an edge of unit capacity between home team $i(g)$ and away team $j(g)$. The coefficient matrix of X is the incidence matrix of this bipartite multigraph, which is totally unimodular (see e.g., Yannakakis 1985). □

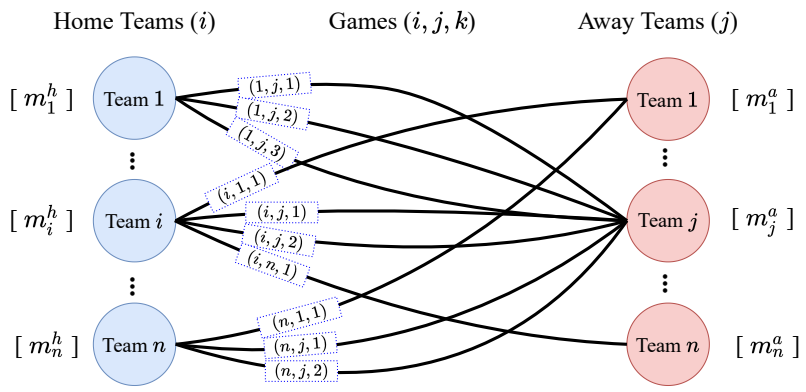


Figure 11 **Set of feasible schedules X corresponds to a bipartite multigraph.**

Appendix C: Analysis of the PW-DQIP Model

Let us take a closer look at the mathematical formulation of the PW-DQIP model, and its objective function in particular. As discussed in §5, PW-DQIP approximates the objective function in equation (8) while formulating a deterministic quadratic integer program. Once, we expand the objective function in PW-DQIP, (12), the resulting expression consists of i) *L2 Norm* of the difference between the vector of win percentages in the shortened season and a target value (i.e., win percentages in the full season), ii) variance of the win percentages in the shortened season multiplied by a constant factor (α), and iii) variance of the win percentages in the full season. Note that the last term $\hat{v}_i$ is constant, as it does not include any decision variables.

$$\mathbb{E}_\xi\left[\sum\nolimits_{i\in T}\left(y_i(\xi)-\hat{y}_i(\xi)\right)^2\right]=\sum\nolimits_{i\in T}\left(\left(\mu_i-\hat{\mu}_i\right)^2+v_i\left(1-\frac{2m}{\hat{m}}\right)+\hat{v}_i\right) \tag{51}$$

$$=\sum\nolimits_{i\in T}\left(\overbrace{\left(\mu_i-\hat{\mu}_i\right)^2}^{\text{L2 Norm}}+\overbrace{\left(1-\frac{2m}{\hat{m}}\right)}^{\text{Constant }\alpha}\overbrace{v_i}^{\text{Variance}}+\overbrace{\hat{v}_i}^{\text{Constant}}\right). \tag{52}$$

Omitting the third (constant $\hat{v}_i$) term in (52), the model PW-DQIP equivalently optimizes the following function:

$$\sum\nolimits_{i\in T}\left(\left(\mu_i-\hat{\mu}_i\right)^2+\alpha\, v_i\right), \tag{53}$$

where:

$$\alpha=1-\frac{2m}{\hat{m}}. \tag{54}$$

Note that the number of games per team in the full regular season is fixed at 82 per the structure of the NBA (i.e., $\hat{m}=82$). Depending on the target number of games in the shortened season (m), parameter α could be positive, negative, or zero. According to (54), parameter α is zero when $m=\frac{\hat{m}}{2}=\frac{82}{2}=41$. In other words, when the target number of games in the shortened season is exactly half of the length of the full season, the coefficient of the variance term in the objective function disappears. The variance term has a positive coefficient when $m<\frac{\hat{m}}{2}$ (shortened season is "short"), and a negative coefficient when $m>\frac{\hat{m}}{2}$ (shortened season is "long"). Given the shortened season is long ($m>\frac{\hat{m}}{2}$) in all of the suspension instances in our

Target # games (m)	Shortened season	Variance coefficient (α)	Desired games by the model PW-DQIP
$m < \frac{\hat{m}}{2}$	Short	Positive	One-sided matchups
$\boldsymbol{m > \frac{\hat{m}}{2}}$	**Long**	**Negative**	**Evenly matched games**

Table 9 Comparing short vs. long shortened seasons and the practical implications in the model PW-DQIP.

experiments ($m \in \{62, 66, 70, 74\}$), the PW-DQIP model favors games with higher variance, as the objective function in (53) is to be minimized and the coefficient of the variance term (α) is negative.

Now, let us revisit our argument focusing on individual games and how the inclusion or exclusion of a game contributes to the variance term in the objective function. The objective function in PW-DQIP can be reformulated as:

$$\sum_{i\in T}\left((\mu_i-\hat{\mu}_i)^2+v_i\left(1-\frac{2m}{\hat{m}}\right)+\hat{v}_i\right)=\sum_{i\in T}\hat{v}_i+\sum_{i\in T}(\mu_i-\hat{\mu}_i)^2+2\left(1-\frac{2m}{\hat{m}}\right)\frac{1}{m^2}\sum_{g\in G}p_g(1-p_g)x_g.$$

For a game $g \in G$, the variance for the predicted probability p_g, following the Bernoulli distribution, is $p_g(1-p_g)$. It is not difficult to see that the variance term is maximized when our estimated p_g is closer to 0.5. We can also define a *sharpness* metric as $\max\{p_g, 1-p_g\}$ to assess how sharp the estimated probabilities are. A higher variance in game outcomes leads to lower sharpness and vice versa. In other words, higher variance in the predicted probabilities will favor more evenly matched games, while a lower variance will favor more one-sided matches. Table 9 summarizes our discussion in this section.

To see the effect of suspension day on the average variance of the outcomes of the selected games compared to the average variance of the excluded games, we further note that:

$$\sum_{i\in T}\hat{v}_i+\sum_{i\in T}(\mu_i-\hat{\mu}_i)^2+2\left(1-\frac{2m}{\hat{m}}\right)\frac{1}{m^2}\sum_{g\in G}p_g(1-p_g)x_g=\sum_{i\in T}\hat{v}_i+\sum_{i\in T}(\mu_i-\hat{\mu}_i)^2+2\left(1-\frac{2m}{\hat{m}}\right)\frac{G_1}{m^2}\bar{v}$$

where G_1 is the number of post-suspension games included in the shortened season (note that $G_1 = \frac{1}{2}\sum_{i\in T}(m_i^a + m_i^h) = \sum_{g\in G} x_g$ for any feasible shortened season $\boldsymbol{x} \in X$), and $\bar{v}$ is the average variance of these games, i.e.:

$$\bar{v}=\frac{1}{G_1}\sum_{g\in G}p_g(1-p_g)x_g.$$

Thus the coefficient $2\left(1-\frac{2m}{\hat{m}}\right)\frac{G_1}{m^2}$ governs the trade-off between the contribution of the average variance of the selected games and the L2 term to the objective function. As depicted in Figure 12, this coefficient increases (its magnitude decreases) as suspension day increases. As a result, relative to the L2-norm component of the objective function, the variance term has a lower weight, and one should expect lower variance in the selected games for later suspension days.

Incidentally, we can find a signature of this behaviour aligned with the summarized conclusions in Table 9 in our computational experiments. Figure 13 illustrates the variance (top row) and the sharpness (bottom row) of the predicted probabilities in i) all remaining games (blue), ii) selected games (green), and iii) excluded games (pink) in our shortened seasons. As we can see, the sharpness of the selected games for suspension days 80 and 100 is typically lower (and the average variance $\bar{v}$ higher) than that of the excluded games. Moreover, for later suspension days (i.e., 120 and 140), the pattern for sharpness and variance dissipates due to the fact that less weight is being put on the variance term later in the season (recall Figure 12).

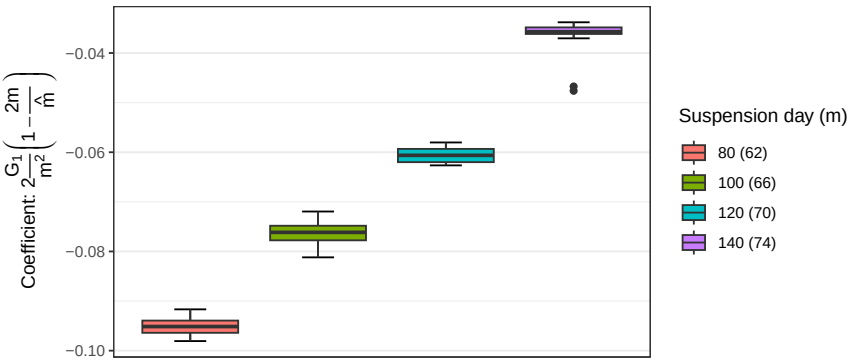


Figure 12 **Distribution of the variance coefficients in model PW-DQIP across 14 NBA seasons**

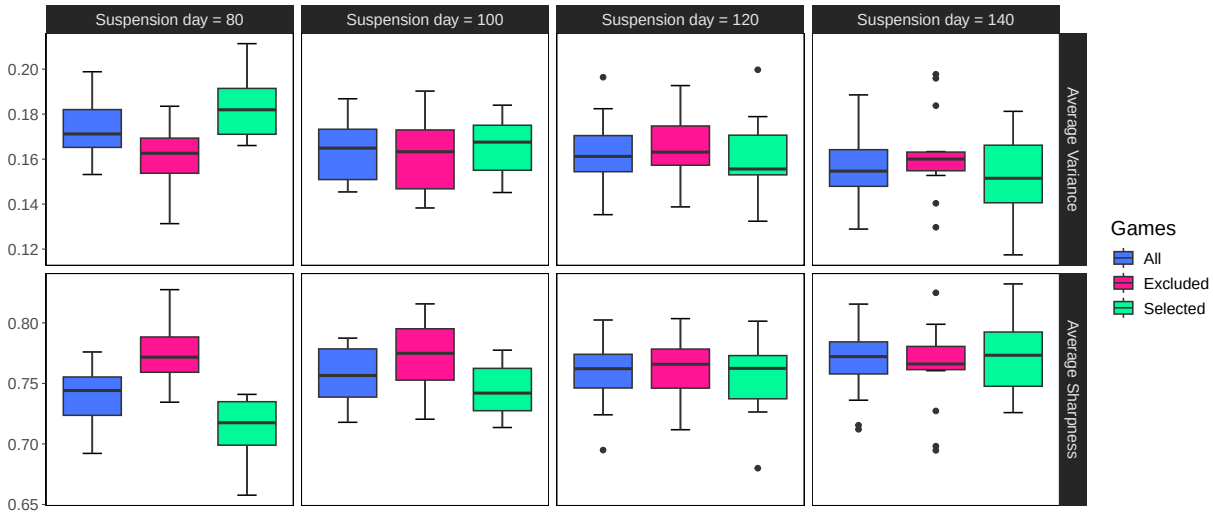


Figure 13 **Comparing the average variance and sharpness of the predicted probabilities between selected/excluded games in the shortened season.**

Appendix D: Frank-Wolfe Algorithm

Algorithm 1 presents the FW algorithm for solving instances of continuous relaxation of PW-DQIP, in which $\bar{X}$ is the continuous relaxation of X, and the objective function f is of the form

$$f(\boldsymbol{x}) = \sum\nolimits_{i \in T} \left((\mu_i(\boldsymbol{x}) - \hat{\mu}_i)^2 + v_i(\boldsymbol{x}) \left(1 - \frac{2m}{\hat{m}}\right) + \hat{v}_i \right), \tag{55}$$

where $\mu_i(\boldsymbol{x})$ and $v_i(\boldsymbol{x})$ are defined in equations (19) and (20), respectively. We remark that, since f is a convex quadratic function, its gradient can be computed easily, and a closed-form optimal solution to the line search problem (57) can be found using the first-order optimality conditions.

Given that the atomic solution $\hat{\boldsymbol{x}}^{(t)}$ produced by solving the transportation problem (56) is a feasible integer solution, it provides an upper bound on the optimal value of PW-DQIP. As the algorithm iterates, $\boldsymbol{x}^{(t)}$ converges to the optimal fractional solution, and $\hat{\boldsymbol{x}}^{(t)}$ yields a tighter upper bound. Consequently, the best atomic solution (i.e., $\hat{\boldsymbol{x}}^*$) can be used as a near optimal integer solution to PW-DQIP. Henceforth, we refer to this procedure of producing the shortened season $\hat{\boldsymbol{x}}^*$ using FW (Algorithm 1) as PW-FW.

Algorithm 1 PW-FW: Frank-Wolfe algorithm for solving continuous relaxation of PW-DQIP

Let $t \leftarrow 0$, and find an integer solution $\boldsymbol{x}^{(0)} \in X$.
Let $\hat{\boldsymbol{x}}^* \leftarrow \boldsymbol{x}^{(0)}$
while not converged **do**
    Compute gradient $d_g^{(t)} = \nabla f_{x_g}(\boldsymbol{x}^{(t)})$ for each $g \in G$
    Find the integer solution $\hat{\boldsymbol{x}}^{(t)}$ by solving the following transportation problem

$$\text{[Transportation Problem]} \quad \hat{\boldsymbol{x}}^{(t)} = \arg\min_{\boldsymbol{x} \in \bar{X}} \quad \sum\nolimits_{g \in G} d_g^{(t)} x_g. \tag{56}$$

    if $f(\hat{\boldsymbol{x}}^{(t)}) < f(\hat{\boldsymbol{x}}^*)$ **then**
        $\hat{\boldsymbol{x}}^* \leftarrow \hat{\boldsymbol{x}}^{(t)}$
    end if
    Compute the step-size $\gamma^{(t)}$ using the following line search

$$\text{[Line Search]} \quad \gamma^{(t)} = \arg\min_{\gamma \in [0,1]} \quad f\left((1-\gamma)\boldsymbol{x}^{(t)} + \gamma \hat{\boldsymbol{x}}^{(t)}\right). \tag{57}$$

    Update $\boldsymbol{x}^{(t+1)} = (1-\gamma^{(t)})\boldsymbol{x}^{(t)} + \gamma^{(t)}\hat{\boldsymbol{x}}^{(t)}$, and set $t \leftarrow t+1$.
end while

We would like to remark that our implementation of FW algorithm (Algorithm 1) solves the continuous relaxation of PW-DQIP to optimality, and produces an integer solution as a byproduct thanks to the feasible region of PW-DQIP being totally unimodular.

Sub-optimality bounds for Algorithm 1. We first note the following property of iterates of FW based on convexity of f:

$$f(\bar{\boldsymbol{x}}^*) \geq f(\boldsymbol{x}^{(t)}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t)})^\top (\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}),$$

which is tight for $\boldsymbol{x}^{(t)} = \bar{\boldsymbol{x}}^*$. Consequently, we can construct a lower bound on the optimal value of the continuous relaxation of PW-DQIP as

$$\underline{f} = \max_t \left\{ f(\boldsymbol{x}^{(t)}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t)})^\top (\boldsymbol{x}^{(t)} - \hat{\boldsymbol{x}}^{(t)}) \right\}.$$

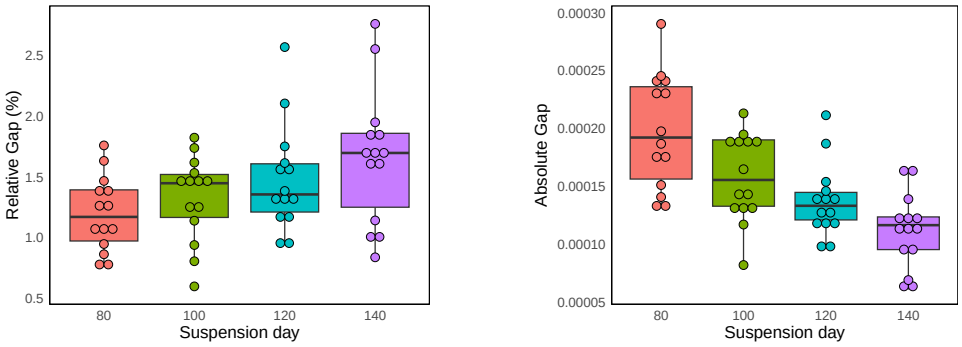
Let $\hat{\boldsymbol{x}}^*$ be the best integer solution produced by FW (i.e., $\hat{\boldsymbol{x}}^* = \arg\min_{\hat{\boldsymbol{x}}^{(t)}} f(\hat{\boldsymbol{x}}^{(t)})$), and $\boldsymbol{x}^*$ be the (unknown) integer optimal solution to PW-DQIP. Noting that $\underline{f} \leq f(\bar{\boldsymbol{x}}^*) \leq f(\boldsymbol{x}^*) \leq f(\hat{\boldsymbol{x}}^*)$ we derive the following relative sub-optimality gap

$$\text{Gap}_{\text{FW}} = \frac{f(\hat{\boldsymbol{x}}^*) - \underline{f}}{\underline{f}}, \tag{58}$$

which upper-bounds the optimality gap with respect to $f(\boldsymbol{x}^*)$. More precisely, rewriting the numerator in Eq. (58) as

$$f(\hat{\boldsymbol{x}}^*) - \underline{f} = \overbrace{f(\hat{\boldsymbol{x}}^*) - f(\boldsymbol{x}^*)}^{\text{Primal gap}} + \overbrace{f(\boldsymbol{x}^*) - \underline{f}}^{\text{Dual gap}}, \tag{59}$$

the gap computed in Eq. (58) captures the compound effect of quality of the solution produced by FW and formulation strength of PW-DQIP. Nonetheless, as illustrated in Figure 14, the sub-optimality gap in Eq. (58) is typically very low. For comparison, we also present the absolute gaps as in Eq. (59). Note that our implementation of FW converges to the optimal continuous solution after a few iterations (i.e., $\underline{f} = f(\bar{\boldsymbol{x}}^*)$).



(a) Relative gap (b) Absolute gap

Figure 14 **Upper bounds on the optimality gap of solutions produced by FW. Left: Relative sub-optimality gap (Eq. 58). Right: Absolute sub-optimality gap (Eq. 59).**

Finally, we remark that the FW algorithm may be extended to produce near-optimal solutions for PW-MMR as well. However, given the non-smooth min-max (i.e., ℓ_∞) objective in PW-MMR, a direct implementation of FW may not converge to an optimal continuous solution of PW-MMR (c.f. Nesterov 2018, for a counterexample); instead, we may minimize a smooth ℓ_p-approximation of ℓ_∞ for some large p or minimize the Moreau envelope of the objective function (c.f., Parikh and Boyd 2014).

Appendix E: Explanatory Features for Predictive Modeling

In this section, explanatory variables are categorized into four groups: overall team performance, basic team-level statistics, advanced team-level statistics, and player-level statistics.

E.1. Overall Team Performance

The most important variable which carries the largest explanatory weight among all the features is *win percentage* of home and guest teams which indicates the relative performance of both teams at the time they play against each other. We also include two variables of the same nature, percentage of home games won by the home team, and percentage of away games won by the guest team, to capture performance variability due to home/away condition. Table 10 lists all four overall performance features.

E.2. Basic Team–Level Statistics

An obvious choice for an explanatory variable to predict the outcome of basketball games is team-level raw statistics. Over a stretch of games, we can consider average team-level statistics by each team (e.g., average number of points, rebounds, assists, blocks, steals) as explanatory features. Table 11 shows the list of such variables for the home team. Note that using a prefix `oppt` before each variable in Table 11 results in the same variable for the guest team, and using a prefix `diff` for the same set of variables results in the difference between performance of home and guest teams with respect to each variable.

Overall Team Performance	Definition
WPCT	home team win percentage
opptWPCT	guest team win percentage
WPCTh	home team win percentage at home
opptWPCTg	guest team win percentage on the road

Table 10 Overall Performance Features for home and guest teams.

Basic Team Features	Definition
PTS	Average number of points per game scored
REB	Average number of rebounds per game
AST	Average number of assists per game
OREB	Average number of offensive rebounds per game
DREB	Average number of defensive rebounds per game
STL	Average number of steals per game
BLK	Average number of blocks per game
TOV	Average number of turnovers per game
PF	Average number of personal fouls per game
FGM	Average number of field goals made per game
FG%	Average field goal made percentage
3PM	Average number of 3-point field goals made per game
3P%	Average 3-point shot percentage
FTM	Average number of free throws made per game
FT%	Average free throw made percentage
PITP	Average number of points in the painted area per game
FBPs	Average number of fast–break points per game
2ndPTS	Average number of second chance points per game
PTSOFFTO	Average number of points off of opponent's turnovers per game
Poss	Average number of possessions per game

Table 11 Team–level statistics used as explanatory features.

E.3. Advanced Team–Level Statistics

These are advanced features calculated based on the raw data shown in Table 11. The goal of introducing and using these advanced features is to highlight strengths and weaknesses of each team adjusted by their style of play (e.g., reliance of each team on 3-point shots, defensive style of play). For instance, FG% and 3P% are two basic statistics while *effective field goal percentage*, denoted by eFG%, computes a weighted average field goal percentage, applying a weight of 2 to regular field goals and a weight of 1 to 3-point shots, scaled by the number of field goal attempts. Table 12 contains the list of advanced team-level statistics. The same set of variables are defined for the guest team (identified by prefix oppt) and the difference of each variable between the two teams (identified by the prefix diff).

E.4. Player-Level Statistics

Each of the features introduced in §E.2 can be defined for an individual player as well. With some adjustments, all the features introduced in §E.3 can also be defined for individual players. Given there are 15 players on the roster for any NBA team, with at least 9 playing considerable minutes each night, the total number

Advanced Team Features	Definition
OffRtg	Offensive rating, which is the number of points scored per 100 possessions
DefRtg	Defensive rating, which is the number of points allowed per 100 possessions
NetRtg	Net rating of a team is calculated by subtracting DefRtg from OffRtg
AST%	An estimate of the percentage of field goals assisted by team players per game
AST/TO	Assists per turnover ration, which is the number of assists per team divided by the number of turnovers the team has committed in a game
ASTRatio	Average number of assists per 100 possessions
OREB%	Offensive rebound percentage, which is an estimate of the percentage of available offensive rebounds a team grabs per game
DREB%	Defensive rebound percentage, which is an estimate of the percentage of available defensive rebounds a team grabs per game
REB%	Total rebound percentage which is an estimate of the percentage of total available rebounds a team grabs per game
TOV%	Turnover percentage, which is the percentage of plays that end in a player or team turnover
eFG%	Effective field goal percentage, which measures field goal percentage adjusting for made 3-point field goals being 1.5 times more valuable than made 2-point field goals.
TS%	True shooting percentage, a measure of shooting efficiency which differentiates between the number of points awarded by a regular field goal, a 3-point field goal, and a free throw.

Table 12 Advanced Team–level statistics calculated based on raw features from Table 11.

of features will grow substantially large, should we choose to define player-level features corresponding to team-level features in Tables 11 and 12. To tackle this issue, there are alternative ways to represent the efficiency of individual players using a combination of raw data. *Efficiency* rating introduced by Manley (1986) is one way to combine individual statistics into a single number. The formula is the following:

$$\texttt{EFF} = \texttt{PTS} + \texttt{REB} + \texttt{AST} + \texttt{STL} + \texttt{BLK} - \text{Missed } \texttt{FG} - \text{Missed } \texttt{FT} - \texttt{TOV}. \tag{60}$$

We calculate the EFF rating for each player according to (60) (using all the games prior to the game under study), and we represent the home and guest teams by their top 10 players, sorted according to their EFF values. Let $\texttt{hp}_i$ denote the EFF rating for the i^{th} best player of the home team, and let $\texttt{gp}_i$ denote the EFF rating for the i^{th} best player of the guest team, $i \in \{1, 2, \ldots, 10\}$. We can use the mean and standard deviation of these 10 numbers to represent overall efficiency of players on each roster and the discrepancy of EFF ratings between players. Let (EFF-mean, EFF-std) and (opptEFF-mean, opptEFF-std) represent the average EFF and standard deviation of EFF values for home and guest teams, respectively.

Lastly, we would like to highlight the importance of including player-level statisics in our model's application to suspension scenarios such as the NBA lockouts, where no games were played initially. To address this, we propose a modified approach using player-level statistics from previous seasons to create team-level features. This method accounts for off-season player movements, ensuring our predictions remain relevant for the upcoming season. By aggregating individual player performances, such as points and assists, we can effectively simulate team capabilities despite the absence of current season games. This adaptation demonstrates our model's flexibility and its ability to provide accurate predictions in a variety of scenarios, including those without any pre-season games.

Appendix F: Supplementary Results

F.1. Cross Simulation

Figure 15 illustrates the simulation results for different choices of predictive models in the prescriptive and simulation phases. Each point corresponds to the concordance value between the shortened and full seasons averaged over 14 seasons.

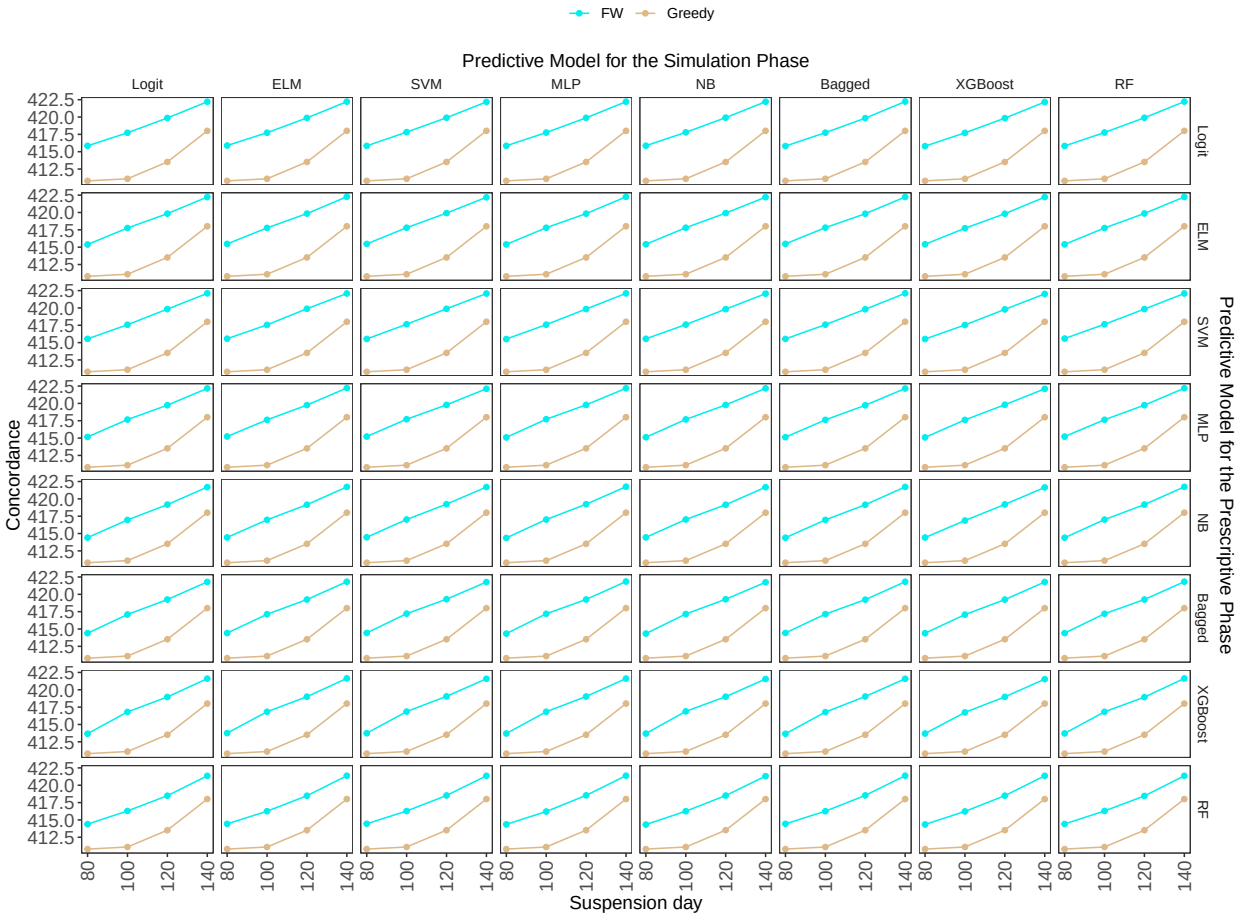


Figure 15 **Cross comparison of simulation results based on different choices of predictive models used for the prescriptive phase (rows) and predictive models used for the simulation phase (columns).**

F.2. Practical Implications: Comparing PW-MMR, Greedy the Status Quo Models

In §6.3.3, we introduced three metrics to gauge the practical implications of various shortened season plans and we compared our best performing prescriptive model (i.e., PW-MMR) with two benchmarks: Greedy and Status Quo in overall agreement. In this section, we plot the agreement distributions for each of the 4 suspension dates for models PW-MMR and the baseline Greedy in Figure 16, and similar distributions for PW-MMR and the baseline Status Quo in Figure 17. Our best-performing solution method, PW-MMR, outperforms both baseline solutions in all three success rate metrics both in terms of average percentage and the variation across 14 NBA seasons. Similar to the simulation results presented in Figure 7, as the suspension day increases, the margin of improvement with respect to the Status Quo model gets smaller.

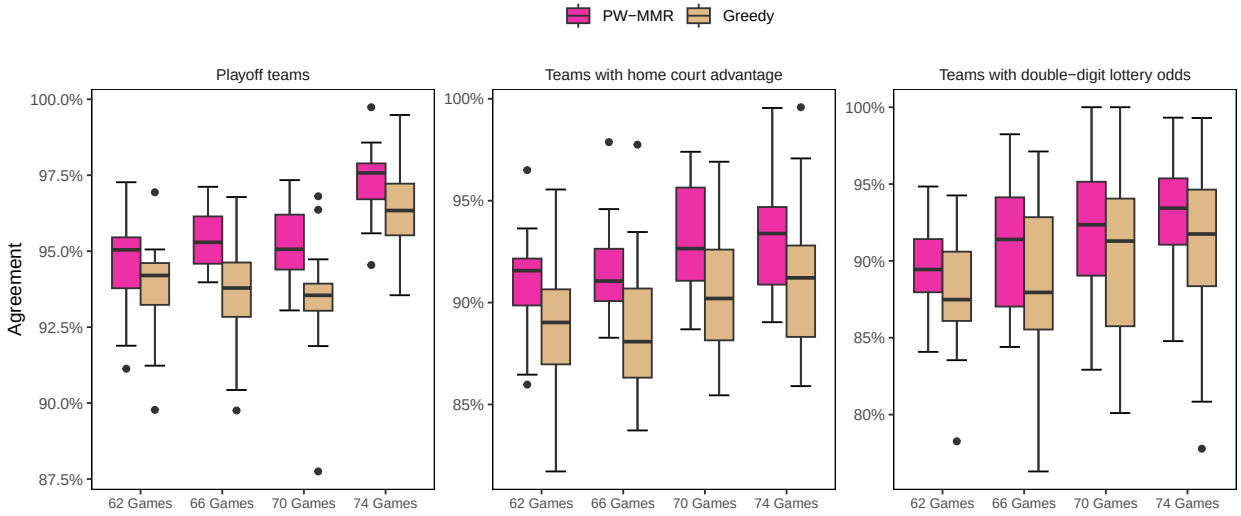


Figure 16 **Agreement distributions for PW-MMR and Greedy.**

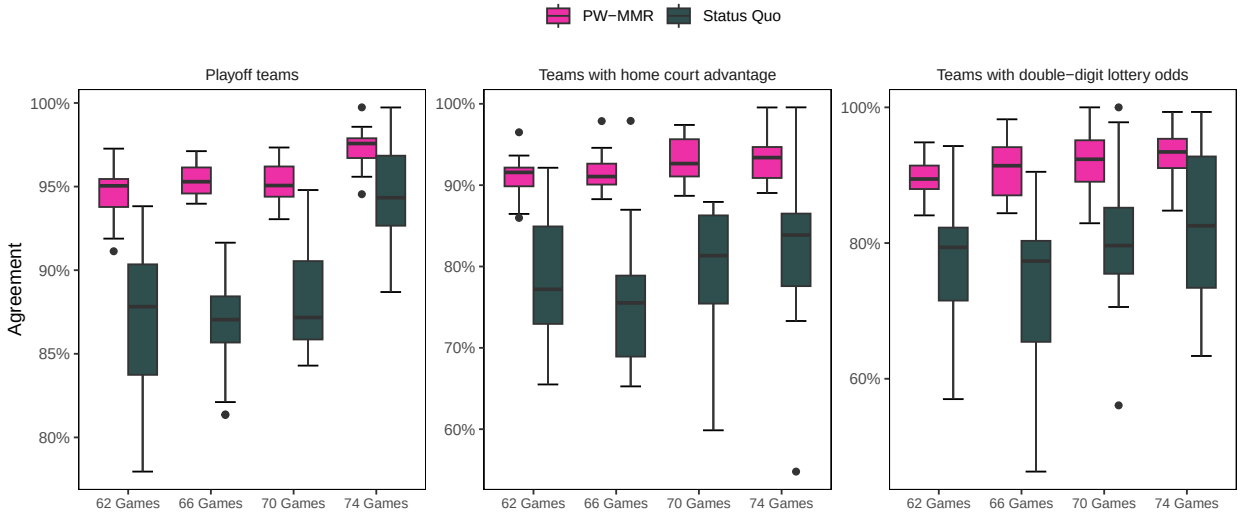
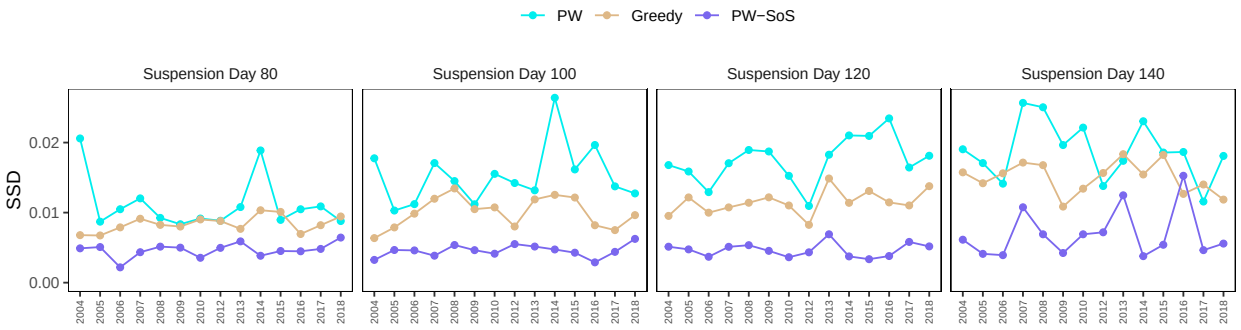


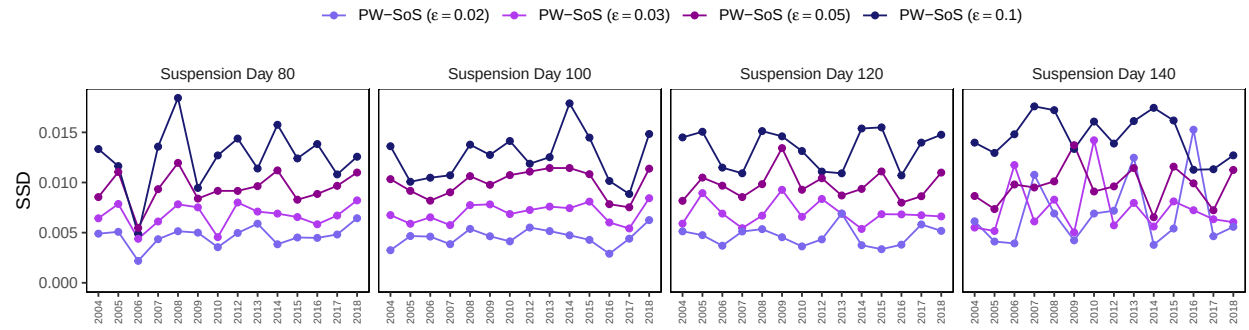
Figure 17 **Agreement distributions for PW and Status Quo.**

F.3. Strength of Schedule Extension: Comparison Across Seasons

In §6.3.4, we introduced model PW-SoS which amends the model PW-DQIP (presented in §4.2) by adding a constraint making sure that the strength of schedule discrepancy (SSD) in the shortened season is no larger than $\epsilon\%$ compared to the SSD metric in the full season. In our experiments, we tested 4 different choices of the parameter ϵ including (0.02, 0.03, 0.05, 0.1). Figure 9 in §6.3.4 illustrates the comparison between PW-SoS models as well as PW and Greedy using values averaged across 14 seasons. The following two figures in this section present the same comparison in terms of the metric SSD for individual seasons as lineplots. Figure 18(a) plots SSD according to PW and the our benchmark Greedy as well as the best choice of PW-SoS (with $\epsilon = 0.02$). Figure 18(b) plots the SSD values in individual seasons for 4 choices of ϵ. We can easily conclude that the model PW-SoS with $\epsilon = 0.02$ has the lowest overall SSD, thus performing better than the other three. Note that we tested the results according to a few other definitions of strength of schedule,



(a) Comparing SSD values in PW-SoS, PW and Greedy based on Figure 9.



(b) SSD values for PW-SoS with various choices for parameter ϵ based on Figure 9.

Figure 18 Comparing Strength of Schedule Discrepancy (SSD) values across 14 NBA seasons.

including Relative Percentage Index (RPI) as defined in NBAstuffer (2023), and the same conclusion stands.

F.4. Suggestions for the 2019–20 Season

In this section, we present the results of our two-phase analytics approach applied to the 2019–20 NBA regular season which was suspended on March 11, 2020 due to the COVID-19 pandemic. We consider 74 games per team as the target length of the shortened season, thus canceling 8 games per team from the remainder of the season. As a result, out of 259 remaining games, we select 139 games to be played in the shortened season using the PW-FW model. The set of selected and canceled games are shown in Figure 19.

According to the NBA's resumption plan announced on June 26, 2020 (NBA 2020), 22 teams (the top 13 teams from the west and top 9 teams from the east) were invited to Orlando, Florida to play 8 more games each to conclude the 2019–20 NBA regular season. In effect, the invited teams will have a shortened season ranging from 71 to 75 games in total, and this variation is a consequence of some teams having played a few more games than others as of the suspension date. We compare the shortened season plan shown in Figure 19 to the NBA's resumption plan. Using the same Monte Carlo simulation approach described in §6.3.2, we evaluated our proposed solution and the NBA's return plan using 1,000 scenarios and computed the average concordance relative to the full season ranking. The concordance for our proposed solution is 404.7, while it is 392.25 for the NBA's resumption plan. As a result, on average, our proposed solution predicts the relative positioning of at least 12 additional pairs of teams correctly, compared the the NBA's return plan.

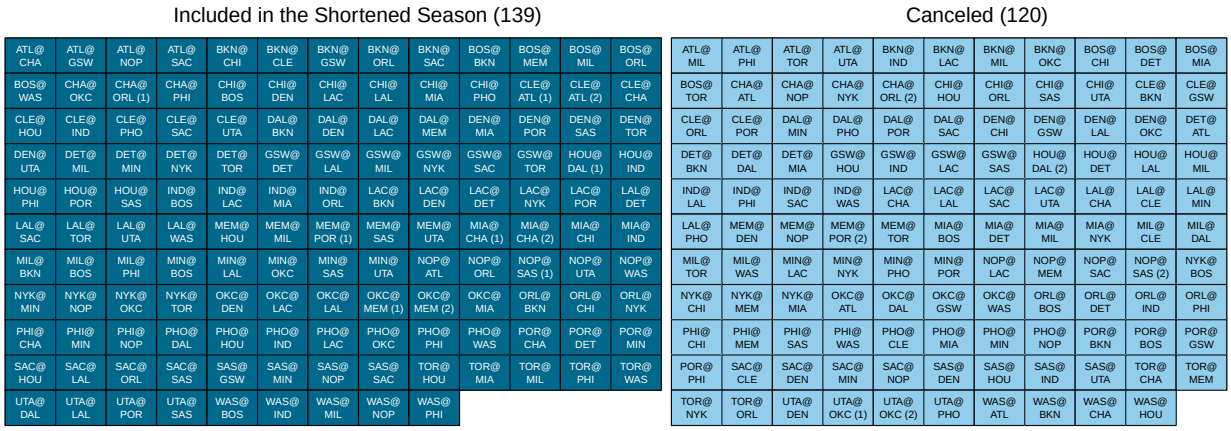


Figure 19 **Selected/canceled games for the remainder of the 2019–20 season according to PW-FW.**

References

Diaconis P, Graham RL (1977) Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39(2):262–268.

Durbin J, Stuart A (1951) Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society: Series B (Methodological)* 13(2):303–309.

Kleywegt AJ, et al. (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502.

Manley M (1986) Efficiency Rating in basketball . Retrieved on March 10, 2022. `martin-manley.eprci.com/sports/efficiency-rating`.

NBA (2020) NBA Articles. Retrieved on June 26, 2020. `nba.com/article/2020/06/26/nba-comeback-schedule-2019-20-seeding-games`.

NBAstuffer (2023) Strength of schedule (sos) Retrieved on October 1, 2023. `https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/`.

Nesterov Y (2018) Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming* 171(1):311–330.

Parikh N, Boyd S (2014) Proximal algorithms. *Foundations and Trends® in Optimization* 1(3):127–239.

Yannakakis M (1985) On a class of totally unimodular matrices. *Mathematics of Operations Research* 10(2):280–304.