Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

SMART SPORTS PREDICTIONS VIA HYBRID SIMULATION: NBA CASE STUDY

Ignacio Erazo

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
755 Ferst Drive NW
Atlanta, GA 30332-0205, USA

ABSTRACT

Increased data availability has stimulated the interest in studying sports prediction problems via analytical approaches, in particular, with machine learning and simulation. We characterize a number of models that have been proposed in the literature, all of which suffer from the same drawback: They cannot incorporate rational decision-making and strategies from teams/players in an effective manner. We tackle this issue by proposing hybrid simulation logic that incorporates teams as agents, generalizing the models/methodologies that have been proposed in the past. We perform a case study on the NBA with two goals: i) study how well it is possible to predict using only one variable, and ii) study how much historical data should be kept to maximize prediction accuracy. Results indicate that there is an optimal range of data quantity, and that studying what data and variables to include is of extreme importance.

1 INTRODUCTION

Over the last decades, data collection, availability and usage has greatly increased. Under this context, simulation is used as a powerful tool that allows to optimize the performance of the system being considered. In the past, discrete-event simulation, system dynamics and agent-based simulation were the main paradigms used, but lately new paradigms better suited to leverage data by incorporating intelligent frameworks (such as Machine Learning) are gaining steam; in particular hybrid models and metamodels. Prediction is a relevant problem in most areas of study, and while having more data available has facilitated its study through Machine Learning (ML) lens, solving some problems requires the combination of methods. In the recent paper “Simulation-based Prediction” (Lim and Glynn 2023), the authors consider a hard system where there is some “observable-data” that is not necessarily aligned with the full state/data of the system, which makes simulation an essential tool to improve their decisions/predictions.

Sports prediction is a really hard problem because: i) a very small amount of “observable-data” from games does not describe all the state of the system (health of players, strategies of teams, interactions between players), ii) correlated data (teams/players against each other), iii) time-varying data in a setting where the “non-observable” variables can change significantly, iv) missing data (some specific matchups). In order to tackle on the sports prediction problem we propose a hybrid simulation approach that allows to incorporate decision-making from teams/players, making the model more realistic, which should improve predictions. Under its most simple configuration, this model generalizes discrete-event models, and we use that setup for a case study on the NBA. The organization is as follows: Section 2 presents a literature review on hybrid simulation, the sport predictions problem, some NBA simulation applications, and our contributions. Section 3 explains our methodology; whereas Section 4 introduces our case study, then shows and discusses our results. Finally, Section 5 provides our conclusions.

2 LITERATURE REVIEW

First, this section goes over recent simulation-optimization using hybrid models and applications where hybrid models enable intelligent decision-making. Then, the literature of sports prediction, and in particular the NBA is presented. Finally, the contributions of this work are presented.

2.1 Simulation-Optimization and Hybrid Simulation with Smart Decision-Making

In general, simulation can be seen as a methodology used to model a system, and (Fu 2002) describes simulation-optimization as the field that looks for the “best” inputs for our decision variables of simulated systems, with “best” being roughly equivalent to optimizing the (or many) metrics of interest. That paper describes the process as a two-step procedure, with the first step focusing on generating candidate solutions, and the second on evaluating them. It follows that procedures allowing to find “good” candidate solutions in quick manner speeds up the process considerably, and that is why metaheuristics and Machine Learning have been embedded in Simulation models, leading to hybrid models. While still accurate for most problems, Fu’s definition does not take into account optimization problems where solutions lie in a “policy-space”, instead of just a “decision-variable” space, (see a more recent panel talk on simulation optimization (Fu et al. 2014)). For these problems the design of the policies is something that needs to be accounted for, just as the optimization of the decision variables (which may vary under different policies). Straightforward applications of “policy-space” problems that can be solved with the help of simulation include: logistic problems (see (Santos and Bispo 2016; Lee et al. 2018; Erazo and De la Fuente 2021)), health-care applications (see (Dorali et al. 2022; Erazo et al. 2022)), reinforcement learning problems, among others.

Hybrid simulation has seen an almost exponential increase of usage in the past two decades, and in the state-of-the-art review (Brailsford et al. 2019), hybrid simulation is referred to as a “modeling approach combining two of the following: discrete-event simulation, system dynamics and Agent-based simulation”. They also mention that the definition has always been flexible, and has become ever looser, since now it tends to also be associated with systems using only one simulation paradigm, but also other analytical tools within it (i.e. machine learning, discrete optimization, metaheuristics, among many others). In fact, the previous works may be considered hybrid models because of the plethora of methods used to get to the desired solutions. Finally, hybrid models incorporating smart-decision making have also been proposed, as those smart-decisions help to represent better what is actually happening in the real systems they are mimicking. In particular, (De la Fuente, Erazo, and Smith 2018) demonstrate how to enable smart processes within simulation models by integrating Machine Learning within the model’s logic; and other applications from digital twins include the use of Neural Networks to improve the representation of the real system (Reed, Löfstrand, and Andrews 2022).

2.2 Prediction of Sport Games in a Nutshell

Sports predictions have always been an area of active developments because of the fact that sports results are intricately tied to betting outcomes. Having said that, the new data availability has helped quantitative research to blossom. In one of the earliest influential works, (Koning et al. 2001) predicted the winner of soccer championships using historical data on scoring, and adjusting for the quality of opponents. The authors recognize a challenge that still exists: there is not enough data to properly assess the strength level of a team. This is related to what was presented on the introduction: the fact that there is still some “unobservable” data is missing, even with all the new sensors and technology. Furthermore, even if one had a numeric “strength index” for each team, there is a need of a probabilistic model that returns, given the strength of both teams in a game, a winning probability for them. Simulation becomes then an attractive solution method for prediction in sports, since different rating models for teams/players and different probabilistic models for outcomes can be tested, and compared to the actual outcomes seen in reality. Recently (Garnica-Caparrós et al. 2022) described a general a simulation framework to analyze predictive models for sports with pair confrontations, such as tennis, soccer, football, rugby, basketball, and many others. Their framework includes four sequential steps: i) The competition network, ii) The rating procedure, iii) The forecasting method and iv) The model validation; and they claim using this methodology may help to provide a better understanding on rating procedures and forecasting techniques for accurate predictions. In the contributions subsection we present how our hybrid approach enhances this framework.

2.3 Simulation for NBA Predictions

The National Basketball Association (NBA) was founded in 1946 and is the main basketball league in the world. Composed by 30 teams with an average value of 2.8 billion USD, it is easy to see its relevance in the sports' world, and why many aspects of the game have been researched. In particular, aspects influencing the games' outcomes have been studied, such as: effects of being "home" or "away" (Ribeiro et al. 2016), factors affecting quality of plays and shots (Rolland et al. 2020), expected value of possessions given tracking data (Cervone et al. 2016), among others. All of the aforementioned aspects affect team performance, which determines games' outcomes.

With respect to simulation towards predicting the games' outcomes, most of the work has been done on using the available data to provide better forecast for plays and games; with the results being compared to betting odds, as it is theorized that they offer a predictor and source of expert advice regarding the outcomes, and they have empirically performed well (Spann and Skiera 2009), which has also been confirmed by state-of-the-art models prediction models not being able to beat the odds (Manner 2016). There is two main approaches, one is where researchers use play-by-play simulation of games such as to predict outcomes; and the other is where game outcomes are predicted directly from the data for teams and their players. With respect to the former approach, (Oh et al. 2015) considered a data-driven graphical model, and (Vračar et al. 2016; Sandholtz and Bornn 2018; Sandholtz and Bornn 2020) used Markov-Decision Processes (MDP) to evaluate policies and also model the transitions between plays. For the latter approach (Song et al. 2020) model scores and predict win probabilities based on a Bivariate normal model depending on 5 performance statistics (with "the four most important factors", see (Kubatko et al. 2007)), and recently deep learning methods have been used for the scores' prediction (Yanai et al. 2022), but with underwhelming results, very likely due to the small amount of "observable" data available.

2.4 Contributions

First, we propose a hybrid simulation model that works for all sports with pair confrontations, and that enhances the state-of-the-art in simulation for sports predictions by:

- Allowing for teams/players to use strategic decision-making by adding agent-based aspects to the hybrid model. This basically helps the simulation to better represent the real system.
- Facilitating the integration of steps ii) and iii) in (Garnica-Caparrós et al. 2022) into a single step that uses Machine Learning (basically eliminating the need of rating the teams and use two models); but also allowing for the former two-step method.

With respect to sports, even with increased data availability, single metrics remain the norm for communication purposes, as it is easier to compare teams based on a single value. Having said that, in many sports these single summary values can be derived in a straightforward manner, such as ATP points/rankings in Tennis, or win percentage in Basketball; however there exists very limited literature focused on their predictive power for single games, and how they should be used to predict outcomes of single matches. Having said so, the contribution of our case study are:

- We use the simplest version of our model to assess the value of single summary statistics for prediction.
- We evaluate the question of: "how much historical/past data should be kept for future prediction?". As will be seen in Section 4, this decision has a large impact.

3 METHODOLOGY

So far in Section 2 we presented many influential works focusing on the prediction of sports games' outcomes. Even though those models are based on different methods and are applied on different sports, they can be classified in four categories: 1) stochastic simulation models, 2) play-by-play simulation models, 3) ML-based models and a 4) tournament model. We start this section by explaining how all of them can be seen as a discrete-event simulation model.

1) In these models each game is an event and outcomes are computed according to statistical distributions (and parameters) that are based out of expert knowledge, historical data, and previous studies.
2) In play-by-play simulation models the plays constitute the events. These models are based on Markov Chains (or MDPs), and the sequence of events is simulated using the transition probabilities. Sequence of games are just simulated by performing the play-by-play model individually for each game, all according to the specified sequence.
3) ML-based methods depend on the data, in particular what variables are being used, and how far into the past the data is accumulated. An event usually corresponds to a game/match, and the algorithms are trained, then tested over the prediction for events. This is a discrete-event model because when the ML model is used for prediction, each game can be seen as an "independent" event, and simulating a sequence of games is just predicting on them sequentially. In reality the events are not "independent" since the columns of a game may contain information from the previous games played, but the ML algorithm just sees the events as independent samples to predict on.
4) In tournament models each game corresponds to an event, and the tournament is just an ordered sequence of games, where the outcome of previous games may (i.e. soccer world cup), or may not determine the following matchups, and even the winning probabilities. Game simulation is just performed by one of the models 1), 2) or 3); leading to a complete discrete-event behavior.

Models may be considered hybrid because of incorporating different methods/algorithms within the simulation, however they all suffer from a large drawback: strategy and game-theory cannot be added to the model, unless it is at the expense of an extremely large increase in the complexity (i.e. distribution parameters, state space for MDPs, number of columns for ML). Some examples highlighting it is important to add team/player logic are:

- The UEFA Champions League features double-game elimination matchups, where strategies play a large role, in particular on the second match where the outcome of the first game is already known. This cup also features a double round-robin style group stage, and on the last few games some teams play knowing they are eliminated or classified to the next stage. Sometimes there even is incentives for both team to draw, which introduces an extra layer of complexity and strategy.
- It is common for tennis players to play both the singles and doubles draw of a tournament. If advancing successfully in both tournaments, then strategy appears: retire from one of the draws to increase chances in the other? Prioritize the effort for the "most-likely-to-win" match?
- On tournaments with unequal rest between matches there is frequent load managing. In the NBA, many players rest one game if there is two on consecutive days, so: what game does the star player play-in? Most of the time that is decided by the team with some goal in mind, usually increasing their win expectations, or making sure start players are available on national TV games.

3.1 A Hybrid Model with an Agent-Based Component

It is easy to see the advantages of allowing for team/player strategy and rational thinking; and it is also clear that the previous models have a hard time doing so. We propose a hybrid simulation model that incorporates agent-based logic such as to tackle this issue in a straightforward manner. Our model allows to test the performance of different strategies, to consider agents that may react differently under the same conditions and to better mimic the real behavior of teams/players. Our simulation model basically behaves as discrete-event model by simulating games sequentially, according to the logic of the tournament (regular season or elimination-based cup), but it is hybrid because instead of just considering "data" to make the predictions, each component of the model can be seen as an agent with properties/data and that interacts with other agents. The type of agents, and their relationships are shown in Figure 1, where the boxes (season, game, team, player) correspond to agents, the blue arrows mean one to many" relationship, and the black arrows mean a "one to one" relationship. Note that some lines and boxes are dashed, and that just means they are not a necessity to get a functional model, and they can (or not) be included.

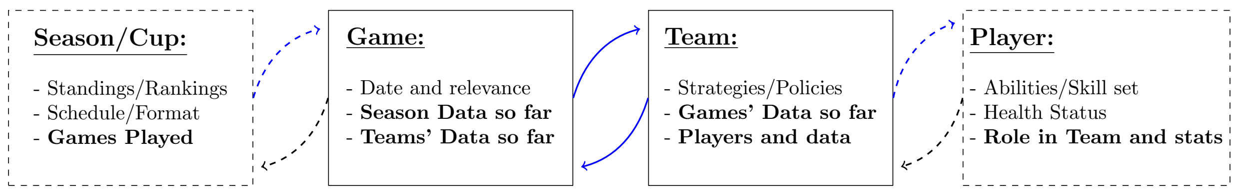

Figure 1: Type of agents for the hybrid model, and their relationships.

The simplest version of this model just consider games to be predicted, with each game being linked to its respective two teams (or players if it is an individual sport). Teams/players may play many games, and there is a constant update of the data for each of them after they play one game. Note that if we just set all the teams/players to have no strategy, then the model just reduces to a discrete-event simulation model that can perform prediction with either a two-step approach (first compute a team rating, then winning probabilities, as explained by (Garnica-Caparrós et al. 2022)) or a one-step approach based on data (distribution or ML-based). It follows that even this simple setup can be seen as a generalization for all of the models categories (i)-(iv), which hints at the modeling power enabled by our framework.

The complete version of our model adds a season/cup agent that keeps track of all the games being played and the overall standings/rankings. Each season/cup is linked to many games, and the games can retrieve the actual information of the season plus the information known (so far) about the future; allowing for teams to react and strategize accordingly. For example, if a soccer team won 6-0 the first leg, their lineup will be strategically modified for the second leg. For team sports the model also includes players as agents. At any point in time a team is linked to many players, but a player only to one team, and the teams can use their players' information to make decisions that will affect the match, all according to their strategy and decision-making logic. The framework continues to allow a two-step or a one-step approach for prediction of games, however the input for such algorithms can now be affected by the agents' actions. We believe that this increased flexibility has no downsides, since it can replicate the current models in place while also adding new modeling tools.

3.2 Two-Step Prediction Approaches: Functions for the Second Step

We just presented our hybrid model and mentioned that it can be used with either: 1) a two-step prediction approach that is based on first computing a "rating" for teams, and then returning win probabilities based on the parameter of both teams; or 2) a one-step approach that predicts directly from the teams' data (i.e. ML algorithm). Our focus will be on the two-step prediction approach, in particular on the second step. The first step seeks to assess the strength or ratings of a team and is a very interesting and widely studied problem, however as it goes outside the scope of our work, we consider this step as a subroutine. We propose two simple functions that can be used for the second step, both being intuitively improving if the first step does so.

For the ease of exposition, we will just refer to teams but they may represent players on individual sports. We will assume that "Team 1" plays against "Team 2" and that we wish to predict the outcome of the game, based on a single numeric input that comes from the first step (rating procedure).

3.2.1 Real-Valued Rating

We start by assuming our input is a real-valued number that represents a rating or "strength" index where more means better. An intuitive and simple procedure corresponds to just returning in deterministic manner the team/player with highest value as the winner (if there is one), or a tie if both teams have the same value (for sports with no ties, return instead a 50% win chance for both teams).

While simple, this method has a desirable property: the accuracy should increase if the rating input is a better representation of the team/match conditions. As example, in soccer this rating could correspond to an approximation of the expected-number of goals that the team will make in the game, and then this procedure just picks the team with the largest value as the winner.

3.2.2 Probability-Valued Rating

Now, let's assume the input is a "probability-valued" index between 0 and 1; for example, the probability of winning versus an "average-team". We define the *Bernoulli Race* between "Team 1" and "Team 2" with parameters $p_1,\ p_2$ (respectively) as the distribution denoted $BR(p_1,\ p_2)$ that returns:

- For sports allowing ties: "Team 1" wins with probability $p_1(1-p_2)$, "Team 2" wins with probability $(1-p_1)p_2$ and there is a tie with probability $(1-p_1)(1-p_2)+p_1p_2$. Intuitively, each team performs a Bernoulli trial according to their parameter, and if the obtained values are equal then it is a tie, otherwise the team that obtains the positive outcome wins.
- For sports without ties: "Team 1" wins with probability $\frac{p_1(1-p_2)}{p_1(1-p_2)+(1-p_1)p_2}$, "Team 2" wins with probability $\frac{(1-p_1)p_2}{p_1(1-p_2)+(1-p_1)p_2}$. Intuitively, "Team 1" and "Team 2" draw Bernoulli samples until one of them gets a value of one (wins) and the other does not (losses).

This distribution is particularly interesting for sports with no ties because of the following property: if we let p be the probability of beating an "average" team, then an average team would have a value of $p = 0.5$. If "Team 1" plays versus an "average Team 2" with $p_2 = 0.5$ then the probability of "Team 1" winning is exactly $\frac{0.5p_1}{0.5p_1+0.5-0.5p_1} = p_1$, which basically ensures the interpretation of values p remains consistent. Also, if both $p_1 = p_2$, both teams have a 0.5 probability of winning their match; again ensuring consistency. Furthermore, the interpretation makes sense on sports with balanced seasons (each teams plays the same amount of games) since the average win probability over such leagues is precisely 0.5 (each game one team wins and another losses).

4 CASE STUDY — NBA

Over our case study we focus on predicting the binary outcome of NBA *regular season* games. In order to do so, we downloaded ten seasons of games' box score data (for both teams), starting on the 2011-2012 season and until season 2021-2022, all done using the Python module nba_api. So far, previous works have achieved an accuracy around 62-69% (Manner 2016; Song et al. 2020) for the binary outcome of games (usually computed over the second half of a NBA season), by using many box score statistics, diverse probabilistic models and/or machine learning algorithms. We will instead consider the simplest version of our proposed hybrid-simulation model (that just generalizes the discrete-event models). Furthermore, we will consider the probabilistic models described in Subsection 3.2 to predict the games' outcomes, and the inputs to such models will be some widely-available and easy-to-compute single statistics. The motivation is two-sided: 1) by using such a simple simulation model we will be able to understand the real value of the single summary statistics and of the two probabilistic models, and 2) this simple setup makes easy to study a relevant but neglected topic in the sports prediction literature: how much *historical* data should we be keeping for predictions? This is relevant because teams' performances are dynamic and keeping more more historical data may hurt the prediction accuracy.

4.1 Value of Simple Summary Statistics and Probabilistic Models

Over this case study we use following two input values:

1. Win percentage: if we let x_1^i be 1 if "Team 1" won its i^{th} game and 0 otherwise, then the win percentage of "Team 1" at game $i+1$ is: $p_1^i = \frac{\sum_{k=1}^{i} x_1^k}{i}$. We denote as the "offline" win percentage the value that is obtained at the end of the season; and we denote as "online" win percentage for game $i+1$ the value $\hat{p_1^i} = \frac{0.5 + \sum_{k=1}^{i} x_1^k}{1+i}$, such as to avoid 0 and 1 probabilities.
2. Net rating: It corresponds to the sum of points scored minus points received over all games played, scaled by number of possessions. It essentially represents the net efficiency of a team, where a larger number is better. We denote the "offline" version as the net rating at the end of the season, and the "online" version as the net rating computed before starting the game.

For both input values, we refer to them as "home-adjusted" if the value used for the team playing at "home" is computed only considering its respective "home" games; while the overall value is used for visiting teams. We consider the following eight methods:

(i) Offline *Bernoulli Race*, based on win percentage.
(ii) Online *Bernoulli Race*, based on win percentage.
(iii) Home-adjusted offline *Bernoulli Race*, based on win percentage.
(iv) Home-adjusted online *Bernoulli Race*, based on win percentage.
(v) Online largest value, based on win percentage.
(vi) Online largest value, based on net rating.
(vii) Home-adjusted online largest value, based on win percentage.
(viii) Home-adjusted online largest value, based on net rating.

Using the downloaded data, we simulated each of the 10 seasons a 1000 times for the 8 different methods (i)-(viii) presented above. Tables 1 and 2 present the average accuracy of predicted games across replications for seasons.

With respect to Table 1, Comparison 1A between methods (i) and (ii) basically shows how much accuracy we "lose" by using the online win percentage data, compared to using the offline data. It makes sense that there is a drop-off in accuracy for each season, since the data used for prediction in the online setting (ii) does not include future information but the offline data (i) does; however the decrease in accuracy is never above 3%, and is just 1.6% on average across the 10 seasons. Comparison 1B between methods (iii) and (iv) also shows the decrease in accuracy between offline and online settings, however this time when using as parameter the home-adjusted win percentage (as explained earlier). Over this comparison, when using the online data there is a more pronounced drop-off effect that averages out to 2.5%, and that tops out at 3.7% for season 2020-2021. Finally, looking at Comparison 1, we see that both the online and offline versions benefit from using home-adjusted parameters, which makes sense since there is a widely-known and reported advantage for teams when playing at home (Ribeiro et al. 2016).

Comparison 2A focuses on using the largest function when the parameters are either the online win percentage (v) or the online net rating (vi). Both parameter choices achieve an average accuracy of over 60% for every season, and an average across seasons that is slightly above 63%. In particular, using net rating seems to provide a slightly better accuracy, with larger average accuracy in 6 out of 10 seasons, and with a larger average across the 10 seasons. With respect to Comparison 2B, we again focus on home-adjusted parameters. Now both parameter choices seem to provide an approximately equal performance, with 4 wins each over the 10 seasons, and two ties; plus the same averaged accuracy across the 10 seasons. When looking at Comparison 2, the home-adjusted parameters (vii)-(viii) seem to provide a very small benefit, and all the methods have between 63 and 64% accuracy.

Table 1: Average accuracy of games across replications for full seasons, depending on the method used.

	Comparison 1				Comparison 2			
	Comparison 1A		Comparison 1B		Comparison 2A		Comparison 2B	
Season	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
2012-2013	58.7%	56.9%	61.7%	59.0%	64.4%	66.2%	65.7%	65.7%
2013-2014	58.9%	57.1%	60.7%	58.1%	64.1%	64.0%	64.0%	63.4%
2014-2015	59.7%	58.7%	61.6%	59.2%	66.5%	66.8%	66.6%	66.4%
2015-2016	60.4%	59.0%	62.5%	60.3%	65.8%	65.7%	68.1%	66.8%
2016-2017	56.8%	55.8%	58.3%	56.1%	60.4%	61.5%	61.9%	62.2%
2017-2018	58.0%	55.9%	59.7%	56.4%	63.1%	62.7%	62.4%	61.9%
2018-2019	57.8%	56.0%	60.2%	57.7%	62.6%	63.1%	63.4%	64.4%
2019-2020	58.3%	57.1%	60.1%	58.3%	62.5%	62.3%	63.8%	63.8%
2020-2021	57.1%	54.4%	58.5%	54.8%	60.7%	61.5%	59.4%	60.4%
2021-2022	57.3%	55.7%	58.1%	55.6%	61.3%	62.4%	62.0%	62.1%
Overall	58.3%	56.7%	60.1%	57.6%	63.1%	63.6%	63.7%	63.7%

Table 2 shows the accuracy results for the same 10 seasons but only computing the values over the second half of each season, such as to be consistent when benchmarking against previous works (they usually use the first half to train models and/or calibrate the models' parameters). It is easy to see that the same discussion for Comparisons 1A, 1B and 1 applies to Comparisons 3A, 3B and 3, respectively. Similarly, the same discussion for Comparisons 2A, 2B and 2 applies for Comparisons 4A, 4B and 4, respectively; however this time it does seem like net rating provides more value for prediction compared to the win percentage parameter. Now comparing across Table 2 and Table 1 it is easy to see that each of the methods (i)-(viii) improves its overall performance when only considering the second half of the season. The improvement is modest when using the Bernoulli Race model (between 0.2 and 0.7%), and larger when using the Largest Value model (between 2.1 and 2.6%), eventually achieving prediction accuracies nearby 66%, right around the performance reported in previous works, but just using a simple deterministic model with single easy-to-compute values. This suggests that even in the presence of a small number of variables, decently good predictions can be obtained if meaningful information is provided.

Table 2: Average accuracy of games in the second half of the season, depending on the method used.

	Comparison 3				Comparison 4			
	Comparison 3A		Comparison 3B		Comparison 4A		Comparison 4B	
Season	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
2012-2013	59.4%	57.7%	62.8%	60.6%	68.0%	69.1%	68.0%	67.2%
2013-2014	59.2%	57.8%	61.0%	58.6%	66.3%	65.5%	66.2%	66.2%
2014-2015	59.2%	57.9%	61.6%	59.3%	67.2%	68.3%	67.8%	67.8%
2015-2016	60.5%	59.6%	62.7%	61.1%	68.5%	69.4%	69.9%	71.5%
2016-2017	56.3%	55.1%	57.8%	56.1%	60.2%	61.5%	62.1%	62.4%
2017-2018	59.1%	57.0%	60.9%	58.1%	65.9%	67.6%	66.7%	65.2%
2018-2019	58.2%	57.1%	60.3%	58.4%	66.7%	65.7%	64.9%	65.2%
2019-2020	57.1%	56.4%	58.8%	57.4%	62.5%	61.7%	64.0%	63.8%
2020-2021	58.3%	56.5%	60.1%	57.5%	66.5%	67.4%	65.4%	67.0%
2021-2022	57.7%	56.0%	58.6%	56.1%	63.4%	66.0%	62.6%	65.5%
Overall	58.5%	57.1%	60.5%	58.3%	65.5%	66.2%	65.8%	66.2%

Now, what is the trade-off between using the largest value method that provides more accuracy, versus the probability-based *Bernoulli Race* method? Intuitively, it is easy to see that in offline settings the *Bernoulli Race* method is close to an unbiased estimator for the number of wins, and that can be seen empirically on Figure 2. On the other hand, the largest value method is deterministic for both offline and online data, and leads to unrealistic records, because teams starting with large net ratings are guaranteed to win, leading to "historic" simulated seasons. Basically, the trade-off for more accuracy is having deterministic outcomes for the simulation, and such outcomes being very unbalanced for good teams (they get more wins that they should) and bad teams (they get less wins that they should).

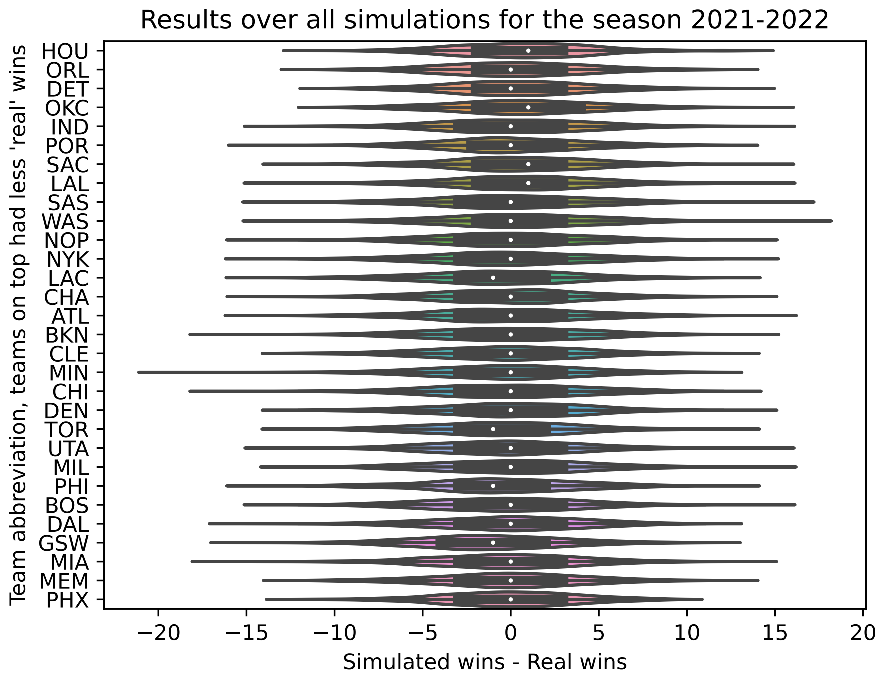

Figure 2: Distribution of Simulated wins - Real wins for method (i).

With respect to the online *Bernoulli Race* method, it is certainly not unbiased, and predicts more wins to teams that start the season winning, compared to teams that have a better second half. Results can be seen in Figure 3, where teams are ordered according to real record, "bad" teams on top. Note that teams that had a poor season seem to be more favored, and that makes sense because of the way the season progresses: NBA preseasons are short, which leads to more unpredictable first few games, usually leading young and inexperienced teams to overachieve at the beginning of the season. Also, teams with bad records rest their star players more often at the end of the year because more losses increase their odds at a better outcome on the nba draft. This analysis clearly reinforces the fact that we should be including agent-based decision-making on the prediction models, such as to account for these phenomenon, but the analysis also helps to motivate why we need to care about how much "historical" data should be considered when making predictions.

4.2 More Temporal Data is not Always Better

In this subsection we study how much historical data should be used for prediction. Figure 4 shows the prediction accuracy (over 1000 replications) for method (iv) over the second half of the 2021-2022 season, depending on the number of previous games that were included on the computation of the online

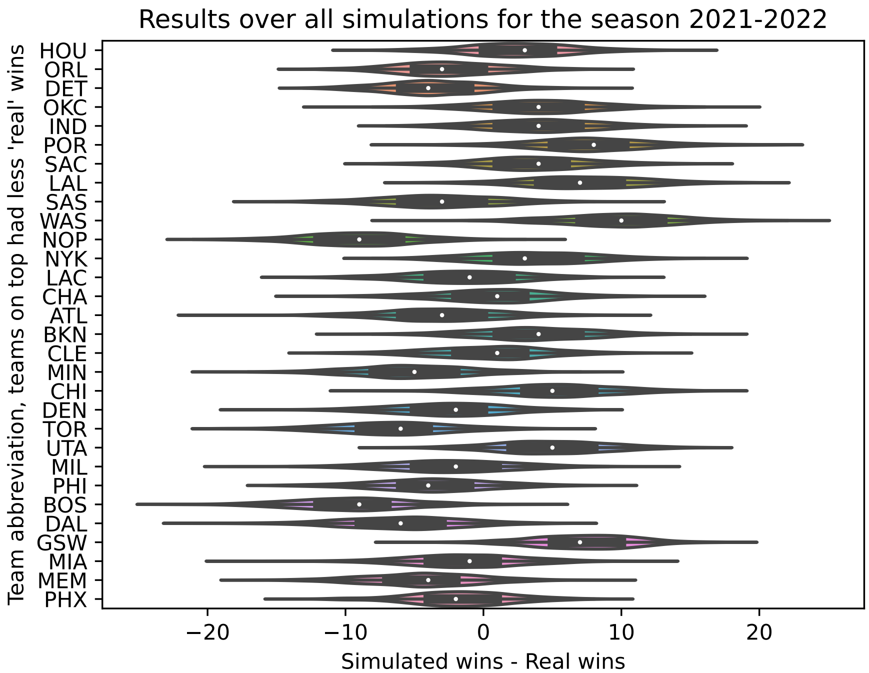

Figure 3: Distribution of Simulated wins - Real wins for method (ii).

home-adjusted winning percentage parameter for teams. We note that the 95% confidence interval is small, and that results indicate that the prediction accuracy initially increases when adding more games as part of the data, then remains in a comparable zone, until it starts declining. In particular, the peak under this scenario seems to be achieved when considering between 8 and 14 games, and the accuracy is around 57.7%, which is greater than the 56.1% reported in Table 2 when using the complete online data for the season at each game. This behavior is consistent across methods and seasons, and so we encourage research on this topic to be a focus in the future, because it has large potential to improve the state of the field.

5 CONCLUSIONS

In this article we described the sports prediction problem and emphasized the complexity of getting accurate prediction results. In particular, we highlighted the methods used to solve this problem and how they fit within what can be loosely considered as discrete-event simulation. We then proposed a hybrid simulation model that combines discrete-event simulation with agent-based logic, allowing to generalize what has been proposed in the literature, and adding new modeling capabilities into the mix. We also presented two simple functions that take a single real-valued input and return a winner for a given game, and used them with summarized single statistics as input over our case study for prediction of NBA regular season games.

Over our case study we first evaluated the value of the single summary statistics and the proposed functions. Results indicate that even under single-valued inputs and simple functions, similar accuracies compared to data-intensive algorithms can be achieved; however it comes at the expense of a large positive (negative) bias towards good (bad) performing teams. This suggests that more effort needs to be done on identifying key measures influencing the success probabilities for teams, rather than using as much data as possible. Then, we evaluated how much historical data to include for prediction-making, and results across the board indicated that more data increases performance only up to a certain point, then there is a zone where performance is constant (while adding more data), and eventually accuracy drops down. This suggests that a relevant design question is to decide how much data to keep for predictions, and that

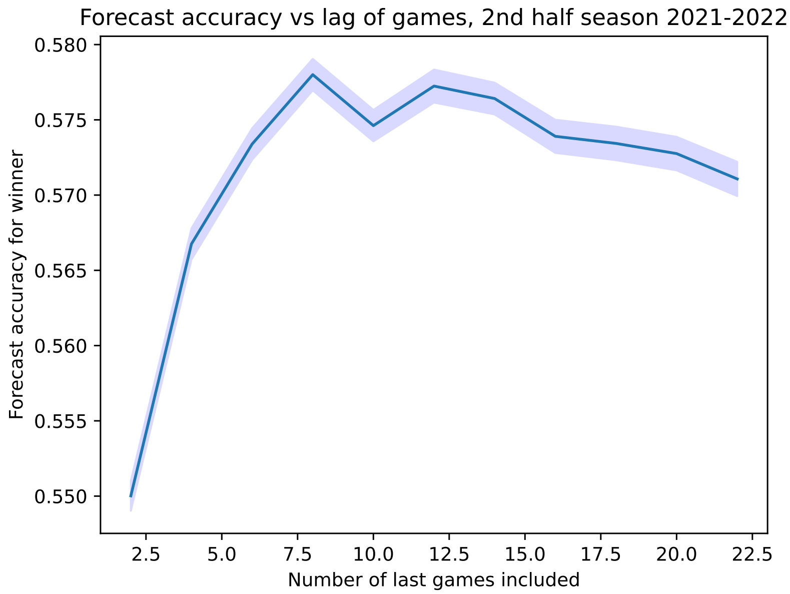

Figure 4: Impact of the number of games included on the estimator for home-adjusted online *Bernoulli Race* (method (iv)) over the second half of the season.

is also something that needs to be studied in the future. Finally, some future research avenues include: expanding the work on creation of single summary statistics with high accuracy power and incorporating the agent-based logic into specific sports, with the agent decision-making based on expert opinion or game-theory, and evaluating the impact of the agent behavior in the forecast accuracy.

ACKNOWLEDGMENTS

The author of this paper thanks Professor David Goldsman for encouragement and support over this work.

REFERENCES

Brailsford, S. C., T. Eldabi, M. Kunc, N. Mustafee, and A. F. Osorio. 2019. "Hybrid simulation modelling in operational research: A state-of-the-art review". *European Journal of Operational Research* 278(3):721–737.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry. 2016. "A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes". *Journal of the American Statistical Association* 111(514):585–599.

De la Fuente, R., I. Erazo, and R. L. Smith. 2018. "Enabling Intelligent Processes in Simulation Utilizing the TensorFlow Deep Learning Resources". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1108–1119. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Dorali, P., R. Limongi, F. K. Torsha, C. Y. Weng, and T. Lee. 2022. "Utilizing Simulation to Update Routine Diabetic Retinopathy Screening Policies". In *2022 Winter Simulation Conference (WSC)*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, 1128–1139. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Erazo, I., and R. De la Fuente. 2021. "A Simulation-Based Approach to Compare Policies and Stakeholders' Behaviors for the Ride-Hailing Assignment Problem". In *Proceedings of the Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Erazo, I., D. Goldsman, P. Keskinocak, and J. Sokol. 2022. "A Simulation-Optimization Framework to Improve the Organ Transplantation Offering System". In *2022 Winter Simulation Conference (WSC)*, edited by B. Feng, G. Pedrielli, Y. Peng,

S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, 1009–1020. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Fu, M. C. 2002. “Feature Article: Optimization for simulation: Theory vs. Practice”. *INFORMS Journal on Computing* 14(3):192–215.

Fu, M. C., G. Bayraksan, S. G. Henderson, B. L. Nelson, W. B. Powell, I. O. Ryzhov, and B. Thengvall. 2014. “Simulation optimization: A panel on the state of the art in research and practice”. In *Proceedings of the Winter Simulation Conference 2014*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. J. Buckley, and J. A. Miller, 3696–3706. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Garnica-Caparrós, M., D. Memmert, and F. Wunderlich. 2022, sep. “Artificial Data in Sports Forecasting: A Simulation Framework for Analysing Predictive Models in Sports”. *Inf. Syst. E-Bus. Manag.* 20(3):551–580.

Koning, R., M. Koolhaas, G. Renes, and G. Ridder. 2001. *A simulation model for football championships*. A65 ed, Volume SOM Research Reports. University of Groningen, SOM research school.

Kubatko, J., D. Oliver, K. Pelton, and D. Rosenbaum. 2007, 02. “A Starting Point for Analyzing Basketball Statistics”. *Journal of Quantitative Analysis in Sports* 3:1–1.

Lee, J., Y. Kim, J. Kim, Y.-B. Kim, H.-J. Kim, B.-H. Kim, and G.-H. Chung. 2018. “A Framework for Performance Analysis of Dispatching Rules in Manufacturing Systems”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3550–3560. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lim, E., and P. W. Glynn. 2023. “Simulation-Based Prediction”. *Operations Research* 71(1):47–60.

Manner, H. 2016. “Modeling and forecasting the outcomes of NBA basketball games”. *Journal of Quantitative Analysis in Sports* 12(1):31–41.

Oh, M., S. Keshri, and G. Iyengar. 2015. “Graphical Model for Baskeball Match Simulation”. In *MIT Sloan Sports Analytics Conference*.

Reed, S., M. Löfstrand, and J. Andrews. 2022. “Modelling stochastic behaviour in simulation digital twins through neural nets”. *Journal of Simulation* 16(5):512–525.

Ribeiro, H. V., S. Mukherjee, and X. H. T. Zeng. 2016, 03. “The Advantage of Playing Home in NBA: Microscopic, Team-Specific and Evolving Features”. *PLOS ONE* 11(3):1–18.

Rolland, G., R. Vuillemot, W. J. Bos, and N. Rivière. 2020. “Characterization of Space and Time-Dependence of 3-Point Shots in Basketball”. In *MIT Sloan Sports Analytics Conference*.

Sandholtz, N., and L. Bornn. 2018. “Replaying the NBA”. In *MIT Sloan Sports Analytics Conference*, 1–13.

Sandholtz, N., and L. Bornn. 2020, 09. “Markov decision processes with dynamic transition probabilities: An analysis of shooting strategies in basketball”. *Annals of Applied Statistics* 14:1122–1145.

Santos, A. F., and C. F. Bispo. 2016. *Simulation Based Optimization Package for Periodic Review Inventory Control*, 2867–2878. IEEE Press.

Song, K., Q. Zou, and J. Shi. 2020. “Modelling the scores and performance statistics of NBA basketball games”. *Communications in Statistics - Simulation and Computation* 49(10):2604–2616.

Spann, M., and B. Skiera. 2009. “Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters”. *Journal of Forecasting* 28(1):55–72.

Vračar, P., E. Štrumbelj, and I. Kononenko. 2016. “Modeling basketball play-by-play data”. *Expert Systems with Applications* 44:58–66.

Yanai, C., A. Solomon, G. Katz, B. Shapira, and L. Rokach. 2022. “Q-Ball: Modeling Basketball Games Using Deep Reinforcement Learning”. *Proceedings of the AAAI Conference on Artificial Intelligence* 36:8806–8813.

AUTHOR BIOGRAPHIES

IGNACIO ERAZO is a Ph.D. Candidate in Operations Research in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. He is focused on efficient data-driven decision making via integer optimization and large-scale simulation-optimization. His email address is ierazo@gatech.edu and his website is https://ierazo.github.io.