Annals of Operations Research (2024) 335:33–58
https://doi.org/10.1007/s10479-023-05725-4

ORIGINAL RESEARCH

A data-driven optimization approach to baseball roster management

Sean Barnes[1] · Margrét Bjarnadóttir[2] · Daniel Smolyak[3] · Aurélie Thiele[4]

Received: 31 July 2021 / Accepted: 22 August 2023 / Published online: 15 January 2024

Abstract
Each year, major league baseball (MLB) teams face complex decisions about which players to retain and which players to recruit. In addition to operational, team and budget constraints, these decisions are further complicated by the fact that an athlete's future performance and its impact on the team are both uncertain. In this paper, we combine prediction modeling with decision optimization to study the MLB free agent market. We develop optimization models for the allocation of a team's recruitment budget using six different metrics that evaluate a player's contributions to a team's success. We consider both an ideal case, where each team can choose among all free agents, and a sequential case, where we assume that teams with stronger appeal (big market) are more successful in attracting talent, while teams with less pull must optimize their rosters over a much smaller pool of remaining players. Using the best-performing metric, which takes into account both players' positions and their positional flexibility, we develop a series of quantitative tools that help teams, especially those with small budgets, identify (1) the players who deliver a key competitive advantage to their teams, appearing in both their ideal and sequential rosters and (2) the players who are in many ideal rosters and thus are likely to be hired by teams with big budgets, perhaps at a substantial salary premium. In order to gain and maintain an edge in the fiercely competitive free agent market, teams need to continuously adapt their strategies, and our models represent a first step towards prescriptive (not just predictive) analytics designed to help them do so. Further, our analysis indicates that a few players are in high demand from many teams (for instance, in every year of the period considered, the ten most in-demand players appear in

✉ Margrét Bjarnadóttir
mbjarnad@umd.edu

Sean Barnes
seanlbarnes@gmail.com

Daniel Smolyak
dsmolyak@umd.edu

Aurélie Thiele
athiele@smu.edu

1 Netflix, Los Angeles, CA, USA

2 Robert H. Smith School of Business, University of Maryland College Park, College Park, MD, USA

3 Computer Science Department, University of Maryland College Park, College Park, MD, USA

4 Engineering Management, Information and Systems, Southern Methodist University, Dallas, TX, USA

the ideal rosters of at least seven teams), while most players appear in one ideal roster or none at all. Our models go beyond players' individual performance metrics to help teams understand which players will be in high demand due to teams' position needs in a given year. The results further emphasize the increasing importance of contract extensions as a strategy to bypass the free agent market.

1 Introduction

Each year, major league baseball (MLB) teams face decisions about which players to retain, release, trade, or recruit and sign. These decisions are complex, as general managers are typically constrained by budget considerations and only have access to a limited market of free agents and minor league players. These decisions must also address specific positional needs for the team's roster, especially those positions that are not likely to be addressed via other mechanisms (e.g., trades, waivers, player promotion from the minor leagues). In choosing players, a team's short-term needs must be balanced with long-term objectives, payroll management, positional flexibility, and the current pipeline of talent in the minor leagues. These roster decisions are further complicated by the fact that any athlete's future performance is highly uncertain; therefore, long-term investments in individual players carry significant risk.

With the introduction of the free agent market in 1976, players gained the freedom to pursue new contracts with any team (Raimondo, 1983). Prior to this change, players had no negotiating power over their salary and team owners could retain players for their entire careers. The free agent market has become the key mechanism through which teams can acquire new players, oftentimes without sacrificing any of their current players under contract (as in a trade). All major league players with at least six years of service are eligible for free agency when their current contract expires, as are players with fewer than six years of service who are not tendered a new contract. In addition to recruiting from the free agent market, teams can also sign players from foreign markets. These contracts are less common and special rules apply in some cases. For example, professional players with at least 10 years of service in the Japanese major leagues are eligible for free agency even with fewer than six years of MLB service.

Since 2006, total spending on the free agent market has exceeded $1 billion each year, with the exception of 2009 ($847 million) (https://www.spotrac.com/mlb/tools/offseason/). Spending on the free agent market is often dominated by a small number of superstar players (i.e., so-called "mega contracts") and large market teams in New York, Los Angeles, and Chicago, with the New York Yankees spending a record $487 million in 2013 with the signings of Masahiro Tanaka, Jacoby Ellsbury, Brian McCann, and Carlos Beltran (https://www.spotrac.com/mlb/tools/offseason/). Since 1999, there have been 97 contracts equal to or in excess of $100 million (https://legacy.baseballprospectus.com/compensation/cots/league-info/highest-paid-players/). In this paper we study the free agent market through a prescriptive lens with the goal of gaining insights that can support teams in gaining a competitive edge in the free agent market.

2 Background

2.1 From sabermetrics to modern analytics

Statistical analysis has a long history in baseball, and basic statistics have been collected since the beginning of organized play. The increasingly widespread availability of this data has fueled the interests of managers, players, academics and baseball enthusiasts alike, and the rise of sabermetrics—named after the Society for American Baseball Research, founded in 1971—has played a major role in ongoing efforts to better quantify the performance of baseball players.

The most basic statistics are simple counts of significant events for each player, such as at bats (AB), hits (H), home runs (HR), and runs batted in (RBI) for batters; stolen bases (SB) and times caught stealing (CS) for base runners; innings pitched (IP), earned runs allowed (ER), strikeouts (K), and wins (W) for pitchers; and putouts (PO), assists (A), and errors (E) for defensive players. In addition to these simple count statistics, more advanced statistics attempt to capture players' efficiency (i.e., how likely they are to generate a specific outcome in a single unit of play such as an AB or IP). These statistics facilitate comparison between batters with a different number of at bats or pitchers with a different number of innings pitched. For example, on base percentage (OBP) captures the proportion of plate appearances in which a batter reaches base (either via a hit, base on balls, or hit by a pitch), whereas earned run average (ERA) captures the average number of earned runs allowed by a pitcher per inning pitched. More recently, more robust performance statistics such as wins above replacement (WAR) and win probability added (WPA) have been developed as a means to capture the overall contribution of a player to the performance of the team. Tymkovich (2012) provides a good overview of WAR and other overall player performance metrics, and Koop analyzes the relative performance of players, measured as a fraction of the best player, using an output aggregator instead of considering absolute performance metrics (Koop, 2002).

The use of analytics to improve teams' performance attracted significant attention after the Oakland Athletics' 2002 season, which was made famous by the book *Moneyball*, published in 2004, Lewis (2004) and the subsequent motion picture. Nowadays, most MLB teams employ an analytics department that provides at least a minimal level of decision support for personnel decisions and in-game strategy. Baumer and Zimbalist (2014) assess the growth of analytics in baseball up to 2014.

A large body of existing literature—including journal articles, books, and various forms of media—analyzes of how player statistics, demographics, and other aspects of the game translate into individual and team performance. These studies range from the effect of managers on individual and team performance (Kahn, 1993) to the effects of racial diversity on team performance (Timmerman, 2000) to testing umpire biases on their game calls (Kim & King, 2014). Since our goal is to use analytical methods to optimize free agent selection in a competitive market, we restrict our attention below to the literature most related to estimating player and team performance and supporting personnel decisions.

2.2 Labor contracts in the MLB

A significant stream of literature studies labor contracts in MLB, and a number of articles have focused on the free agent market in particular. Around the time the free agent market was introduced, studies focused on estimating the marginal value of players and on estimating the economic loss of players due to contract restrictions (Scully, 1974). For instance, Krautmann

(1990) investigated player performance after signing long-term contracts, debunking the myth that player performance declines after signing long-term contracts, while Turvey (Turvey, 2013) explored the benefits and risks of long-term contracts for younger players. A recent study focuses on the magnitude of exploitation for MLB players, showing that the level of exploitation is much lower for free agents than for rookies and arbitration-eligible players (i.e., between three and six years of MLB service) (Humphreys & Pyun, 2017). Krautmann (2016) focused on contract extensions and their impact on the free agent market, finding that the increasing use of contract extensions (i.e., signing pre-free-agency-eligible players to multi-year contracts that extend beyond the first eligible year of free agency) correspondingly reduces the size of the free agent market. Gross and Link (2017) analyzed the use of options in free agent contracts, that is, the right of a player or team to extend the contract beyond its guaranteed term, concluding that the use and pricing of these options are consistent with option theory.

2.3 Player and team evaluations

A player's performance evolves throughout his career in the major leagues. Schultz and Curnow (1988) estimated that both pitchers and non-pitchers peak in their late twenties (with a few exceptions; for instance, the ability to steal bases requires sprinting and peaks earlier). By comparison, the median age of rookie position players (i.e., non-pitchers) is 23 years and of rookie pitchers is 27 years, which suggests that the timing of a player's first free agent contract is around the same time as their peak or is sometime during their decline. Aging curves were introduced by Bill James as a representation of how a player's ability increases, peaks, and then declines. James later introduced the idea of comparing current players to past players to estimate relative performance (Bendtsen, 2017). Combining these ideas with a nearest neighbor approach, Silver (2012) developed the PECOTA system to predict player performance in upcoming seasons. In a more recent article, Bendtsen (2017) applied a probabilistic graphical model to analyze performance regimes over a player's career. His finding shows that players' careers transition through different regimes and that players may revisit previous performance states.

It is common to use player statistics to characterize performance, although there is some debate over whether the statistics commonly used are up to the task. For example, Albert (2006) analyzed batting average and concluded that it is a relatively poor measure of a batter's true abilities, suggesting other more efficient statistics. Similarly, in an earlier paper, Schall and Smith (2000) argued that because baseball statistics capture true abilities only imperfectly, player performance tends to regress to the mean, and that prediction models for future performance can be improved by incorporating this fact. Yet despite the shortcomings of the available statistics, multiple efforts have successfully focused on predicting the future performance of MLB players (see Barnes and Bjarnadóttir, 2016; Koseler and Stephan, 2017 and citations therein).

Researchers have also analyzed how player performance translates into compensation. Several efforts have focused on how on-base percentage and slugging percentage align with player salaries (Farrar & Bruggink, 2011; Hakes & Sauer, 2006; Lewis, 2004), inspired in part by *Moneyball*. Rockerbie (2009) developed a more robust regression model of free agent salaries based on auction theory. Barnes and Bjarnadóttir (2016) developed multiple data-driven models of free agent salaries and future performance to identify over- and underpaid free agents. Relatedly, Lackritz (1990) developed models to estimate a player's value to his club by measuring the impact of individual players' statistics on the revenues generated by

the clubs, and Elitzur (2020) analyzes whether teams and general managers who rely on analytics have enjoyed any pay-performance advantage.

The performance of a baseball team is not simply a sum of the performances or abilities of individual players, but rather a complex combination of these skills. Humphrey et al. (2009) argued that the roles of pitchers and catchers are more important than other roles on the team. Chan and Fearing (2013) applied a robust optimization approach to team selection, estimating that positional flexibility contributed 3–15% of overall team performance. Chan and Fearing (2019) subsequently formalized the concept of process flexibility in sports analytics to protect against injury risk and player unavailability. A key insight is that teams whose performance is driven by one or two star players are over four times as fragile as the most robust teams, while top teams can attribute at least one to two wins per season to flexibility alone.

Although few additional studies explore how individual player statistics and other factors translate into team performance and overall wins, studies of compensation and of racial diversity are the exception. A 2002 study (Hall et al., 2002) analyzed the correlation between payroll and team performance from 1980 to 2000 and found that the correlation intensified during the 1990s. A 2003 study confirmed these findings, investigating the relationship between team payroll and overall team performance from 1985 through 2002 (Wiseman & Chatterjee, 2003). The authors found that during this period the relationship between payroll and performance intensified, but teams also performed better when the payroll was more equally distributed across the team. Further studies have validated the negative effects of wage disparity on team performance (Depken, 2000; Frick et al., 2003; DeBrock et al., 2004), citing factors such as perceived fairness, even distribution of talent, and disproportional costs of star talent relative to the marginal contributions to team performance as potential explanations for these effects. Additionally, studies have found that racial diversity does not affect baseball teams' overall performance (Timmerman, 2000).

The literature to date has focused on insights gained by predictive rather than prescriptive models. Our goal in this paper is to investigate whether prescriptive analytics can provide teams with a substantial competitive advantage through the use of optimization models. To this end, we develop multiple team performance models to understand the contribution of individual players to team success. Building on these empirical investigations and the performance of our predictive models, we develop optimization models to allocate the team's recruiting budget for free agent players. We consider both an ideal case, where each team can choose from all the free agents, and a sequential case, where we assume that teams with stronger appeal (big market) are more successful in attracting talent and teams with less pull must optimize their rosters over a much smaller pool of remaining players. In any given year, our models allow general managers to identify the positions for which competition in the free agent market is fierce, the players who are likely to be sought after by many teams, possibly leading to salary bidding wars, and the good players who are less likely to be in high demand. These models indicate that teams need to continuously adapt their strategies in order to gain an edge in the market. Our results further emphasize the increasing importance of contract extensions as a strategy to bypass the free agent market.

The rest of the paper is organized as follows. In Sect. 3, we lay out our approach to modeling team performance, describing the various metrics we consider, and formulate the optimization problems. In Sect. 4, we solve the optimization models for the data sets considered and discuss the outputs in detail, including the insights that can generate competitive advantage for a team. Section 5 contains concluding remarks.

3 Modeling approach

In order to formulate and solve the player selection problem as an optimization problem, we need to estimate each player's contribution to their team's success; however, different teams may define success differently. A large market team (for example the East Coast rivals, the New York Yankees and the Boston Red Sox) may define success as winning the World Series, while a smaller team's goal may be to make the playoffs. Our modeling approach maximizes the expected number of wins during the regular season: if high enough, this number secures a place in the playoffs, and it is further correlated with success in the playoffs.

We apply a two-step approach. First we build an empirical model that links team composition, i.e., player rosters, to team success using predictive analytics (Sect. 3.1). Next, we introduce two prescriptive analytics setups, one focused on identifying players who would maximize a team's success if the team could sign any player (what we call the ideal model) and the other focused on rosters that take into account the teams' different degrees of market power (what we call the sequential model). These models are described in more detail below. In each case we use the empirical models of teams' regular seasons wins to formulate an optimization model maximizing their chances of success (Sect. 3.2).

The data was obtained from baseball-reference.com for the years 1952–2014 and included all team rosters and information on over 6000 players. The detailed player data contains basic and advanced statistics, although in our final models we primarily focused on WAR and WPA. In addition we collected information on age, positions, and salaries. Free agent contract information was also collected from Baseball-Reference. When the contract information was missing we attempted to manually collect the missing information. Finally, the data collection included team-specific information for each season, including the win-loss record, playoff results, payroll and attendance. All empirical modeling was done using Python and R. Optimization analysis was conducted using the AMPL software and the CPLEX solver.

3.1 Modeling teams' performance

We connect the expected number of wins to team composition through various predictive models. In our experiments, machine learning models such as trees and forests did not outperform regression models. We therefore report on six increasingly sophisticated performance models, all estimated with OLS, connecting individual players' performance to a team's expected number of wins. (In "Appendix B" we discuss the observed heteroskedasticity of these models.) Below, we summarize these models in order of increasing complexity. We distinguish between two main groups of models: *homogeneous models*, which are not position-specific, i.e., the impact of a specific player's predicted performance on the objective is independent of his position on the team, and *heterogeneous models*, which are position-specific.

We further introduce positional flexibility into the models, as having players who can play multiple positions increases the rosters' robustness (e.g., with respect to injury) and therefore the chances of success. We define positional flexibility using the information provided on baseball-reference.com, which lists the positions a player has held in decreasing order of number of games played. For instance, a listing of "3/5D" indicates the player's main position is first base (3) but he has also played games—although fewer than ten—as third baseman (5) and designated hitter (D). In some cases, a player has multiple main positions: for instance, in 2009 Adam Dunn had positions 379/DH, meaning his main positions were first base (3), left field (7) and right field (9), but he had also played as designated hitter (DH). The narrow

definition of positional flexibility only uses the numbers of the positions most frequently played (the ones provided by baseball-reference.com before the "/" sign) and the broad definition uses the numbers of all positions played.

The first two models are provided as a benchmark and only use the traditional WPA and WAR metrics. The next four models are the predictive analytics models we developed. In all cases, we define a team by its *extended* roster of 40 players. Importantly, for all models, we use each player's predicted WPA and WAR for the upcoming season, as the aim of these models is to predict next year's performance and the actual realized performance is unknown. We summarize the models below and provide the full regression results in Tables 1 and 2.

1. WPA-only model (**Model 1**) where performance is measured by WPA only, with an R^2 of 0.378:

$$\text{Wins} = 76.5864 + 1.0386 \cdot \text{Predicted Total WPA}$$

2. WAR-only model (**Model 2**) with an R^2 of 0.360:

$$\text{Wins} = 60.0391 + 0.7415 \cdot \text{Predicted Total WAR}$$

3. Homogeneous (position-independent) model (**Model 3**) where performance is measured by a linear combination of WPA and WAR independently of players' positions, with an R^2 of 0.391:

$$\text{Wins} = 68.779 + 0.3311 \cdot \text{Predicted Total WAR} + 0.6879 \cdot \text{Predicted Total WPA}$$

4. Heterogeneous (position-dependent) model with narrow definition of positional flexibility (**Model 4**), with an R^2 of 0.428:

$$\begin{aligned}\text{Wins} = {} & 83.4818 + 0.6701 \cdot \text{PP WAR} + 0.6572 \cdot \text{RP WAR} + 0.6386 \cdot \text{CL WAR} \\ & +0.3607 \cdot \text{PP WPA} + 0.9796 \cdot \text{SP WPA} + 0.4754 \cdot \text{RP WPA} \\ & -0.6189 \cdot \text{Number Position Players} - 0.6894 \cdot \text{Number Relief Pitchers} - 0.9048 \\ & \cdot \text{Positional Flexibility}\end{aligned}$$

where positional flexibility is the number of positions that players have played significantly during the season

5. Heterogeneous (position-dependent) model with broad definition of positional flexibility (**Model 5**), with an R^2 of 0.429:

$$\begin{aligned}\text{Wins} = {} & 84.4667 + 0.6767 \cdot \text{PP WAR} + 0.6620 \cdot \text{RP WAR} + 0.6426 \cdot \text{CL WAR} \\ & +0.3459 \cdot \text{PP WPA} + 0.9754 \cdot \text{SP WPA} + 0.4846 \cdot \text{RP WPA} \\ & -0.6235 \cdot \text{Number Position Players} - 0.6901 \cdot \text{Number Relief Pitchers} - 1.5352 \\ & \cdot \text{Positional Flexibility}\end{aligned}$$

6. Heterogeneous (position-dependent) model with nonnegative coefficients (**Model 6**), with an R^2 of 0.396:

$$\begin{aligned}\text{Wins} = {} & 70.8183 + 0.417005 \cdot \text{PP WAR} + 0.551917 \cdot \text{CL WAR} \\ & +0.677056 \cdot \text{PP WPA} + 1.094199 \cdot \text{SP WPA} + 0.645073 \cdot \text{RP WPA}\end{aligned}$$

This model is the same for the broad and narrow definitions of positional flexibility, which both have a zero coefficient. The motivation for this model is that we may not want to assign negative coefficients to certain positions such as relief pitchers, because they may come into the game when the team is already losing the game, rather than causing the team to lose.

Table 1 Regression models − all coefficients $p < 0.001$

	Model 1	Model 2	Model 3
Const	76.5864	60.0391	68.779
	(8.304)	(8.479)	(8.368)
Predicted Total WPA	1.0386		0.6879
	(0.054)		(0.089)
Predicted Total WAR		0.7415	0.3311
		(0.041)	(0.067)
N	1234	1234	1234
R^2	0.379	0.361	0.391
Adj. R^2	0.378	0.360	0.390
MAE	7.983	7.827	7.785

Mean absolute error (MAE) is estimated with 5-fold cross validation

Table 2 Regression models ($^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$)

	Model 4	Model 5	Model 6
Const	83.4818***	84.4667***	70.8183***
	(8.588)	(8.637)	(8.928)
PP WAR	0.6701***	0.6767***	0.4170***
	(0.086)	(0.086)	(0.088)
RP WAR	0.6572**	0.6620**	
	(0.328)	(0.328)	
CL WAR	0.6386*	0.6426*	0.5519
	(0.387)	(0.387)	(0.398)
PP WPA	0.3607***	0.3459***	0.6771***
	(0.119)	(0.119)	(0.122)
SP WPA	0.9796***	0.9754***	1.0942***
	(0.133)	(0.133)	(0.137)
RP WPA	0.4754*	0.4846*	0.6451**
	(0.278)	(0.278)	(0.285)
# Position Players	−0.6189***	−0.6235***	
	(0.122)	(0.122)	
# Relief Pitchers	−0.6894***	−0.6901***	
	(0.135)	(0.135)	
Positional Flexibility	−0.9048	−1.5352**	
	(0.587)	(0.760)	
N	1234	1234	1234
R^2	0.428	0.429	0.396
Adj. R^2	0.423	0.424	0.391
MAE	7.622	7.613	7.676

MAE is estimated with 5-fold cross validation

For position-dependent Models 4 and 5, we observe that the WAR of the position players (PP) contributes more to the predicted number of wins than WAR of relief pitchers or closing pitchers, but the WPA of the starting pitchers and relief pitchers contributes more than the WPA of the position players. For position-dependent Model 6, the WAR of the closing pitchers contributes more than that of the position players, and the WPA of the starting pitchers contributes more to the predicted number of wins than that of the position players and relief pitchers. In essence, our approach allows the prediction models to emphasize the WAR or the WPA metric depending on players' positions.

Models 4 and 5 notably have negative coefficients for the non-WAR and non-WPA variables. The negative coefficients for the number of position players and the number of relief pitchers is consistent with the hypothesis that teams need to bring in and experiment with more position players and relief pitchers when their starters are injured or otherwise unable to play. When losing their preferred starter, teams may add two to three players per position to their roster to determine which player is the best fit for the team. The negative coefficient for positional flexibility is likely due to the fact that stronger players are kept in fewer positions, both to maximize their contribution to the team and in order to build their consistency in those positions - thus the fewer strong players a team has, the more they are likely to shuffle players between positions, linking greater positional flexibility to weaker teams.

While of course none of the models provide perfect predictions, they capture to varying degrees the variability in a team's success, reflecting the high degree of uncertainty in baseball outcomes. The extensive numerical analysis we perform in Sect. 4 will allow us to determine whether the use of position-dependent models leads to meaningful improvement in player selection and predicted number of wins.

3.2 Optimization formulation

We consider two cases in our formulation:

Ideal case: Each team is able to select its preferred players, i.e., each team proceeds first in the player selection process. This is useful when understanding which players would maximize the team's expected performance. We note this scenario with i, indicating the ideal case.

Sequential case: Teams select players sequentially in decreasing order of their player signing power, which we assume is in the same order as their payroll for that same season. In this setting, once a team has selected a player, that player is no longer available to the other teams that follow. We note this scenario with s, indicating the sequential case.

3.2.1 The decision variables

Let I be the set of free agent players on the market, who are available to sign a new contract with a major league team during the season considered. Further, let J be the set of possible metrics $\{1, \ldots, 6\}$ (as calculated by Models 1–6), K be the set of teams and P be the set of positions. In this setting we define two groups of decision variables for each case $c \in \{i, s\}$, metric $j \in J$, and team $k \in K$.

1. The *main decision variables* are binary variables x_i^{cjk} where $i \in I$, with $x_i^{cjk} = 1$ if player i is selected as part of the extended roster for case $c \in \{i, s\}$, metric j and team k and 0 otherwise.
2. The *auxiliary decision variables* are integer slack variables y_p^{cjk} where $p \in P$ and $c \in \{i, s\}$, which denote the number of players of a position p who must be brought

up from the minor leagues under metric j for team k in order for the team to satisfy all its position requirements. Such variables enable teams to only recruit a couple of strong (and expensive) players in the ideal case, and they are required in the sequential case to prevent the lowest ranking teams from being unable to recruit players given their position requirements and their budgets. For example, the optimization problem may have recommended that better ranked teams hire more affordable players, leaving the lower-ranked teams with expensive players they cannot afford, or in some cases all free agents in a certain position may have been recruited.

For notational clarity, in what follows we drop the indices referring to the case (ideal or sequential) and the team from the mathematical formulations.

3.2.2 The objective and constraints for the ideal case

We now translate each of our regression models into performance metrics. The decision maker seeks to maximize the total metric, which is the sum of the individual metrics of the players selected to the team. Let Met_i be the contribution of player i to each of the metrics considered.

In Metrics 1–3 and 6, the objective is to maximize

$$\sum_{i \in I} Met_i x_i$$

for the appropriate metric. Let a_{ip} be a binary parameter that is equal to 1 if player i plays in position p and 0 otherwise.

Metric 4 maximizes

$$\begin{aligned} &\textstyle\sum_{i \in I} Met_i\, x_i - 0.6189 \cdot \sum_{i \in I} (1 - (a_{i,RP} + a_{i,SP} + a_{i,CL}))x_i \\ &-0.6894 \cdot \textstyle\sum_{i \in I} a_{i,RP}\, x_i - 0.9048 \sum_{i \in I} aflex_i^1\, x_i, \end{aligned}$$

where $aflex_i^1$ refers to the narrow definition of positional flexibility, i.e., the number of positions in which the player has played a substantial number of games.

Then, Metric 5 maximizes

$$\begin{aligned} &\textstyle\sum_{i \in I} Met_i\, x_i - 0.6189 \cdot \sum_{i \in I} (1 - (a_{i,RP} + a_{i,SP} + a_{i,CL}))x_i \\ &-0.6894 \cdot \textstyle\sum_{i \in I} a_{i,RP}\, x_i - 0.9048 \sum_{i \in I} aflex_i^2\, x_i, \end{aligned}$$

where $aflex_i^2$ refers to the broad definition of positional flexibility, i.e., the total number of positions in which the player has played any number of games.

In the ideal case, each team is able to sign the players it is most interested in and can afford, subject to the following constraints of the decision maker:

Budget $\sum_{i \in I} sal_i x_i \leq B$, with sal_i the salary of player i and B the budget actually used by the team to secure its roster of new players. This means that the amount of money allocated for acquiring the roster of new players is bounded from above by the amount of money actually spent.

Roster Size $\sum_{i \in I} x_i = ntot$, with $ntot$ the actual number of new players selected by the team that season.

Position Requirements for any position p, $\sum_{i \in \mathcal{P}_p} x_i + y_p \geq r_p$, where $\mathcal{P}_p$ denotes the set of players who can play at position p and r_p represents the number of players at that position who were actually signed by the team that season.

3.2.3 Objective and constraints for the sequential case

The formulation in the sequential case is similar to that in the ideal case, with additional constraints to remove from consideration players already signed by higher-ranked teams. The opportunity to sign additional players from the minor leagues ensures the feasibility of the player signing problem; otherwise, feasibility issues may arise when the teams that sign players last face a substantially reduced pool of players. In this case, the decision maker determines how many untested, minor league players in each position to sign at minimum salary in order to meet the roster requirements. Reflecting the lack of information about players from the lower leagues, those players have WPA and WAR metric predictions of zero and we assume that their actual metrics, if they had played, would have also been zero. We seek to sign as many players from the existing pool as possible; hence, we penalize the use of minor league players through the use of a Big-M penalty coefficient. This is because we assume an existing major league player, even with currently negative performance metrics (which may be due to past injury or team dynamics), would bring benefits to a new team such as experience or name recognition that are not captured in the WAR and WPA performance metrics.

The teams' optimization problems are solved in the decreasing order of their payroll for that year as a proxy of their ability to sign the players they want. In Metrics 1–3 and 6, the objective is to maximize $\sum_{i\in I} Met_i x_i - M \sum_{p\in P} y_p$ for the appropriate metric, with M an appropriately large coefficient. Metrics 4 and 5 maximize $\sum_{i\in I} Met_i\, x_i - M \sum_{p\in P} y_p - 0.6189 \cdot \sum_{i\in I}(1-(a_{i,RP}+a_{i,SP}+a_{i,CL}))x_i - 0.6894 \cdot \sum_{i\in I} a_{i,RP}\, x_i - 0.9048 \sum_{i\in I} aflex_i^1\, x_i$ and $\sum_{i\in I} Met_i\, x_i - M \sum_{p\in P} y_p - 0.6189 \cdot \sum_{i\in I}(1-(a_{i,RP}+a_{i,SP}+a_{i,CL}))x_i - 0.6894 \cdot \sum_{i\in I} a_{i,RP}\, x_i - 0.9048 \sum_{i\in I} aflex_i^2\, x_i$, respectively, reflecting the different definitions of positional flexibility.

In the sequential case, the constraints become:

Budget $\sum_{i\in I} sal_i x_i + S \sum_{p\in P} y_j \le B$, with sal_i the salary of player i, S the minimum salary for that year and B the budget actually used by the team to secure its roster of new players. This means that the amount of money allocated for acquiring the roster of new players is bounded from above by the amount of money actually spent.

Roster Size $\sum_{i\in I} x_i + \sum_{p\in P} y_p = ntot$, with $ntot$ the actual number of new players selected by the team that season.

Position Requirements for any position p, $\sum_{i\in\mathcal{P}_p} x_i + y_p \ge r_j$, where $\mathcal{P}_p$ denotes the set of players who can play at position p and r_p represents the number of players at that position who were actually signed by the team that season.

Player Availability for all $i \in I$, $x_i \le u_i$ where u_i is a binary parameter equal to 1 if player i remains available in the pool when the team solves its optimization problem and 0 if he has already been signed by another team. The u_is are updated after each team's optimization problem has been solved, based on the optimal solution for that team.

4 Numerical results

For each year, we solve the roster optimization problems for each team, in decreasing order of each team's signing power. In the ideal case, the solution is not affected by the order in which

we solve the roster optimization problems. In the sequential case, once we have solved the problem for each team considered and for each objective function (each performance metric), we remove from the pool all the players selected by that solution. When a team's name changed during the time period considered (for instance, the Florida Marlins became the Miami Marlins in 2012), we provide all results using the most recent name. All salaries are inflation adjusted to 2014.

4.1 Model evaluation

The performance of the optimized teams is not available, as teams' real-world recruitment differs from our optimized solutions. However, in contrast to a sport like basketball were players' performance is highly dependent on the team, baseball has a sequential discrete event style of play in which players' performance is less dependent on their teammates. We can therefore compare and contrast the proposed rosters with the actual rosters using a regression model that regresses players' realized WAR and WPA on the number of team wins:

$$\text{Actual Wins} = 74.1923 + 5.6169 \cdot \text{Average WAR} + 33.8945 \cdot \text{Average WPA} \quad (1)$$

This linear regression model, which we refer to as the predicted number of actual wins, has an R^2 of 0.958. We note that to create this model, we use players' realized performance, in stark contrast to our metrics, which are based on predicted future performance. We further note that this model's granularity is at the team level, meaning that it focuses on each team and does not enforce the league-level fact that there must be 2,430 wins per season (because there are 30 teams and 162 games, in which only one team out of two wins, i.e., $30 \cdot 162/2 = 2430$).

The goal of our numerical experiments is to investigate the benefits of a quantitative approach based not only on predictive analytics but also on prescriptive analytics to optimize team rosters. These experiments particularly highlight the ability of mathematical models to zero in on the depth or shallowness of the player pool for a certain position in a given year, which significantly influences recruitment outcomes: for instance, a shallow pool means that certain players will be in high demand due to the market conditions rather than their own performance metrics.

We run the ideal and sequential optimization models described in Section 2.2 for each of the five years of the 2009–2013 time period. After solving the models, we first output, for each metric in each case for each team, the optimal player selections and the predicted number of actual wins given in Eq. (1). This allows us to compare the different performance metrics to the roster of actual players. We further study what drives the most extreme (best or worst) divergence in performance between our optimized model and the actual roster. Second, for each metric and each player we output the percentage of teams that select that player in the ideal case, which quantifies how in-demand each player is on a given year, and which may also serve as proxy for that player's negotiating power and the upward pressure on his salary as various teams attempt to sign him. Finally, for each metric for each team we study the proportion of players from the ideal roster who are also selected in the sequential roster. This helps us better understand for which teams the ideal solution can serve as a guide.

4.2 Results of optimization models

We first compare the predicted number of actual wins of the actual roster to that of the ideal roster suggested for each team. Specifically, to select the best metric, we compute the

Table 3 Comparison of metrics for the optimization model. For each metric in each year, the table displays the counts of the number of times that a roster had the highest predicted actual wins

Case	Metric	1	2	3	4	5	6
Ideal	2009	6	8	8	11	12	4
	2010	11	2	8	3	9	2
	2011	5	4	9	7	2	11
	2012	10	7	5	3	5	6
	2013	14	2	4	5	8	3
	All Years	**46**	23	34	29	36	26
Sequential	2009	5	6	5	5	10	4
	2010	4	7	6	3	8	4
	2011	7	3	4	6	6	4
	2012	4	7	3	7	7	3
	2013	3	6	5	8	7	2
	All Years	23	29	23	29	**38**	17
Combined	Total	69	52	57	58	**74**	43

number of times a given metric achieves the maximum predicted number of actual wins for each year. The results are presented in Table 3. The best metric for each case (ideal, sequential or combined) is highlighted in bold. Both for the sequential case and the sum of both ideal and sequential cases, Metric 5 (position-dependent model with a broad definition of positional flexibility) emerges as the best. For the ideal case, Metric 1 (WPA only) is best. Because the sequential case is a more realistic approximation of the real-life decision-making process, while the ideal case where each team selects its players first cannot be implemented in practice, we select Metric 5. The worst two metrics are Metric 2 (WAR only) and Metric 6 (position-dependent model with non-negative coefficients).

We then compare the performance of each team for each year using our best metric, Metric 5 *in the sequential case*, to the performance of the actual roster, as measured by the predicted number of actual wins in Eq. (1). We compute the number of years, out of five, in which our analytical approach would lead to a higher number of wins for that team. The results are provided in Table 4, with the teams that would have obtained more wins three, four or five years using our approach highlighted in bold. Chicago White Sox (CHW), Houston Astros (HOU), Milwaukee Brewers (MIL), New York Mets (NYM), Pittsburgh Pirates (PIT) and Washington Nationals (WSN) would have won more games in four years out of five. Interestingly, as we analyze the highlighted teams they are large, medium and small cap teams. For example, St Louis Cardinals (STL) ranked tenth in terms of payroll in 2013. Note that it is not possible to increase the number of wins for all teams, since a team must lose its game for another team to win.

Table 5 shows the difference, per year and team, in games won by our optimized roster (again in the sequential case) versus the actual roster, as well as the average difference over all five years for each team. We also show the average rank of each team in the sequential case. A positive number means our optimized roster would have won more games. It is again important to note that in the sequential case, each player can only be selected once, and as a result in the approach some teams will be winners and others will not benefit. The 16 teams with positive average difference are shown in bold in Table 5. Washington Nationals (WSN) would have won an average of 3.78 more games per year, Atlanda Braves (ATL) 2.58 and New York Mets (NYM) 1.21. The teams that would have suffered the worst

Table 4 Number of years, out of five, in which our optimized team outperforms the actual roster, in the sequential case

Team	Nb	Team	Nb
ARI	2	**MIL**	4
ATL	3	MIN	1
BAL	2	**NYM**	4
BOS	3	**NYY**	3
CHC	2	**OAK**	3
CHW	4	PHI	2
CIN	3	**PIT**	4
CLE	2	SDP	2
COL	3	**SEA**	3
DET	2	SFG	2
HOU	4	STL	1
KCR	2	TBR	2
LAA	2	TEX	2
LAD	2	TOR	2
MIA	2	**WSN**	4

performance degradation on average in 2009–2013 when using our analytical method are Arizona Diamondbacks (ARI), San Francisco Giants (SFG) and St Louis Cardinals (STL), with an average of 2.24, 1.56 and 1.58 more games lost per year, respectively.

Table 5 highlights some important differences in number of wins, and we therefore study the teams and years corresponding to the ten largest positive and negative differences, when the optimized roster is obtained in the sequential case using Metric 5. Our approach far outperforms the actual roster, with the ten teams/years that fare best being Atlanta Braves (ATL) in 2009 and 2013, Boston Red Sox (BOS) in 2011, Chicago White Sox (CHW) in 2011, Kansas City Royals (KCR) in 2009, Milwaukee Brewers (MIL) in 2013, New York Mets (NYM) in 2009, New York Yankees (NYY) in 2011, Washington Nationals (WSN) in 2011 and 2012. The ten teams/years that fare worst under our metric are Arizona Diamondbacks (ARI) in 2012, Baltimore Orioles (BAL) in 2012, Detroit Tigers (DET) in 2011, Houston Astros (HOU) in 2010, Los Angeles Angels (LAA) in 2009, Miami Marlins (MIA) in 2011, Philadelphia Phillies (PHI) in 2009, Seattle Mariners (SEA) in 2013, San Francisco Giants (SFG) in 2009 and 2013. The years with the most extreme differences (best or worst) are 2009 and 2011, which have six out of 20 largest differences, and the year with the fewest large differences was 2010, with one out of 20. The years 2012 and 2013 have, respectively, three and four out of 20.

Below and in the "Appendix A", we consider the teams/years that fared best and worst under our metric. This analysis suggests that the optimization approach is most beneficial when it changes the high-level composition of the roster by moving it away from the very highly paid players in favor of signing more *not as* highly paid players. When the optimization approach does not fundamentally affect the team composition in terms of very highly paid players, highly paid players and base-salary players, the outcome tends to be closer to that of a game of chance in a highly uncertain environment, so the optimization approach has less to contribute in that setting. The optimization approach tends to particularly under-perform the actual roster in cases when there are large differences in the estimates of the different performance metrics.

Table 5 Number of games, per year, for which our optimized team outperforms the actual roster, in the sequential case

Team/Year	2009	2010	2011	2012	2013	Aver. diff	Aver. rank
ARI	0.54	1.76	−3.51	−9.26	−0.74	−2.24	21
ATL	9.46	−1.91	0.18	−0.81	6.00	2.58	15
BAL	3.56	−0.38	2.39	−5.42	−2.67	−0.50	16.2
BOS	1.40	1.25	4.96	−3.70	−1.39	0.50	3.4
CHC	−1.74	4.29	−1.99	2.93	−0.39	0.62	7.8
CHW	−3.11	0.43	4.70	1.26	0.55	0.77	8.4
CIN	0.83	−0.24	−3.20	1.75	1.45	0.12	18.6
CLE	−0.92	1.35	−0.41	3.11	−3.30	−0.04	23.2
COL	1.72	−4.21	−1.49	1.47	2.73	0.04	17.8
DET	2.47	−1.31	−6.18	−2.52	0.81	−1.35	7
HOU	3.72	−5.03	0.58	0.86	3.65	0.75	20.6
KCR	6.65	2.89	−1.41	−0.53	−3.17	0.89	25.6
LAA	−8.64	−1.14	−3.09	4.52	1.20	−1.43	4.6
LAD	−4.76	3.92	−3.58	0.51	−0.13	−0.81	7.4
MIA	−3.04	−1.91	−6.40	1.39	3.56	−1.28	24.2
MIL	0.12	1.19	0.83	−4.21	4.90	0.57	17.4
MIN	−3.40	−0.05	−3.19	−2.82	1.94	−1.50	14.2
NYM	5.29	−2.71	2.25	0.24	0.97	1.21	13
NYY	2.84	2.17	5.92	−4.20	−4.33	0.48	1.2
OAK	0.11	2.92	0.21	−0.18	−4.27	−0.24	24.2
PHI	−6.96	−3.16	−0.87	2.77	2.16	−1.21	3.6
PIT	0.05	0.00	−0.27	2.15	1.06	0.60	25.8
SDP	−0.05	−3.71	2.13	1.92	−0.04	0.05	27.8
SEA	−1.89	2.34	4.52	2.67	−5.96	0.34	18.4
SFG	−5.07	−1.72	2.13	3.44	−6.62	−1.56	9.2
STL	−1.98	−1.09	−3.47	0.08	−1.42	−1.58	11.4
TBR	2.15	−0.82	0.92	−1.37	−4.34	−0.69	26.8
TEX	0.20	−2.92	2.84	−1.83	−0.53	−0.45	14.6
TOR	−0.12	1.84	−0.04	−1.11	3.46	0.81	17.2
WSN	−0.41	4.43	5.02	7.73	2.14	3.78	19.4

Best performance of our metric.

- Atlanta Braves (ATL) in 2009: In the actual roster of seven new players, Derek Lowe is the most expensive new player signed, at an adjusted salary of $16.6m, but he underperformed his prediction with an actual WAR of 0.9 and WPA of −1.3. The second highest-paid player in the actual roster is Kenshin Kawakami at $9.2m. The next highest paid players are paid $2.8m and $1.6m, respectively. All other players are paid less than $1 m, with two players at base salaries. In contrast, the optimized roster does not select very highly paid players, with the highest adjusted salary being $10.0m (Ryan Dempster). This allows the optimized team to sign four highly paid players in the $3.6-$8.0m salary range, who are predicted to perform well.

- New York Mets (NYM) in 2009: Our optimized roster of 13 new players outperformed the actual roster because three of the four highest paid players on the actual team performed poorly. These top three players had salaries in the $10.2-$15.5m range, and the fourth highest paid player had a $2.5m adjusted salary. All other players were at base salaries except one, at $6.7m. Our roster had 11 new players with an adjusted salary above $1 m, but the maximum salary was “only" $11.1m and the second highest was $7.8m. This freed up the team’s budget to recruit players in the $1.1m-$5.5m salary range.
- Boston Red Sox (BOS) in 2011: Both the actual and optimized rosters of 14 new players include Carl Crawford, who is the highest paid player with an adjusted salary of $15.6m. The actual roster has four other players paid above $1 m, and the optimized one has five. The second highest salary in the actual roster is $6.3m, while in the optimized roster it is “only" $4.2m, which supports the signing of a player paid $4.2m (Jesse Crain, with an actual WAR of 2.2 and WPA of 1.9.)
- Chicago White Sox (CHW) in 2011: This is a case where the highest salary of the optimized roster is higher than the actual team’s highest salary ($15.5m vs $12.6m). However, the combined salaries of the top two highest paid players are $25.2m for the actual team and $26.0m for the optimized team. The superior performance of the optimized roster is due to the signing of Carlos Pena, the second highest paid player on the optimized roster, with an actual WAR of 2.5 and WPA of 3.2, and to the underperformance of Adam Dunn (the highest paid player on the actual roster), with an actual WAR of -2.9 and WPA of -2.6.
- New York Yankees (NYY) in 2011: In the actual roster, three players out of 11 receive an adjusted salary over $10 m with a maximum of $15.5m, while in the optimized roster, two players out of 11 receive an adjusted salary over $10 m with a maximum of $14.7m. For the optimal team, this frees up additional funds to sign eight players who are paid over $1 m, while the adjusted salaries of only five players in the actual roster exceed $1 m.

Worst performance of our metric.

- Los Angeles Angels (LAA) in 2009: Both the actual roster and the optimized roster sign five new players. The real-world team signs one player at base salary; regarding the other four, the highest actual salary is $9.4m with the next three in the $3.6-$5.5m range. On the optimized roster, the highest acquired player, Francisco Rodriguez, has a $10.2m salary, followed by a second player with a salary of $4.7m and the next two highest paid players earning $1.1-$1.4m, and the fifth one being paid a base salary. LAA did particularly well in the 2009 season thanks to Bobby Abreu and Juan Rivera, whose performance resulted in WPA above 3 for both players. Meanwhile, while on the optimized roster, Francisco Rodriguez underperformed, with a predicted WPA of 1.32 and an actual WPA of 0.1 at a salary of $10.2m.
- Philadelphia Phillies (PHI) in 2009: PHI signed eight new players in both the optimized case and in real life; there are actually no large differences between the actual and optimized rosters in terms of distribution between top talent, middle range and base players. However, on the actual team, Raul Ibanez significantly surpassed expectations, resulting in a WPA of 3.6, while the highest actual WPA of the players on the optimized roster was 0.1.
- San Francisco Giants (SFG) in 2009: On both the actual and optimized rosters, six of the 11 newly signed players were paid more than $1 m. The highest paid recruited player on both is Edgar Renteria, who was paid $8.9m but performed poorly, with a WPA of -0.9. The best player on the actual team, Jeremy Affeldt, had a WPA of 3.1. The highest WPA

on the optimized roster was 2 (Darren Oliver); this difference in performance was further exacerbated by the very poor performance of Braden Looper, who had the optimized team's worst WPA of −1.8, while the worst WPA on the actual team was only −0.9.

- Detroit Tigers (DET) in 2011: In this team's scenario, the optimization approach only uses \$30.0m of the budget while the actual roster used a budget of \$37.6m. This was likely a contributing factor to the optimized team's poor performance: although DET was ranked ninth in the sequential approach, the leftover funds suggest that the real-world team managed to recruit a couple of expensive players that under the sequential optimization approach were assigned to different teams.
- Miami Marlins (MIA) in 2011: This case is similar to DET in 2011 in the sense that the optimization approach uses only \$9.9m while \$15.1m is spent on the actual roster (MIA was ranked 24th in the sequential approach). This indicates that MIA's real-world recruitment secured players beyond the expected capability of the team's overall payroll."

4.3 Insights for competitive advantage

The outputs of the models solved above can be used to generate key insights for general managers, especially managers of teams with small budgets who must be especially strategic. Below, we continue to use Metric 5 (broad positional flexibility), as we explore some of these practical applications.

First, we consider the number of players in a team's ideal roster who also appear in the team's sequential roster for a given year, i.e., the number of players the team would have signed if they had first choice in the market. This gives us a measure of the team's dominance over its competition. For example, Table 6 shows the teams that would have had at least three players from their ideal roster in their sequential roster. Unsurprisingly, the teams with the highest total numbers of players from their ideal roster also on their sequential roster are the New York Yankees, followed by (in order) the Boston Red Sox, the Philadelphia Phillies, the New York Mets and the Detroit Tigers.

We observe that most teams have a maximum of two players from their ideal roster in their sequential roster; many do not have any players from their ideal roster in their sequential one. Table 7 shows the number of teams, out of 30, that have at least one player from their ideal roster also in their sequential roster.

Beyond identifying which players from their ideal rosters each team may find it easier to recruit, the proposed methodology provides a team's general manager with insight into which players will be in highest demand on the free agent market in a given year. This indicates to the general manager which players are the most likely to be high-value players, with higher WAR-to-expected-salary ratio (most commonly, at lower salary levels), and therefore offer efficiency in the optimization problem. Some players may negotiate their salary upward if they are in high demand. For instance, the 10 players in highest demand for each given year are provided in Table 8. In every year of the period considered, these top 10 players appeared in the ideal rosters of at least seven teams. The year's most popular player appeared in the roster of 25 teams out of 30 in 2009 (Will Ohman), 22 teams in 2010 (Juan Rinson), and 19 teams in 2011, 2012, and 2013 (Joel Peralta, Guillermo Mota and Jason Grilli). As an example, Will Ohman was probably popular because he was a relatively low-cost relief pitcher with a solid WAR projection, something desirable to many teams, which also would have had the roster and budget space to pursue that transaction. The results appear to be pitcher-heavy, which matches our intuition because there are a lot of spots available for pitchers (typically

Table 6 Teams with at least three players from their ideal roster in their sequential roster, per year

Teams	Year	Nb players
BOS	2009	3
BOS	2010	6
BOS	2011	6
BOS	2012	3
BOS	2013	6
DET	2012	3
NYM	2009	6
NYM	2010	3
NYY	2009	7
NYY	2010	7
NYY	2011	11
NYY	2012	7
NYY	2013	9
PHI	2010	4
PHI	2011	3
PHI	2012	7

Table 7 Number of teams (out of 30) with at least one player from their ideal roster in their sequential roster

Year	2009	2010	2011	2012	2013
Nb teams	9	10	12	12	13

5–6 starting pitchers and 7–8 relief pitchers) relative to position players (for instance, a team may seek only 1–2 players at first baseman position.)

The general manager can also use our approach to filter by expected salary, focusing on players with a salary higher than a threshold, for instance $5 m. Those would be the most competitive players, for whom salary negotiation may be more critical. The manager can also decide, among expensive players at a certain position, whether to pay a premium to get another, more expensive player, even if that means exceeding the budget that was expended in practice when the actual team was signed.

On the other end of the spectrum are players, that are not part of many teams' rosters. Table 9 summarizes the number of players who appear in the ideal rosters of no teams, one team, two teams, three teams, four teams or five or more teams, broken down by year. We observe that about half the free agents do not appear in any team's ideal roster, i.e., they are only being signed in our sequential model because a team could not get the players they actually wanted. In a given year, between 31% and 44% of the players are split roughly equally between appearing in only one team and appearing in five or more teams. The remaining 13%-19% of players, depending on the year, appear in the ideal rosters of two, three or four teams.

In other words, it is useful for the general manager to know which free agents never appear in any team's ideal and sequential rosters or who appear only once: compared to the more popular free agents who appear in both rosters of many teams, these players are less likely to face much upward salary pressure during negotiation. Given this information, managers could better prioritize their fallback choices, avoiding some of these lower-ranking players while targeting others who may offer good performance at an affordable salary. Table 10

Table 8 Top 10 most popular free agents per year

Year	2009	2010	2011	2012	2013
Rank 1	Will.Ohman	Juan.Rincon	Joel.Peralta	Guillermo.Mota	Jason.Grilli
Rank 2	Chris.Richard	Yorvit.Torrealba	Jose.Veras	Dioner.Navarro	Chris.Snyder
Rank 3	Pedro.Martinez	Rafael.Betancourt	Hisanori.Takahash	Peter.Moylan	Roy.Oswalt
Rank 4	Damaso.Marte	Brendan.Donnelly	Ronny.Paulino	Dan.Wheeler	Scott.Atchison
Rank 5	Michael.Barrett	Brent.Clevlen	Vicente.Padilla	Chris.Young	Jose.Contreras
Rank 6	Juan.Cruz	Jim.Thome	Sean.Burroughs	Mike.Costanzo	Andres.Torres
Rank 7	Julian.Tavarez	Rich.Harden	Cliff.Lee	Rick.Ankiel	Angel.Pagan
Rank 8	David.Eckstein	Rocco.Baldelli	Casey.Kotchman	Adam.Greenberg	Freddy.Guzman
Rank 9	Cliff.Floyd	Russ.Springer	Joaquin.Benoit	Albert.Pujols	Gil.Velazquez
Rank 10	Justin.Huber	Dustin.Moseley	Donny.Lucy	Gil.Velazquez	Joel.Peralta

Table 9 Counts of free agents who appear in the ideal rosters of zero, one, two, three, four, or five or more teams

Year	2009	2010	2011	2012	2013
0 team	47	48	53	45	45
1 team	13	14	19	24	26
2 teams	4	12	6	6	7
3 teams	8	4	8	3	2
4 teams	3	3	7	6	5
5+ teams	19	17	18	15	20

Table 10 Players who appear in only one team's ideal and sequential roster in any given year

Year	Name	Team	Year	Name	Team
2009	So Taguchi	CHC	2012	Scott Moore	HOU
2009	David Eckstein	DET	2012	Albert Pujols	LAA
2009	Chris Richard	LAA	2012	Ross Ohlendorf	LAD
2009	Cliff Floyd	PHI	2012	Travis Buck	NYM
2009	Eric Munson	SEA	2012	Jimmy Rollins	STL
2010	Rocco Baldelli	CHW	2013	Kensuke Tanaka	TOR
2010	Vladimir Guerrero	OAK	2013	Juan Pierre	BAL
2010	Kevin Mench	SEA	2013	Nick Swisher	CHC
2011	Jeff Francoeur	BAL	2013	Blake DeWitt	DET
2011	Sean Burroughs	CHW	2013	Ed Lucas	MIA
2011	Vladimir Guerrero	DET	2013	Omar Quintanilla	NYM
2011	Geoff Blum	LAA	2013	Melky Cabrera	SFG
2012	Chris Young	TEX	2013	Casey Kotchman	TEX

lists the players who appeared in only one team's ideal and sequential rosters in a given year, along with the name of the team. Note that these players are generally non-pitchers, and are signed by a team with a specific positional need. We suspect that the positional constraints are driving this behavior, even if the player's value is poor. For instance, David Eckstein in 2009 was a second baseman with a $0.94m salary and Cliff Floyd that same year was a designated hitter with a $0.83m salary. Few teams were willing to pay those salaries at those positions. Other players like Chris Richard were untested minor league players; hence, there was not a strong case for multiple teams to be interested in them.

In summary, the proposed methodology provides insights into how competitive the free agent market is, which players are much sought after and which may be easier to sign, allowing teams to better tailor their recruitment strategies to their needs and constraints.

5 Discussion and conclusion

In this paper we have proposed a modeling framework that optimizes each team's ideal roster based on the team's needs and budgets. This framework allows teams to understand which players may maximize their expected number of wins and better reach their goals - whether that means making the playoffs or winning the World Series. Our modeling approach also

explains the competition for top talent, which may be contributing to the recent trend of very large contracts for the best players.

While our study was based on retrospective data, teams can apply it prospectively in several stages. First they can run the roster optimization for the ideal case in order to understand which set of players maximizes their expected success. They can further expand the analysis by running a sensitivity analysis on each player in the ideal roster, which would show how the team's performance may be impacted by not securing a given player. In other words, this would show which player(s) deliver a key competitive advantage. Further sensitivity analysis can quantify the advantage of additional budget: How many additional expected wins does each $1 m bring? Then, by collecting information about their competitors' positional needs and approximate budgets for the upcoming season, teams can understand the competitiveness of the market. This better understanding of their competition will allow them to develop more competitive recruiting strategies.

Our study therefore takes the first step in translating predictive analytics into a prescriptive strategy. The next steps could explore several interesting research avenues. First, as previously discussed, different teams may have different objective functions: anything short of winning the World Series may be a failure for one team, while another considers making the playoffs a huge success. Yet for such smaller teams aiming to make the playoffs, maximizing the expected number of wins may not correctly capture this objective. If making the playoffs is a low probability event, then maximizing the probability of making the playoffs may include a more risky strategy. Adjusting the objective function for different team objectives is therefore an interesting future line of inquiry.

Second, while in this paper we ultimately focus on a single performance model out of six initial possibilities, a general manager could also aggregate the six rosters obtained from our six optimization models, noting in how many rosters each player appeared. This would give the manager a broader set of players to choose from as he starts negotiations, along with the corresponding knowledge of which players would be most critical to the team's performance. Additionally, a roster that is robust across multiple performance metrics may be more likely to succeed than a roster that maximizes one metric.

Third, the model we have developed can easily be modified to account for additional considerations, both operational constraints and market dynamics. For less competitive teams in the free agent market, augmenting the models with additional constraints, such as eliminating the most sought after talent from the feasible set, may provide more actionable roster suggestions that can be the foundation of an optimized recruiting strategy.

Fourth, the model we have developed assumes that players will provide the same level of performance on any team, whereas there are many team-specific factors that may influence performance, including city quality of life, team environment, or player complementarity. While these factors are certainly important, this assumption is a useful simplification for this paper, and future work should incorporate research into these important factors (Brave et al., 2019).

Fifth, in this paper we developed different linear regression models to link predicted individual player performances to teams' success. There are ample opportunities to improve these models, including accounting for prediction errors in the decision models. Additionally, including special considerations for key roles may prove beneficial (although in our experiments including more detailed player information did not improve the team performance models).

Finally, the fact that our approach allows managers to develop methodological recruitment strategies is particularly important for teams that end up with no free agent from their ideal roster on their sequential roster: it helps them make the best available choice in this situation.

However, it is also relevant to any team that faces budget constraints and competition for players, as in practice teams do not decide sequentially. Therefore, we see the opportunity to weave game theoretical considerations and the modeling of competition into our approach. For example, a team may consider recruiting a player not simply because it maximizes their expected number of wins, but also because it leaves other teams worse off.

Recruitment decisions in Major League Baseball, as in other professional or high-level sports, are fraught with uncertainty. Although this context is rich with statistics that track individual and team performance, the future of any particular athlete or team is difficult to predict. Although tremendous advances have been made in applying analytics to improve team performance, there is still opportunity to more fully leverage the available data to help general managers make optimal decisions in complex and shifting recruitment situations. While predictive analytics may give us some idea of the future performance of a given roster of players, our work moves towards a prescriptive analytic approach that can guide general managers in selecting a roster optimized for their team's needs and constraints. By using these prescriptive tools, teams will be able to get the most out of their budget, maximize their competitive advantage, and make good long-term investments in their athletes - whether those are the sought-after superstars or the lesser-known players who will make valuable contributions to the team.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10479-023-05725-4.

Appendix A: Additional examples: ideal rosters exceeding and falling short of actual rosters

Best performance of our metric.

- Kansas city Royals (KCR) in 2009: The actual roster is more balanced than the optimized one with a maximum salary of \$4.7m and five players paid over \$1 m, while in the optimal roster, the maximum is \$9.2m, the next highest salary is \$1.1m, and the other 10 players are all paid less than \$1 m, with nine of them at base salary. The overperformance of the optimization approach is due to the outstanding performance of Kenshin Kawakami, who was selected by the optimization model despite his WAR performance prediction of 0. Hence, in this case the overperformance may be due to chance rather than systematic advantages.
- Washington Nationals (WSN) in 2011: Five players in the actual roster and four in the optimized roster are paid over \$1 m, but in the actual roster, two players are paid at least \$10 m and the third highest salary drops to \$1.6m. In contrast, in the optimized roster, the maximum salary is \$10.5m and two next highest are \$5.5m and \$3.7m. While the highest paid player in the optimized roster had an actual WPA of -0.9, the second and third highest paid players had WPAs of 2.1 and 1.2, respectively. In contrast, in the actual roster, the highest WPA was 1.1 and the second highest was 0.4.
- Washington Nationals (WSN) in 2012: This is a case where the maximum salary is higher in the optimized roster than in the actual one (\$13.4m, Carlos Beltran, vs. \$11.3m, Edwin Jackson). In this case, Beltran did very well with an actual WAR of 3.9 and an actual WPA of 2.4. The optimized roster was also helped by the presence of Fernando Rodney, with an actual WAR of 3.8 and WPA of 5.1 (with an adjusted salary of \$1.8m).
- Atlanta Braves (ATL) in 2013: The main reason the optimized roster of five new players performs better than the actual one is that instead of signing B.J. Upton at \$12.7m,

who underperformed (actual WAR −1.3, actual WPA −2.8), the optimization approach signs two players in the $6.1–6.7m adjusted salary range, who both performed quite well. In addition, the optimization approach signs Dioner Navarro, who also exceeded predictions, at an $1.8m adjusted salary.

- Milwaukee Brewers (MIL) in 2013: Both the optimized and actual rosters sign Kyle Lohse, who has the highest adjusted salary at $11.2m and had an actual WAR of 3.3 and WPA of 1.1. However, in the actual roster, none of the other WPAs are positive while four of the other WPAs in the optimized roster are positive, leading to a cumulative WPA of −0.3 in the optimized case vs. −4.9 in the real world. Because the salary distribution is not fundamentally altered, the overperformance of the optimization approach might to some degree be due to luck.

Worst performance of our metric.

- Houston Astros (HOU) in 2010: The optimization approach results in a roster of 11 new players with a maximum of $7.6m in adjusted salary, two players in the $0.71-$0.76m range and the remaining eight at base salary, while the actual roster has two players in the $3.3-$4.6m range, two in the $0.76–0.87m range, and the remaining seven at base salary. Hence, the star in the optimized roster is Carl Pavano with a salary of $7.6m, with all the other salaries being much lower, while the actual roster splits his salary over two players. With an actual WAR of 4 and WPA of 0.6, Pavano did quite well, but his performance is counterbalanced by that of Rodrigo Lopez, with a WAR of −0.7 and WPA of −3.2. The worst WPA of the actual roster is −1.1 (Gustavo Chacin).
- Arizona Diamondbacks (ARI) in 2012: The maximum salary in this roster of 11 new players is $7.7m in the actual roster and $8.2m in the optimized one. The second highest salary in the actual roster is $5.6m, with the third highest dropping to $2.0m. Six players in the actual roster are paid over $1 m. In the optimized roster, four players were paid above $1 m, with all of those being paid at least $2 m. The cumulative WPA of the players paid over $1 m was 3.4 in the actual roster and -5 in the optimized roster. Particularly detrimental to the performance of the optimized roster was the selection of Francisco Rodriguez, who is the highest paid player but had an actual WAR of −0.2 and WPA of −1.3.
- Baltimore Orioles (BAL) in 2012: This is another case where the optimization approach leads to an overemphasis on very expensive players that backfires. In this case, the optimized approach signs Casey Kotchman at $3.1m, but his actual WAR was −0.9 and his WPA was −2.8.
- Seattle Mariners (SEA) in 2013: The underperformance of the optimization approach is due to the signing of Hisashi Iwakuma to the actual team, who far exceeded predictions with a WAR of 7 and WPA of 3.5.
- San Francisco Giants (SFG) in 2013: The underperformance of the optimization approach is due to the signing of B.J. Upton, who underperformed, to the optimized team at an adjusted salary of $12.7m. The maximum adjusted salary of the actual roster was $8.4m, allowing two other salaries in the $6.1–6.8m range. In the optimization approach, the next highest salaries are $8.1m and $1.4m.

Appendix B: Heteroskedasticity in team performance models

We investigate potential model misspecification in our models in Sect. 3.1 with tests for heteroskedasticity. We use the Breusch-Pagan Lagrange Multiplier test (Breusch & Pagan,

1979) for heteroskedasticity on each model, the results of which are shown in Supplementary Table 1. All of the p-values are below 0.01 for Models 1–3, indicating the presence of heteroskedasticity. In Supplementary Figure 1 we highlight the heteroskedasticity of Model 1. The models tends to predict closer to the mean, causing a pattern of under-prediction for high performing teams and over-prediction for low performing teams.

Heteroscedasticity commonly results in inconsistent estimates of standard errors of linear regression models, leading to confidence intervals that are either too wide or too narrow. To investigate this effect we reran Models 1–3 with robust standard errors, using the HC1 estimator (MacKinnon & White, 1985). The robust standard errors for each model were in fact close to or lower than the original standard errors; In all three models the intercept standard error decreased and the standard error for WPA and/or WAR increased by less than 10% (with p-values remaining highly significant). More importantly, as in this paper we are using the models as predictive inputs to the decision models it is important to note the the regression estimates are not affected when using robust errors.

References

Albert, J. (2006). Pitching statistics, talent and luck, and the best strikeout seasons of all-time. *Journal of Quantitative Analysis in Sports*, 2(1).

Barnes, S. L., & Bjarnadóttir, M. V. (2016). Great expectations: An analysis of major league baseball free agent performance. *Statistical Analysis and Data Mining, 9*(5), 295–309.

Baumer, B., & Zimbalist, A. (2014). *The sabermetric revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press.

Bendtsen, M. (2017). Regimes in baseball players' career data. *Data Mining and Knowledge Discovery, 31*, 1580–1621.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization*. Princeton series in applied mathematicsPrinceton University Press.

Bertsimas, D., & Sim, M. (2003). Price of robustness. *Operations Research, 52*, 35–53.

Brave, S. A., Butters, R. A., & Roberts, K. A. (2019). Uncovering the sources of team synergy: Player complementarities in the production of wins. *Journal of Sports Analytics, 5*(4), 247–279.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica, 47*(5), 1287–1294.

Busing, C., Koster, A., & Kutschka, M. (2011). Recoverable robust knapsacks: The discrete scenario case. *Optimization Letters, 5*, 379–392.

Chan, T. C. Y., & Fearing, D. S. (2013). The value of flexibility in baseball roster construction. In *MIT sloan sports analytics conference*.

Chan, T. C. Y., & Fearing, D. S. (2019). Process flexibility in baseball: The value of positional flexibility. *Management Science, 65*(4), 1642–1666.

Chung, D. J. (2017). How much is a win worth? An application to intercollegiate athletics. *Management Science, 63*, 548–565.

Cot's Baseball Contracts. Highest paid players. https://legacy.baseballprospectus.com/compensation/cots/league-info/highest-paid-players/

DeBrock, L., Hendricks, W., & Koenker, R. (2004). Pay and performance. The impact of salary distribution on firm-level outcomes in baseball. *Journal of Sports Economics, 5*(3), 243–261.

Depken, C. A. (2000). Wage disparity and team productivity: Evidence from major league baseball. *Economics Letters, 67*, 87–92.

Elitzur, R. (2020). Data analytics effects in major league baseball. *Omega, 90*, 102001. https://doi.org/10.1016/j.omega.2018.11.010

Farrar, A., & Bruggink, T. H. (2011). A new test of the Moneyball hypothesis. *The Sport Journal, 14*(1), 1–9.

Frick, B., Prinz, J., & Winkelmann, K. (2003). Pay inequalities and team performance: Empirical evidence from the North American major leagues. *International Journal of Manpower, 24*(4), 472–488.

Gross, A., & Link, C. (2017). Does option theory hold for Majorl League Baseball contracts? *Economic Inquiry, 55*(1), 425–433.

Hakes, J. K., & Sauer, R. D. (2006). An economic evaluation of the Moneyball hypothesis. *Journal of Economic Perspectives, 20*(3), 173–185.

Hall, S., Szymanski, S., & Zimbalist, A. S. (2002). Testing causality between team performance and payroll. The cases of major league baseball and English soccer. *Journal of Sports Economics, 3*, 149–168.
Humphrey, S. E., Morgenson, F. P., & Mannor, M. J. (2009). Developing a theory of the strategic core of teams: A role composition model of team performance. *Journal of Applied Psychology, 94*(1), 48–60.
Humphreys, B. R., & Pyun, H. (2017). Monopsony exploitation in professional sport: Evidence from major league baseball position players, 2000–2011. *Managerial and Decision Economics, 28*, 676–688.
Kahn, L. M. (1993). Managerial quality, team success, and individual player performance in major league baseball. *ILR Review, 46*, 531–547.
Kasperski, A., & Zielinski, P. (2016). Robust discrete optimization under discrete and interval uncertainty: A survey. In *Robustness analysis in decision aiding, optimization and analytics*. Springer.
Kim, J. W., & King, B. G. (2014). Seeing stars: Matthew effects and status bias in major league baseball umpiring. *Management Science, 60*(11), 2619–2644.
Koop, G. (2002). Comparing the performance of baseball players. *Journal of the American Statistical Association, 97*(459), 710–720. https://doi.org/10.1198/016214502388618456
Koseler, K., & Stephan, M. (2017). Machine learning applications in baseball: A systematic literature review. *Applied Artificial Intelligence, 31*(9–10), 745–763. https://doi.org/10.1080/08839514.2018.1442991
Krautmann, A. C. (1990). Shirking or stochastic productivity in major league baseball? *Southern Economic Journal, 5*(4), 961–968.
Krautmann, A. C. (2016). Contract extensions: The case of major league baseball. *Journal of Sports Economics, 19*, 1–16.
Lackritz, J. R. (1990). Salary evaluation for professional baseball players. *The American Statistician, 44*(1), 4–8. https://doi.org/10.1080/00031305.1990.10475682
Lesaege, C., & Poss, M. (2016). The partial choice recoverable knapsack problem. *Computational Management Science, 1*, 189–194.
Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. W. W. Norton & Company.
Liebchen, C., Lubbecke, M., Mohring, R., & Stiller, S. (2009). The concept of recoverable robustness, linear programming recovery, and railway applications. In *Robust and online large-scale optimization* (pp. 1–27). Springer.
MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305–325.
Monaci, M., Pferschy, U., & Serafini, P. (2013). Exact solution of the robust knapsack problem. *Computers and Operations Research, 40*, 2625–2631.
Nasrabadi, E., & Orlin, J. (2013). *Robust optimization with incremental recourse*. Technical report. MIT Sloan School of Management.
Raimondo, H. J. (1983). Free agents' impact on the labor market for baseball players. *Journal of Labor Research, 4*(2), 183–193.
Rockerbie, D. W. (2009). Strategic free agency in baseball. *Journal of Sports Economics, 10*(3), 278–291.
Schall, T., & Smith, G. (2000). XXX double check the first name XXX. Do baseball players regress toward the mean? *The American Statistician, 54*(4), 231–235.
Schultz, R., & Curnow, C. (1988). Peak performance and age amount superathletes: Track and field, swimming, baseball, tennis and golf. *Journal of Gerontology, 43*(5), 113–120.
Scully, G. W. (1974). Pay and performance in major league baseball. *The American Economic Review, 64*(6), 915–930.
Silver, N. (2012). *The signal and the noise*. Penguin.
Spotrac. MLB offseason spending. Online tool. https://www.spotrac.com/mlb/tools/offseason/
Timmerman, T. A. (2000). Racial diversity, age diversity, interdependence, and team performance. *Small Group Research, 13*(5), 592–606.
Turvey, J. (2013). The future of baseball contracts: A look at the growing trend in long-term contracts. *The Baseball Research Journal, 42*(2), 101–107.
Tymkovich, J. L. (2012). A study of minor league baseball prospects and their expected future value. CMC Senior Theses (p. 442). http://scholarship.claremont.edu/cmc_theses/442
van den Akker, J., Bouman, P., Hoogeveen, J., & Tonissen, D. (2014). Decomposition approaches for recoverable robust optimization problems. Technical report, Utrecht University, Utrecht.
Wiseman, F., & Chatterjee, S. (2003). Team Payroll and team performance in major league baseball: 1985–2002. *Economics Bulletin, 1*(2), 1–10.