Expert Systems With Applications 44 (2016) 58–66

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Modeling basketball play-by-play data

Petar Vračar*, Erik Štrumbelj, Igor Kononenko

Faculty of Computer and Information Science, Večna pot 113, Ljubljana 1000 Slovenia

ARTICLE INFO

Keywords:
Forecasting
NBA
Logistic regression
Decision tree
Markov process

ABSTRACT

We present a methodology for generating a plausible simulation of a basketball match between two distinct teams as a sequence of team-level play-by-play in-game events. The methodology facilitates simple inclusion into any expert system and decision-making process that requires the performance evaluation of teams under various scenarios. Simulations are generated using a random walk through a state space whose states represent the in-game events of interest. The main idea of our approach is to extend the state description to capture the current context in the progression of a game. Apart from the in-game event label, the extended state description also includes game time, the points difference, and the opposing teams' characteristics. By doing so, the model's transition probabilities become conditional on a broader game context (and not solely on the current in-game event), which brings several advantages: it provides a means to infer the teams' specific behavior in relation to their characteristics, and to mitigate the intrinsic non-homogeneity of the progression of a basketball game (which is especially evident near the end of the game). To simplify the modeling of the transition distribution, we factorize it into terms that can be estimated with separate models. We applied the presented methodology to three seasons of National Basketball Association (NBA) games. Empirical evaluation shows that the proposed model outperforms the state-of-the-art in terms of forecasting accuracy and in terms of the plausibility of the generated simulations.

1. Introduction

Statistics and mathematical modeling have become an important part of sports and a lot of effort is dedicated to predicting the outcomes of sporting events (Percy, 2015; Stekler, Sendor, & Verlander, 2010). In this paper we focus on a more general task of sports outcome forecasting, where the goal is to predict not only the outcome, but also a more detailed evolution of the sporting event.

One of the most significant changes in the past decade has been the introduction and growing public availability of play-by-play data. In particular, in basketball. Compared to the more traditional box-score summary statistics, play-by-play data offer a richer description of within-match events (see Tables 1 and 2).

Often, such data are used for modeling the outcome of sports matches. Previous experience suggests that bookmaker odds are the best source of probabilistic forecasts for sports matches (Forrest, Goddard, & Simmons, 2005; Song, Boulier, & Stekler, 2007; Spann & Skiera, 2009). It appears that predictions from statistical and other types of prediction models are less accurate than those of fixed-price bookmakers and betting exchanges (which perform better not only because they use all publicly available information but also because of the wisdom-of-the-crowds effect). On the other hand, statistical models can be used to generate credible simulations of the likely progression of a sports match between two specific opponents.

The current state-of-the-art methods produce simulations that reflect the actual teams' win probabilities. However, the simulations are limited to reproducing only one aspect of the game (e.g. scoring statistics) or, in the case of more detailed simulations, exhibit some flaws in their credibility. The main reason for this disadvantage is that the models do not take into account the current context of the game, which affects the dynamics of the game.

We propose a methodology for learning from play-by-play data that deals with the issues outlined above. We assume the Markov property and model state transitions with a Logistic regression model, similar to Štrumbelj and Vračar (2012). We deal with non-homogeneity of the progression of a basketball game by incorporating relevant variables (time, point difference, …) into the state space of the model. The model also includes the transition time, which we model conditional to the predicted transition type. We show that our approach leads to a more credible simulation and more accurate forecasts of game outcomes.

The presence of such a detailed simulator can be beneficial for the further development of expert systems in sports. Having already

* Corresponding author. Tel.: +38631328933.
E-mail addresses: petar.vracar@fri.uni-lj.si (P. Vračar), erik.strumbelj@fri.uni-lj.si (E. Štrumbelj), igor.kononenko@fri.uni-lj.si (I. Kononenko).

http://dx.doi.org/10.1016/j.eswa.2015.09.004
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

Table 1
An excerpt from the first quarter of an NBA game between Denver (Away) and Orlando (Home) 2010-03-28.

Time	Team	Player	Action	Score	Quarter
7:21	DEN	N Hilario	2 Point Field Goal	DEN 10–8	1
7:05	ORL	M Barnes	2 Point Miss	DEN 10–8	1
7:04	DEN	N Hilario	Defensive Rebound	DEN 10–8	1
6:54	DEN	J Petro	2 Point Miss	DEN 10–8	1
6:54	DEN	Team	Offensive Rebound	DEN 10–8	1
6:51	DEN	N Hilario	2 Point Miss	DEN 10–8	1
6:48	DEN	A Afflalo	Offensive Rebound	DEN 10–8	1
6:40	ORL	D Howard	Foul	DEN 10–8	1
6:40	DEN	N Hilario	1 Point Free Throw	DEN 11–8	1
6:40	DEN	N Hilario	Missed Free Throw	DEN 11–8	1
6:38	ORL	M Barnes	Defensive Rebound	DEN 11–8	1

Table 2
Box-score data for the NBA game from Table 1. FG - field goals, FGA - field goal attempts, FG% - field goal percentage, 3P - 3-point field goals, 3PA - 3-point field goal attempts, 3P% - 3-point field goal percentage, FT - free throws, FTA - free throw attempts, FT% - free throw percentage, ORB - offensive rebounds, DRB - defensive rebounds, TRB - total rebounds, AST - assists, STL - steals, BLK - blocks, TOV - turnovers, PF - personal fouls, PTS - points.

Team	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%
DEN	42	80	.525	5	16	.313	8	11	.727
ORL	39	79	.494	11	29	.379	14	22	.636
Team	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
DEN	9	30	39	18	3	3	9	21	97
ORL	10	27	37	25	3	4	9	14	103

been used in several decision-making processes (Ballı & Korukoğlu, 2014; Dadelo, Turskis, Zavadskas, & Dadeliene, 2014; Papić, Rogulj, & Pleština, 2009), such expert systems can incorporate the simulator to provide coaches with insight into the performance of their teams during a game. The simulator would act as a sandbox for studying scenarios which may arise in different circumstances on the playing field/court. This functionality will help coaches understand how a team can increase their odds of winning, how individual playing skills affect team performance, and what performance can be expected using different approaches. Through further analysis of the simulation models (e.g. using the method of Štrumbelj & Kononenko (2010)), an expert system can provide its users with an explanation of how different factors affect the possible outcomes of a sequence of events, which could extract new knowledge about basic principles of the sport in question.

The rest of the paper is structured as follows. In the next section we briefly present the related work. In Section 3 we describe our approaches to modeling NBA Play-by-Play data. Section 4 provides the results of the experimental evaluation. The paper concludes with Section 5 where we discuss the results and give some ideas for further work.

2. Related work

Play-by-Play data are now routinely captured for most major team sports and competitions. However, the appropriate tools for modeling and simulating play-by-play data have not been developed. Most related work on modeling the progression of a sports game was done in basketball, which is one of the sports most suitable for such analyses, because of a high frequency of relevant events and better availability of historical statistical data. Stern (1994) used a Brownian motion process to model the evolution of a basketball score between the Home and Away teams, using 1st quarter, half-time, 3rd quarter and final scores. Goldman and Rao (2012) used a similar model to study the effects of 'pressure' on players' performance. They modeled the sensitivity of the win probability to small changes in the score and used it as a proxy for the importance of a particular situation within a game.

The above studies model the progression of a basketball score but not within-game events. Shirley (2007) proposed a more detailed model. He used a homogeneous Markov model with states based on within-game events. In particular, which team has possession (Home or Away), how the possession was obtained, and how many points were scored on the previous possession. Štrumbelj and Vračar (2012) included team-specific variables to Shirley's to account for individual teams strength. The proposed model was good at predicting the final score of a game. However, the game of basketball is neither time-homogeneous nor points-difference homogeneous (see (Štrumbelj & Vračar, 2012) for details). Therefore, a homogeneous model is not realistic.

The duration of transitions between states (that is, of the time that passes between two within-match events) has received little attention. Shirley (2007) and Štrumbelj and Vračar (2012) did not model time explicitly. Instead, they used the average number of state transition to determine the length of a basketball game. Gabel and Redner (2012) built a computational random-walk model to describe several statistical properties of scoring in basketball games. They showed that the distribution of time elapsed between scoring events has an exponential tail. They also argued that the intrinsic strengths of teams play a small role in the random-walk picture of scoring. Merritt and Clauset (2014) modeled the scoring dynamics between featureless teams using two stochastic processes. The first process produces scoring events, while the second process determines which team wins the points. Relying exclusively on the observed patterns in scoring events, they fitted a generative model that accurately reproduces the observed dynamics in lead-sizes over the course of games in several team sports (American football, hockey, and basketball). They also found that the model is able to make highly accurate predictions of game outcomes, after observing only the first few scoring events of the game.

The recent availability of optical player tracking data has opened new prospects in sport modeling. Cervone, D'Amour, Bornn, and Goldsberry (2014) used the player locations on the court to provide a real-time estimation of the expected number of points (EPV) obtained by the end of a possession. They also discussed some potential applications of EPV in revealing novel insights into players' decision-making tendencies. Oh, Keshri, and Iyengar (2015) used diverse data sources (player-tracking data, team lineup data, and play-by-play game log data from the matches played in the 2013–2014 NBA season) to model the progression of a basketball match on individual player level. They treated a game as a random sequence of transitions between discrete states that represent individual players on the court, their actions, and event outcomes. The model simulates the ball movement of every play and subsequent game events based on the player level interaction. However, the simulations are limited to the observed lineups only, since the methodology provides no means of inferring transition probabilities for a hypothetical lineup of players from different teams.

Therefore, while there are similar related works, our contribution is unique in that we facilitate the modeling of in-game events for two distinct teams. As such, a direct comparison is only possible with our previous work (Štrumbelj & Vračar, 2012), which we extend by proposing a model that allows us to incorporate broader in-game context (time, points difference) into the Markov state space, making the assumption of homogeneity more plausible.

3. Modeling NBA basketball data

A basketball game is a realization of a random process that depends on the underlying fundamentals of the sport and the characteristics of the competing teams. We can approximate this process and produce simulations by using a historical set of play-by-play data. We

Table 3
Set of events.

Events label	Description
AJB, HJB	the Away / Home team won the jump ball.
AINB, HINB	the Away / Home team inbounds.
AORB, HORB	the Away / Home team got an offensive rebound.
ADRB, HDRB	the Away / Home team got a defensive rebound.
AREB, HREB	the Away / Home team got a team rebound.
A2PA, H2PA	the Away / Home team missed a 2 point field goal attempt.
A2PM, H2PM	the Away / Home team made a 2 point field goal attempt.
A3PA, H3PA	the Away / Home team missed a 3 point field goal attempt.
A3PM, H3PM	the Away / Home team made a 3 point field goal attempt.
AFT*i*A, HFT*i*A	the Away / Home team missed the first out of *i* remaining free-throw attempts ($i \in [1 \ldots 3]$).
AFT*i*M, HFT*i*M	the Away / Home team made the first out of *i* remaining free-throw attempts ($i \in [1 \ldots 3]$).
ATO, HTO	the Away / Home team committed a turnover.
APF, HPF	the Away / Home team committed a foul.

used play-by-play data from 3 regular NBA seasons (seasons 2008/09 to 2010/11). The data were obtained from msn.foxsports.com/.

We want to model the progression of a basketball game as a Markov process $\{\mathbf{X_i}, i \in \mathbb{N}_0\}$ with state space $\mathcal{S}$. Therefore we want a state space that contains all the variables to the immediate progression of the process. We encode each state as a vector

$$<Evt, Qtr, Time, PtsDiff, \mathbf{a}, \mathbf{h}>$$

where *Evt* is the current game related event, *Qtr* and *Time* the current quarter and seconds left in the quarter, *PtsDiff* is the points difference (from the home team's perspective), **a** and **h** are vectors with the two team's characteristics. Now we describe each component of the state space and the rationale behind this design.

3.1. Set of events

The set of events is limited by what is available in standard basketball play-by-play data: 2 point field goal attempts (missed and made), 3 point field goal attempts (missed and made), free-throw attempts (missed and made), rebounds (offensive, defensive, and team), turnovers, fouls, inbound passes, and jump balls. In order to distinguish between teams, events should be labeled not only with the event type, but also with the team to which the event relates (home or away). Finally, in order to distinguish between different attempts in a single free throws procession, the corresponding events are labeled with the number of remaining attempts (the value of 1 stands for the last attempt in the series). Events are summarized in Table 3.

3.2. Situational variables

Teams play differently when the game is tied or when trailing by 15 points. Furthermore, there is a difference between trailing by 15 points 3 min left in the game or trailing by the same difference at half time. For this reason, we included into a state description the current points difference, the current quarter, and seconds left in the current quarter.

3.3. Team characteristics

The following considered are the most important in determining a basketball team's success: shooting, turnovers, rebounding, and free throws (see (Kubatko, Oliver, Pelton, & Rosenbaum, 2007) and (Oliver, 2004) for details). In total, 16 variables are used as a proxy for team characteristics, 8 for each team (see Table 4).

Table 4
Summary statistics used as a proxy for the team's characteristics. They are based on standard box-score statistics: Field goals made (*FG*), field goals attempted (*FGA*), three point shots made (3*P*), free throws made (*FT*), free throws attempted (*FTA*), turnovers (*TOV*), offensive rebounds (*ORB*), and defensive rebounds (*DRB*). The *o* prefix denotes that the average of the team's opponents is used.

Effective Field Goal %	**Turnover Ratio**
$EFG\% = \frac{FG+\frac{1}{2}3P}{FGA}$	$TOVr = \frac{TOV}{FGA+TOV+0.44*FTA}$
Off. Rebound Ratio	**Def. Rebound Ratio**
$ORBr = \frac{ORB}{ORB+oDRB}$	$DRBr = \frac{DRB}{DRB+oORB}$
Free Throw Factor	**opponents EFG%**
$FTF\% = \frac{FT}{FTA}$	$oEFG\% = \frac{oFG+\frac{1}{2}o3P}{oFGA}$
opponents TOV	**opponents FTF**
$oTOVr = \frac{oTOV}{oFGA+oTOV+0.44*oFTA}$	$oFTF = \frac{oFT}{oFTA}$

Table 5
A sequence of states that correspond to the play-by-play sequence presented in Table 1. For brevity, only two out of 16 team specific variables are included (AEFG and HEFG).

AEFG%	HEFG%	Qtr	Time	PtsDiff	Evt
0.501	0.536	1	441	−2	A2PM
0.501	0.536	1	441	−2	HINB
0.501	0.536	1	425	−2	H2PA
0.501	0.536	1	424	−2	ADRB
0.501	0.536	1	414	−2	A2PA
0.501	0.536	1	414	−2	AREB
0.501	0.536	1	411	−2	A2PA
0.501	0.536	1	408	−2	AORB
0.501	0.536	1	400	−2	HPF
0.501	0.536	1	400	−3	AFT2M
0.501	0.536	1	400	−3	AFT1A
0.501	0.536	1	398	−3	HDRB
0.501	0.536	1	383	−3	H3PA

Before each game, the latest statistics are calculated for both participating teams. A team's statistics are calculated by averaging the statistics from all previous games. We do this for home and away games separately, to take into account the home team advantage.

3.4. Dataset preparation

Now we can transform each game's play-by-play description into a sequence of states (see Table 5 for an example). Note that some states in the resulting sequence describe game events that do not appear explicitly in the play-by-play game description, but are a logical consequence of other events (for example, the HINB event immediately follows the A2PM event, since the home team inbounds after the away team made a 2 point field goal without a foul).

3.5. Modeling state transitions

To simulate a match, we sample a Markov chain $(\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x}_\omega)$, where $\mathbf{x_0}$ is start of the match and $\mathbf{x}_\omega$ the end. Let $f(\mathbf{y}|\mathbf{x})$ be the distribution of $\mathbf{X_{n+1}}$ given $\mathbf{X_n}$. Among all the components in the description of each state $\mathbf{x_i}$, there are only two random variables: the game related event which is about to happen next (a discrete variable *Evt*), and the duration of that event (in playing time, a continuous variable *Dur*), since the score difference and time left in the game are a direct consequence of these two variables, while the statistics (although they vary during the season depending on the team's performance) remain constant during a single game. This allows us to simplify modeling by factorizing the conditional distribution:

$$f(\mathbf{y}|\mathbf{x}) = f_{\mathbf{Y}^{(Evt)}|\mathbf{X}}(y^{(Evt)}|\mathbf{x})\, f_{\mathbf{Y}^{(Dur)}|\mathbf{X},\mathbf{Y}^{(Evt)}}(y^{(Dur)}|\mathbf{x}, y^{(Evt)}) \quad (1)$$

The conditional distribution from Eq. 1 can be estimated using two models, model M_{Evt} that estimates the marginal conditional distribution for the next event given the current state description, and model

M_{Dur} that estimates the time between two events given the current state description and conditional on the next event. The models can be inferred separately. To generate a sample from the joint distribution f (that is, to simulate the next state) we, in sequence, draw samples from these two marginal distribution models.

In general, we could use any parametric or non-parametric model to estimate the desired conditional distributions. However, the progression of a sports match is also governed by a set of discrete sports rules that must not be violated when simulating a match. Due to these rules, every state can only be followed by certain states. This leads to a natural division of states into disjoint subsets of states, depending on what can follow.

Therefore, to avoid invalid state transitions, we have to hierarchically partition the space and model each partition separately. Partitioning can be performed either automatically or manually by introducing background knowledge (e.g. by enforcing a particular tree structure to partition the state space).

In accordance with these observations, we use a decision tree (recursive partitioning) and a separate regression model for each partition. When modeling a continuous distribution, we sample from the empirical distribution of instances in the corresponding leaves of a tree structure. In the case of discrete distributions (that is, nominal target variables), we fit a multinomial logistic regression model for each leaf separately. As an additional benefit, these models are easy to interpret and verify by domain experts.

The historical sequences of states from the play-by-play data can be used as a learning set for the models M_{Evt} and M_{Dur}. Any successive states $\mathbf{x_t}$ and $\mathbf{x_{t+1}}$, which are within the same quarter, define one learning example. In the case of M_{Evt}, the components of $\mathbf{x_t}$ define the values of independent variables, while the $\mathbf{x_{t+1}^{(Evt)}}$ defines the value of the target variable. In the case of M_{Dur}, both the components of $\mathbf{x_t}$ and $\mathbf{x_{t+1}^{(Evt)}}$ define the values of independent variables, while the value of the target variable is defined as $\mathbf{x_t^{(Time)}} - \mathbf{x_{t+1}^{(Time)}}$.

3.6. Simulating basketball games

Once the two models are learned, we can simulate the progression of a basketball game between two specific teams. We generate the progression of a game one quarter at a time, using the final state of the previous quarter as the initial state for the next quarter (see Algorithm 1).

Algorithm 1 Simulate single quarter.

```
Require:
    s - initial state,
    $M_{Evt}$ - a model for $f_{\mathbf{Y}^{(Evt)}|\mathbf{X}}$,
    $M_{Dur}$ - a model for $f_{\mathbf{Y}^{(Dur)}|\mathbf{X},\mathbf{Y}^{(Evt)}}$
Ensure:
    outSeq - a generated sequence of states.

function SIMULATEQUARTER($\mathbf{s}, M_{Evt}, M_{Dur}$)
    $outSeq \leftarrow \emptyset$
    $\mathbf{x} \leftarrow \mathbf{s}$
    while $\mathbf{x}^{(Time)} \geq 0$ do
        $\mathbf{y} \leftarrow \mathbf{x}$
        $\mathbf{y}^{(Event)} \leftarrow$ sample from $M_{Evt}(\mathbf{x})$
        $dur \leftarrow$ sample from $M_{Dur}(\mathbf{x}, \mathbf{y}^{(Event)})$
        $\mathbf{y}^{(PtsDiff)} \leftarrow update(\mathbf{y}^{(PtsDiff)}, \mathbf{y}^{(Event)})$
        $\mathbf{y}^{(Time)} \leftarrow update(\mathbf{y}^{(Time)}, dur)$
        $outSeq \leftarrow append(outSeq, \mathbf{y})$
        $\mathbf{x} \leftarrow \mathbf{y}$
    end while
    return $outSeq$
end function
```

4. Experimental verification

We compared several models with respect to how realistic (credible) are generated simulations and how accurate they are in predicting the winner of a game.

4.1. Models for $\mathrm{M_{Evt}}$

EvtRelFreq is the simplest of the considered models for M_{Evt}. The probability of the next event is modelled using the relative frequencies of the possible outcomes given the event that has just happened (other variables are ignored).

EvtHomModelTree is a M_{Evt} model in the form of a decision tree with a single internal node and logistic regression functions at the leaves. The root node performs a non-binary split of the state space according to the current event. The logistic models in the leaves take into account only the teams' statistics and ignore the context of the game (time left and the current point difference). For that reason, the modelled transition probabilities are time- and score-homogeneous.

EvtModelTree is a M_{Evt} model similar to *EvtHomModelTree*, except that its root node can have full grown tree-like structures as children. Like in the previous model, the root node performs a non-binary split with the respect to the current event, while its subtrees are constructed with a recursive partitioning algorithm using the complete set of variables. At each node the best binary split test is chosen according to MDL criterion (Kononenko, 1995). We used a pre-pruning strategy to stop the growth of subtrees: when the number of instances is below 500 or none of the binary splits are compressive according to the MDL criterion, a leaf node with a logistic regression model is formed (using the complete set of variables).

4.2. Models for $\mathrm{M_{Dur}}$

The simplest model for M_{Dur} that we have considered is *DurSample*. It models the duration of the transition between two states by sampling from the population of all transitions between the observed pair of events (the event that has just happened and the predicted next event).

DurTreeSample is a M_{Dur} model in the form of a regression tree. We forced the structure of the first two levels of the tree in order to stratify the regression according to the event that has just happened (the root node) and the predicted next event (the child nodes of the root node). The rest of the tree structure is constructed with a recursive partitioning algorithm using the complete set of variables. We used the least squares split criterion. As pre-pruning criterion, we set the minimum number of instances per leaf to 500. The result that is returned by the tree is sampled from the values in the corresponding leaf node.

4.3. Evaluating the credibility of simulations

Since the generation of simulations requires two models (M_{Evt} and M_{Dur}), we compared the following composite models. *PROPOSED* is a model that uses *EvtModelTree* to predict the next transition and *DurSample* to predict the duration of that transition. *PROPOSED** is a model similar to the previous one, except that *DurTreeSample* is used to predict the duration of the next transition. *BASELINE* generates predictions based on *EvtHomModelTree* and *DurSample*. In its essence, it is identical to the model from Štrumbelj and Vračar (2012). *RELFREQ* is the simplest model involved in the experiment, based on *EvtRelFreq* and *DurSample*.

First we measured how well the models predict the next event that will take place given the current description of the game. All models were learned using the games from one season and evaluated on the games from the following season. For each record in the Play-by-Play data the models returned the probability distribution of

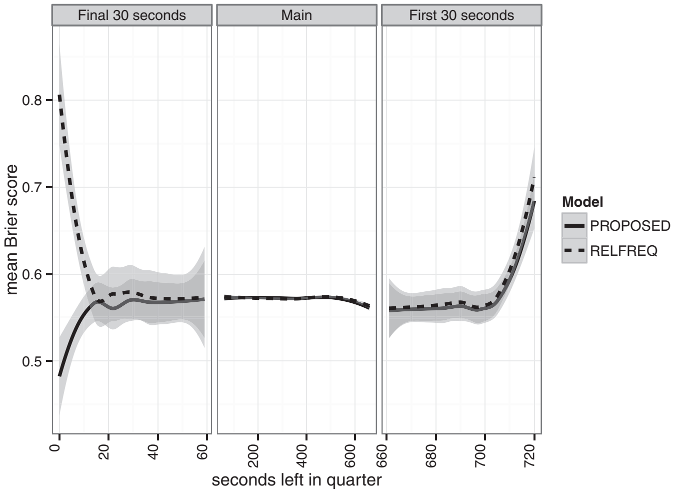


Fig. 1. Mean Brier score as a function of seconds left in quarter. Results were post-processed with loess smoothing. Shaded region represents the standard error.

Table 6
Mean Brier score for next transitions probability forecasts, measured on 2009/10 and 2010/11 NBA seasons (with standard errors in parentheses). *PROPOSED* and *PROPOSED** are equal in performance since the experiment does not include the prediction of a time component.

	2009/10	2010/11
PROPOSED	0.5709	0.5715
	(5.6×10^{-4})	(5.6×10^{-4})
BASELINE	0.5803	0.5786
	(5.6×10^{-4})	(5.6×10^{-4})
RELFREQ	0.5759	0.5775
	(5.5×10^{-4})	(5.5×10^{-4})
	$N = 569185$	$N = 566393$

an event that will occur next in the game. We compared the obtained results with the actual outcomes of events using the quadratic (Brier) score (Brier, 1950). Table 6 shows the results.

PROPOSED and *PROPOSED** are on average significantly better forecasters of the next transition's probability distribution that the other two tested models. Fig. 1 shows the mean Brier score of *PROPOSED* and *RELFREQ* models as a function of seconds left in a quarter. The models are substantially different only in the final 20 s of a quarter. *BASELINE* is not shown in Fig. 1 as it is almost identical to *RELFREQ*.

The next experiment was motivated by Gabel and Redner (2012), who presented several basic features of scoring statistics. We reproduced those features using *PROPOSED* model and compared them to observed values. We focus on the statistical properties based on ball possession. A possession is defined as the period of time that begins when one team gains control of the ball and ends when the opposing team gains control of the ball. A single possession can be composed of several plays if missed shots are followed by offensive rebounds. A scoring possession is a possession during which points were scored.

As in the previous experiments, we learned a model on the games from one season and then used that model to generate 10 simulations of each game from the following season. The presented results were produced using the pooled data from the 2009/10 and 2010/11 seasons.

The distribution of point values per basket in generated simulations corresponds to the empirical data. The production rate, defined as the number of scored points per second, is shown in Fig. 2. The number of points scored remains on a similar level throughout a quarter, with the significant deviations near the start and end of each quarter. *PROPOSED** method follows that trend fairly well. However, a detailed view at the last 60 s of the first three quarters reveals that the methods do not capture the significant drop in production rate that occurs when teams intentionally delay their final shot in order to prevent the opposing team from having another shot opportunity. This rate drop is not present at the end of the fourth quarter (see Fig. 3).

An interesting feature of basketball games is that the scoring probability is affected by the current score. The probability that the leading (trailing) team will score during the possession decreases

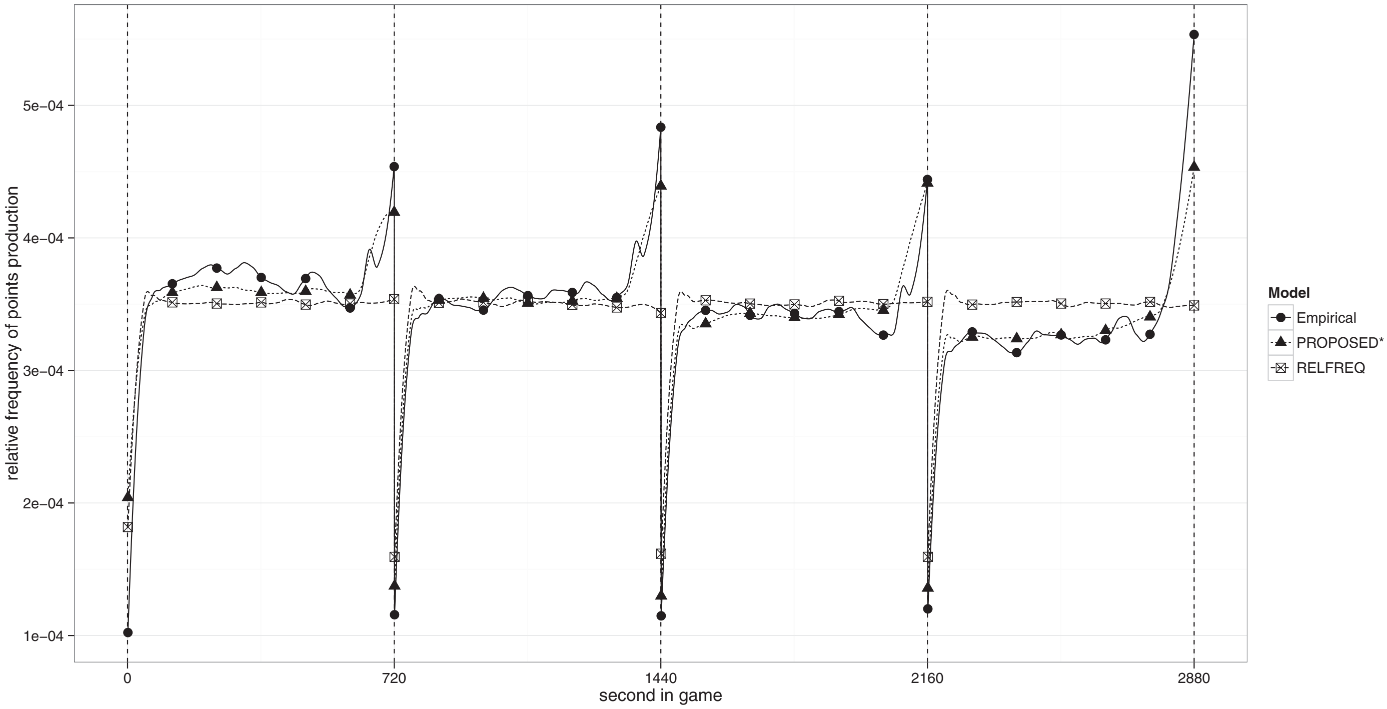


Fig. 2. Average production rate as a function of time. *PROPOSED* and *BASELINE* are similar in performance to *RELFREQ* and therefore not shown.

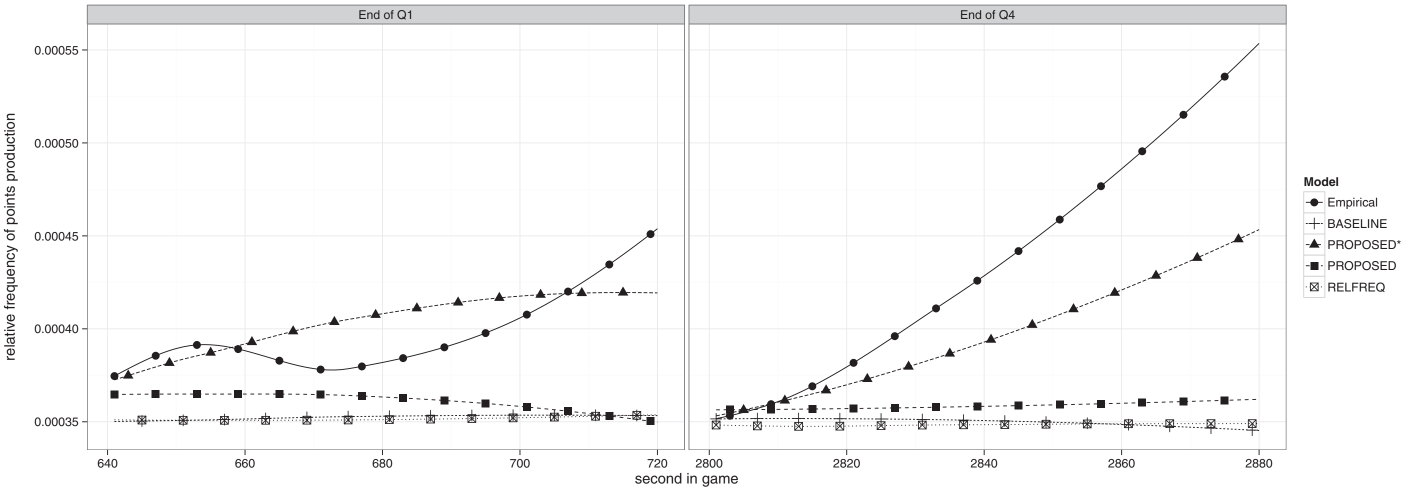


Fig. 3. Detailed view of the average production rate near the end of the first and the fourth quarter.

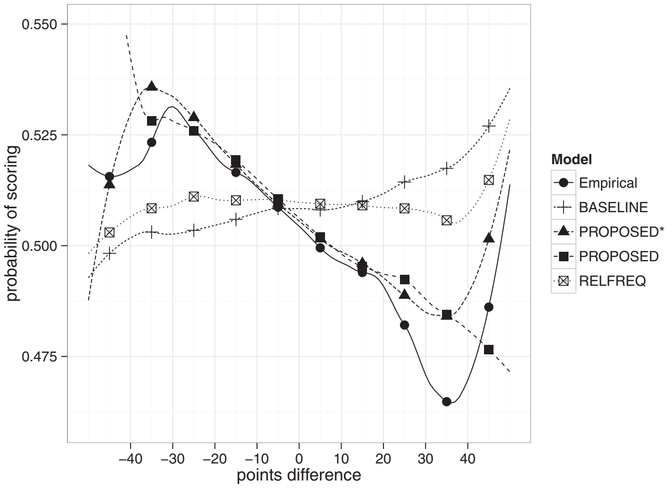


Fig. 4. Probability that a team will end its possession with a basket given a lead.

(increases) with its lead (deficit) size. As Fig. 4 shows, both proposed methods preserve that effect.

Fig. 5 shows how models capture the evolution of the score difference. The variance in the score difference as a function of game time is presented in Fig. 5(a). Both *PROPOSED* and *PROPOSED** methods generate more stable simulations than other methods discussed in this evaluation. Fig. 5(b) presents the probability distribution of time intervals between successive scoring possessions for either team (only *PROPOSED** is shown since other methods produce almost identical results). The simulations contain an adequate number of quick successful plays where the interval between successive points is less than 10 s. Finally, the distribution of the number of lead changes during a game, and the probability for a team to lead for a given game time are shown in Fig. 5(c) and (d), respectively. Both proposed models are more accurate with a lower number of lead changes. The generated simulations also capture that a single team is more likely to lead for the most of the game as opposed to equally sharing the lead.

We used the Kullback–Leibler divergence (Kullback & Leibler, 1951) to measure the dissimilarity between distributions of different statistical features of basketball and the observed data. Results are shown in Table 7.

Table 7
Kullback–Liebler divergence between several distributions (estimated from the generated simulations) and the observed evidence.

Feature	PROPOSED	PROPOSED*	BASELINE	RELFREQ
Possession duration	0.0142	0.0147	0.0156	0.0158
Time intv. between succ. scores for either team	0.0080	0.0089	0.0078	0.0077
Time intv. between succ. scores for the same team	0.0063	0.0061	0.0061	0.0061
Total score in a single game	0.0773	0.0714	0.0592	0.0907
Consecutive point streak	0.0014	0.0014	0.0017	0.0015
Max lead in a single game	0.0155	0.0128	0.0597	0.0362
Number of lead changes	0.0110	0.0099	0.0276	0.0272
Time leading in a single game	0.4004	0.4091	0.4247	0.4187

4.4. Evaluating the forecasting accuracy of simulations

We learned all models using the games from one season and evaluated them with respect to how well they can predict the winners of the games in the following season. An individual game is a test case represented by the characteristics of the competing teams. We used each forecaster to calculate the home team's win probability by first generating 1000 simulations of the game between given teams and then returning the fraction of simulations the home team has won. As in the previous experiment, we used the Brier score to evaluate the quality of predictions as the sum of the squared differences between the calculated home team's win probabilities and the actual outcomes. Results are shown in Table 8.

PROPOSED model is the best forecaster. We tested the significance of differences in performance between *PROPOSED* and *BASELINE* models using one-tailed t-test. The results shown in Table 9 suggest that *PROPOSED* model is significantly better source of forecasts (applies to forecasts provided at the beginning of the game and the half-time, while the difference in the quality of forecasts made before the fourth quarter is not sufficiently large).

Fig. 6 shows the moving average of the mean Brier score measured on forecasts provided at the beginning of the game over the 2009/10 and 2010/11 NBA regular seasons. The graphs show that

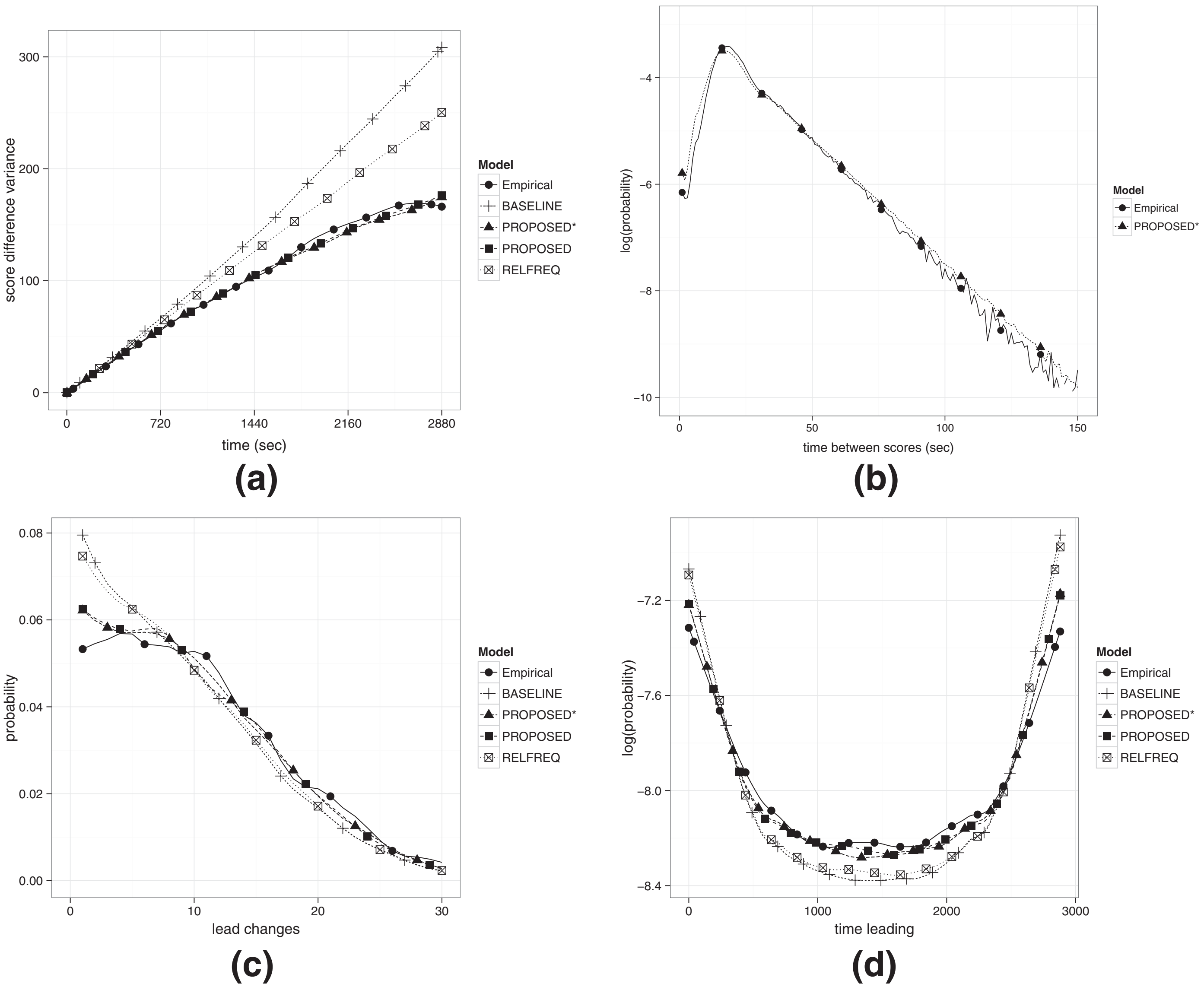
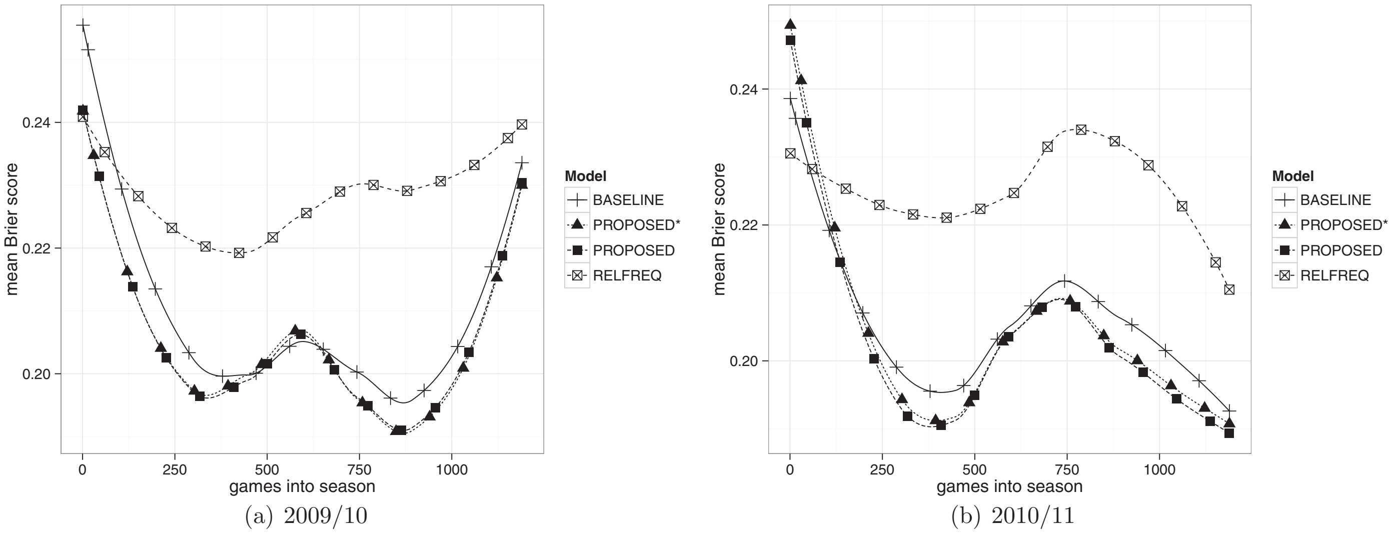


Fig. 5. Evolution of the score difference. (a) Variance in the score difference as a function of time. (b) Probability distribution of time interval between successive scores for either team. (c) Probability distribution of the number of lead changes. (d) Probability distribution of the time leading in a single game.



(a) 2009/10

(b) 2010/11

Fig. 6. The moving average of the mean Brier score measured on forecasts provided at the beginning of the game (window size = 200 games).

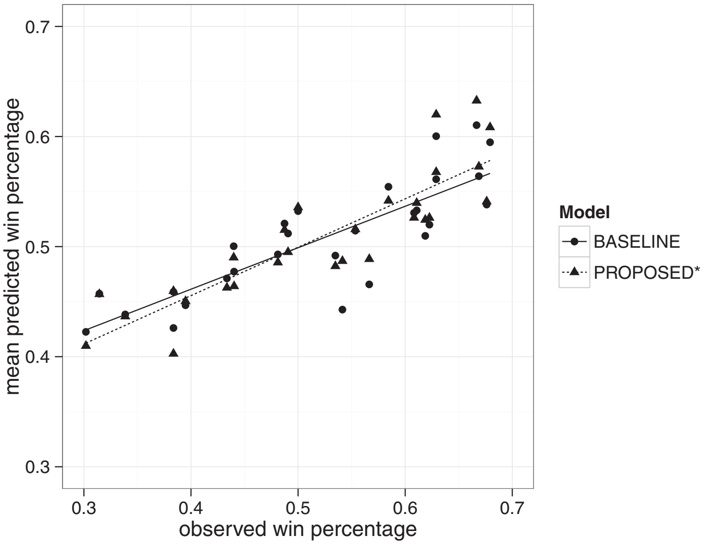


Fig. 7. Obtained vs estimated win percentages for all games in the 2009/10 and 2010/11 seasons. *PROPOSED* is similar in performance to *PROPOSED** and therefore not shown.

Table 8
Mean Brier score (with standard error in parentheses) for the home team's win probability forecasts, measured on 2009/10 and 2010/11 NBA regular seasons, starting from the beginning of the game (Q1), from the actual result at the half-time (Q3), and from the actual result at the start of the fourth quarter (Q4). The best model is shown in bold ($N = 1190$).

	BASELINE	PROPOSED	PROPOSED*	RELFREQ
09/10, Q1	0.2074 (4.39×10^{-3})	**0.2029** (4.33×10^{-3})	**0.2029** (4.31×10^{-3})	0.2276 (2.82×10^{-3})
09/10, Q3	0.1627 (6.11×10^{-3})	**0.1577** (5.54×10^{-3})	0.1582 (5.57×10^{-3})	0.1682 (5.84×10^{-3})
09/10, Q4	0.1089 (5.38×10^{-3})	0.1081 (5.10×10^{-3})	**0.1078** (5.14×10^{-3})	0.1114 (5.27×10^{-3})
10/11, Q1	0.2050 (3.55×10^{-3})	**0.2015** (3.95×10^{-3})	0.2027 (3.99×10^{-3})	0.2256 (2.53×10^{-3})
10/11, Q3	0.1586 (5.47×10^{-3})	**0.1558** (5.07×10^{-3})	0.1567 (5.10×10^{-3})	0.1653 (5.45×10^{-3})
10/11, Q4	0.1148 (5.36×10^{-3})	**0.1137** (5.17×10^{-3})	0.1147 (5.26×10^{-3})	0.1185 (5.41×10^{-3})

Table 9
PROPOSED vs *BASELINE* comparison. One-tailed *t*-test is used to compare the difference between the loss of individual forecasts ($N = 1190$).

	Diff	*t*	*p*-value
09/10, Q1	4.45×10^{-3}	2.22	0.013
09/10, Q3	5.03×10^{-3}	2.90	0.002
09/10, Q4	7.82×10^{-4}	0.78	0.217
10/11, Q1	3.49×10^{-3}	3.26	0.001
10/11, Q3	2.81×10^{-3}	2.77	0.003
10/11, Q4	1.06×10^{-3}	1.64	0.051

Table 10
Transition probabilities from event HINB in an imaginary game from the 2009/10 season between the Phoenix Suns (the home team) and the Golden State Warriors (the away team). Phoenix is trailing by 4 points with 500 s left in the fourth quarter. The transition probabilities for other events are negligible and are omitted to conserve space.

TEAM	APF	H2PA	H2PM	H3PA	H3PM	HTO	HPF
PROPOSED	0.159	0.260	0.267	0.119	0.081	0.098	0.014
BASELINE	0.145	0.261	0.276	0.120	0.077	0.104	0.014
RELFREQ	0.145	0.298	0.258	0.108	0.062	0.109	0.017

Table 11
Transition probabilities from event HINB in an imaginary game from the 2009/10 season between the Phoenix Suns (the home team) and the Golden State Warriors (the away team). Phoenix is trailing by 4 points with 10 s left in the fourth quarter. The transition probabilities for other events are negligible and are omitted to conserve space.

TEAM	APF	H2PA	H2PM	H3PA	H3PM	HTO	HPF
PROPOSED	0.418	0.105	0.102	0.184	0.141	0.046	0.005
BASELINE	0.145	0.261	0.276	0.120	0.077	0.104	0.014
RELFREQ	0.145	0.298	0.258	0.108	0.062	0.109	0.017

PROPOSED and *PROPOSED** models are better forecasters than the other two models practically over the entire season.

We pooled data from the 2009/10 and 2010/11 seasons and estimated the teams win percentage against an average opponent. Fig. 7 shows that *PROPOSED** model is a good estimator of the teams actual win percentages.

4.5. An illustrative example

We conclude this section with an illustrative example to show the difference between predictions of individual models. We used *PROPOSED, BASELINE*, and *RELFREQ* models trained on the 2008/9 season to simulate a fictional game between the Phoenix Suns as the home team and the Golden State Warriors as the away team. We used the teams' 2009/10 regular season summary statistics. During the 2009/10 campaign, the Phoenix Suns was the top scoring team in the league with the highest field goal percentage and 3-point field goal percentage, while the Golden State Warriors had committed the most personal fouls. Suppose that the home team is trailing by 4 points and has just started a new play with an inbound pass. Table 10 shows predicted transition probabilities if there are 500 s left in the fourth quarter, while Table 11 shows predictions if there are only 10 s left in the fourth quarter. Observe how *PROPOSED* model takes into account the current context of the game and, as the time is running out, increases the probability that the home team will try to reduce the deficit with a 3-point field goal. The model also increases the probability that the away team will commit a personal foul to prevent their opponents from scoring a field goal. The other two models predict a fixed probabilities in both situations, which undoubtedly reduces the credibility of the simulation and can affect the final outcome of the game. *BASELINE* model takes into account the characteristics of teams which can be seen in higher probabilities for the home team to score a filed foal, while *RELFREQ* model returns probabilities corresponding to a game between the average home team and the average away team.

5. Discussion and conclusion

Empirical evaluation shows that *PROPOSED* model exceeds the current state-of-the-art in terms of the accuracy of forecasting the winner and the credibility of the generated simulations.

The analysis of the results showed that the progression of a basketball game is, in a large part, a homogeneous process, with the exception of a few seconds at the beginning and, even more so, at the end of each quarter (at least in terms of modeling with use of the described state space). However, modeling these non-homogeneous parts significantly improves the quality of the probabilistic forecasts and produces simulations that better capture the dynamics of the evolution of a basketball game.

Although *PROPOSED* model is a good estimator of individual team's win percentages, it is also evident that it systematically overestimates weaker teams and underestimates stronger teams. The same result is observed in Štrumbelj and Vračar (2012), where it was hypothesized that this can be attributed to its homogeneity. In this work we tried to mitigate the model homogeneity by taking into account the current context of the game. Even though the result is slightly better, the bias of the model remains. This leads us to the conclusion that the summary statistics used do not describe well enough the teams' characteristics and more research to explain and overcome this flaw needs to be done.

Our methodology facilitates simple inclusion into any expert system and decision-making process that requires the ranking and evaluation of teams and players. For example, Dadelo et al. (2014) describe a multi-criteria decision-making system based on the TOPSIS approach to select the starting five of a basketball team. The win probabilities and other performance indicators produced by our simulations could be included as another truly objective criteria, thus complementing the functionality of the expert system.

It can also be used as a primary or supplementary source of predictions in systems involved in sports betting (Spann & Skiera, 2009; Sung & Lessmann, 2012). In particular, in systems for in-play betting, where probabilities need to be updated as the game evolves, and exotic bets on the number of particular events (rebounds, turnovers, personal fouls, etc.), which will occur within a specified period (in the third quarter, before the first points are scored, etc.) given the current game situation.

The proposed approach easily be extended to other sports. The main focus of our further work will be on generalizing the approach to other sports where play-by-play data are available and develop a method for automated construction of the state space. In the current form, our methodology requires the user to provide features that describe team characteristics. It would be practical to develop a method for automatic discovery of informative features using historical play-by-play data. In particular, for sports where a useful set of features has not been identified. Another path for further research is incorporating richer data that are now becoming more readily available. First, the data from optical player tracking technologies, such as shot locations, passing patterns, and player positioning, which would enable a more detailed simulation. And second, probabilistic forecasts from (online) bookmaker odds, which are arguably the most accurate source of predictions of sports outcomes, and would, because they contain most, if not all, publicly available information related to the outcome, increase the forecasting quality of the model. Finally, and although the current data do not support such a model, the ultimate goal of the research is to extend the model to the player level. In order to facilitate the implementation of such a system, it is necessary to identify appropriate set of features descriptive enough to allow modeling each individual players their own specific style of play.

References

Ballı, S., & Korukoğlu, S. (2014). Development of a fuzzy decision support framework for complex multi-attribute decision problems: a case study for the selection of skilful basketball players. *Expert Systems, 31*, 56–69.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 75*, 1–3.

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014). A multiresolution stochastic process model for predicting basketball possession outcomes. arXiv preprint arXiv:1408.0777. Retrieved from http://arxiv.org/abs/1408.0777

Dadelo, S., Turskis, Z., Zavadskas, E. K., & Dadeliene, R. (2014). Multi-criteria assessment and ranking system of sport team formation based on objective-measured values of criteria set. *Expert Systems with Applications, 41*, 6106–6113.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-setters as forecasters: the case of english football. *International Journal of Forecasting, 21*, 551–564.

Gabel, A., & Redner, S. (2012). Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports, 8*, 1–20.

Goldman, M., & Rao, J. M. (2012). Effort vs. concentration: the asymmetric impact of pressure on nba performance. In *Proceedings of the mit sloan sports analytics conference* (pp. 1–10).

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *Ijcai* (pp. 1034–1040).

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports, 3*, 1–22.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86.

Merritt, S., & Clauset, A. (2014). Scoring dynamics across professional team sports: tempo, balance and predictability. *EPJ Data Science, 3*, 1–21.

Oh, M.-h., Keshri, S., & Iyengar, G. (2015). Graphical model for basketball match simulation. In *MIT sloan sports analytics conference*.

Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis*. Potomac Books, Inc.

Papić, V., Rogulj, N., & Pleština, V. (2009). Identification of sport talents using a web-oriented expert system with a fuzzy module. *Expert Systems with Applications, 36*, 8830–8838.

Percy, D. F. (2015). Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes. *Journal of the Operational Research Society*, 1–10.

Shirley, K. (2007). A markov model for basketball. In *Proceedings of the New England symposium for statistics in sports*.

Song, C., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of american football games. *International Journal of Forecasting, 23*, 405–413.

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

Stekler, H., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting, 26*, 606–621.

Stern, H. S. (1994). A brownian motion model for the progress of sports scores. *Journal of the American Statistical Association, 89*, 1128–1134.

Sung, M.-C., & Lessmann, S. (2012). Save the best for last? the treatment of dominant predictors in financial forecasting. *Expert Systems with Applications, 39*, 11898–11910.

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research, 11*, 1–18.

Štrumbelj, E., & Vračar, P. (2012). Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting, 28*, 532–542.