Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

Pubblished: August 2023

Cleup sc
‘Coop. Libraria Editrice Università di Padova’
via G. Belzoni, 118/3 – Padova Italy
Phone +39 049 8753496 Fax +39 049 9865390
info@cleup.it – www.cleup.it – www.facebook.com/cleup

Summary

Statistica Applicata
ITALIAN JOURNAL OF APPLIED STATISTICS
Vol. 35, Number 2

doi.org/10.26398/IJAS.0035-007

THE RELATIONSHIP BETWEEN PLAYERS’ AVERAGE MARGINAL CONTRIBUTIONS AND SALARIES: AN APPLICATION TO NBA BASKETBALL USING THE GENERALIZED SHAPLEY VALUE

Francesco Biancalani[1],
Laboratory for the Analysis of CompleX Economic Systems (AXES), IMT School for Advanced Studies, Lucca, Italy
Giorgio Gnecco[2],
Laboratory for the Analysis of CompleX Economics Systems (AXES) and Game Science Research Center, IMT School for Advanced Studies, Lucca, Italy
Rodolfo Metulini[3],
Department of Economics (DSE), University of Bergamo, Bergamo, Italy

Abstract *Measuring players’ importance in basketball is allowed by many proposed advanced measures based on play-by-play data, such as the adjusted plus-minus, the wins above replacement, and the generalized Shapley value. In this paper we focus on the latter one in order to study whether, for a player, obtaining a large salary can be explained by his/her average marginal contribution to the team performance. In order to explore this issue, a linear regression model strategy where the logarithm of salary (Y) depends on the generalized Shapley value (X) is proposed and applied to players of selected National Basketball Association (NBA) teams over selected seasons. A leave-one-out cross validation shows that the accuracy in predicting whether free-agent players will obtain a more profitable contract solely basing on their generalized Shapley value is generally fairly good.*

Keywords: *Players’ performance, Salary, Sports statistics, Cooperative game theory, National Basketball Association.*

1. INTRODUCTION

To attribute the right salary to players in basketball, as well as in other professional sports, is a critical issue for the team managers, who generally have to

[1] francesco.biancalani@imtlucca.it

[2] giorgio.gnecco@imtlucca.it (Corresponding author)

[3] rodolfo.metulini@unibg.it

form the team with limited economic resources. A quite recent strand of research is dedicated to find whether features related to players' performance have the potential to explain players' salaries. According to Papadaki and Tsagris (2022), years of experience in the league and minutes played are the players' variables that mostly explain salaries.

A wide variety of synthetic indices have been proposed in the literature (in parallel with the above-mentioned studies on salaries), to measure each player's contribution to the team win. These include Plus-Minus (PM) and its generalizations (see, e.g., Kubatko et al. (2007)), Win-Shares (WS), Wins Above Replacement Player (WARP) and their advances (see, for a review, Sarlis and Tjortjis (2020)). Given that all such indices are "composite" measures that aim to evaluate the importance of a player (generally, in terms of winning the game) within his/her team, it is reasonable to think that a player may be rewarded with a salary that is proportional to the value assumed by one of such indices. Recently, Metulini and Gnecco (2023) developed a new measure of each player's value, based on a statistical and game theoretical approach. Such a measure adopts a combination of two-step logistic regression and the concept of generalized Shapley value (Nowak and Radzik, 1994), in order to determine players' values in their team. The reader is referred to Section 2.1 for a comparison of such a measure with the above-mentioned industry-standard measures.

The goal of this work is to explore the predictive power of the generalized Shapley value in explaining players' salaries. In doing so, we present an application to National Basketball Association (NBA) professional basketball league, where the law of "salary cap" imposes a constraint on the sum of players' salaries of a team and so the issue of attributing the right salary is a relevant aspect.

Schematically, our approach is based on two steps. In the first step, after having computed players' generalized Shapley values as in Metulini and Gnecco (2023), appropriate log-linear regression models (Christensen, 2006) are proposed to measure the role of players' generalized Shapley values to their salaries. In a second step, to validate our approach, just for "free-agents" (i.e., players that are not bound by a contract at the end of the season and are eligible to sign with other teams), deviations of estimated salaries (according to log-linear models) from the true salaries are analyzed with respect to the salaries obtained by such players in the following season. We show that the proposed instrument might be adopted by team managers for decision-making (especially during the summer trade season), as it allows predicting the deal of a good contract after the end of the current season based on the generalized Shapley value in that season.

Section 2 introduces the state of the art in this topic, Section 3 outlines the methodological approach, Section 4 is dedicated to present the adopted data. The application to a case study is discussed in Section 5, while Section 6 concludes.

2. STATE OF THE ART

2.1. MEASURES OF PLAYER'S IMPORTANCE

A large variety of industry-standard measures were developed in the literature to evaluate each single player's contribution to a team. These are based, for instance, on the difference between the points scored by a player's team and the ones scored by his/her opponent team during the time that specific player was on the court. Such measures include, e.g.:

- simple Plus-Minus (PM);
- regression-based versions of the PM metric that aim to measure each player's contribution by taking into account the other players on the court (Adjusted PM, or APM, see Rosenbaum (2004));
- modifications of the APM that include statistics of other players among the explanatory variables and that control also for the strength of the team (box-score PM, BPM, see, e.g., Grassetti et al. (2021); Ilardi (2007); Kubatko et al. (2007));
- other modifications of the APM that try to take into account the presence of multicollinearity in that measure (see, e.g., Engelmann (2017); Sill (2010)), and are based on ridge regression regularization (Regularized APM);
- Real Plus Minus (RPM), which normalizes the PM measure by taking into account the numbers of offensive and defensive possessions.

Overall, although recent PM versions moved in the direction of i) not just considering only scoring factors, and ii) solving for multicollinearity problems, those issues still need attention (Terner and Franks, 2021).

Other measures of players's importance, based on different approaches, are reported in the following list.

- Beside PM and its variations, Win-Shares (WS), computed by taking into account player, team and league statistics, attempts to measure the contribution for team success of its individuals. Moreover, WS48 (or WS per 48 minutes) expresses the WS values in a per-minute basis.

- Wins Above Replacement (WAR), also referred to as WAR Player (WARP), was firstly developed for baseball with the aim of measuring each player's contribution in terms of how many additional wins he/she brings to the team. Such a measure seeks to evaluate a player by contrasting the performance of a team which is made up of him/her and four average players with the one of a team which is composed of four average players and one replacement-level player. Nevertheless, as remarked in Sarlis and Tjortjis (2020), although WS has the advantage of being defined in terms of the marginal utility of a single player to the victory (by comparing him/her to an average replacement level), it turns out that a player's WS score is positively influenced both by being part of a good team and by the amount of time he/she is on the court. Moreover, WARP and WS48 outperform WS as they are expressed on a per-minute basis.

- Value Over Replacement Player (VORP), defined as an estimate of the points per 100 team possessions that a player contributed above a replacement level player, aims at combining the advantages of both BPM and WARP. However, in a similar way as WARP, VORP suffers from problems of multicollinearity (Sarlis and Tjortjis, 2020).

- Finally, the method developed in Metulini and Gnecco (2023) to measure players' contribution in a basketball team, which has already been mentioned in the Introduction, gathers most of the advantages (and reduces disadvantages) of the above-mentioned industry-standard measures. In fact, in a similar way as BPM, that new method presents the advantage of being based on both scoring and non scoring, offensive and defensive factors. Furthermore, it takes into account game winning probabilities, which are estimated – with an extremely high goodness of fit – based on a long time span of box-score synthetic measures (the so-called four Dean's factors, see Kubatko et al. (2007); Oliver (2004)). Moreover, similarly to what WARP and VORP do by introducing the replacement level player, the approach proposed in Metulini and Gnecco (2023) accounts for marginal utilities of players when considering lineups. This is achieved by accounting explicitly for all the lineups in which each player has played with (so, adopting a more "holistic" approach, which is the expression of a general solution concept coming from cooperative game theory). In doing this, there is no need for considering a proper level for the replacement player, and problems of multicollinearity are avoided (Mishra, 2016). Another advantage of this method is that the generalized Shapley value, on which it is based, has a well-known axiomatic characterization, which is expressed in terms of simple properties (Michalak et al., 2014). Such properties can be easily transferred to team sports and especially to basketball. Furthermore, in case one wants to include more features to increase the goodness of fit of the model to predict the winning probabilities, he/she has to change only the specific definition of the generalized characteristic function considered in the method, letting every other part of the method unaltered. It is worth remarking that the method proposed in Metulini and Gnecco (2023) presents also some limitations. In particular, in order to be estimated with high precision, the generalized Shapley value of a player needs the observation of a large number of different lineups containing that player. Indeed, the variance of its estimate is inversely proportional to the number of these lineups.

2.2. LITERATURE REVIEW ON SALARIES

An emerging stream of literature is focused on salaries not in conventional jobs where a PhD degree may represent a plus, but in a sport context. A vast part of this literature analyzes salaries in the baseball US major league, such as the seminal work by Scully (1974), and the works by Annala and Winfree (2011) and Holmes (2011). The relationship between players' performance and salaries is an emerging topic that has attracted the interest of numerous researchers (Garris and Wilkes, 2017; Vincent and Eastman, 2009; Wiseman and Chatterjee, 2010; Yilmaz and Chatterjee, 2003). The interest percolated to several Bachelor, Master and PhD students, as demonstrated by many theses about the aforementioned research question (Hentilä, 2019; Huang, 2016; Li, 2014; Zhu, 2019).

The general question of interest is whether players deserve their salaries based on their performance statistics. From a theoretical point of view, salaries should be equal to marginal contributions. For many reasons, in practice, this is often not the case. In part, one can expect salary to be explained by box-score and play-by-play features. Nevertheless, as a matter of fact, being able to correctly quantify how a player must be rewarded for his/her contribution to the team performance is a complex task that requires sophisticated techniques. This is mainly due to the presence of teammates and opponents.

Papadaki and Tsagris (2022) found, based on reviewing the state of the art on this topic, that the relationship between salary and player's performance is nonlinear. Hence, linear models are bound to fail in capturing the underlying true association (unless they contain, e.g., a final nonlinear transformation of the output). An additional concern, separate from nonlinearity, is model predictability, for which internal evaluation has limitations and leads to an over-optimistic estimate of the performance. Specifically for the NBA, Sigler and Sackley (2000) studied the task of salary prediction using data from the 1997-1998 season but with only three predictor variables: rebounds, assists and points per game. Ertug and Castellucci (2013) related the players' salaries to a set of predictor variables, most of which were not associated with the players' performance on court. Their data were gathered from the 1989-1990 up to the 2004-2005 period. More recently, Xiong et al. (2017) performed a similar analysis using more predictor variables measuring the players' performance on court for the 2013-2014 season. Sigler and Compton (2018) studied the 2017-2018 season but related the salaries to predictor variables exposing the players' abilities on court. Papadaki and Tsagris (2022) found, by using LASSO (Tibshirani, 1996), random forest and, as the response variable, the player's share of team's salary, that the most important variables in explaining player's salary are experience and minutes played, number of games played, points scored, defensive rebounds, and field goal attempts.

The relationship between salary and the generalized Shapley value has not been addressed yet in basketball or, to the best of our knowledge, in team sports in general. However, the aspect of correctly rewarding an individual based on his/her contribution to the team performance is addressed, e.g., in Yan et al. (2020), in terms of the (classical, i.e., not generalized) Shapley value, which has also several applications, e.g., in political science (such as measuring the power of parties, see Shapley and Shubik (1954)) and in machine learning (such as ranking features, see Štrumbelj and Kononenko (2014)). As a non-cooperative foundation of the Shapley value, it is also worth mentioning the classical game-theoretic model of bargaining between a firm and multiple employees considered in Stole and Zwiebel (1996). In that work, it was proved that workers' salaries and the firm's profit in the stable bargaining outcome of the model coincide with the respective Shapley values. Recently, Shapley values have been used also as an alternative to classical measures of importance (or centrality) of vertices and edges in graphs (see, e.g., Gnecco et al. (2019); Hadas et al. (2017); Michalak et al. (2013); Passacantando et al. (2021)). Such an approach could be used also in the context of basketball data analysis, by modeling basketball players as vertices of a suitable graph (constructed, e.g., as in Buldú et al. (2018) for the case of soccer).

3. METHODS

Loosely speaking, the generalized Shapley value of a player in a generalized coalitional game with *n* players represents his/her measure of importance in the team. This is expressed as his/her average marginal utility to a suitably randomly formed ordered coalition of players. It is similar to the well-known Shapley value for a coalitional game (Maschler et al., 2013), but it is based on a generalized characteristic function instead of a characteristic function (Michalak et al., 2014). Details about the specific definition of the generalized Shapley value are provided later in this section.

To obtain the generalized Shapley value for a player in the case of a basketball team, we adopt the following three steps strategy recently proposed in Metulini and Gnecco (2023).

1. The first step deals with computing the coefficients of a logit model (in the specific case, using data coming from all the NBA seasons between 2004-2005 and 2020-21). In the model, the dependent variable (called *Outcome*, win=1, defeat=0) expresses the result of the investigated team, whereas the explanatory variables are represented by suitable synthetic measures evaluated using play-by-play statistics related to both the teams participating in the game.

2. In the second step, the estimated coefficients of the logit model are exploited to express the winning probability associated with each lineup, which is later used to determine the value of the generalized characteristic function.

3. In the third step, one considers two different versions (*unweighted* and *weighted*) of the generalized characteristic function, hence of the generalized Shapley value for each player. As detailed in the Appendix, these two versions differ with respect to taking/not taking into account the amount of time players are on the court.

In the following, it is recalled from Nowak and Radzik (1994) that the generalized Shapley value (also known in the game-theoretical literature as Nowak-Radzik value) of player $i = 1, \dots, n$ in a generalized coalitional game is expressed by the next formula[4]:

[4] In the following, a similar notation as the one adopted in Michalak et al. (2014) is used. First, one denotes the elements of each ordered coalition $T \in \mathcal{T}$ as $T_1, \dots, T_{|T|}$. In this notation, the index refers to the order according to which a player enters the ordered coalition T. For simplicity, the ordered coalition which is made only by the element i is denoted by i itself. For every two disjoint ordered coalitions $T^{(1)}$ and $T^{(2)} \in \mathcal{T}$, $(T^{(1)}, T^{(2)})$ represents the ordered coalition constructed by the concatenation of $T^{(1)}$ and $T^{(2)}$, i.e., it is the ordered coalition in which all the elements of $T^{(1)}$ (which are ordered as in $T^{(1)}$) precede the ones belonging $T^{(2)}$ (which are ordered as in $T^{(2)}$).

$$\varphi_i^{NR}(N,\upsilon) = \frac{1}{n!} \sum_{T \in \mathcal{T} \text{ with } |T|=n} (\upsilon((T(i),i)) - \upsilon(T(i))). \quad (1)$$

In the above, $\mathcal{T}$ refers to the set of all ordered coalitions of players, $T(i)$ represents the ordered (sub)coalition made by the predecessors of i in the permutation T, whereas $(T(i),i)$ is the ordered (sub)coalition made by $T(i)$ followed by i. Finally, $\upsilon : \mathcal{T} \to \mathbb{R}$ (such that $\upsilon(\emptyset) = 0$) is called generalized characteristic function.

In their application of the generalized Shapley value to basketball data analysis, Metulini and Gnecco (2023) described two possible choices for the generalized characteristic function $\upsilon(.)$ appearing in Equation (1). They were denoted therein respectively by $\upsilon_1(.)$ and $\upsilon_2(.)$. The generalized characteristic function $\upsilon_1(.)$ is related to the probability $P(Win)$ of winning the game for every specific quintet (lineup) of players. Instead, the definition of the generalized characteristic function $\upsilon_2(.)$ takes into account not only the probability $P(Win)$ of winning the game for every specific quintet, but also the probability of occurrence $P(Occ)$ of that quintet on the court. More details about the definitions of the two generalized characteristic functions $\upsilon_1(.)$ and $\upsilon_2(.)$ and about an approximate method for computing the corresponding generalized Shapley values are provided in the Appendix. In the following, such generalized Shapley values are called, respectively, unweighted generalized Shapley value of a player (UWGS), and generalized Shapley value of a player (WGS).

4. DATA

Data to compute the UWGS and WGS values were extracted from the play-by-play of all NBA games (both regular seasons and play-offs were considered). These data were made available to us thanks to a friendly agreement with BigDataBall Company (UK) (`www.bigdataball.com`). BigDataBall collected and provided us with the play-by-play of all the NBA regular season and play-off games for all the seasons from 2004/2005 to 2020/2021 (for a total of 17 seasons). For each game and for both home and away teams, the available data include detailed information about the type of each event (e.g., start/end of the period, made/missed 2 points shot, made/missed 3 points shot, made/missed free throw, offensive/defensive rebound, assist, steal, block, foul), the precise moment in which that event occurs, and also the associated lineups of both the two teams. When the event refers to a shot (made or missed) we also have at our disposal the position on the court, expressed in terms of $x-$axis and $y-$axis coordinates.

These are respectively related to court length and court width. Information on player's income was recovered from the website `basketballinsiders.com`, which represents one of the top online newspapers on the NBA. Finally, the values of players' performance (WS, WS48, VORP48, BPM) that are used in this work for comparison purposes were retrieved from the website `www.basketball-reference.com`.

5. APPLICATION

Adopting the strategy reported in Section 3, we compute the value assumed by the winning probability for each lineup (considering both regular season and playoffs). Then, we determine the estimates of the UWGS and WGS values, as in Equation (8) in the Appendix, for all the players of three teams (Milwalkee Bucks, Phoenix Suns, and Utah Jazz) for the seasons 2019/20 and 2020/21[5].

The dataset is reported in Table 1. The table reports not only the UWGS and WGS values for each of the 73 considered players (of the three teams in the two seasons), but also the name of the team associated with each player, the salary (in dollars) received by the player in the current season (t) and in successive season ($t+1$), and the information on the free agency status of that player at the end of the current season.

It is worth remarking that data on salaries could not always match the ones reported in a successive updated version of the website. In particular, we retrieved such data before the end of the 2021/2022 season. Players with unguaranteed contracts might have increased their salaries during that season. Finally, data on other websites may differ because of another way of expressing salaries.

[5] We have decided to analyze close-by seasons in such a way that our results are not affected by the average increase of salaries. The choice of these three teams is motivated by the need of considering teams having similar strength, since generalized Shapley values of players coming from teams with different strength are not comparable. Bucks concluded seasons 2019/20 and 2020/21, respectively, with a record of 56-17 (1^{st} in the Eastern Conference) and 46-26 (3^{rd} in the Eastern Conference). Suns concluded seasons 2019/20 and 2020/21, respectively, with a record of 34-39 (10^{th} in the Western Conference, but with a strong improvement at the end of the season) and 51-21 (2^{nd} in the Western Conference). Jazz concluded seasons 2019/20 and 2020/21, respectively, with a record of 44-28 (6^{th} in the Western Conference) and 52-20 (1^{st} in the Western Conference).

Player	salary (dollars)$_t$	UWGS	WGS ($\times$100)	season	team	salary (dollars)$_{t+1}$	free agent
Brook Lopez	12,093,024	0.0415	0.0340	2019/20	MIL	12,697,675	No
Giannis Antetokounmpo	25,842,697	0.0376	0.0346	2019/20	MIL	27,528,088	No
Eric Bledsoe	15,625,000	0.0440	0.0376	2019/20	MIL	16,875,000	No
Khris Middleton	30,603,448	0.0394	0.0316	2019/20	MIL	33,051,724	No
Wesley Matthews	2,564,753	0.0398	0.0376	2019/20	MIL	3,623,000	Yes
Donte DiVincenzo	2,905,800	0.0234	0.0124	2019/20	MIL	3,044,160	No
George Hill	9,133,907	0.0286	0.0101	2019/20	MIL	6,109,082	No
Robin Lopez	4,767,000	0.0677	0.0125	2019/20	MIL	7,300,000	Yes
Ersan Ilyasova	7,000,000	0.0072	0.0241	2019/20	MIL	1,194,542	No
Marvin Williams	604,278	0.0428	0.0022	2019/20	MIL	retired	Yes
Pat Connaughton	1,723,050	0.0418	0.0128	2019/20	MIL	4,938,273	Yes
Kyle Korver	1,620,564	0.0601	0.0246	2019/20	MIL	retired	Yes
Giannis Antetokounmpo	27,528,088	0.0370	0.0407	2020/21	MIL	39,344,900	No
Brook Lopez	12,697,675	0.0326	0.0352	2020/21	MIL	13,302,325	No
Khris Middleton	33,051,724	0.0395	0.0439	2020/21	MIL	35,500,000	No
Donte DiVincenzo	3,044,160	0.0364	0.0397	2020/21	MIL	4,675,830	No
Jrue Holiday	25,876,111	0.0403	0.0451	2020/21	MIL	32,431,333	No
P. J. Tucker	7,969,537	0.0408	0.0453	2020/21	MIL	7,000,000	Yes
Bryn Forbes	2,337,145	0.0252	0.0232	2020/21	MIL	4,500,000	Yes
Pat Connaughton	4,938,273	0.0410	0.0442	2020/21	MIL	5,333,334	No
Thanasis Antetokounmpo	1,701,593	0.0909	0.0709	2020/21	MIL	1,729,217	Yes
Bobby Portis	3,623,000	0.0500	0.0445	2020/21	MIL	4,347,600	Yes
D. J. Augustin	2,694,064	0.0545	0.0496	2020/21	MIL	7,000,000	No
Devin Booker	27,285,000	0.0408	0.0245	2019/20	PHX	29,467,800	No
Tyler Johnson	19,245,370	0.0280	0.0121	2019/20	PHX	2,028,594	No
Ricky Rubio	16,200,000	0.0418	0.0255	2019/20	PHX	17,000,000	No
Kelly Oubre Jr.	15,625,000	0.0442	0.0265	2019/20	PHX	14,375,000	No
Deandre Ayton	9,562,920	0.0393	0.0216	2019/20	PHX	10,018,200	No
Aron Baynes	5,453,280	0.0515	0.0368	2019/20	PHX	7,000,000	Yes
Frank Kaminsky	4,767,000	0.0226	0.0110	2019/20	PHX	1,620,564	Yes
Mikal Bridges	4,161,000	0.0431	0.0228	2019/20	PHX	4,359,000	No
Cameron Johnson	4,033,440	0.0256	0.0138	2019/20	PHX	4,235,160	No
Dario Saric	3,481,986	0.0356	0.0221	2019/20	PHX	9,250,000	Yes
Jevon Carter	1,416,852	0.0278	0.0124	2019/20	PHX	3,925,000	Yes
Elie Okobo	1,416,852	0.0450	0.0176	2019/20	PHX	other league	Yes
Cameron Payne	196,288	0.0000	0.0000	2019/20	PHX	1,977,011	No
Chris Paul	41,358,814	0.0335	0.0345	2020/21	PHX	30,800,000	Yes
Devin Booker	29,467,800	0.0340	0.0301	2020/21	PHX	31,650,600	No
DeAndre Ayton	10,018,200	0.0340	0.0320	2020/21	PHX	12,632,950	No
Jae Crowder	9,258,000	0.0347	0.0364	2020/21	PHX	9,720,900	No
Dario Saric	9,250,000	0.0371	0.0131	2020/21	PHX	8,510,000	No
Mikal Bridges	4,359,000	0.0350	0.0329	2020/21	PHX	5,557,725	No
Jalen Smith	4,245,720	0.0000	0.0000	2020/21	PHX	4,458,000	No
Cameron Johnson	4,235,160	0.0346	0.0146	2020/21	PHX	4,437,000	No
Jevon Carter	3,925,000	0.0286	0.0072	2020/21	PHX	3,650,000	No
Cameron Payne	1,977,011	0.0353	0.0195	2020/21	PHX	6,500,000	Yes
Abdel Nader	1,752,950	0.0000	0.0000	2020/21	PHX	2,000,000	Yes
Frank Kaminsky	1,620,564	0.0212	0.0230	2020/21	PHX	2,089,448	Yes
Langston Galloway	1,620,564	0.0054	0.0024	2020/21	PHX	257,418	Yes
E'Twaun Moore	1,620,564	0.0261	0.0060	2020/21	PHX	2,641,691	Yes
Torrey Craig	1,620,564	0.0340	0.0129	2020/21	PHX	1,654,051	Yes
Rudy Gobert	25,008,427	0.0454	0.0331	2019/20	UTA	27,525,281	No
Royce O'Neale	1,618,520	0.0463	0.0342	2019/20	UTA	8,500,000	No
Donovan Mitchell	3,635,760	0.0458	0.0365	2019/20	UTA	5,195,501	No
Mike Conley	32,511,624	0.0462	0.0366	2019/20	UTA	34,502,132	No
Bojan Bogdanovic	17,000,000	0.0467	0.0353	2019/20	UTA	17,850,000	No
Joe Ingles	11,954,546	0.0438	0.0317	2019/20	UTA	10,863,637	No
Jordan Clarkson	13,437,500	0.0328	0.0146	2019/20	UTA	11,500,000	Yes
Tony Bradley	1,962,360	0.0133	0.0171	2019/20	UTA	3,542,060	No
Emmanuel Mudiay	1,620,564	0.0414	0.0245	2019/20	UTA	other league	Yes
Jeff Green	1,620,564	0.0422	0.0110	2019/20	UTA	2,564,753	No
Georges Niang	1,645,357	0.0251	0.0061	2019/20	UTA	1,783,557	No
Rudy Gobert	27,525,281	0.0514	0.0526	2020/21	UTA	35,344,828	No
Royce O'Neale	8,500,000	0.0447	0.0420	2020/21	UTA	8,800,000	No
Donovan Mitchell	5,195,501	0.0453	0.0441	2020/21	UTA	28,103,500	No
Mike Conley	34,502,132	0.0583	0.0635	2020/21	UTA	21,000,000	Yes
Bojan Bogdanovic	17,850,000	0.0439	0.0418	2020/21	UTA	18,700,000	No
Joe Ingles	10,863,637	0.0517	0.0454	2020/21	UTA	13,036,364	No
Jordan Clarkson	11,500,000	0.0385	0.0323	2020/21	UTA	12,420,000	No
Derrick Favors	9,258,000	0.0338	0.0242	2020/21	UTA	9,720,900	Yes
Miye Oni	1,517,981	0.0001	0.0000	2020/21	UTA	799,106	No
Trent Forrest	470,690	0.0000	0.0000	2020/21	UTA	8,558	No
Georges Niang	1,783,557	0.0506	0.0526	2020/21	UTA	3,300,000	Yes

Table 1: Unweighted and weighted generalized Shapley values for the 73 considered players along with name of the team, season, salary (in dollars) of the current season (t) and successive season ($t+1$), information on the free agency status.

5.1. RELATIONSHIP BETWEEN SALARY AND GENERALIZED SHAPLEY VALUE

In order to evaluate the relationship between salary and the unweighted and weighted generalized Shapley value, we run several linear regression models with ordinary least squares (OLS) estimation, in which a suitably normalized (or rescaled) generalized Shapley value is one of the explanatory variables, and the natural logarithm of salary is the dependent variable. It is worth saying that, in all these regressions, the UWGS (and WGS) values are normalized in such a way that the summation of such normalized values of all the players of a single team in a single season is equal to 1. Indeed, despite the three teams have been chosen of similar strength, by doing this way, we further ease the comparison of the (normalized) UWGS or WGS values of players belonging to different teams and/or playing in different seasons, and we are better allowed to estimate regression models with players coming from different teams or evaluated in different seasons. It is also worth mentioning that a logarithmic transformation of the dependent variable is adopted here in order to ease the classical assumption of residual normality of OLS to be guaranteed. As discussed in Section 2.2, the state of the art of the analysis of the relationship between salaries and players' performance considers this relationship to be nonlinear. So, nonlinear models should be used in principle. However, as far as the classical assumptions for a linear model are guaranteed, we believe that the use of OLS is appropriate for an exploratory study as ours.

First, the battery of linear regressions presented in Table 2 shows the dependence of the natural logarithm of salary on the normalized UWGS value. By using the full dataset with players belonging to both seasons and all teams, we first consider a simple linear regression model in which the normalized UWGS value is the only explanatory variable, then we also control (by means of a set of dummy variables) for the average level of salaries in different teams and/or different seasons. It looks like the goodness of fit (expressed in terms of R^2) is not so high. The intercept and the UWGS coefficients are, anyways, always significant. The coefficients associated with dummy variables are not significant.

The results reported in Table 3 are related to the regressions where the interaction of the normalized UWGS value with variables team and season is considered in a unique solution that preserves the degrees of freedom (dummy variables are not considered). All in all, the goodness of fit does not change considerably by including interaction terms. Moreover, the interaction effects of the normalized UWGS value with team and with season, as displayed in columns 2–4 of Table 3, turn out to be not significantly different from zero. Overall, by looking at the results reported in both Table 2 and Table 3, we do not reject, as the best model, the null model with just the main effect of the normalized UWGS value.

The battery of linear regressions presented in Table 4 shows the dependence of salary on the normalized WGS value. By using the full dataset with players belonging to both seasons and all teams, we first consider a simple linear regression model in which the normalized WGS value is the only explanatory variable, then we also control (by means of a set of dummy variables) for the average level of salaries in different teams and/or different seasons. The goodness of fit in each of these cases is larger than the one obtained in each of the corresponding regression models in which the normalized UWGS value is used instead of the normalizedWGS value. The intercept and the WGS coefficients are also always significant. The coefficients associated with dummy variables are not significant also in the WGS case.

The results reported in Table 5 are related to the regressions where the interaction of the normalized WGS value with variables team and season is considered in a unique solution that preserves the degrees of freedom. Overall, the goodness of fit does not increase considerably with the inclusion of the interaction terms. About the interaction effects of the normalized WGS value with team and with season, according to the results reported in columns 2–4 of Table 5, we can see that the related coefficients are not significantly different from zero. All in all, based on the results listed in both Table 4 and Table 5, even in the case of the normalized WGS value, we do not reject the null hypothesis that the best model is the one with just the main effect of the normalized WGS value as explanatory variable. Moreover, due to the higher R^2, we retain that the WGS measure is a better predictor for the salary, if compared to UWGS. From now on, we consider the model with just the main effect of the normalized WGS value (first regression of Table 4) as the best model.

Table 6 displays the Pearson correlation coefficient among various players' performance measures reported in the specialized literature – including the normalized WGS value – and the logarithm of salary. More precisely, in the table, the correlations of the normalized WGS value, WS, WS48, VORP48, and BPM with the logarithm of salary are reported. We can notice that, according to the Pearson correlation, WS and VORP48 are comparable to the normalized WGS value. The WS48 and BPM measures have low association with salaries. For the case of BPM, the β coefficient is only slightly significant.

Overall, we are confident that our sample of 73 observations (with a number of degrees of freedom always larger than 65) is large enough to use R^2 as an accurate goodness-of-fit measure. However, in future developments of this study, especially in case of analyses in which one would be forced to work with really small samples, we may consider to use nonparametric tests for testing the goodness of fit and the significance of the regression coefficients in univariate or multivariate models. These include the nonparametric testing method to detect possible causal effects in the case of bivariate regression models (Bonnini and Cavallo, 2021) and the combined permutation test proposed in Bonnini and Borghesi (2022).

In order to further explain the obtained outcomes of our analysis, let us consider as an example the regression containing just the intercept and the normalized WGS value as explanatory variable, which is the one whose results are displayed in the second column of Table 4. The estimated β coefficient for the normalized WGS value stands to 15.420. If one wants to interpret the coefficients of the regression results, he/she has to take into consideration that a β coefficient related to the normalized WGS value equal to 15.420 quantifies the increase in the prediction of the natural logarithm of the salary by an increase in the normalized WGS value equal to 1. Now, let us consider a player whose normalized WGS value is equal to the median across the full sample, which is 0.0822. An increase in the normalized WGS value of that player to the value corresponding to the third quartile (0.1165) explains an average increase in the salary of 3,801,409 dollars. Similarly, a decrease in the normalized WGS value of that player to the value corresponding to the first quartile (0.0499) explains an average decrease in the salary of 2,139,040 dollars.

Moreover, to justify the use of a linear modeling strategy, diagnostic checks on the residuals have been made. Considering for instance the first regression of Table 4 (which we consider as the best model), its Q-Q plot, displayed in the left chart of Figure 1, does not exclude the validity of the normality assumption. In support to this evidence, both Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests have been performed on the same regression. Results, displayed in Table 7, provide evidences against the rejection of the null hypothesis of a normal distribution.

Moreover, the plot of residuals versus fitted values (right chart of Figure 1) displays the absence of residuals' heterogeneity and the absence of correlation between residuals and explanatory variables. In light of these results, the linear regression model assumptions turn out to be satisfied.

The use of the player's share of team's salary as the dependent variable of the regression, as suggested by Papadaki and Tsagris (2022) and of the log of the player's share of team's salary * 100, have been also tried, but no particular differences with respect to the previous results have been found.

Variables	ln(salary)	ln(salary)	ln(salary)	ln(salary)
Norm. UWGS	11.171**	11.232**	11.005**	11.051**
	(3.786)	(3.808)	(3.964)	(3.987)
Team:PHX	-	-	-0.078	-0.084
	-	-	(0.325)	(0.327)
Team:UTA	-	-	-0.047	-0.050
	-	-	(0.339)	(0.340)
Season:2020/21	-	0.130	-	0.132
	-	(0.263)	-	(0.267)
intercept	14.593***	14.522***	14.651***	14.583***
	(0.338)	(0.369)	(0.418)	(0.442)
R^2	0.109	0.112	0.110	0.113
Observations	73	73	73	73

Note: $^{\cdot}$ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 2: Ordinary least squares (OLS) log-in-linear regressions, full sample. “Normalized” unweighted generalized Shapley (UWGS).

Variables	ln(salary)	ln(salary)	ln(salary)
Norm. UWGS	9.094*	10.382*	8.313·
	(4.313)	(4.215)	(4.699)
Norm. UWGS:PHX	3.057	-	3.085
	(3.757)	-	(3.780)
Norm. UWGS:UTA	3.477	-	3.460
	(3.523)	-	(3.544)
Norm. UWGS:2020/21	-	1.293	1.291
	-	(2.964)	(2.984)
intercept	14.585***	14.605***	14.596***
	(0.348)	(0.341)	(0.351)
R^2	0.124	0.112	0.126
Observations	73	73	73

Note: $^{\cdot}$ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 3: Ordinary least squares (OLS) log-in-linear regressions, full sample. “Normalized” unweighted generalized Shapley (Norm. UWGS). Interaction terms with season and team variables.

Variables	ln(salary)	ln(salary)	ln(salary)	ln(salary)
Norm. WGS	15.420***	15.463**	15.495***	15.525***
	(2.682)	(2.695)	(2.777)	(2.791)
Team:PHX	-	-	-0.008	-0.016
	-	-	(0.283)	(0.284)
Team:UTA	-	-	-0.065	-0.068
	-	-	(0.296)	(0.298)
Season:2020/21	-	0.139	-	0.139
	-	(0.230)	-	(0.234)
intercept	14.244***	14.170***	14.260***	14.191***
	(0.249)	(0.278)	(0.318)	(0.340)
R^2	0.318	0.321	0.318	0.322
Observations	73	73	73	73

Note: $^{\cdot}$ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 4: Ordinary least squares (OLS) log-in-linear regressions, full sample. "Normalized" weighted generalized Shapley (Norm. WGS).

Variables	ln(salary)	ln(salary)	ln(salary)
Norm. WGS	15.135***	15.171***	14.909***
	(3.340)	(3.027)	(3.608)
Norm. WGS:PHX	0.515	-	0.492
	(3.135)	-	(3.160)
Norm. WGS:UTA	0.332	-	0.313
	(3.052)	-	(3.076)
Norm. WGS:2020/21	-	0.455	0.437
	-	(2.493)	(2.532)
intercept	14.244***	14.246***	14.246***
	(0.253)	(0.250)	(0.255)
R^2	0.318	0.318	0.319
Observations	73	73	73

Note: $^{\cdot}$ p<0.1; * p<0.05; ** p<0.01; *** p<0.001

Table 5: Ordinary least squares (OLS) log-in-linear regressions, full sample. "Normalized" weighted generalized Shapley (Norm. WGS). Interaction terms with season and team variables.

Variables	ln(salary)
Norm. WGS	0.564
WS	0.589
WS48	0.388
VORP48	0.593
BPM	0.238

Table 6: Pearson correlations between ln(salary) and "normalized" weighted generalized Shapley (Norm. WGS), Win-share (WS), Win-share per 48 minutes (WS48), Value over replacement player per 48 minutes (VORP48), Box plus-minus (BPM). Full sample (n=73).

Test	Statistic	p-value
Shapiro-Wilk	W = 0.9774	0.2084
Kolmogorov-Smirnof	D = 0.0648	0.8996
Anderson-Darling	A = 0.4216	0.3147

Table 7: Tests for the normality assumption of residuals, based on the estimated residuals of the first regression of Table 4.

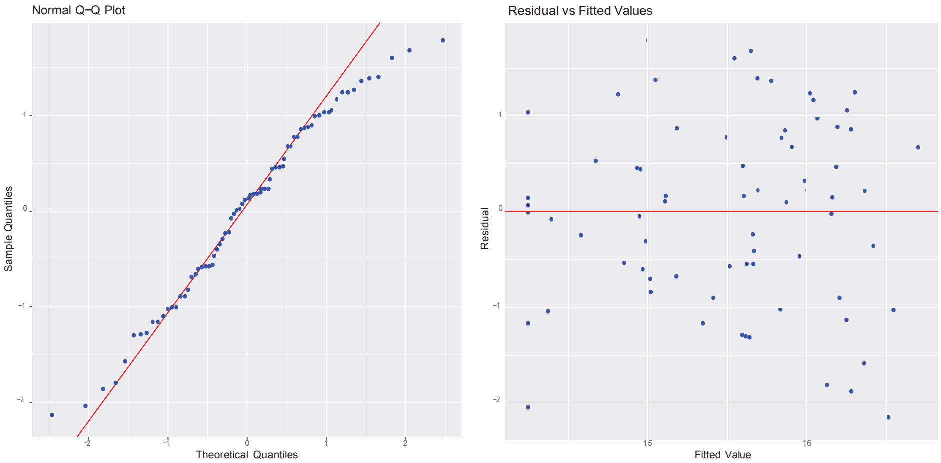


Figure 1: Q-Q plot for normality of residuals (left). Plot of the residuals versus fitted values (right). First regression of Table 4.

5.2. ACCURACY EVALUATION

In this section we aim at evaluating the accuracy of our regression strategy based on using either the unweighted or weighted (and normalized) generalized Shapley value as explanatory variable. We do so by studying the performance in predicting for which free agent players the salary will increase, by analyzing the deviations of the true salaries from those estimated by the model. By looking at the scatterplot reported in Figure 2, we aim at finding those players whose estimated salary – based on the normalized unweighted generalized Shapley value (top chart) or the normalized weighted generalized Shapley value (bottom chart) – is larger (smaller) compared to their actual salary. Specifically, if the point related to that player in that season is below the regression line of his/her team in that season (according to the Y-axis), then the estimated salary of that player in that season is larger than the actual salary. On the contrary, if the point related to that player in that season is above the regression line of his/her team in that season (according to the Y-axis), then the estimated salary of that player in that season is lower than the actual salary. Table 8 reports the list of players of the considered teams/seasons who were free-agents at the end of the season. For each one, in the fourth and the fifth column we report the information whether the estimated salary ($salary_t$), according, respectively, to the normalized UWGS value and to the normalized WGS value, was larger than the actual salary ($salary_t$). In estimating the salary of these players, we adopt a leave-one-out cross validation (LOOCV) strategy (Hastie et al., 2009) in which in each step of the process the model is trained on all the observations but the one related to a specific player/year, where that observation represents the one to be validated. To perform our LOOCV we use, in the two cases, the normalized UWGS value along with the model with just the main effect for the normalized UWGS value, whose results are displayed in second column of Table 2, and the normalized WGS value along with the model with just the main effect for the normalized WGS value, whose results are displayed in the second column of Table 4. The choice of LOOCV instead of a k-fold cross validation or a leave-p-out cross validation is motivated by the moderate dimension of our sample. In the sixth column of Table 8 it is reported whether the player's salary after the free agency ($salary_{t+1}$, i.e., in the successive season) was larger than the current salary ($salary_t$) increased by the 4%. Table 9 and Table 10 report, by team and over the full sample, the confusion matrix obtained by crossing the two information, respectively for the case of the normalized UWGS value and of the normalized WGS value. The *hit rate* (HR, Bensic et al. (2005)) of these confusion matrices is in all cases quite high.

This highlights that, by evaluating players in terms of how their estimated salary deviates from the actual value, it is possible to predict (with fairly good accuracy) whether a free agent will obtain a new more profitable contract or not just on the basis of his/her normalized (unweighted or weighted) generalized Shapley value. Interestingly, the hit rates obtained for the UWGS case are higher than or equal to the ones obtained for the WGS case. This might be explained by the fact that (except for few cases) the managerial staff would like to test free agents' abilities before employing them on a regular basis, so they are expected to be part of a long-lasting lineup less often than other players. In this case, their normalized UWGS value may better evaluate their ability (then predict their future salary) with respect to their normalized WGS value.

6. DISCUSSION AND CONCLUSIONS

This work belongs to the stream of literature whose aim is to study how the salaries are linked with the marginal utilities of people within a group. From a quite theoretical economical point of view, salaries should be equal to marginal contributions. For many reasons, in practice this is often not the case, and some deviations are observed (although the presence of a significant positive correlation between salaries and marginal contributions is still expected). The reasons of that are several, e.g.: the presence of agreements with trade unions, market imperfection, moral hazard, asymmetric information. In general, salaries are determined ex-ante, but the outcome due to a person's behavior will appear only ex-post. Salaries in team sports form a rather peculiar case, since contracts are quite short (they last typically no more than 4 years, sometimes 1 year only) and bargaining is common. Often in sport disciplines, players are mostly paid through a relevant fixed salary which is determined in advance. Nevertheless, a variable part of the salary (linked, e.g., to personal performance or to team performance) is allowed, even though this is not very common. It is worth mentioning that, in sport disciplines, one portion of the marginal contribution of each team member might appear relatively simple to quantify through the box-score and the play-by-play features. As a matter of facts, this is not the case, mainly because of the presence of the other teammates (and opponents), that makes the correct quantification of the marginal contribution of the player a rather complex issue.

[6] According to the following source : https://runrepeat.com/salary-analysis-in- the-nba-1991-2019, the average salary increased by about 4% between season 2019/20 and season 2020/21.

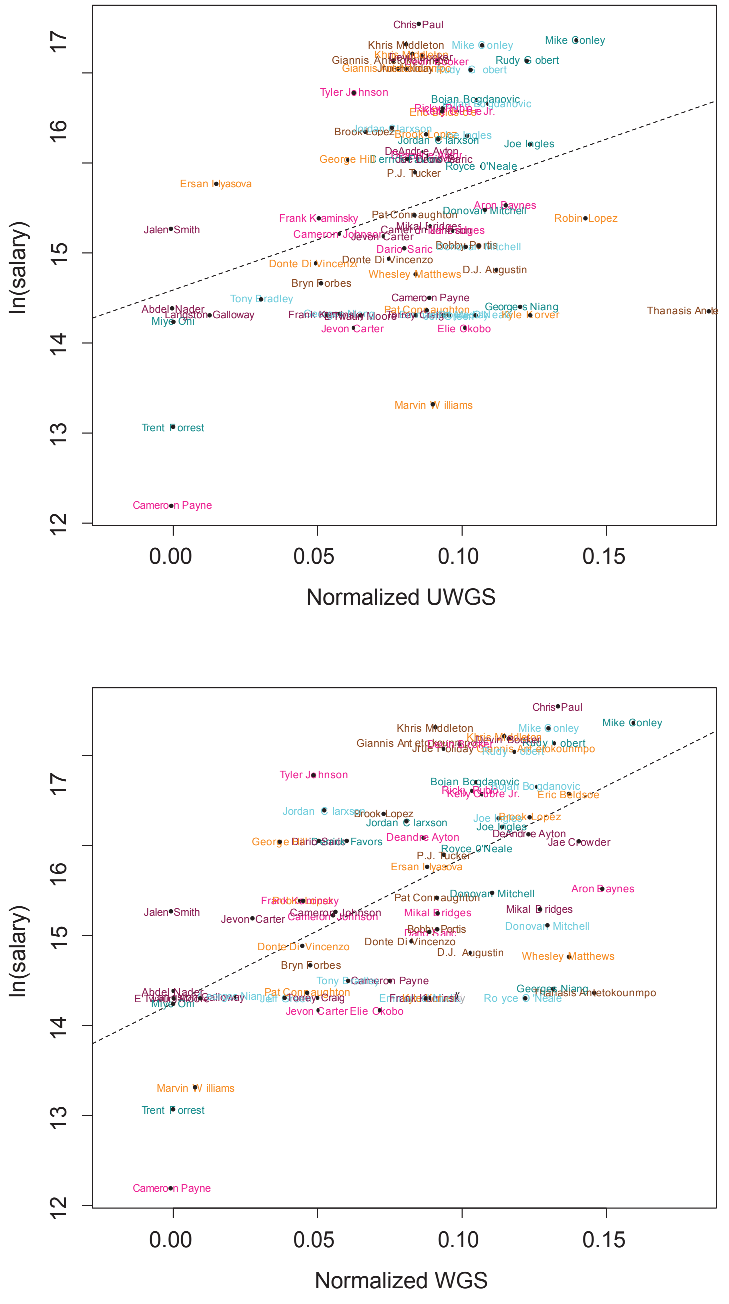


Figure 2: Scatter plot of the natural logarithm of salary (*ln(salary)*, on the Y-axis) and, on the X- axis: (top chart) the *normalized UWGS value*, with regression lines according to the regression in the second column of Table 2; (bottom chart) the *normalized WGS value*, with regression lines according to the regression in the second column of Table 4.

Player	team	season	$\hat{salary}_t > salary_t$ (Norm. UWGS)	$\hat{salary}_t > salary_t$ (Norm. WGS)	$salary_{t+1}$ $>1.04{*}salary_t$
Wesley Matthews	MIL	19/20	Yes	Yes	Yes
Robin Lopez	MIL	19/20	No	No	Yes
Marvin Williams	MIL	19/20	Yes	Yes	(retired)
Pat Connaughton	MIL	19/20	Yes	Yes	Yes
Kyle Korver	MIL	19/20	Yes	Yes	(retired)
P. J. Tucker	MIL	20/21	No	No	No
Bryn Forbes	MIL	20/21	Yes	Yes	Yes
Thanasis Antetokounmpo	MIL	20/21	Yes	Yes	No
Bobby Portis	MIL	20/21	Yes	Yes	Yes
Aron Baynes	PHX	19/20	Yes	Yes	Yes
Frank Kamisnky	PHX	19/20	No	No	No
Dario Saric	PHX	19/20	Yes	Yes	Yes
Jevon Carter	PHX	19/20	Yes	Yes	Yes
Elie Okobo	PHX	19/20	Yes	Yes	No
Chris Paul	PHX	20/21	No	No	No
Cameron Payne	PHX	20/21	Yes	Yes	Yes
Abdel Nader	PHX	20/21	Yes	No	Yes
Frank Kamisnky	PHX	20/21	Yes	Yes	Yes
Langston Galloway	PHX	20/21	Yes	Yes	No
E'Twaun Moore	PHX	20/21	Yes	No	Yes
Torey Craig	PHX	20/21	Yes	Yes	No
Jordan Clarxson	UTA	19/20	No	No	No
Emmanuel Mudiay	UTA	19/20	Yes	Yes	No
Mike Conley	UTA	20/21	No	No	No
Derrick Favors	UTA	20/21	No	No	Yes
Georges Niang	UTA	20/21	Yes	Yes	Yes

Table 8: List of free agents along with information on team, season, whether $\hat{salary}_t > salary_t$ according to the leave-one-out cross validation with the model in the second column of Table 2 (fourth column) and to the leave-one-out cross validation with the model in the second column of Table 4 (fifth column), and whether $salary_{t+1} > 1.04 * salary_t$.

Phoenix Suns	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ "Yes"	7	0
$salary_{t+1} > 1.04 * salary_t$ = "No"	3	2
Hit Rate = 0.750		
Milwaukee Bucks	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	4	1
$salary_{t+1} > 1.04 * salary_t$ = "No"	1	1
Hit Rate = 0.714		
Utah Jazz	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	2	1
$salary_{t+1} > 1.04 * salary_t$ = "No"	1	1
Hit Rate = 0.600		
Full sample	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	12	2
$salary_{t+1} > 1.04 * salary_t$ = "No"	6	4
Hit Rate = 0.708		

Table 9: Confusion matrices for free agents of Milwaukee Bucks (MIL), Phoenix Suns (PHX), and Utah Jazz (UTA). Normalized UWGS.

Phoenix Suns	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ "Yes"	5	2
$salary_{t+1} > 1.04 * salary_t$ = "No"	3	2
Hit Rate = 0.583		
Milwaukee Bucks	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	4	1
$salary_{t+1} > 1.04 * salary_t$ = "No"	1	1
Hit Rate = 0.714		
Utah Jazz	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	1	1
$salary_{t+1} > 1.04 * salary_t$ = "No"	1	2
Hit Rate = 0.600		
Full sample	$\widehat{salary}_t > salary_t$ = "Yes"	$\widehat{salary}_t > salary_t$ = "No"
$salary_{t+1} > 1.04 * salary_t$ = "Yes"	10	4
$salary_{t+1} > 1.04 * salary_t$ = "No"	5	5
Hit Rate = 0.625		

Table 10: Confusion matrices for free agents of Milwaukee Bucks (MIL), Phoenix Suns (PHX), and Utah Jazz (UTA). Normalized WGS.

In this work we have focused on the player's average marginal contribution formalized by the generalized Shapley value because, by separately considering all the different lineups in which the player has played with, it is based on an "holistic" approach which is the expression of a general solution concept coming from cooperative game theory. The player's salary should be quite related to the generalized Shapley value: A higher generalized Shapley value should be associated with a higher salary.

Our findings are in line with these theoretical arguments, as demonstrated by the regressions performed using both the unweighted and weighted (and suitably normalized) generalized Shapley value reported in Tables 2, 3, 4 and 5, for which the coefficients associated with the main effect of the normalized generalized Shapley value are always positive and statistically significant.

A limitation of our analysis may reside on the fact that players might have signed their contract few seasons before based on their performance in past years. It is also worth noting that contract rules are rather complex in NBA. Said that, a player may not currently have the salary he/she deserves based on his/her current performance. By analyzing deviations from the model, we turn the above-mentioned limitation into an advantage, as the proposed approach may be used by the player's manager to realize if the current remuneration can be increased and, on other side, by the team managers, to avoid less strong players to receive too high salaries.

As future work we would like to include constraints on the players' roles to define their generalized Shapley values, then investigate their relationship with salaries. Moreover, it would be worth studying the distribution of generalized Shapley values inside a team, then relate such a distribution with team performance. Finally, it could be interesting to extend the available dataset to include the possibility of computing and exploiting different movement-related features – associated, e.g., with spacing (Metulini et al., 2018), with the cohesion of a group of people (Glowinski et al., 2015), or with the origin of movement in an action (Kolykhalova et al., 2020) – with the aim of estimating the probability of winning within the model adopted for the generalized characteristic function.

7. APPENDIX: UNWEIGHTED AND WEIGHTED GENERALIZED SHAPLEY VALUES

In order to provide the definitions of the two generalized characteristic functions $\upsilon_1(.)$ and $\upsilon_2(.)$, the next steps are followed. First, one considers the case in which the ordered coalition, which is the argument of the generalized characteristic functions $\upsilon_1(.)$ and $\upsilon_2(.)$, has cardinality $m = 5$. In particular, when $|(T(i), i)| = 5$, one denotes by

$$\upsilon_1((T(i), i)) = P(Win)_{(T(i),i)} \tag{2}$$

the probability of winning the game for the ordered coalition of players $(T(i), i)$ (which contains player *i*). Analogously, when $|T(i)| = 5$, one denotes by

$$\upsilon_1(T) = P(Win)_{T(i)} \tag{3}$$

the probability of winning the game for the ordered coalition of players $T(i)$ (which does not contain player *i*). Similarly, for an ordered coalition made of 5 players, the values assumed by the other generalized characteristic function $\upsilon_2(.)$ are obtained by replacing Equations (2) and (3), respectively, with

$$\upsilon_2((T(i), i)) = P(Occ)_{(T(i),i)}P(Win)_{(T(i),i)} \tag{4}$$

and

$$\upsilon_2(T) = P(Occ)_{T(i)}P(Win)_{T(i)}\,. \tag{5}$$

In the above, $P(Occ)_{(T(i),i)}$ and $P(Occ)_{T(i)}$ represent the probabilities of occurrence on the court of the ordered coalitions of players $(T(i), i)$ and $T(i)$, respectively, and are estimated from the available data as in Metulini and Gnecco (2023).

Finally, one extends as follows the definitions of the two characteristic functions $\upsilon_1(.)$ and $\upsilon_2(.)$ to all the other ordered coalitions, having cardinality different from 5:

$$\upsilon_1(T) = \begin{cases} 0 & \text{if}\,|T| < m = 5 \\ \upsilon_1(\{T_1, T_2, T_3, T_4, T_5\}) & \text{if}\,|T| > m = 5, \end{cases} \tag{6}$$

and

$$\upsilon_2(T) = \begin{cases} 0 & \text{if}\,|T| < m = 5 \\ \upsilon_2(\{T_1, T_2, T_3, T_4, T_5\}) & \text{if}\,|T| > m = 5. \end{cases} \tag{7}$$

It is worth observing that the exact determination of the generalized Shapley value through Equation (1) can be computationally expensive (depending on the total number of players *n* of the generalized coalitional game), since it requires the evaluation of all the terms in its summation. Moreover, some of those terms may be even not available in practice. This justifies approximating the generalized Shapley value. A possible approximate evaluation can be obtained according to the following procedure, detailed in Metulini and Gnecco (2023). The average marginal utility in Equation (1) is substituted therein by an empirical average marginal utility. This is constructed by taking into account the observed quintets, and is based also on the simplifying assumption that each player has probability $\frac{5}{n}$ of entering in one of the first 5 positions, i.e., of being part of a lineup. In summary, denoting by L_i the set of observed (unordered) lineups (or quintets) in which player i occurs, one obtains the following estimate of his/her generalized Shapley value, for $k = 1, 2$:

$$\hat{\varphi}_i^{NR}(N, \upsilon_k) = \frac{5}{n}\frac{1}{5|L_i|}\sum_{L\in L_i}(\upsilon_k(L) - 0) = \frac{1}{n|L_i|}\sum_{L\in L_i}\upsilon_k(L). \qquad (8)$$

The right-hand side of Equation (8) is proportional to the average value of a quintet in which player i occurs. The proportionality factor $\frac{1}{5}$ is due to the fact that, for every specific quintet, each player has the same probability of being the last player to join all the other members of that quintet (conditional on his/her presence in the quintet). Under the stated assumptions, the estimate (8) is unbiased, and its variance is inversely proportional to $|L_i|$. In practice, different players may have distinct probabilities of being part of a lineup, so that estimate may become biased without that assumption. Still, in this case the estimate above could be used as a first approximation of the generalized Shapley value, based on the observed quintets. It is worth taking into account that this possible non-uniform sampling issue is partially compensated by the fact that the second characteristic function $\upsilon_2(.)$ takes implicitly into account that different players may have distinct probabilities of being part of a lineup. Moreover, the estimate (8) takes partially into account the same issue by averaging over possibly different numbers of quintets for distinct players.

REFERENCES

Annala, C.N. and Winfree, J. (2011). Salary distribution and team performance in major league baseball. In *Sport Management Review*, 14 (2): 167–175.

Bensic, M., Sarlija, N. and Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. In *Intelligent Systems in Accounting, Finance & Management: International Journal*, 13 (3): 133–150.

Bonnini, S. and Borghesi, M. (2022). Relationship between mental health and socio-economic, demographic and environmental factors in the covid-19 lockdown period - A multivariate regression analysis. In *Mathematics*, 10 (18): 3237.

Bonnini, S. and Cavallo, G. (2021). A study on the satisfaction with distance learning of university students with disabilities: Bivariate regression analysis using a multiple permutation test. In *Statistica Applicata - Italian Journal of Applied Statistics*, 33 (2): 143–162.

Buldú, J.M., Busquets, J., Martínez, J.H., Herrera-Diestra, J.L., Echegoyen, I., Galeano, J. and Luque, J. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. In *Frontiers in Psychology*, 9. Article no. 900.

Christensen, R. (2006). *Log-linear Models and Logistic Regression*. Springer Science & Business Media.

Engelmann, J. (2017). Possession-based player performance analysis in basketball (adjusted+/–and related concepts). In *Handbook of Statistical Methods and Analyses in Sports*, 231–244. Chapman and Hall/CRC.

Ertug, G. and Castellucci, F. (2013). Getting what you need: How reputation and status affect team performance, hiring, and salaries in the NBA. In *Academy of Management Journal*, 56 (2): 407–431.

Garris, M. and Wilkes, B. (2017). Soccernomics: Salaries for world cup soccer athletes. In *International Journal of the Academic Business World*, 11 (2): 103–110.

Glowinski, D., Dardard, F., Gnecco, G., Piana, S., and Camurri, A. (2015). Expressive non-verbal interaction in a string quartet: An analysis through head movements. In *Journal on Multimodal User Interfaces*, 9: 55–68.

Gnecco, G., Hadas, Y. and Sanguineti, M. (2019). Some properties of transportation network cooperative games. In *Networks*, 74 (2): 161–173.

Grassetti, L., Bellio, R., Di Gaspero, L., Fonseca, G. and Vidoni, P. (2021). An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. In *IMA Journal of Management Mathematics*, 32 (4): 385–409.

Hadas, Y., Gnecco, G. and Sanguineti, M. (2017). An approach to transportation network analysis via transferable utility games. In *Transportation Research Part B: Methodological*, 105: 120–143.

Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Springer.

Hentilä, D. (2019). *The Link between Salary and Performance: Are NBA Players Overpaid?* Bachelor's thesis, Tallin University of Technology.

Holmes, P. (2011). New evidence of salary discrimination in major league baseball. In *Labour Economics*, 18 (3): 320–331.

Huang, J. (2016). Salary in the national basketball association. Joseph Wharton Scholars. Available at `https://repository.upenn.edu/joseph_wharton_scholars/`.

Ilardi, S. (2007). Adjusted plus-minus: An idea whose time has come. In *82games.com*.

Kolykhalova, K., Gnecco, G., Sanguineti, M., Volpe, G. and Camurri, A. (2020). Automated analysis of the origin of movement: An approach based on cooperative games on graphs. In *IEEE Transactions on Human-Machine Systems*, 50: 550–560.

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

Li, N. (2014). *The Determinants of the Salary in NBA and the Overpayment in the Year of Signing a New Contract*. Ph.D. thesis, Clemson University.

Maschler, M., Solan, E. and Zamir, S. (2013). *Game Theory*. Cambridge University Press.

Metulini, R. and Gnecco, G. (2023). Measuring players' importance in basketball using the generalized Shapley value. In *Annals of Operations Research*, 325 (1): 441–465."

Metulini, R., Manisera, M. and Zuccolotto, P. (2018). Modelling the dynamic pattern of surface area in basketball and its effects on team performance. In *Journal of Quantitative Analysis in Sports*, 14 (3): 117–130.

Michalak, T.P., Aadithya, K.V., Szczepan´ski, Ravindran, B. and Jennings, N.R. (2013). Efficient computation of the Shapley value for game-theoretic network centrality. In *Journal of Artificial Intelligence Research*, 46: 607–650.

Michalak, T.P., Szczepański, P.L., Rahwan, T., Chrobak, A., Brânzei, S., Wooldridge, M. and Jennings, N.R. (2014). Implementation and computation of a value for generalized characteristic function games. In *ACM Transactions on Economics and Computation (TEAC)*, 2 (4): 1–35.

Mishra, S.K. (2016). Shapley value regression and the resolution of multicollinearity. In *Journal of Economics Bibliography*, 3 (3): 498–515.

Nowak, A.S. and Radzik, T. (1994). The Shapley value for n-person games in generalized characteristic function form. In *Games and Economic Behavior*, 6 (1): 150–161.

Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*. Potomac Books, Inc.

Papadaki, I. and Tsagris, M. (2022). Are NBA players' salaries in accordance with their performance on court? In *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies: Techniques and Theories*, 405–428.

Passacantando, M., Gnecco, G., Hadas, Y. and Sanguineti, M. (2021). Braess' paradox: A cooperative game-theoretic point of view. In *Networks*, 78 (3): 264–283.

Rosenbaum, D.T. (2004). Measuring how NBA players help their teams win. 82Games.com (`http://www.82games.com/comm30.htm`).

Sarlis, V. and Tjortjis, C. (2020). Sports analytics-Evaluation of basketball players and team performance. In *Information Systems*, 93: 101562.

Scully, G.W. (1974). Pay and performance in major league baseball. In *The American Economic Review*, 64 (6): 915–930.

Shapley, L.S. and Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. In *American Political Science Review*. 48: 787-792.

Sigler, K. and Compton, W. (2018). NBA players' pay and performance: What counts? In *Sport Journal*, 24.

Sigler, K.J. and Sackley, W.H. (2000). NBA players: Are they paid for performance? In *Managerial Finance*, 26 (7): 46–51.

Sill, J. (2010). Improved NBA adjusted+/-using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*. 7 pages.

Stole, L. and Zwiebel, J. (1996). Intrafirm bargaining under non-binding contracts. In *Review of Economic Studies*. 63: 375-410.

Terner, Z. and Franks, A. (2021). Modeling player and team performance in basketball. In *Annual Review of Statistics and Its Application*, 8: 1–23.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. In *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (1): 267–288.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. In *Knowledge and Information Systems*. 41, 647-665.

Vincent, C. and Eastman, B. (2009). Determinants of pay in the NHL: A quantile regression approach. In *Journal of Sports Economics*, 10 (3): 256–277.

Wiseman, F. and Chatterjee, S. (2010). Negotiating salaries through quantile regression. In *Journal of Quantitative Analysis in Sports*, 6 (1). 14 pages.

Xiong, R., Greene, M., Tanielian, V. and Ulibarri, J. (2017). Research on the relationship between salary and performance of professional basketball team (NBA). In *Proceedings of the 8th International Conference on E-business, Management and Economics*, 55–61.

Yan, T., Kroer, C. and Peysakhovich, A. (2020). Evaluating and rewarding teamwork using cooperative game abstractions. In *Advances in Neural Information Processing Systems*, 33: 6925–6935.

Yilmaz, M. and Chatterjee, S. (2003). Salaries, performance, and owners' goals in major league baseball: A view through data. In *Journal of Managerial Issues*, 15 (2): 243–255.

Zhu, X.J. (2019). Strategies to raise the first contract price for new unrestricted free agents in the NBA. University of Ottawa, Economics - Research Papers.

doi.org/10.26398/IJAS.0035-008

EVALUATION OF OFF-THE-BALL ACTIONS IN SOCCER

Lucas Wu, Tim Swartz[1]

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC, Canada V5A1S6

Abstract *Whereas there is no shortage of statistics that have been proposed and reported for invasion sports, almost all of the widely reported statistics are based on actions involving the ball. Yet, in football (soccer), it is well-known that players typically possess the ball for less than three minutes during a 90-minute match. In this paper, we develop automatic methods that analyze the activities of players that are "off-the-ball" in soccer. Specifically, a metric is introduced which measures defensive anticipation in soccer. The approach is conceptually straightforward: Using roughly four million spatio-temporal instances, we utilize machine learning techniques to predict the velocity (two-dimensional directional vector and speed) of a defensive player in a given situation. A metric is then developed which compares the player's actual velocity with the predicted velocity of a typical player in this situation. The interpretation of the defensive anticipation metric is based on the tenet that fast is better than slow. The analysis is facilitated through the availability of player tracking data which records the position of players at frequent and regular intervals throughout matches. The metric is calculated for players based on a season of soccer data, where validity and reliability are demonstrated. The metric also conforms to common sense where it is expected and observed that there is a reduction in defensive anticipation as players tire. The proposed approach is applicable and can be tailored to all invasion sports where player tracking data are available.*

Keywords: *OR in sports, Big data, Machine learning, Model validation, Player tracking data.*

ACKNOWLEDGMENT:

Swartz has been partially supported by the Natural Sciences and Engineering Research Council of Canada. The work was carried out with support from the CANSSI (Canadian Statistical Sciences Institute) Collaborative Research Team (CRT) in Sports Analytics. The authors thank Daniel Stenz, former technical director of Shandong Luneng Taishan FC who provided the data used in this paper. We also thank three anonymous reviewers whose comments helped improve the paper.

[1] Corresponding Author: Tim Swartz, tswartz@sfu.ca

1. INTRODUCTION

In the sport of football (soccer), it has been estimated that on average, throughout a 90-minute match, individual players have possession of the ball for less than two minutes (Link and Hoernig 2017). Therefore, traditional “on-the-ball” statistics such as goals, tackles, assists, shots and pass completion percentages examine only a snapshot of overall player performance. Encouraged by the “moneyball” phenomena (Lewis 2013), player evaluation via statistical analysis has become widespread across sports (Albert et al. 2017). In particular, there have been significant contributions to the sport of football as described and reviewed by Cefis (2022). For example, Cefis and Carpita (2022) utilize 29 key performance indicators to create composite indicators of performance quality. This paper considers a particular aspect of player evaluation in the context of “off-the-ball” activity in soccer.

This paper introduces novel methods and a metric that evaluates a fundamental defensive objective in soccer, namely defensive anticipation. When a defender anticipates quickly, the defender denies the offensive team both time and space, and this contributes to winning. Defensive awareness is important and is not always recognized. For example, by moving quickly, the defensive player may prevent a valuable pass which is never realized and hence, never recorded. We apply our methods to an actual dataset, where the validity and reliability of the metric are demonstrated.

There are currently no automatic methods (i.e. computer code) that produce metrics for defensive anticipation. For an analyst (e.g. coach) to assess the defensive anticipation of a player, there are two overriding difficulties. First, the analyst would need to monitor the player for the entire 90 minutes of a match, and repeat this over many matches. This is both time consuming and expensive. Second, the analyst would need to objectively evaluate the player’s actions, sometimes in contexts where it is not obvious what the player ought to do. The purpose of this paper is to develop automatic methods which objectively evaluate defensive anticipation. With these methods, information on defensive anticipation could be made available for players from various leagues across the world. Therefore, we believe that our methods may be beneficial when teams are seeking a replacement player.

Our investigation is made possible by the availability of player tracking data. Player tracking data in soccer consists of the Cartesian coordinates of the ball and the 22 players on the pitch recorded at regular and frequent time intervals. With player tracking data, we know the locations of all players at all times during a match, and this facilitates off-the-ball evaluation. Gudmundsson and Horton (2017) provide a review paper on spatio-temporal analyses used in invasion sports (including soccer) where player tracking data are available. The visualization of team formations in soccer is a problem that has received particular attention (Wu et al. 2019). The analysis of player tracking data has also been prominent in the sport of basketball; see for example, Miller et al. (2014).

The study of off-the-ball activity is a new research area of great potential. Historically, a limiting factor for such research has been the availability of tracking data. Tracking data are necessary because we need to know what all players are doing at all times - this is the basis for off-the-ball studies. There have been some off-the-ball analyses in basketball and soccer that are based on the concept of "ghosting" (Lowe 2013, Le et al. 2016, Le et al. 2017 and Seidl et al. 2018). The rationale behind ghosting is that there are optimal and expected paths for defensive players. In the ghosting work (which is proprietary), a main contribution is the claim that if defensive players can replicate the optimal ghosting paths, then outcomes would improve for the defensive team in terms of lower expected points/goals by the offensive team. Also, coaches may be able to assess what-if scenarios. That is, if a given play is drawn up, the expected ghost paths may indicate how the defensive team ought to respond. In the ghosting approach, actual match sequences are studied from a given frame where observed defensive positions are established. Then time frames are allowed to advance where the offensive players continue on their observed path and the ghosts react to the offensive movement. A limitation is that in real matches, offensive players move and react according to the defence. Therefore, the offensive movements that were observed cannot be utilized as responses to the ghosting paths. Spearman (2018) also used tracking data to investigate off-the-ball activity through positioning. Goal scoring probabilities were estimated at player locations using expected goal (xG) considerations and the probabilities of making successful passes to the player locations. This interesting line of research is instructive in identifying optimal positioning from an offensive perspective.

A major challenge in off-the-ball research is the evaluation of actions. Our approach is conceptually straightforward: Using roughly four million spatio-temporal instances, we utilize machine learning techniques to predict the velocity (two-dimensional directional vector and speed) of a defensive player in a given situation. A defensive anticipation metric is then developed which compares the player's actual velocity with the predicted velocity of a typical player in this situation. The interpretation of the defensive anticipation metric is based on the tenet that fast is better than slow (Blank 2012). Of course, "playing fast is better than slow" is a general principle that may not apply absolutely in every situation. Players that excel in this trait may be thought of as energetic and quick-thinking, and they provide a particular benefit to teams. Importantly, this type of analysis is amenable to other invasion sports for which tracking data are available.

In Section 2, we describe the dataset. In Section 3, we develop the methods used to evaluate defensive anticipation. The work is highly computational and we describe our approach which is based on the use of a tree-based boosting algorithm. In Section 4, the methods are then applied to an analysis of players from the Chinese Super League where validity and reliability of the approach are demonstrated. We conclude with a short discussion in Section 5.

2. DATA

Our data consists of matches from the 2019 season of the Chinese Super League (CSL). The league involved 16 teams where each team played every opponent twice, once at home and once away. From these potential 240 matches, we have three missing matches.

From these 237 matches, event data and tracking data were collected independently where event data consists of occurrences such as tackles and passes, and these were recorded along with auxiliary information whenever an "event" takes place. The events were manually recorded by technicians who view film. Both event data and tracking data have timestamps so that the two files can be compared for consistency. For example, the times for the event data are recorded to two decimal points whereas the tracking data times are recored to one decimal point. We dealt with this by rounding the times for the event data. In the CSL dataset, tracking data were obtained from video and the use of optical recognition software. The tracking data consists of roughly one million rows per match measured on 7 variables where the data are recorded every 1/10th of a second. The 7 variables are the time, the x coordinate, the y coordinate, a player identifier, the player jersey number, an indicator variable for the ball, and the half of play. The data were collected by Stats Perform which is a leading service provider of tracking data and operates the Opta data platform. The Stats Perform cameras are high resolution and are placed in various locations for the extrapolation of 2-d pictures to 3-d images. The accuracy and reliability of optical tracking data is high (Mara et al. 2017). Wu and Swartz (2022) have investigated the accuracy of the Stats Perform data in the context of player velocities in soccer. Each row of tracking data corresponds to a particular player at a given instant in time. Therefore, we have a big data problem where both event data and player tracking data are available based on 237 regular season matches. Although the inferences gained via our analyses are specific to the CSL, we suggest that the methods are applicable to any soccer league which collects tracking data.

3. METHODS

3.1. Rationale of the Approach

Consider a defender at a particular instant in time during a match. Our approach begins with the prediction of a velocity vector $(y\hat{}_1,y\hat{}_2)$ for the defender. It is important to emphasize that the two-dimensional velocity vector contains both a directional component and magnitude (i.e. speed). The prediction is facilitated through the availability of tracking data associated with the 2019 season of the CSL. With this massive dataset, there exist “similar” circumstances in a spatio-temporal sense to the spatio-temporal state of the defender whose velocity that we are attempting to predict. For example, the defenders in each circumstance may have an opponent with the ball directly in front of them and who is stationary. Therefore, the prediction represents the velocity (i.e. speed and direction) of a typical player in the situation of interest. Of course, the observed velocity (y_1,y_2) of the defender will not be exactly the same as the predicted velocity $(y\hat{}_1,y\hat{}_2)$. The observed velocity is calculated from the tracking data as the change in location given by the positional coordinates by the change in time over a short time period (Wu and Swartz 2022). The change in location is calculated as Euclidean distance. We posit that the defender will have performed above average if they move quicker than predicted in the predicted direction. The quantification of performance is formalized in Section 3.4. The desirability of moving quickly is a tenet of many sports, including soccer, and is discussed in Chapter 1 of Blank (2012).

3.2. Prediction of Velocities

Given a snapshot of the spatio-temporal state of the match which includes player locations, player velocities, possession and the location of the ball, it is possible for subject matter experts to predict where players ought to move. However, such assessments are subjective. Alternatively, formulating a parametric predictive model is a formidable task due to the complexity of spatio-temporal configurations.

A rationale for machine learning methods in prediction is that complex phenomena are often difficult to model explicitly. We may have a response variable y and a high-dimensional explanatory vector $x = (x_1, x_2, \dots, x_k)$ where we have little apriori knowledge about the relationship between y and x. For example, the relationship may only involve a subset of the variables x, the components of x may be correlated, and most importantly, the relationship $y \approx f(x)$ involves an unknown and possibly complex function f. In addition, the stochastic aspect of the relationship is typically unknown and big data sets may introduce computational difficulties.

In our problem, we face all of the challenges mentioned above. The response variable y is the velocity (speed and direction) that a player moves in a specific off-the-ball situation. We emphasize that y is a two-dimensional response. The explanatory variable x is the state of the match as described by the player tracking data. The idea is that the state of the match x is predictive of movement y. For each observation (x, y), the covariate x and the response y are each measured at a specific point in time t.

A restriction that we introduce is that we consider off-the-ball actions only for defensive players. Whereas offensive reactions are also important, we find this to be a more challenging prediction problem since offensive players may choose from multiple potential paths.

A first step in the data analysis is the determination of ball possession which then defines the defensive and offensive teams. In addition to player tracking data, we are also provided with tagged event data that provides the timing of passes, dribbles, shots, etc. A possession is retained if the same team maintains the control of the ball by either passing, dribbling or attempting a shot, and the possession ends when the opponent gains control of the ball, a penalty occurs, the ball goes out of bounds, etc. For a pass, the team that made the pass is deemed to be in possession until the ball is intercepted.

To make the prediction problem more tractable, we introduce two data reductions. First, we analyse match states every $\varepsilon = 1$ seconds. This is a tremendous data reduction (reduction by a factor of 10) since tracking data are recorded every 1/10th of a second. However, over a 90-minute match this still leaves us with 5,400 potential observations per player per match. With 11 defensive players on the pitch and the 237 regular season matches, this provides us with over 14 million records. We view $\varepsilon > 0$ as a tuning parameter which we can increase or decrease to adjust the total number of observations. The data reduction is advantageous in the sense that player actions are essentially independent for larger values of ε. In soccer, a player's objectives at a given point in time are different and independent from his objectives ε seconds later for sufficiently large ε. Our intuition is that player options change considerably over states separated by $\varepsilon \geq 1$ second.

Another data reduction decision involves the covariate vector x provided by the tracking data. Based on our soccer knowledge, we posit that a player's actions are mostly dependent on the spatio-temporal characteristics of the ball and the players within their immediate vicinity. Of course, there are long passes in soccer, but we exclude these considerations as they are the exception rather than the rule. We therefore introduce the following covariates for a given defensive player in a particular state:

- x_1 - location of the player (2-dim)
- x_2 - player velocity at time $t-\Delta$ (2-dim)
- x_3 - location of the ball (2-dim)
- x_4 - distance of the player to the ball (1-dim)
- x_5 - distance of the player to offensive goal (1-dim)
- x_6 - angle of the player to offensive goal (1-dim)
- x_7 - location of the goalkeeper (2-dim)
- x_8 - distance of the player to goalkeeper (1-dim)
- x_9 - indicator for player on offensive or defensive side of the field (1-dim)
- x_{10} - indicator for player belonging to the home or away team (1-dim)
- x_{11} - seconds remaining in the half of the game (1-dim)
- x_{12} - seconds remaining in the full game (1-dim)
- for each of the player's three nearest teammates:
  - x_{13} - location of the teammate (2-dim)
  - x_{14} - velocity of the teammate (2-dim)
  - x_{15} - distance of the player to teammate (1-dim)
  - x_{16} - distance of the ball to teammate (1-dim)
  - x_{17} - relative angle of the teammate to the player of interest (1-dim)
- for each of the player's three nearest opponents:

- x_{18} - location of the opponent (2-dim)
- x_{19} - velocity of the opponent (2-dim)
- x_{20} - distance of the player to opponent (1-dim)
- x_{21} - distance of the ball to opponent (1-dim)
- x_{22} - expected possession value EPV of opponent (1-dim)
- x_{23} - relative angle of the opponent to the player of interest (1-dim)

Therefore, even though we have dramatically reduced the dimensionality of the tracking data, we have retained a 61-dimensional covariate which we hope captures the main drivers of how a player responds in a given situation. We note that the covariates contain a great amount of information which is related to y in complex ways. For example, if a player is close to goal, they may behave differently than if they are near midfield. Also, the movements and space of nearby players naturally impact decisions. We experimented with different numbers of nearest teammates and opponents (i.e. covariates x_{13} through x_{23}). However, we found little improvement in prediction beyond using the three nearest teammates and opponents.

The variable x_2 and the associated tuning parameter $\Delta \geq 0$ require additional discussion. We cannot include x_2 as a covariate with $\Delta = 0$ as this would render $y = x_2$ at all times t, and consequently, any fitting algorithm would yield the useless prediction $\hat{y} = y$. That is, our predicted velocity would not be a typical velocity given the circumstances, but instead, the observed velocity of the player of interest. However, the observed velocity y of the player of interest at time t depends on his movement prior to time t. For example, if a player is moving forward at speed s, it is easier for him to quickly transition to speed $s+\delta$ moving forward than speed $s+\delta$ moving backward. In summary, we ought to know about a player's movement before time t as this impacts movement at time t. In Section 3.3, we investigate the selection of Δ.

The expected possession value EPV (feature x_{22}) was made publicly available by Shaw (2019). Given the spatial state of a match, EPV provides a measure of the attacking value of each location on the field. We modify the EPV covariate of a player by setting it equal to zero if the offensive player is offside. This is an important covariate in our analysis since defenders should be cautious of balls being played to high EPV positions.

There is some redundancy in our covariates. For example, if we know the Cartesian coordinates of two objects, then the distance between these two objects is a function of their positions. However, to assist the machine learning algorithm of Section 3.3, we provide some of these derived covariates. We have limited the covariates to the three nearest teammates and three nearest opponents. In most cases, these are the players who most influence the movement of the player of interest. The player of interest cannot intervene in locations that are too distant.

3.3. Computational Overview

Recall that our fundamental problem is the development of a metric for defensive anticipation. This metric requires the prediction of the velocity y corresponding to the features x which describe the spatio-temporal state of the match. We constructed a design matrix where we stepped through each $\varepsilon = 1$ seconds of time over all matches to determine team possession. If the time t is part of a possession sequence, then one row of the design matrix is generated for each defender. The columns consisted of all of the features x for a defender as described in Section 3.2. The procedure resulted in a design matrix with 3,770,289 rows and required just under 100 hours of computation on a laptop computer. Note that the construction of the design matrix consists of tasks that can be divided according to matches. Therefore, this data management component is amenable to parallel processing.

For the prediction problem, we used a fast and efficient gradient boosting model, LightGBM, which is based on tree-based learning algorithms (Ke et al. 2017). We trained two LightGBM models, one for predicting the horizontal velocity component and one for predicting the vertical velocity component based on the field orientation. We partitioned the 20-week data into training and test datasets, where the training data included all the even weeks (eg. weeks $2,4,\ldots,20$) and the test data included all the odd weeks (eg. weeks $1,3,\ldots,19$). For model training, we used leave-one-week-out cross validation to select the best tuning parameters which minimized the mean absolute error of the response variables. The training procedure using LightGBM required approximately 1.5 hours of running time on a laptop computer. The LightGBM procedure assumes independent data which should approximately be the case with player velocities measured at one second intervals.

Recall that we are interested in setting the parameter $\Delta \geq 0$ which provides the velocity covariate x_2 of the player of interest at time $t - \Delta$. We want to set Δ so that it assists prediction of the velocity of a typical player at time t. Again, Δ cannot equal zero; otherwise we simply obtain predictions that are the actual velocities of the players of interest. On the other hand, if we choose large Δ, then the velocity covariate x_2 is too distant in time from the current time t to facilitate the prediction of velocity at time t. In Figure 1, we plot the correlation of the predicted speed at time t and the actual speed at time $t - \Delta$. For $\Delta = 0$ seconds, the correlation is perfect (i.e. $r = 1$) as expected. We wish to choose a time lag Δ so that the model provides a good but not a perfect predictor. From Figure 1, we choose the tuning parameter $\Delta = 0.5$ seconds where the correlation $r \approx 0.9$. We experimented with other choices of Δ up to $\Delta = 1$ second, and found that our results were robust to the choice. The fitted model from LightGBM provides a mean absolute error of 0.319 m/sec in the x-coordinate velocity and 0.398 m/sec in the y-coordinate velocity.

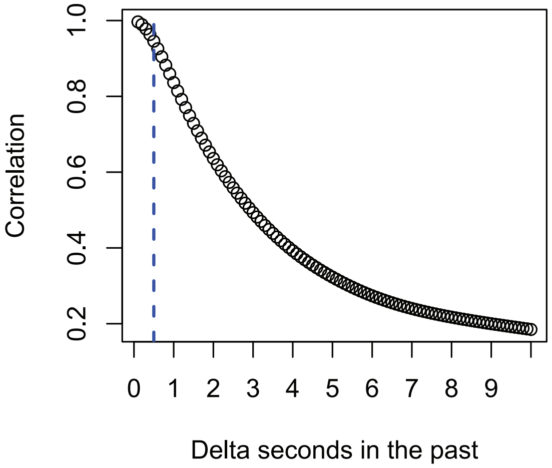


Figure 1: Correlation of predicted speed at time t and actual speed at time $t - \Delta$ where time is measured in seconds. The dashed line corresponds to the selected value $\Delta = 0.5$ seconds.

3.4. Derivation of a Metric for Defensive Anticipation

We return to the motivation for off-the-ball player evaluation. Recall that the core idea from Chapter 1 of Blank (2012) is that doing things quickly in soccer is better. For example, one could imagine a defender moving towards a forward who is about to receive a pass. In this case, getting there early increases the chance of intercepting the pass or preventing the forward from creating a goal scoring opportunity. Now, there are many instances during a match where moving quickly makes no sense (e.g. the ball is at the opposite end of the field). In cases where the predicted response is to move slowly, we will not use these cases for the purpose of player evaluation.

Consider time t where the match state x is recorded and the predicted velocity for a defensive player is $(\hat{y}_1, \hat{y}_2)$. Recall that velocity is two-dimensional as it involves both speed and direction in the plane. Again, we only consider observations where the predicted speed exceeds a specified threshold speed. The predicted velocity $(\hat{y}_1, \hat{y}_2)$ is obtained by the machine learning prediction methods of Section 3.2. We let $(y_{\text{obs1}}, y_{\text{obs2}})$ denote the corresponding observed velocity of the player under evaluation. Then, we define the player's off-the-ball performance at time t by

$$p = \begin{cases} \left(\sqrt{v_1^2+v_2^2} - \sqrt{\hat{y}_1^2+\hat{y}_2^2}\right) / \sqrt{\hat{y}_1^2+\hat{y}_2^2} & v_1\hat{y}_1 \geq 0 \\ \left(-\sqrt{v_1^2+v_2^2} - \sqrt{\hat{y}_1^2+\hat{y}_2^2}\right) / \sqrt{\hat{y}_1^2+\hat{y}_2^2} & v_1\hat{y}_1 < 0 \end{cases} . \tag{1}$$

A geometric interpretation of p is provided in Figure 2. The statistic p in (1) is based on the projection (v_1, v_2) of the observed performance $(y_{\text{obs1}}, y_{\text{obs2}})$ onto the velocity line defined by the predicted velocity $(\hat{y}_1, \hat{y}_2)$. The line $k(\hat{y}_1, \hat{y}_2)$ for $k > 0$ emanates from the origin and is given by $y_2 = (\hat{y}_2/\hat{y}_1)y_1$ and the projection is calculated by $(v_1, v_2) = ((\hat{y}_1^2 y_{\text{obs1}} + \hat{y}_1\hat{y}_2 y_{\text{obs2}})/(\hat{y}_1^2+\hat{y}_2^2),\ (\hat{y}_1^2\hat{y}_2 y_{\text{obs1}} + \hat{y}_1\hat{y}_2^2 y_{\text{obs2}})/(\hat{y}_1^3+\hat{y}_1\hat{y}_2^2))$. Therefore, "good" performance according to (1) takes into account moving quicker in the predicted direction. Longer projections on the yellow velocity line in Figure 2 are better in terms of player performance, and lead to larger values of p. Values of $p > 0$ are interpreted as above average performance and values of $p < 0$ are interpreted as below average performance.

The player's season long performance is then given by the defensive anticipation metric

$$P = \left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) 100\% \tag{2}$$

where the summation is taken over all instances where the predicted velocity exceeds the threshold speed and the index $i = 1, \ldots, N$ corresponds to the cases involving the player during the season. In (2), it is sensible to calculate an average since the observations can be regarded as independent due to the time spacing provided by the tuning parameter $\varepsilon = 1$ seconds introduced in Section 3.2.

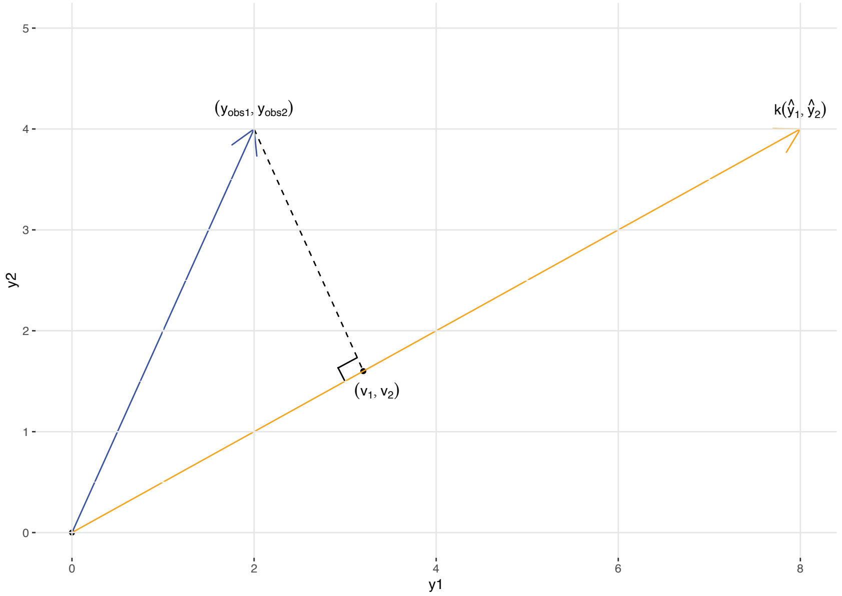


Figure 2: Geometric diagram which illustrates the components of the statistic p in equation (1). Imagine a player who is located at the origin $(0,0)$. The observed velocity of the player is shown by the blue vector pointing towards $(2,4)$. The predicted velocity of an average player is shown by the yellow vector pointing towards $(8,4)$. The perpendicular line indicates the projection of the observed velocity vector on the predicted velocity vector. Using equation (1), the defensive anticipation value, p, is equal to -0.6, which can be interpreted as a 60% reduction compared to the average player.

We can think of the metric (2) as a measure of defensive anticipation. It also possibly assesses aspects of player energy and quickness of thought. The multiplicative factor 100% in (2) permits a nice interpretation; a P-score of $+x$ describes a player whose defensive anticipation is $x\%$ above the average player whereas a P-score of $-x$ describes a player whose defensive anticipation is $x\%$ below the average player. The reported scores for players (Table 2) are relative to players from the 2019 season of the CSL. Although the metric P is unbounded by it's construction, it appears that reported values are surely contained in the interval $(-10, 10)$.

4. RESULTS AND ASSESSMENT

Of course, with new metrics such as defensive anticipation, there is no truth against which results can be compared. For example, we simply don't know which players are best at defensive anticipation. In this section, we look at the defensive anticipation metric from various angles with an attempt to establish validity and reliability.

First, to get a sense of the prediction results, Figure 3 provides a plot of the predicted velocities and the observed velocities for all 20 players on the field (not including the keepers) at a given instant in time. In most cases, the predicted and observed velocity vectors tend to point in roughly the same direction. For illustration, consider defensive player #16. His movement is directly towards the ball. However, the model predicts that he ought to move a little bit more towards his own goal at roughly the same speed. The predicted movement may be viewed as cautious and preferable since offensive player #35 (who is in possession) is moving downfield and may pose a risk. We observe that for some players, the velocity vectors are short, and this suggests that little is happening in their immediate surroundings. For example, defensive player #7 is barely moving, and this appears sensible as there are no threatening offensive players in the vicinity. For the evaluation of the defensive anticipation metric (2), we removed observations for which the predicted speed $\sqrt{\hat{y}_1^2 + \hat{y}_2^2}$ is less than the threshold speed of 0.20 m/sec which corresponds to 0.72 km/hour. In the test dataset, 1.8% of the observations were removed due to the threshold constraint.

Based on the examination of many frames such as given in Figure 3, we did not find predicted velocities that contradicted our soccer intuition. This provides an indication that in a given situation, the predicted velocity of a typical player is sensible. This may be expected because the predicted velocity is based on the fitting of a massive dataset, where on average, professional athletes make good decisions. We note that the model was assisted by the inclusion of covariate x_{22} (previously discussed). The recognition of players in offside positions improved prediction.

4.1. Reliability

With respect to a metric, *reliability* refers to the consistency of the measure. In other words, reliability addresses reproducibility. For example, it would be undesirable if our defensive anticipation metric (2) identified a player as having great defensive anticipation for half of the matches and terrible defensive anticipation in the other matches. Since we expect some consistency in professional athletes,

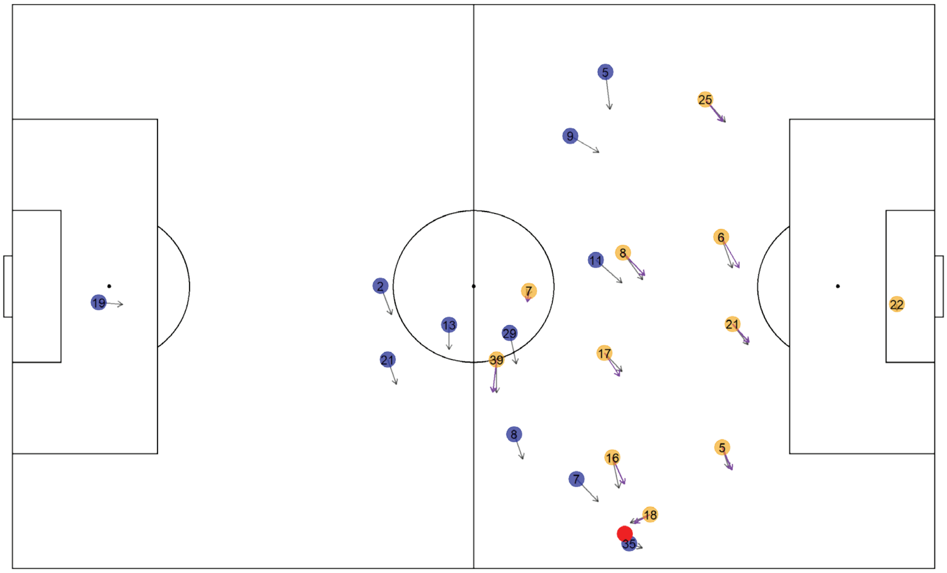


Figure 3: Plot of predicted velocities (purple arrows) and observed velocities (black arrows) at a given instant in time. The blue team is in possession, the yellow team is defending and the red dot corresponds to the ball.

this would suggest that there is little value in the metric.

To investigate this, we divided the 2019 CSL season into even and odd weeks. The premise is that the metric (2) measures an aspect of playing style, and that style should not differ greatly between the two sets of weeks. In Table 1, we provide results for the 10 players on Shandong Luneng for whom the number of instances $N > 10,000$ in (2) for both sets of weeks. Shandong Luneng is an interesting CSL team as two of the international players (Fellaini and Pelle) are well known to those who follow the English Premier League. We observe that there is consistency in the player metrics across the two sets of weeks. In fact, the ranks of the 10 players are the same across the two weeks. This suggests that the defensive anticipation metric (2) is reliable and is capturing an aspect of playing style. The standard errors for P_{even} and P_{odd} are small for all players, lying between 0.31 and 0.55.

Player	N_{even}	N_{odd}	P_{even} (rank)	P_{odd} (rank)
Marouane Fellaini	17,146	17,340	2.8 (1)	2.4 (1)
Zhang Chi	16,647	17,235	2.4 (2)	2.0 (2)
Liu Yang	19,556	19,845	1.6 (3)	1.8 (3)
Wang Tong	13,955	20,034	0.4 (4)	0.2 (4)
Hao Junmin	16,050	16,696	-0.3 (5)	-1.4 (5)
Zheng Zheng	14,582	10,849	-1.6 (6)	-2.6 (6)
Dai Lin	14,030	18,423	-2.1 (7)	-3.1 (7)
Graziano Pelle	19,337	18,302	-3.7 (8)	-4.1 (8)
Gil	10,159	13,306	-4.1 (9)	-5.1 (9)
Roger Guedes	14,067	16,737	-5.5 (10)	-5.7 (10)

Table 1: The defensive anticipation metric *P* calculated during even and odd weeks for players on Shandong Luneng during the 2019 season.

4.2. Validity

With respect to a metric, *validity* refers to the accuracy of measure. In our investigation, we are interested whether the metric P in (2) really measures defensive anticipation.

To investigate validity, we first consider the defensive anticipation metric (2) for all 438 outfield players in the CSL dataset. The players are categorized according to the five broad playing positions as follows: wide midfielder ($n = 79$) wide defender ($n = 77$), and forward ($n = 86$), central midfielder ($n = 110$) and central defender ($n = 86$). Density plots of (2) corresponding to each of the playing positions are shown in Figure 4. We observe that there is little difference in (2) across the playing positions. We note that central midfielders have slightly larger values of (2) than other players on average (as might be expected). This may be related to the defensive aggressiveness required at that position. We also observe that there is more variability in (2) amongst the forwards than the other playing positions.

Recall that a difficulty in assessing the validity of the proposed metric (2) is that there is no gold standard for the truth. We do not know with certainty which players play with more and less defensive anticipation (combination of energy and quick-thinking). Therefore, we took the same players from Shandong Luneng as in Table 1, and ranked these players according to their P-scores (2) from the entire 2019 season. The results are provided in Table 2. In Table 2, we made compar-

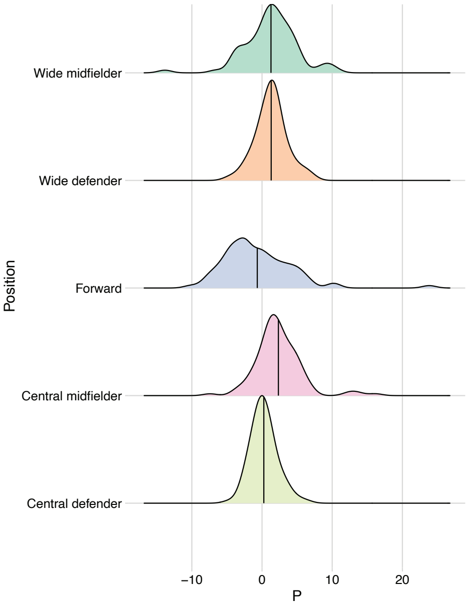


Figure 4: Density plots of (2) based on playing position. For each player, the defensive anticipation metric (2) was calculated for all matches in the 2019 CSL season. We observe that central midfielders have slightly larger defensive anticipation values than other players on average, and there is more variability amongst the forwards than the other playing positions.

isons with various measures of aggression. We provide season long data on fouls, successful tackles and interceptions. We excluded card accumulation as cards are relatively rare events. We observe that the aggressiveness inherent in fouls, successful tackles and interceptions correlates with our defensive anticipation metric. For example, the correlation coefficients between P and these three statistics are 0.58, 0.65 and 0.74, respectively. The corresponding 95% confidence intervals are $(-0.08, 0.88)$, $(0.03, 0.91)$ and $(0.21, 0.93)$, respectively.

In Table 2, we explored the relationship between P with player interceptions and tackles in the context of Shandong Luneng. We expanded this investigation by considering all players in the CSL who had played at least 500 minutes during the 2019 season. Figure 5 provides scatterplots relating P to interceptions and tackles. Of course, we do not expect extraordinarily high correlations between P, interceptions and tackles since these measurements consider different aspects of play. What these statistics have in common is an element of aggression. We

Player	P (rank)	Fouls (rank)	Tackles (rank)	Interceptions (rank)
Marouane Fellaini	2.64 (1)	46 (1)	21 (5.5)	23 (4)
Zhang Chi	2.20 (2)	32 (2.5)	21 (5.5)	29 (2)
Liu Yang	1.71 (3)	26 (4.5)	33 (1)	6 (8)
Wang Tong	0.26 (4)	15 (9)	19 (7)	27 (3)
Hao Junmin	-0.85 (5)	25 (6)	23 (4)	22 (5)
Zheng Zheng	-1.99 (6)	17 (8)	29 (2)	12 (7)
Dai Lin	-2.67 (7)	32 (2.5)	24 (3)	33 (1)
Graziano Pelle	-3.91 (8)	26 (4.5)	6 (10)	2 (9.5)
Gil	-4.65 (9)	6 (10)	13 (8)	13 (6)
Roger Guedes	-5.63 (10)	21 (7)	7 (9)	2 (9.5)

Table 2: The defensive anticipation metric P given by (2) for 10 players on Shandong Luneng who received the most playing time during the 2019 CSL season. We also provide comparison metrics involving aggression during the 2019 season, namely the total number of fouls committed, tackles made and the number of interceptions.

observe that interceptions and tackles correlate positively with P leaguewise.

We investigated the validity of our metric further by calculating the average P-score for all CSL players where we divided matches into 10-minute intervals. The plot is provided in Figure 6. We observe that P decreases as the match progresses. Since players tire as the game proceeds (both physically and mentally), it makes sense that our metric (2) decreases. There appears to be a big drop after the 70-th minute of the match.

It is interesting that amongst CSL players with regular minutes, the two players with the highest P-scores are Chang Feiya of Wuhan Zall ($P = 5.71$) and Yang Shiyuan of Shanghai SIPG ($P = 5.33$). Feiya is primarily a midfielder and does not have remarkable statistics; he scored only one goal in the 2019 season. Interestingly, the website https://www.allfamousbirthday.com/chang-feiya/ describes Feiya as one of the most popular Chinese football players. Shiyuan is a midfielder who also does not have remarkable statistics; he did not score during the 2019 season. Interestingly, the website https://www.whoscored.com/Players/143864/Show /Yang-Shiyuan describes Shiyuan as a player who likes to tackle and commits fouls often.

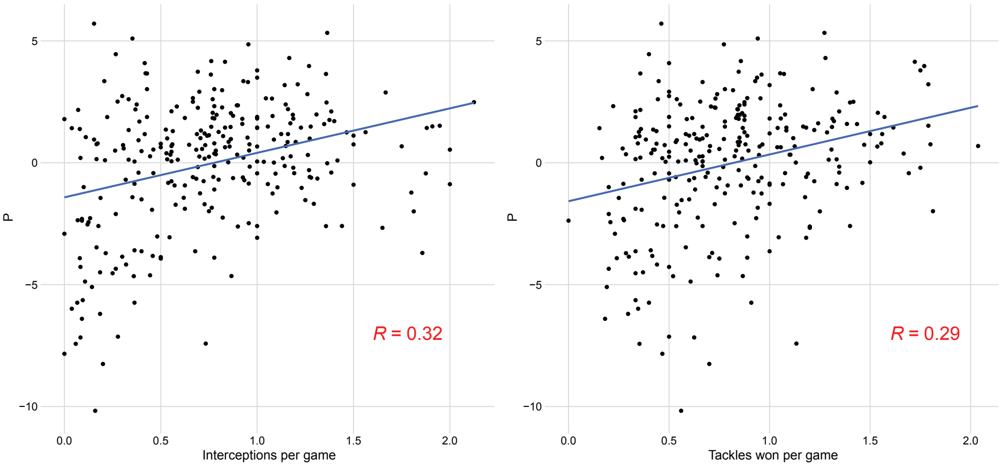


Figure 5: Scatterplots of the defensive anticipation metric (2) plotted against player interceptions and tackles made during the 2019 CSL season.

5. DISCUSSION

We have introduced an important and seminal area of research where automatic and objective methods have been developed to assess a particular defensive characteristic of off-the-ball behaviour. We have referred to the proposed metric (2) as defensive anticipation. The methods can be adapted to any invasion sport where tracking data are available.

The evaluation of off-the-ball performance is viewed in a narrow context where fast is considered better than slow. Even if speed is not the ultimate metric in off-the-ball evaluation, the metric (2) developed here may uncover insights into aspects of play. Perhaps players with high evaluations may be thought of as "high motor" players whose skills are useful to teams. An important aspect of the research is that our metric measures aspects of industry, laziness, anticipation and quick-thinking; these are characteristics that have not been previously quantified.

Some other notable aspects of our work include the following: the proposed metric is seen as reliable in the sense that it truly captures intrinsic player tendencies (Table 1), the metric adheres to expected results such as the positive correlation between the metric and other statistics related to aggression (Figure 5), and decreasing defensive anticipation as players tire (Figure 6).

A possible application involves the evaluation of P on a game by game basis. Managers would like to know how players have performed in terms of defensive

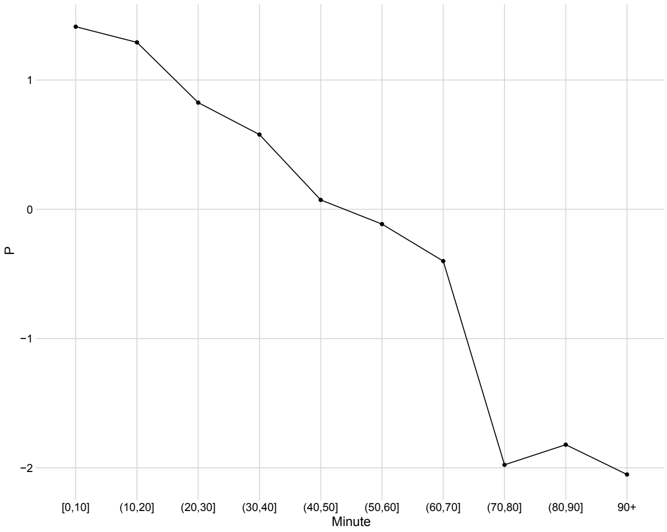


Figure 6: Plot of the defensive anticipation metric (2) averaged over all CSL players during 10-minute intervals.

anticipation. For example, perhaps *P*-scores may form a reason for future inclusion in the lineup. Also, it may be possible to detect the effects of illness.

5.1. Connections to Existing Literature

Whereas there does not seem to be any previous work on defensive anticipation in soccer, there are various recent papers that attempt to assess off-the-ball performance. Of course, this is a relatively new research topic since tracking data has only recently become available. We discuss two substantive contributions.

Dick and Brefeld (2019) use a reinforcement learning approach to evaluate player positioning. Unlike our investigation that has a defensive focus, Dick and Brefeld (2019) are interested in offensive configurations that are more likely to lead to goals. The approach takes into account current positioning and movement vectors that enable the consideration of future formations. A scoring function is learned from past data that maps game states to values that assess the benefit to the attacking team. Also, in contrast to our work, the proposed measures correspond to the team level rather than the performance of individual players.

In his PhD thesis, Fernández (2022) develops a framework for the investigation of various problems in soccer. The comprehensive approach is predicated on the development of an expected possession value model that decomposes the sport into components. For example, the action space is composed of passes, drives and shots where each component has its own set of estimation procedures. Applications in the thesis which are related to our work concern off-the-ball performance. For example, in Chapter 7, Fernández (2022) provides insights as to how teams can defend against buildup play, and how to calculate a player's optimal offensive positioning.

5.2. Future Research

There are at least three avenues for future research. First, there are many alternative predictive models that could be investigated. Given that we are predicting average player movement, the determination of which model provides better predictions is not straightforward. Second, the approach only considers off-the-ball actions for players when the opponent has possession. Naturally, player movement for players on the team in possession is also important. This is a more difficult prediction problem since there appears to be more viable options for offensive players. Third, we hope to gain access to tracking data from other leagues. Assessing the proposed metric and evaluating a wider pool of players with respect to defensive anticipation are topics of great interest.

REFERENCES

Albert, J.A., Glickman, M.E., Swartz, T.B. and Koning, R.H., Editors (2017). In *Handbook of Statistical Methods and Analyses in Sports*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, Boca Raton.

Blank, D. (2012). *Soccer IQ*, www.soccerpoet.com

Cefis, M. (2022). Football analytics: A bibliometric study about the last decade contributions. In *Electronic Journal of Applied Statistics*, 15(1), 232-248.

Cefis, M. and Carpita, M. (2022). The higher-order PLS-SEM confirmatory approach for composite indicators of football performance quality. In *Computational Statistics*, https://doi.org/10.1007/s00180-022-01295-4

Dick, U. and Brefeld, U. (2019). Learning to rate player positioning in soccer. In *Big Data*, 7, 71-82.

Fernández, J. (2022). A framework for the analytical and visual interpretation of complex spatiotemporal dynamics in soccer. Department of Computer Science, Polytechnic University of Catalonia. Accessed March 16, 2022 at https://upcommons.upc.edu/handle/2117/363073

Fernández, J., Bornn, L. and Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. *13-th MIT Sloan Sports Analytics Conference*, Accessed September 21, 2020 at http://www.lukebornn.com/papers/fernandez_sloan_2019.pdf

Gudmundsson, J. and Horton, M. (2017). Spatio-temporal analysis of team sports. In *ACMComputing Surveys*, 50(2), Article 22.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30, 3146¿3154.

Le, H.M., Carr, P., Yue, Y. and Lucey, P. (2016). Data-driven ghosting using deep imitation learning. *10-th MIT Sloan Sports Analytics Conference*, Accessed November 30, 2020 at https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5fee0a8b9838792227ec7fa5_Data-Driven

Le, H.M., Yue, Y., Carr, P. and Lucey, P. (2017). Coordinated multi-agent imitation learning. *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia.

Lewis, M. (2013). *Moneyball: The Art of Winning an Unfair Game*, WW Norton, New York.

Link, D. and Hoernig, M. (2017). Individual ball possession in soccer, In *PLoS One*, 12(7): e0179953. https://doi.org/10.1371/journal.pone.0179953

Lowe, Z. (2013). Lights, cameras, revolution. In *Grantland*, Accessed August 25, 2020 at https://grantland.com/features/the-toronto-raptors-sportvu-cameras-nba-analytical-revolution/

Mara, J., Morgan, S., Pumpa, K. and Thompson, K. (2017). The accuracy and reliability of a new optical player tracking system for measuring displacement of soccer players. In *International Journal of Computer Science in Sport*, 16(3), 175-184.

Miller, A., Bornn, L., Adams, R.P. and Goldsberry, K. (2014). Factorized point process intensities: A spatial analysis of professional basketball. In *Proceedings of the 31st International Conference on Machine Learning* - Volume 32, JMLR.org, Beijing, 235-243.

Seidl, T., Cherukumudi, A., Hartnett, A., Carr, P. and Lucey, P. (2018). Bhostgusters: Realtime interactive play sketching with synthesized NBA defenses. *12-th MIT Sloan Sports Analytics Conference*, Accessed August 25, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2018/02/1006.pdf

Shaw, L. (2019). Friends-of-Tracking-Data-FoTD/LaurieOnTracking Accessed November 20, 2021 at https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Spearman, W. (2018). Beyond expected goals. *12-th MIT Sloan Sports Analytics Conference*, Accessed September 21, 2020 at http://www.sloansportsconference.com/wp-content/uploads/2018/02/2002.pdf

Wu, L.Y. and Swartz, T.B. (2022). The calculation of player speed from tracking data. In *International Journal of Sports Science and Coaching*, https://doi.org/10.1177/17479541221124036

Wu, Y., Xie, X., Wang, J., Deng, D., Liang, H., Zhang, H., Cheng, S. and Chen, W. (2019). ForVizor: Visualizing spatio-temporal team formations in soccer, In *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 65-75.

doi.org/10.26398/IJAS.0035-009

A SURVIVAL ANALYSIS STUDY TO DISCOVER WHICH SKILLS DETERMINE A HIGHER SCORING IN BASKETBALL

Ambra Macis [1], **Marica Manisera, Paola Zuccolotto**
Department of Economics and Management, University of Brescia, Italy

Marco Sandri
Big and Open Data Innovation Laboratory, University of Brescia, Italy

Abstract *Over the years data analytics for sports has developed consistently. Survival analysis is a method that allows to study the occurrence of a particular event during a period of follow-up. This work aims* at *studying the main achievements associated to the probability of reaching a certain amount of points during a NBA season segment. A stepwise Cox regression model and a Lasso Cox regression were used to select the most important variables. Two settings were examined, with* 20% *and* 50% *censoring. Results showed that attempting more shots, gaining more achievements (double doubles) and having been selected for the All-Star game increase the probability of success, i.e. exceeding the given threshold of points. Moreover, a higher number of steals seems to decrease the probability of reaching a certain amount of points. Thus, players more involved in this fundamental are penalized in terms of scored points.*

Keywords: *Stepwise Cox regression; Lasso Cox regression; Basketball analytics; Performance.*

1. INTRODUCTION

Sport analytics has developed consistently over the years. For what concerns basketball, data science has been widely used to answer different questions and several studies have been carried out with a wide variety of aims. Just as an example, contributions in the literature deal with the analysis of players' performance and of the impact of high pressure game situations, the prediction of the outcomes of a game or a tournament, the identification of factors that distinguish successful and unsuccessful teams and the monitoring of playing patterns with reference to roles (Zuccolotto and Manisera, 2020).

In this work we deal with survival analysis, a class of statistical methods devoted to the study of the occurrence of an event during a given observation time. This kind of analysis was firstly introduced for applications in medicine

[1] Ambra Macis ambra.macis@unibs.it

for studying, for example, disease recurrence or death; however, it has also been widely used in sport analytics. Survival analysis has been used in the context of sports with several different aims, such as, for example, for studying the relationship between specific features and dropout of young athletes in many sports (Back et al., 2022; Moulds et al., 2020; Pion et al., 2015; Smith and Weir, 2022), or for evaluating the career length of professional basketball players (Fynn and Sonnenschein, 2012). Other studies analyzed the criterion determining the decision of a football coach of doing the first substitution during a match (Del Corral et al., 2008), the effect of team performance in the dismissal of coaches (Tozetto et al., 2019; Wangrow et al., 2018), the duration of Olympic success (Csurilla and Fertő, 2022; Gutiérrez et al., 2011), whether Olympic medalists live longer than the general population (Clarke et al., 2012). Furthermore, many studies dealt with injury prevention and the prediction of risk factors for injury (Beynnon et al., 2005; Buist et al., 2010; Ekeland et al., 2020; Hopkins et al., 2007; Lu et al., 2022; Mahmood et al., 2014; Venturelli et al., 2011; Zumeta-Olaskoaga et al., 2021) and recovery after injuries and sport-related concussions (Dekker et al., 2017; Howell et al., 2019; Jack et al., 2019; Kontos et al., 2019; Lawrence et al., 2018; Mai et al., 2017; Nelson et al., 2016; Sochacki et al., 2019). Finally, other works analyzed the impact of performance indicators on the time when the first goal is scored or the effect of that time on the following goal in football (Nevo and Ritov, 2013; Pratas et al., 2016) or studied times between goals in ice hockey (Thomas, 2007).

Up to now, to the best of our knowledge, survival analysis has not been used for studies in which the event of interest is a measure of the overall performance of a player. This work aims indeed to study the offensive performance of the National Basketball Association (NBA) players in a novel way, using survival analysis. In details, the player's performance has been measured in terms of exceeding of a given amount of points during a season segment, and the interest has been focused on the identification of the main achievements related to the occurrence of this event. Thus, from a statistical point of view this means performing a well-defined variable selection with only few variables selected. For this reason, the Lasso Cox has been chosen because it allows to select only few variables from all those taken into account. Moreover, the stepwise Cox regression has been used as additional method to have a term of comparison.

The article is organized as follows. The following section reports the methodological framework. Then, Sections 3 and 4 show respectively the data used for carrying out the study and the obtained results. The paper ends with the final discussion.

2. METHODS

Survival analysis aims to study the occurrence of a particular event during an observed period of time. The main feature of this kind of data is censoring. A subject is censored when for him/her the event of interest has not been observed during the observation time, so that the only known thing is the last time he/she did not experience the event (Collett, 2015). In this context a subject is denoted by three elements: (i) a time point τ that can be the observed time t or the censoring time c; (ii) an event indicator δ that equals 1 if the subject experienced the event and 0 if he/she is censored; and (iii) a vector of observed covariates $\boldsymbol{x}$. More in detail, for the i^{th} subject:

$$\tau_i = \min(t_i, c_i) = \begin{cases} t_i & \text{if } \delta_i = 1 \\ c_i & \text{if } \delta_i = 0 \end{cases} . \tag{1}$$

The actual survival time can be seen as the observed value of a non-negative random variable T with density function $f(t)$, such that $f(t) \geq 0$ and $\int_0^{+\infty} f(t)dt = 1$. Key elements in survival analysis, which allow to specify the probability distribution of T, are the survival and hazard functions.

The survival function $S(t)$ measures the probability that an individual survives (does not experience the event) beyond a given timepoint t and can be defined as

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du . \tag{2}$$

This function, that is the complementary to one of the cumulative distribution function $F(t)$, is non-increasing and right-continuous with $S(0) = 1$ and $\lim_{t\to+\infty}$ (t) =0.

The hazard function measures the probability that an individual has the event of interest at time t conditional that the event has not occurred until that time. Formally, it is defined as

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} . \tag{3}$$

It specifies the instantaneous rate at which events occur for subjects that are surviving at time t.

2.1. COX REGRESSION MODEL

The Cox proportional hazards (PH) regression model (Cox, 1972) is one of the most used classical methods for analyzing survival data. It allows to estimate the hazard of a subject depending on a set of covariates. The main assumption of the model is the proportionality of hazards, implying that the hazards of two groups of subjects, $h_1(\cdot)$ and $h_2(\cdot)$, are proportional, so that their ratio is constant over time:

$$\Psi = \frac{h_1(t)}{h_2(t)} \qquad \forall t, \tag{4}$$

where Ψ is a constant called *hazard ratio* (HR) or *relative hazard*.

The Cox PH model can then be expressed as

$$h_i(t) = h_0(t)\Psi = h_0(t)e^{\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_K X_{Ki}} = h_0(t)e^{\sum_{k=1}^{K}\beta_k x_{ki}} ,$$

where $h_i(t)$ represents the hazard function for the i^{th} subject; $h_0(t)$ is the baseline hazard, that is the risk for a subject whose values of all the independent variables are equal to zero; x_{ki} is the observed value of the k^{th} covariate for the i^{th} subject and β_k is the related coefficient.

Due to the presence of censoring, only a partial likelihood can be considered. The partial log-likelihood function can be expressed as

$$\ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\delta_i\left[\boldsymbol{\beta}'\boldsymbol{x}_i - \ln\left(\sum_{l\in R_j} e^{\boldsymbol{\beta}'\boldsymbol{x}_l}\right)\right] = \sum_{i=1}^{n}\delta_i\boldsymbol{\beta}'\boldsymbol{x}_i - \ln\left(\sum_{l\in R_j} e^{\boldsymbol{\beta}'\boldsymbol{x}_l}\right)\sum_{i=1}^{n}\delta_i . \tag{5}$$

where n is the number of subjects in the sample, $\boldsymbol{x}_i$ is the observed covariate vector for the i^{th} subject who experienced the event at the j^{th} ordered event time $t_{(j)}$ and R_j is the set of subjects at risk (risk set) at time $t_{(j)}$. According to this expression, only uncensored subjects ($\delta_i = 1$) have a direct effect on (5); on the contrary, censored observations do not directly contribute to the likelihood, but indirectly enter in the likelihood function because all the subjects are included in the risk set.In presence of ties, approximations of the two functions are needed (Collett, 2015).

Then, the β coefficients are estimated maximizing the partial log-likelihood, using iterative methods as the Newton-Raphson algorithm (Collett, 2015). Each coefficient represents the estimated change in the logarithm of the hazard ratio in correspondence of a change of the corresponding covariate. Usually, their exponential is considered, measuring the hazard ratio. A value of e^{β_k} greater (lower) than 1 indicates that for a one-unit increase in the continuous variable X_k, the hazard increases (decreases) by e^{β_k}, or, in an analogous way, if X_k is categorical, that a subject in group k has a higher (lower) hazard (equal to e^{β_k}) relative to a subject in the reference group.

The Cox PH model is called semiparametric because it is based on a nonparametric component (the baseline hazard - no distributional assumption is made for survival times) and on a parametric term.

Once the model has been fitted, stepwise variable selection can be used, based on different information criteria, as the Akaike information criterion (AIC). This selection method can be an efficient way to select a parsimonious model, because of the limited computation time and the possibility of tracking the variable selection process easily, allowing also to have further information on the excluded variables. However, it has also some disadvantages. Among these, there are (i) multiple comparisons problems and (ii) biased regression coefficient estimates (Harrell, 2015). Moreover, the obtained results may depend on the criterion used and on the ordering of the selected variables. Furthermore, another disadvantage concerns the fact that it is not possible to carry out an exhaustive analysis of all the possible combinations of the K predictors. Finally, in many situations, variable selection using stepwise regression shows a high unstability, especially when the sample size is small compared to the number of candidate variables, because many variable combinations can fit the data in a similar way (Derksen and Keselman, 1992).

2.2. REGULARIZED COX REGRESSION MODEL

In high-dimensional data contexts, usually, the interest is on variable selection in order to identify, among the many available variables, the most important ones. Thus, variable selection helps determining all the (informative) variables that are strictly related to the outcome, removing uninformative variables that decrease the precision and increase the complexity of the model. So, variable selection provides a balance between parsimony and goodness of fit of the model.

To this extent, regularized models are a good choice because they can allow to obtain a sparse model with many estimated coefficients equal to zero. In particular, they are based on the minimization of a loss function under a constraint that penalizes the flexibility of the model. Also, for Cox PH regression model different regularizations have been proposed (e.g. Lasso, Ridge, Elastic net). Among these, the Lasso (Least absolute shrinkage and selection operation), firstly proposed by Tibshirani in 1997 (Tibshirani, 1997), is one of the most used if the aim is variable selection. This because the Lasso is based on the use of a ℓ_1-norm penalty, that allows to obtain a well-defined solution with few nonzero coefficients β_k (Simon et al., 2011). Therefore, the Lasso is advantageous in terms of interpretation of the model and computational convenience (Hastie et al., 2015). Moreover, it is an interesting and useful method because it simultaneously performs feature selection among all the covariates and estimates the regression coefficients. The only requirement is to have standardized variables. However, the Lasso has also some limitations related to the possibility of obtaining biased estimates; so, it is not possible, for example, to combine the estimates with standard errors and to make inference through the estimation of confidence intervals and hypothesis testing.

Regularized parameters can be obtained by minimizing the negative partial log-likelihood $l(\beta)$ (see Equation (5)) under the constraint that the sum of the absolute values of the parameters is bounded by a constant (the Lasso penalty):

$$\hat{\beta} = argmin_{\beta} - l(\beta) \, \text{subject to} \sum_{k=1}^{K} |\beta_k| \leq s \text{ , with } s > 0.$$

The regularization parameter s is a non-negative tuning parameter that controls the impact of the penalty. The larger the value, the lower the amount of shrinkage (Ekman, 2017).

K-fold cross-validation is used for identifying the best parameter s; the optimal regularization parameter is the one that minimizes the cross-validation error. In survival analysis one of the most used performance measures is the Harrell's concordance index (C-index), a ranking measure based on the concordance of observed and predicted values (Harrell et al., 1982). This index is therefore used for measuring the cross-validation error during the estimation of the regularized parameters. As a higher C-index value means a better performance, the cross-validation error is measured as $1-C$ (Tay et al., 2022).

The Lasso technique for variable selection in the Cox model is a worthy competitor to stepwise selection (Tibshirani, 1997), a variable selection procedure usually performed on the basis of AIC. From the simulation studies performed by Tibshirani in 1997 (Tibshirani, 1997) it emerged that the Lasso is less variable than the stepwise Cox, still yielding interpretable models.

In the case study that will be presented in the next section, analyses have been performed by using *R* (R Core Team, 2021). In particular, `survival` and `glmnet` packages have been used for estimating the Cox and the Lasso Cox regression models. Finally, `riskRegression` has been used for evaluating the time-dependent area under the curves (AUCs).

3. DATA AND STUDY DESIGN

NBA data were analyzed. In particular, we considered the 2020-2021 regular season and divided it in two segments, the pre- and the post-All-Star (AS) game. The pre-AS game data have been used for extracting the baseline covariates (retrieved from the NBA website), while, play-by-play data have been used for the follow-up. This dataset has been kindly made available by BigDataBall (`www.bigdataball.com`), a reliable source of validated and verified data for the NBA, the Major League Baseball (MLB), the National Football League (NFL) and the Women's NBA (WNBA). All the variables included in this study have been chosen to characterize the performance abilities of each player.

A sample of $n = 359$ players has been considered. For each player a set of $K = 34$ baseline covariates $\boldsymbol{X} = (X_1,X_2,...,X_K)$ corresponding to the main achievements gained in the pre-AS game has been observed. Let's denote with $\boldsymbol{x}_i = (x_{1i},x_{2i},...,x_{Ki})$ the vector of observed baseline covariates for the i^{th} player. The full set of covariates is listed in Table 1. Besides the main players' achievements and some statistics of the relative team, two categorical variables were created: All-Star game (if the player was selected or not for playing at this competition) and G-League (if the player also played in the young championship).

Play-by-play data of the post-AS game season segment have been analyzed for extracting the needed information relative to the outcome variable. In detail, for each player, the minutes played until different time points (time referred to the second season segment - $s_1,...,s_J$) and the corresponding scored points were collected. Let's denote with M_j and P_j, respectively, the variables relative to the minutes played until time s_j ($j = 1,2,...,J$) and the corresponding scored points. For each player, we recorded the amount of minutes $m_{i\,j}$ played at time s_j, and the points $p_{i\,j}$ gained after having played $m_{i\,j}$ minutes (see Table 2). So, the time variable M_j was treated as player-time: $m_{i\,j}$ increases when the player is in the court and remains constant when he is not playing.

Then, we fixed a given threshold P of scored points and we defined the event of interest as the exceeding of that threshold. Censoring occurred when the player did not exceed the fixed amount of points at the end of the post-AS game regular

Table 1: Baseline variables

Variable	Type
FGM - Field Goals Made	Numeric
FGA - Field Goals Attempted	Numeric
FG% - Percentage of Field Goals Made	Numeric
3PM - Three-Point Shots Made	Numeric
3PA - Three-Point Shots Attempted	Numeric
3P% - Percentage of Three-Point Shots Made	Numeric
2PM - Two-Point Shots Made	Numeric
2PA - Two-Point Shots Attempted	Numeric
2P% - Percentage of Two-Point Shots Made	Numeric
FTM - Free Throws Made	Numeric
FTA - Free Throws Attempted	Numeric
FT% - Percentage of Free Throws Made	Numeric
OREB - Offensive Rebounds	Numeric
DREB - Defensive Rebounds	Numeric
REB - Rebounds	Numeric
AST - Assists	Numeric
TOV - Turnovers	Numeric
STL - Steals	Numeric
BLK - Blocks	Numeric
PF - Personal Fouls	Numeric
FP - Fantasy Points	Numeric
DD2 - Double Doubles	Numeric
TD3 - Triple Doubles	Numeric
$+/-$ - Plus/Minus	Numeric
Age	Numeric
GP - Games Played	Numeric
Percentage Won matches (player)	Numeric
Percentage Loss matches (player)	Numeric
MIN - Minutes played	Numeric
Percentage won matches (team - per game)	Numeric
PTS (player) - Points gained (player)	Numeric
PTS (team) - Points gained (team - per game)	Numeric
All-Star game	Categorical (Yes-No)
NBA G-League	Categorical (Yes-No)

season segment. Two different settings with percentages of censoring equal to 20% and 50% were analyzed, corresponding to threshold values equal to 99 and 255 points, respectively. Moreover, an exploratory analysis has been performed for other thresholds and corresponding censoring percentages.

Finally, the response variable can be defined as follows. We denote with $j_i^*(P) = \text{argmin}_{j=1,\dots,J} p_{ij} > P$ the index j corresponding to the first time that the i^{th} player exceeds the threshold P. The time at which the player exceeds the threshold is therefore $s_{j_i^*(P)}$. Thus, the outcome of the study is composed of (i) the time-to-event $\tau_i = \min(m_{ij_i^*(P)}, m_{iJ})$, where $t_i = m_{ij_i^*(P)}$ is the amount of minutes played by the player for exceeding the threshold and $c_i = m_{iJ}$ corresponds to the amount of minutes played at the end of the season segment, and of (ii) the event indicator $\delta_i = I[p_{ij} > P]$. Then, the survival outcome for the i^{th} subject is

$$\tau_i = min(t_i, c_i) = \min(m_{ij_i^*(P)}, m_{iJ}) = \begin{cases} m_{ij_i^*(P)} & \text{if } \delta_i = I[p_{ij} > P] = 1 \\ m_{iJ} & \text{if } \delta_i = I[p_{ij} > P] = 0 \end{cases} . \tag{6}$$

Table 2: Example of collected data. Each row refers to a player, and each column to a time point. In each cell the overall amount of minutes played and points gained until that given time point are recorded

s_1	s_2	$\cdots$	s_J
(m_{11}, p_{11})	(m_{12}, p_{12})	$\cdots$	(m_{1J}, p_{1J})
$\vdots$	$\vdots$	$\ddots$	$\vdots$
(m_{n1}, p_{n1})	(m_{n2}, p_{n2})	$\cdots$	(m_{nJ}, p_{nJ})

4. RESULTS

4.1. PRELIMINARY ANALYSIS

The overall sample included 359 players, after having excluded those who changed team during the season and those who played less than 48 minutes in all the post-AS game season segment. Two distinct cases have been analyzed, with percentage of censoring 20% and 50% respectively, in order to examine if the covariates have a different impact on the outcome. The points' thresholds that allowed the desired censoring percentage were 99 and 255 for the setting with

20% and 50% censoring, respectively. Moreover, an exploratory analysis has been performed for other thresholds and corresponding censoring percentages.

The game variables denoting the total number of achievements of each player in the analyzed period have been normalized dividing by his minutes played (to have comparable results). Then, all the covariates have been standardized (to have mean 0 and standard deviation 1).

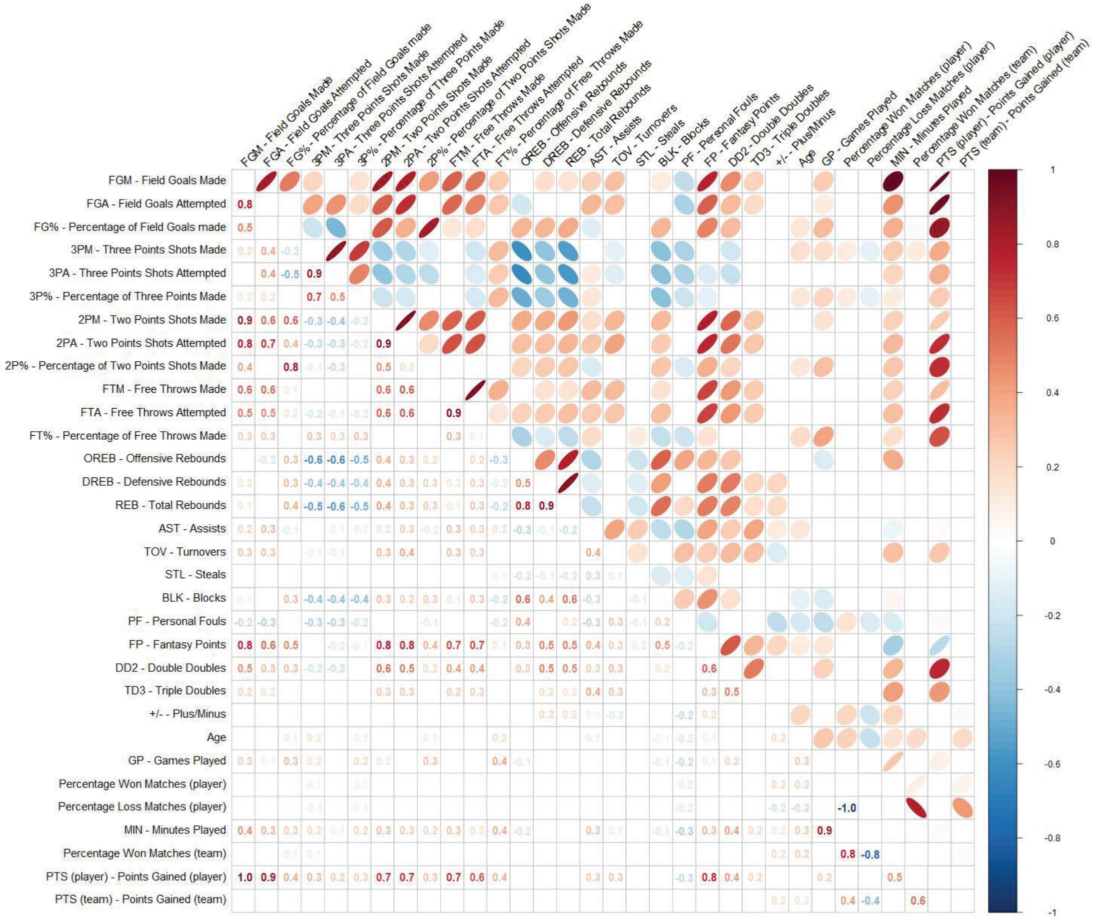
The first step of variable selection was carried out on the basis of prior knowledge and through the examination of correlations analysis. Indeed, considering all the NBA statistics would imply a high risk of multicollinearity, due to the presence of highly correlated variables (see Figure 1a). Therefore, we excluded some redundant variables (Figure 1b). After this step, the set of baseline covariates passed from 34 to 23. All these variables, as already pointed out, refer to the first season segment, i.e. the part of the regular season before the All-Star game. Among the excluded variables there is also the amount of fantasy points, due to the high multicollinearity found in another study (Macis et al., Submitted).

The following step of variable selection involved the use of stepwise Cox regression model and its regularized version through the Lasso, for selecting the most important variables.

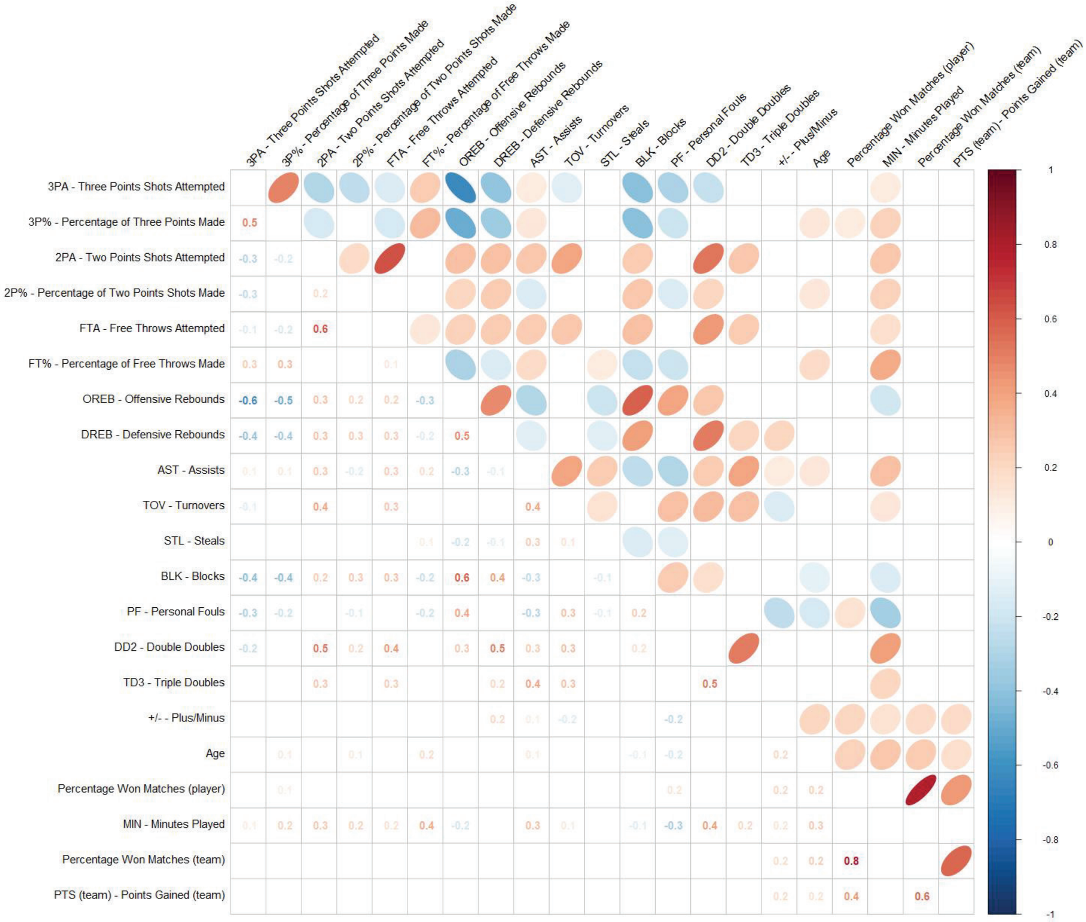
4.2. REGRESSION ANALYSIS

PERFORMANCE OF NBA PLAYERS

Using as a threshold 99 points, 287 players exceeded the given amount of points (corresponding to the 80% of the sample). Table 3 shows the results obtained fitting the two models. The stepwise Cox model and the Lasso identified almost the same variables (with the exception of the All-Star game variable). More in detail, both models selected the amount of minutes played in the pre-AS season segment, the number of attempted shots (free throws -FTA-, two- -2PA- and three- -3PA- pointers), the percentage of two-point shots made (2P%) and the number of gained double doubles (DD2). All these variables resulted positively associated with the outcome: An increase in the number of these achievements is associated to a higher probability of exceeding the threshold. Moreover, the stepwise Cox regression model and the Lasso Cox also identified the number of steals (STL), even if the estimated hazard ratio was not statistically significant in stepwise Cox model ($p = 0.144$) and in the Lasso it seems to have a low estimated impact on the outcome (HR approximately equal to 1.00). Finally, the Lasso also identified the All-Star game variable, even if with an estimated HR approximately equal to one.



(a) Full set of covariates



(b) Set of covariates after the first step of variable selection

Figure 1: Correlation plot of the a) full set of covariates b) set of covariates obtained after the first step of variable selection. The game variables denoting the total number of achievements of each player have been normalized dividing by his minutes played; all the covariates have been standardized. Ellipses towards left (right) indicate negative (positive) correlations

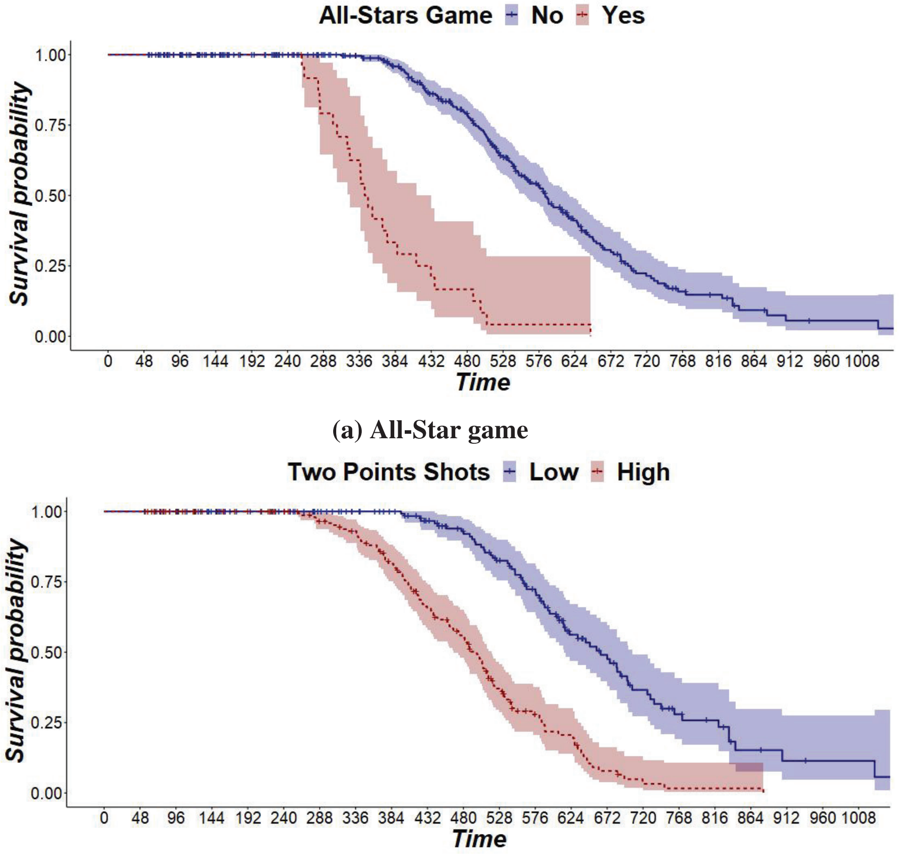
Increasing the threshold to 255 points, only one half (179 players) of the sample exceeded the points' cut-off. In this setting almost all the variables identified in the previous one were selected by the two models, but with some differences (Table 3). More in detail, the amount of minutes played in the previous season segment and the number of FTA were only identified by the Lasso with an estimated HR close to 1.00, suggesting a lower impact on the outcome. The 2P%, instead, was only identified by the stepwise Cox model. Moreover, increasing the threshold of points, the number of STL was found negatively associated with the outcome. The estimated coefficients resulted lower than 1 (equal to 0.77 and 0.95 in stepwise Cox and Lasso respectively), indicating that this achievement is negatively associated with the outcome: a unit increase of it leads to a lower probability of exceeding the threshold. Finally, with a higher threshold, the All-Star game resulted to be a relevant feature identified by both the models: having been selected for playing at the All-Star game doubles the probability of gaining more than 255 points.

Interestingly, most of the variables identified in this setting have a higher effect (higher HR) than in the setting with a lower threshold.

Figure 2 shows two examples of the estimated survival curves of the sample stratifying for two of the most important variables for the setting with a 50% of censoring. The number of 2PA (normalized and standardized) has been dichotomized in two categories with respect to the median. Figure 2a shows that players selected for the All-Star game reach the fixed amount of points very earlier than those who have not been selected for the match. Similarly, Figure 2b shows that players who attempted a higher number of two-point shots in the first part of the season have a higher probability of gaining the given threshold earlier than those who attempted a lower number of shots.

COEFFICIENTS AND PERCENTAGE OF CENSORING

Finally, an analysis of the pattern of the estimated coefficients as the percentage of censoring varies has been carried out. The results are shown in Figure 3. Each subfigure reports the estimated hazard ratios (i.e. $e^{\hat{\beta}}$) of each variable for different censoring settings (x-axis) for both the stepwise Cox and the Lasso Cox. In details, we analyzed the censoring percentages ranging from 10% to 75% with a step of 5%. It can be seen that, almost always, the hazard ratios estimated with the stepwise Cox are greater than those obtained by the Lasso Cox. The most important variables are the number of attempted two- and three-point shots (2PA and 3PA), the percentage of two-point shots (P2%), the number of double dou-



(a) All-Star game

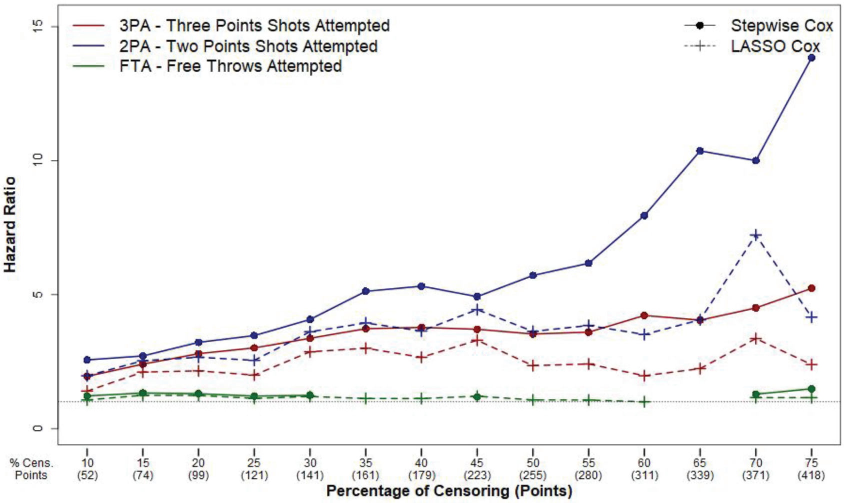
(b) Two-Point Shots

Figure 2: **Survival curves of the sample stratified for: a) All-Star game and b) Number of attempted two-point shots. The 2PA variable (normalized and standardized) has been dichotomized with respect to the corresponding median. The dashed line refers to the selection of the All-Star game and to the high category of 2PA, the solid line refers to having not been selected for the All-Star game and to the low category of 2PA.**

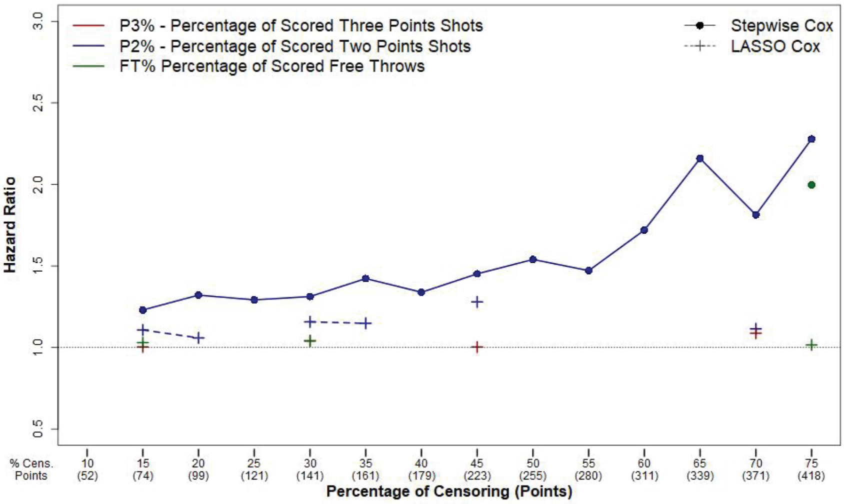
Table 3: Results of the variable selection procedure in both the two settings (20% **and** 50% **of censoring)**

	20% censoring		50% censoring	
Variables Included in the Model	Stepwise HR (p)	Lasso HR	Stepwise HR (p)	Lasso HR
3PA - 3-Point Shots attempted	2.80 (< 0.001)	2.16	3.53 (< 0.001)	2.35
3P% - % 3-Point Shots made				
2PA - 2-Point Shots attempted	3.22 (< 0.001)	2.66	5.72 (< 0.001)	3.63
2P% - % 2-Point Shots made	1.32 (0.004)	1.06	1.54 (0.003)	
FTA - Free Throws attempted	1.30 (0.002)	1.24		1.07
FT% - % Free Throws made				
OREB - Offensive Rebounds				
DREB - Defensive Rebounds				
AST - Assists				
TOV - Turnovers				
STL - Steals	0.90 (0.144)	1.00	0.77 (0.018)	0.95
BLK - Blocks				
PF - Personal Fouls				
DD2 - Double Doubles	1.27 (0.001)	1.21	1.27 (0.003)	1.21
TD3 - Triple Doubles				
$+/-$ - Plus/Minus				
Age				
% Won Matches (by the player)				
MIN - Minutes played	1.24 (0.004)	1.20		1.08
% Won Matches (by the team)				
Points Gained (by the team)				
All-Star Game (Yes/No)		1.05	2.27 (0.004)	1.86
NBA G-League (Yes/No)				

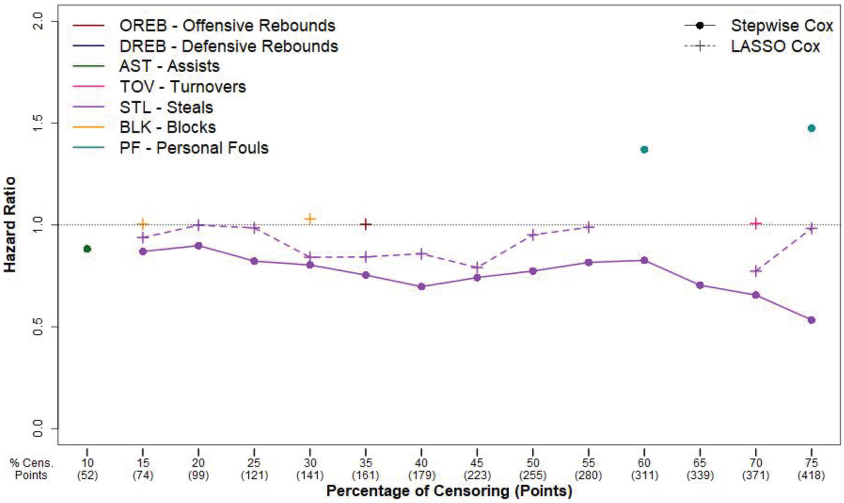
bles (DD2) and the number of steals (STL). These variables are selected in almost all the settings by the two models. Moreover, the All-Star game and the number of minutes played are selected many times by the stepwise Cox and always by Lasso Cox. As the percentage of censoring increases, the estimated hazard ratios increase. On the other hand, the estimated hazard ratio for STL is always negative and its value decreases as the fixed threshold (and consequently the percentage of censoring) increases.



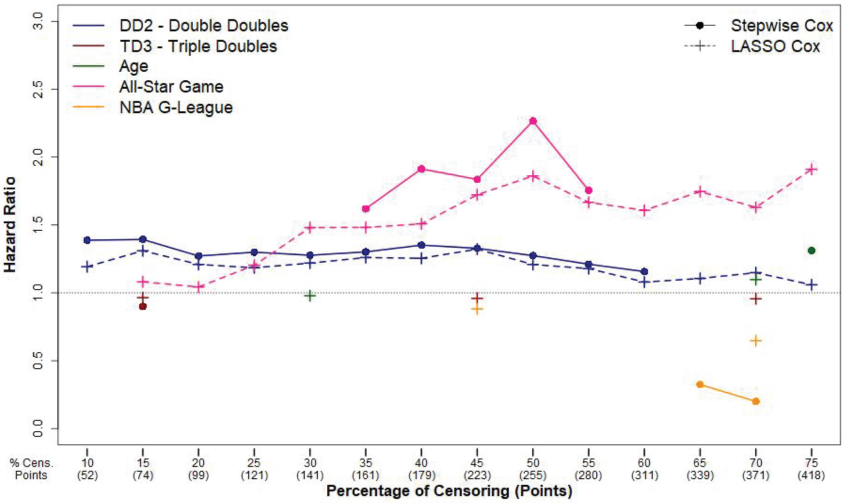
(a) Attempted shots



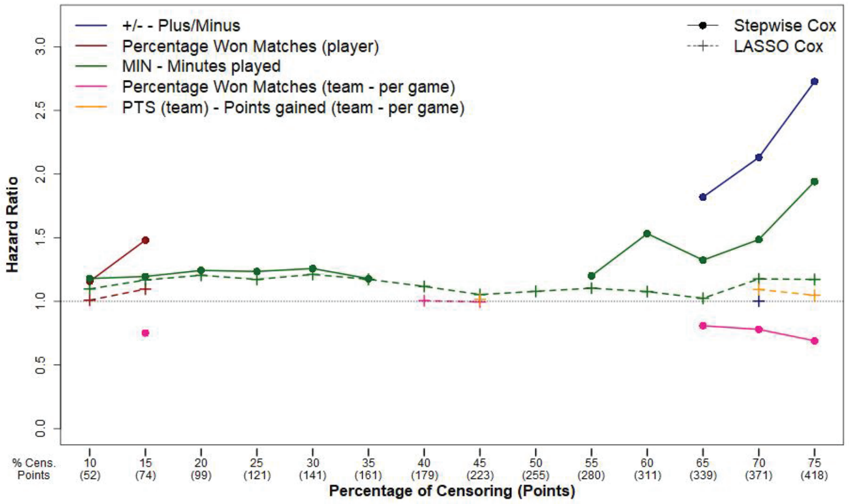
(b) Percentage of realized shots



(c) Defense and game construction



(d) Achievements



(e) Points and Winning Percentage

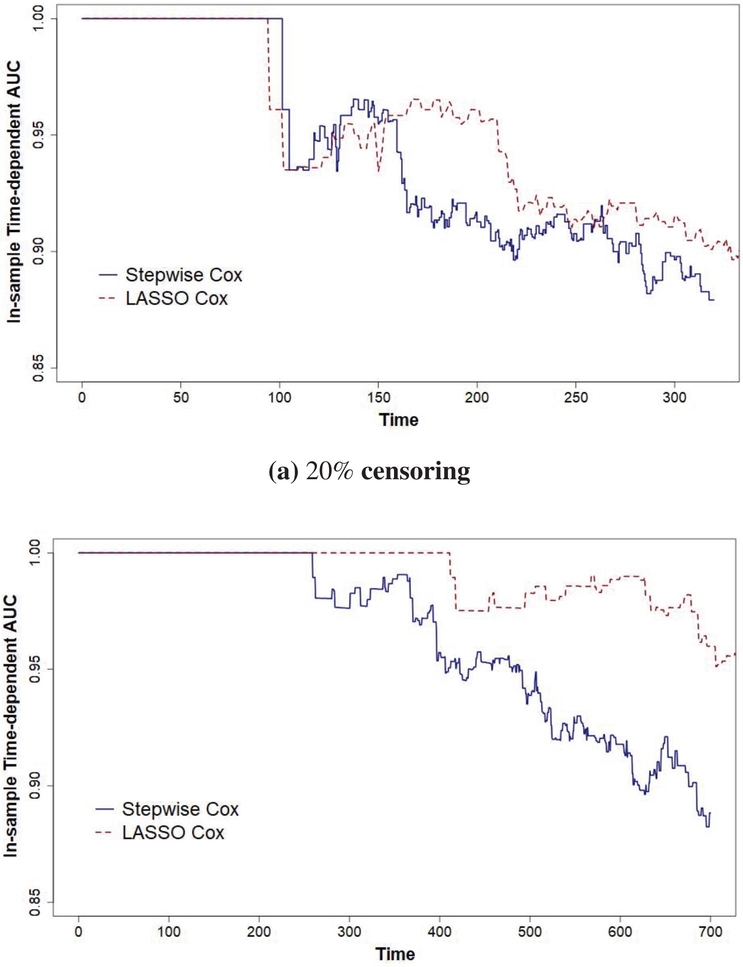
Figure 3: Estimated hazard ratios for different censoring percentages

4.3. EVALUATION OF MODELS' PERFORMANCE

In order to evaluate the performance of the Lasso Cox model and test the assumption of proportionality of hazards, a Cox model with the variables selected by Lasso was fitted.

The two models satisfied the assumption of proportionality of hazards (null hypothesis) in both settings, as measured by the statistical test based on Shoenfield residuals ($p = 0.072$ and 0.077, respectively, for stepwise and Lasso Cox models, when the percentage of censoring was equal to 20%, and $p = 0.093$ and 0.077 when the percentage of censoring was equal to 50%).

Then, model's performance has been evaluated through the evaluation of time-dependent AUCs. The time-dependent AUC measures the area under the ROC curve evaluated at different timepoints; thus, it differs from classical ROC analysis because the outcome of an observation can change over time and because of the presence of censoring. The time-dependent AUC assesses the ability of the model to discriminate the binary outcome (event/non event) at different timepoints. Values close to one indicate a good performance. The performance assessment has been made in-sample, so the performance could have been overestimated. It can be seen that in both the settings the performance of the two models is good (greater than 0.85). Moreover, it can be observed that the Lasso seems to slightly overperform stepwise Cox in both the settings (Figure 4).



(a) 20% **censoring**

(b) 50% **censoring**

Figure 4: Time-dependent AUC for stepwise Cox and Lasso Cox in the two settings. a) 20% **of censoring. b)** 50% **of censoring.**

5. DISCUSSION

Survival analysis has been already used for sport analytics for many aims; however, up to now, to the best of our knowledge, it has never been used for evaluating players' performance. This study shows the use of a classical method of survival analysis, as stepwise Cox regression, and of a more recent extension, regularized Cox regression through Lasso, for identifying the achievements that are highly associated to the offensive performance of NBA players, measured in terms of exceeding of a given threshold of points. Two settings were analyzed, with thresholds equal to 99 and 255 points, corresponding to censoring percentages of 20% and 50%, respectively, for examining whether there is a different impact of the considered variables on the outcome. Other reasonable values for the percentage of censoring that can be investigated are those ranging from 10% to about 75%, as shown in Figure 3. Values higher than 75% could instead lead to possible non-meaningful results, because in these cases it is more likely that the follow-up is too short with respect to the event under analysis. However, besides the meaning of the results obtained from the analyses, attention has to be paid to the proportional hazard assumption, which may not be respected in all the settings.

Summarizing, in the two examined settings almost all the same variables were selected by both the models. In particular, from both sáepwise Cox regression and Lasso, it emerged that (as expected) players who attempted more shots, free throws, 2− and 3−point shots in the pre-AS game season segment have a higher probability of exceeding the fixed amount of points in the second part of the season. In particular, it emerged that the most important variable, i.e. the one with the highest estimated hazard ratios, is the number of 2PA, followed by the number of 3PA, and that their impact increases when the threshold is higher. Moreover, both the two models suggest that gaining more achievements (as measured by the number of DD2) is associated to an increase of the probability of success in a shorter time. In addition, the number of STL was identified as negatively associated to the outcome. Thus, it seems that this variable decreases the probability of reaching the two thresholds (HR $<$ 1). Its importance is relevant when the threshold is high, while it is not statistically significant ($p > 0$ .05) when the fixed threshold of points is low. This is an attractive result because it is the only variable related to defense included in the models. This may suggest that who is more involved in defense is then penalized in terms of scored points. On the other hand, variables like rebounds and blocks have not been selected by the models and, from this point of view, defense seems to be not influential on the scored points.

These remarks about the role of defense should be deepened with some research specifically devoted to answer this question.

Finally, an interesting finding emerged from the comparison of the results of the two settings; indeed, All-Star game becomes an important factor when considering a higher threshold. Therefore, when the fixed amount of points is higher, being a very good player (and so having been selected for playing at the All-Star game) almost doubles the probability of reaching that threshold (HR=2.3 and 1.9 for stepwise Cox and Lasso respectively).

All these results are also confirmed by those shown in Figure 3 for different percentages of censoring.

An interesting idea is to also consider the features of the opponent teams, in order to verify whether the opponent teams have an impact on the players' performance. This could be done, for example, by weighting some of the covariates (e.g. 2PA and 2PM) with respect to the ranking of the corresponding opponent team, in order to also consider the possible impact of the team against which the achievements have been gained.

The study has some limitations associated to possible issues related to the assumption of *random censoring*. Random censoring occurs when the subjects who are censored at time t are representative of all the study subjects who remained at risk at time t, with respect to their survival experience (Kleinbaum et al., 2012). This hypothesis may be not respected due to the fact that it is likely that censored players (i.e. players who didn't exceed the threshold) have not the same abilities of players who manage to exceed that amount of points. On the other side, the assumption of *independent censoring* can be retained valid. Indeed, it is reasonable to assume that within any subgroup of interest, the subjects who are censored at time t are representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. So, random censoring could be assumed conditional on each level of covariates (Kleinbaum et al., 2012). For this reason, once having taken into account the abilities of each player, as measured through the covariates introduced in the model, the probability of being censored can be considered independent of the probability that the event of interest occurs. Future research will deepen this issue.

Moreover, due to the relatively low number of subjects, the full original sample has been used for fitting the model and performance has been evaluated in-sample. Future work will consider the use of data relative to 2021-2022 season as test set.

Finally, some improvements include considering the possible presence of interactions among covariates, non-linear effects of the predictors and threshold effects. Non-parametric and machine learning methods will be used to examine this point.

ACKNOWLEDGMENTS

The authors thank the reviewers for their valuable comments, which greatly improved the paper, the editor and the guest editors. Thanks go to BigDataball (`www.bigdataball.com`) for supplying us high quality sports data for scientific research.

This study was carried out in collaboration with the Big&Open Data Innovation Laboratory at the University of Brescia (project 'BDsports: Big Data analytics in sports'; `bdsports.unibs.it`).

REFERENCES

Back, F.A., Hino, A.A.F., Bojarski, W.G., Aurélio, J.M.G., de Castro Moreno, C.R. and Louzada, F.M. (2022). Evening chronotype predicts dropout of physical exercise: A prospective analysis. In *Sport Sciences for Health*, 19(1): 309-319.

Beynnon, B.D., Vacek, P.M., Murphy, D., Alosa, D. and Paller, D. (2005). First-time inversion ankle ligament trauma: The effects of sex, level of competition, and sport on the incidence of injury. In *The American Journal of Sports Medicine*, 33 (10): 1485–1491.

Buist, I., Bredeweg, S.W., Bessem, B., Van Mechelen, W., Lemmink, K.A. and Diercks, R.L. (2010). Incidence and risk factors of running-related injuries during preparation for a 4-mile recreational running event. In *British Journal of Sports Medicine*, 44 (8): 598–604.

Clarke, P.M., Walter, S.J., Hayen, A., Mallon, W.J., Heijmans, J. and Studdert, D.M. (2012). Survival of the fittest: retrospective cohort study of the longevity of Olympic medallists in the modern era. In *British Medical Journal*, 345.

Collett, D. (2015). *Modelling Survival Data in Medical Research.* CRC press.

Cox, D.R. (1972). Regression models and life-tables. In *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 (2): 187–220. URL `http://www.jstor.org/stable/2985181`.

Csurilla, G. and Fertő, I. (2022). How long does a medal win last? Survival analysis of the duration of Olympic success. In *Applied Economics*, 1–15.

Dekker, T.J., Godin, J.A., Dale, K.M., Garrett, W.E., Taylor, D.C. and Riboh, J.C. (2017). Return to sport after pediatric anterior cruciate ligament reconstruction and its effect on subsequent anterior cruciate ligament injury. In *Journal of Bone and Joint Surgery*, 99 (11): 897–904.

Del Corral, J., Barros, C.P. and Prieto-Rodriguez, J. (2008). The determinants of soccer player substitutions: A survival analysis of the Spanish soccer league. In *Journal of Sports Economics*, 9 (2): 160–172.

Derksen, S. and Keselman, H.J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. In *British Journal of Mathematical and Statistical Psychology*, 45 (2): 265–282.

Ekeland, A., Engebretsen, L., Fenstad, A.M. and Heir, S. (2020). Similar risk of ACL graft revision for alpine skiers, football and handball players: The graft revision rate is influenced by age and graft choice. In *British Journal of Sports Medicine*, 54 (1): 33–37.

Ekman, A. (2017). Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, Lasso and bootstrap approach. Master’s Thesis. Umea University. https://www.diva-portal.org/smush/get/diva2:1067479/FULLTEXT01.pdf.

Fynn, K.D. and Sonnenschein, M. (2012). An analysis of the career length of professional basketball players. In *The Macalester Review*, 2 (2): 3.

Gutiérrez, E., Lozano, S. and González, J.R. (2011). A recurrent-events survival analysis of the duration of Olympic records. In *IMA Journal of Management Mathematics*, 22 (2): 115–128.

Harrell, F.E. (2015). *Regression Modeling Strategies.* Spinger Series in Statistics, Springer, Cham.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. and Rosati, R.A. (1982). Evaluating the yield of medical tests. In *Journal of the American Medical Association*, 247 (18): 2543–2546.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). Statistical learning with sparsity. In *Monographs on Statistics and Applied Probability*, 143: 143.

Hopkins, W.G., Marshall, S.W., Quarrie, K.L. and Hume, P.A. (2007). Risk factors and risk statistics for sports injuries. In *Clinical Journal of Sport Medicine*, 17 (3): 208–210.

Howell, D.R., Potter, M.N., Kirkwood, M.W., Wilson, P.E., Provance, A.J. and Wilson, J.C. (2019). Clinical predictors of symptom resolution for children and adolescents with sport-related concussion. In *Journal of Neurosurgery: Pediatrics*, 24 (1): 54–61.

Jack, R.A., Sochacki, K.R., Hirase, T., Vickery, J., McCulloch, P.C., Lintner, D.M. and Harris, J.D. (2019). Performance and return to sport after hip arthroscopic surgery in major league baseball players. In *Orthopaedic Journal of Sports Medicine*, 7 (2): 2325967119825835.

Kleinbaum, D.G., Klein, M. et al. (2012). *Survival Analysis: A Self-Learning Text*, vol. 3. Springer.

Kontos, A.P., Elbin, R., Sufrinko, A., Marchetti, G., Holland, C.L. and Collins, M.W. (2019). Recovery following sport-related concussion: Integrating pre-and postinjury factors into multidisciplinary care. In *The Journal of Head Trauma Rehabilitation*, 34 (6): 394–401.

Lawrence, D.W., Richards, D., Comper, P. and Hutchison, M.G. (2018). Earlier time to aerobic exercise is associated with faster recovery following acute sport concussion. In *PLoS One*, 13 (4): e0196062.

Lu, Y., Jurgensmeier, K., Till, S.E., Reinholz, A., Saris, D.B., Camp, C.L. and Krych, A.J. (2022).Early ACLR and risk and timing of secondary meniscal injury compared with delayed ACLR or nonoperative treatment: A time-to-event analysis using machine learning In *The American Journal of Sports Medicine*, 03635465221124258.

Macis A, Manisera M, Sandri M, et al (2023) Which achievements are associated with a better offensive performance in NBA? A survival analysis study. In: *13th World Congress of Performance Analysis of Sport & 13th International Symposium on Computer Science in Sport (IACSS2022), vol. 1448 of Advances in Intelligent System and Computing*. https://doi.org/10.1007/978-3-031-31772-9.

Mahmood, A., Ullah, S. and Finch, C. (2014). Application of survival models in sports injury prevention research: A systematic review. In *British Journal of Sports Medicine*, 48 (7): 630–630.

Mai, H.T., Chun, D.S., Schneider, A.D., Erickson, B.J., Freshman, R.D., Kester, B., Verma, N.N. and Hsu, W.K. (2017). Performance-based outcomes after anterior cruciate ligament reconstruction in professional athletes differ between sports. In *The American Journal of Sports Medicine*, 45 (10): 2226–2232.

Moulds, K., Abbott, S., Pion, J., Brophy-Williams, C., Heathcote, M. and Cobley, S. (2020). Sink or swim? A survival analysis of sport dropout in Australian youth swimmers. In *Scandinavian Journal of Medicine & Science in Sports*, 30 (11): 2222–2233.

Nelson, L.D., Tarima, S., LaRoche, A.A., Hammeke, T.A., Barr, W.B., Guskiewicz, K., Randolph, C. and McCrea, M.A. (2016). Preinjury somatization symptoms contribute to clinical recovery after sport-related concussion. In *Neurology*, 86 (20): 1856–1863.

Nevo, D. and Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. In *Journal of Quantitative Analysis in Sports*, 9 (2): 165–177.

Pion, J., Lenoir, M., Vandorpe, B. and Segers, V. (2015). Talent in female gymnastics: A survival analysis based upon performance characteristics. In *International Journal of Sports Medicine*, 94 (11): 935–940.

Pratas, J.M., Volossovitch, A. and Carita, A.I. (2016). The effect of performance indicators on the time the first goal is scored in football matches. In *International Journal of Performance Analysis in Sport*, 16 (1): 347–354.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. In *Journal of Statistical Software*, 39 (5): 1.

Smith, K.L. and Weir, P.L. (2022). An examination of relative age and athlete dropout in female developmental soccer. In *Sports*, 10 (5): 79.

Sochacki, K.R., Jack, R.A., Hirase, T., Vickery, J. and Harris, J.D. (2019). Performance and return to sport after hip arthroscopy for femoracetabular impingement syndrome in National Hockey League players. In *Journal of Hip Preservation Surgery*, 6 (3): 234–240.

Tay, K., Simon, N., Friedman, J., Hastie, T., Tibshirani, R. and Narasimhan, B. (2022). *Regularized Cox Regression.* https://cran.r-project.org/web/packages/glmnet/vignettes/coxnet.pdf.

Thomas, A.C. (2007). Inter-arrival times of goals in ice hockey. In *Journal of Quantitative Analysis in Sports*, 3 (3).

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. In *Statistics in Medicine*, 16 (4): 385–395.

Tozetto, A.B., Carvalho, H.M., Rosa, R.S., Mendes, F.G., Silva, W.R., Nascimento, J.V. and Milistetd, M. (2019). Coach turnover in top professional Brazilian football championship: A multilevel survival analysis. In *Frontiers in Psychology*, 10: 1246.

Venturelli, M., Schena, F., Zanolla, L. and Bishop, D. (2011). Injury risk factors in young soccer players detected by a multivariate survival model. In *Journal of Science and Medicine in Sport*, 14 (4): 293–298.

Wangrow, D.B., Schepker, D.J. and Barker III, V.L. (2018). Power, performance, and expectations in the dismissal of NBA coaches: A survival analysis study. In *Sport Management Review*, 21 (4): 333–346.

Zuccolotto, P. and Manisera, M. (2020). *Basketball Data Science: with Applications in R.* CRC Press.

Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J., Bikandi, E., Setuain, I., Lekue, J., Küchenhoff, H. and Lee, D.J. (2021). Prediction of sports injuries in football: A recurrent time-to-event approach using regularized Cox models. In *Advances in Statistical Analysis*, 1–26.

doi.org/10.26398/IJAS.0035-010

FOOTBALL ANALYTICS BASED ON PLAYER TRACKING DATA USING INTERPOLATION TECHNIQUES FOR THE PREDICTION OF MISSING COORDINATES

Christos Kontos and Dimitris Karlis
Department of Statistics, Athens University of Economics and Business, Athens, Greece

Abstract. *In recent days we have seen an increasing interest in using tracking data for sports and especially for football. Such data can reveal the location of the players and the ball many times per second allowing for examining tactics, efficiency of players, formations, and other characteristics of the game. Unfortunately, such systems are still expensive and their data are not widely available. Alternatively, limited tracking data can be obtained from broadcasting videos. They are of less quality and of course they are censored in the sense that they do not provide information for all players but only those in the frame taken. Within this framework, the primary aim of this paper is the exploration of the most suitable method for retrieving the missing information of players' and ball's positions, rectifying as much as possible the effect of censoring which leads to discontinuous player tracks and unreliable player identification. In this paper we explore and compare different interpolation methodologies. Moreover, we distinguish possible differences between the actual data, as they were tracked from the camera and the interpolated data that have been estimated from our best selected method, by extracting insights to support tactical analyses as well as players' performance evaluation.*

Keywords: *Football analytics, Player tracking data, Missing values, Interpolation techniques, Regression and time series algorithms*

1. INTRODUCTION

Due to the rapid development of tracking technologies, processing software and more powerful data storages, analytics in sports industry and especially in football, has become increasingly popular in the last decade. Nowadays, teams and sports organizations are increasingly interested on data to inform players,

kontos.christos@yahoo.com, karlis@aueb.gr
Corresponding Author: Dimitris Karlis (ORCID: 0000-0003-3711-1575)

coaches and other stakeholders, facilitate decision making both during and prior to sporting events. Major fields of applications of sports analytics include scouting, revenue increase, performance analysis (Mohr et al., 2003), team tactics (Rein and Memmert, 2016), match outcomes (Karlis and Ntzoufras, 2003), player development (Rampinini et al., 2007), injuries prevention and rehabilitation (Dvorak et al., 2000) among others.

With the advancement of technology, a new data source that is commonly used nowadays has been created, the so-called tracking data (Rein and Memmert, 2016). In our days a huge amount of tracking data is being collected for nearly every popular professional team sport. More and more companies are now offering their services, using either camera-based systems, or wearable technology. Over the last few years, initiatives and techniques have been implemented around the area of player tracking data. One of the main ingenuities are the different implementations that are based on analyses of coordinate data that are generated by converting a traditional broadcast video feed to player locations, utilizing different computer vision and artificial intelligence techniques (Lu et al., 2013). This type of data could be easily interpreted as broadcast tracking data. In contrast with the traditional multicamera tracking data, the broadcast tracking data are currently distinguished from their advancement of availability they provide (websites of Skillcorner and Sportlogiq for example).

As the collection of event and tracking data has been rapidly evolved, different statistical methodologies have been applied in various areas of interest. For example, Link et al., (2016) presented an approach to quantify the attacking performance in football, characterizing as dangerousity the probability of a goal being scored for every point in time at which a player is in possession of the ball. In addition, Bialkowski et al. (2016) an unsupervised method has been examined in order to learn a formation template which allows to align the tracking data at the frame level. An innovative hierarchical clustering method has also been developed by Diquigiovanni and Scarpa, (2018) in order to divide a sample of undirected weighted networks into groups. (Horton et al., 2014) presented a model that constructs numerical predictor variables from spatiotemporal match data using feature functions based on methods from computational geometry. Another research has been conducted, which estimated how both the risk and reward of a pass across tracking data can be measured (Power et al., 2017). Finally, Shaw and Glickman, (2019) developed an innovative model for the computation of the formation of any team at any moment of a match, by taking the position of team members relative to their teammates. The examples above are indicative of the increasing interest about tracking data and of course is far from being a complete list of the existing applications.

Broadcast tracking data contain less information as we collect data only for the players in the video frame. One major challenge, for applying existing methods for multicamera tracking data to broadcast tracking data is the issue of censoring and the related effect of missingness (Mortensen, 2020). In addition, the main information that is currently extracted from this type of data, is based at the location of events (passes, shots on target, goals, etc.) and players. Therefore, any action that happens outside the recorded events is not easily inferable and cannot be taken into consideration for future decisions. Another great challenge, is the ability to derive powerful insights from frame-by-frame tracking data and provide all the necessary information needed for the event of interest. Many sports organizations nowadays are interested in relationships that can occur from specific outcomes or when they are using a specified type of formation when they are attacking or defending. By using this type of data, often leads to millions of possible locations of players and thus results in undesirable results due to the complexity of data.

Existing methods for tracking data when capturing players' movements, are widely used by professional football clubs to provide insights into activity demands during training sessions and competitive matches. Global positioning systems (GPS), local positioning systems (LPS) trackers and accelerometers, are some of the specific methods. In recent years, computer vision techniques can be applied directly to broadcast video in order to identify the precise location and movement of players and ball. Automated broadcast tracking data, have made a huge leap forward, as they enable locations of objects to be accurately collected from a standard television broadcast, without the need for any capital investment in stadiums or human operator costs (SkillCorner, 2020). However, despite their undisputed value, broadcast tracking data, lag behind on a very basic issue, and that is they are only available for a player and the ball as long as they are observed inside the main broadcast camera shoot, which pans left and right across the pitch (see Figure 1).

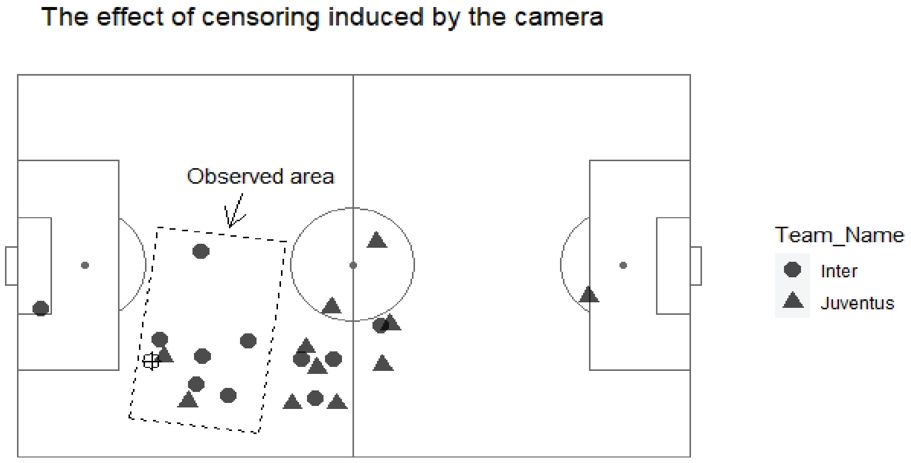


Figure 1: The effect of censoring induced by the camera. The rectangular area refers to the video frame, all players outside the rectangular cannot be positioned in the field and hence their locations are unknown.

Hence, our purpose in the current work is to estimate the missing coordinates of the players and the ball that are produced from the censoring effect of the broadcast camera. In that way, we tried to predict the exact position of each object during the live broadcast of a game. To make this clear we want to "estimate" the positions of the players not seen in the frame based on the existing information from the frame and past frames. We proceeded also with the exploration and discovery of the best and most accurate method for filling the missing coordinates of players and rectify as much as possible the effect of censoring, that is mainly produced from broadcast tracking data and often leads to discontinuous player tracks and unreliable player identification. We explore and compare different approaches, by using a number of various interpolation algorithms, non-linear regression models and a time series forecasting approach. We also try to examine possible differences generated between the interpolated data and the actual data as they are derived from the broadcast camera, by extracting important insights that are based on the tactical analyses of teams as well as on the players' performances on different types of scenarios. Finally, as an extra step of further research, we try to derive important insights by effectively estimate teams' formations and players' consistencies based on their initial formations, as well as possible correlations when a team is attacking or defending, using an appropriate pitch control model that quantifies the probability for a player to control the ball assuming it is at that location.

The remaining of the paper is organized as follows: In Section 2, we present the data sources used on this work. The methodological approach is described in Section 3. Section 4 applies the methodology to the data. We present a comparison of the different approaches but also certain applications for tracking data based on the interpolated data so as to illustrate the potential of the methodology. Finally, concluding remarks, limitations and further research taken into consideration are also discussed in Section 5.

2. DATA USED

The dataset used for this work, retrieved from an open-source joint initiative between SkillCorner and Friends of Tracking. The specific repository consists of 9 football games of broadcast tracking data, collected through computer vision and machine learning out of the broadcast video and are referring to the 2019/2020 league matches between the champions and runners up in English Premier League, French L1, Spanish LaLiga, Italian Serie A and German Bundesliga. In the current paper, a match from the Italian Serie A, between Inter and Juventus was selected (around 600.000 observations). Each match consists of two files. The first file, contains all the necessary information about the lineup formations, players, referee and ball characteristics, pitch size, coaches etc. The second file, includes all the tracking data for the players, the referee and the ball. For the spatial coordinates, the unit of the field modulization is the meter and the center of the coordinates is at the center of the pitch (0,0). The initial predetermined dimensions of the field have a size of 105m x 68m and the x axis is the long side of the field while the y axis is the short side of the field. The frame of the video of the data comes from at 10 frames per second and the timestamp in the match has a precision of 1/10 seconds. The broadcast shows an average of 14 players out of 22 at each frame and during replays or close up views, the data is not included. Therefore, the actual data present missing information for all players throughout the duration of the game at different intervals each time. Finally, 95% of the player identity that is provided is accurate while for the rest 5% the algorithm could not identify for sure the player.

For the data cleansing and feature engineering part, several procedures were implemented before running our analyses. Firstly, we had to exclude all the unwanted list of elements existing in the tracking data file, as they did not correspond to the actual playing time and thus did not provide us with any important information. The elements of the tracking data file were related with information regarding the possession, the frames of the video, the coordinates of the players, as well as the period and the time of the game. Therefore, we tried to match every single player with the elements of the tracking data file, by using the unique identifier of each player as exists in the first file which included all the necessary information about the players characteristics. As a next step, we treat each element of the list as a unique data frame, in order to include all the necessary information existed at each unique element, resulting in five (5) unique data frames of a total 70.299 observations each. Also, by unlisting all the elements of the specific list (≈ 600.000 observations), we took care that each frame had to be repeated equal times as the number of elements that is included at each unique list of the initial list. The (x, y) coordinates of each object (players and ball) have been also rescaled, in order to be aligned with the common size of a field.

Finally, after a thorough check on the data, we decided to increase the number of players that were visible on each frame (14 players out of 22 were visible at each frame on average). At some instances of the game, if a player was not easily observable, i.e., it was not possible from the broadcasting to identify him, the column referring to the unique identifier of a player ascribes as values the "home team" or "away team" attributes accordingly. For the specific implementation, we relied on a variable acting like a unique identifier when a player is observed on consecutive frames. It should be mentioned that for a number of slim picking instances during the game, some players had also the same unique identifier with a different player. On that account, we tried to remove the players that were observed in the same frame, by keeping the player who appeared first in the frame and had the correct unique identifier with what had been stated to him from the beginning. This issue provoked from the fact that during the game the positions of two (2) or more players, were so close to each other that the technique of detecting a player eventually ended up detecting more than one.

Before the implementation of any of the proposed methods that are discussed in Section 3, we had to properly define first the proportions of missing observations for each one of the objects (players and ball), that were induced due to the censoring effect of the camera. Afterwards, we resample the data from unevenly time data to equi-spaced time data, as between time increments the data were missing without any specific ordinance or pattern (replays, close-up views, goals, etc.). This gives an idea of the kind of data destruction that we can have (Figure 2). In Figure 2 one can see the full data for a player at the left and those with missing values for the same player at the right, while we have randomly omitted some of the time stamps pretending that the data were censored. This resampling approach allows to check different proportion of missingness to investigate the effect. In principle what we want to be able to do is to derive the full path the left using the available data at the right. Note that for each player we have a different proportion of missingness and also different patterns of consecutive missing points. A simple example is the goalkeeper who can be unobserved for a long time if his teams is in offense.

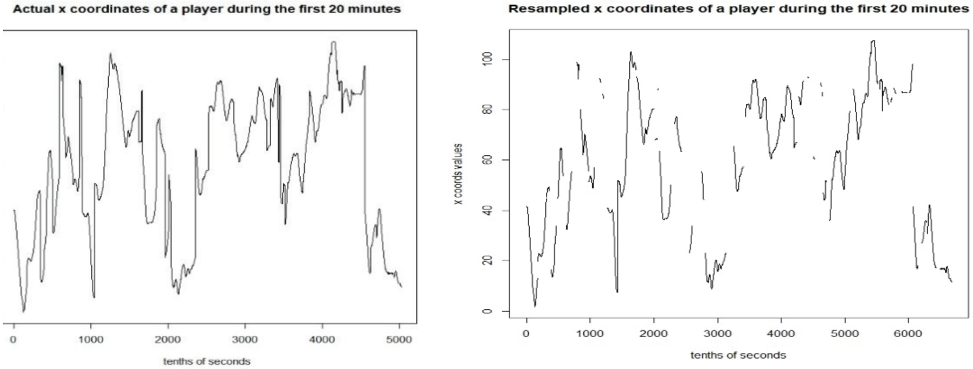


Figure 2: An example of data: the left-hand side refers to the full data for a player for a large time interval and the right one, data sampled from the full data to have missingness.

3. METHODOLOGY

3.1 INTERPOLATION OF MISSING DATA

In order to successfully address the problem, we compared several methods and we tried to retrieve the method whose predicted values were as close as possible to the actual positional data that we had in our disposal. For the implementation of this attempt, we used some interpolation algorithms known for their strong predictive ability in matters of interpolation of missing values using imputation techniques. Following that, we employed three non-linear models by using imputation techniques via regression. Due to the structure of our data, we decided also to observe the predictive interpolation power of a time series forecasting technique, on the specific type of data. Finally, it should be stressed that, for all the regression type of models we have used some covariates, namely

- player's position,
- timestamp,
- player's distance from the ball at each frame and
- distance of the player travelled from the last few frames, which is a proxy of his speed.

Note that several other variables have been also tried but we have decided to use only those listed above based on our findings. Such variables not used in the analysis are speed and acceleration of the player, ball possession, ball position and referee position.

Regarding the interpolation algorithms, we based our approaches on methods related for indexed totally ordered observations (Grothendieck and Zeileis, 2005) and on cases associated with univariate time series imputation (Moritz and Beielstein, 2017). We have used different approaches, including linear interpolation and splines methodology.

3.1.1. LINEAR INTERPOLATION

Regarding the linear interpolation approach, we replaced each missing value with linear interpolated values via approximation. Specifically, linear interpolation is a method of curve fitting using linear polynomials to construct new data points within the range of a discrete set of known data points. In our case, if the two known points are given by the coordinates (x_0, y_0) and (x_1, y_1), the linear interpolation is the straight line between these points. For a value x in the interval (x_0, x_1), the value y along with the straight line is given from the equation of slopes $\frac{y-y_0}{x-x_0} = \frac{y_1-y_0}{x_1-x_0}$ and consequently by solving the equation for y, which is the unknown value at x gives the following which denotes the formula for linear interpolation for the interval (x_0, x_1): $y = y_0 + (x - x_0)\frac{y_1-y_0}{x_1-x_0} = \frac{y_0(x_1-x)+y_1(x-x_0)}{x_1-x_0}$.

3.1.2. SPLINES INTERPOLATION

For the spline interpolation approach that we followed, we replaced each missing value with a spline interpolation approach. Spline interpolation is a method of using piecewise polynomials. That is, instead of fitting a single, high-degree polynomial to all of the values at once, spline interpolation fits low-degree polynomials to small subsets of the values such that they fit smoothly together. More specifically, with cubic spline interpolation we tried to construct a spline $f\colon [x_1, x_{n+1}] \rightarrow R$, which consists of n polynomials of degree three, referred to as f_1 to f_n. Opposed to regression, the interpolated function traverses all $n+1$ pre-defined points of a dataset D. The resulting function has the following structure:

$$f(x) = \begin{cases} a_1x^3 + b_1x^2 + c_1x + d_1, & if\ \ x \in [x_1, x_2] \\ a_2x^3 + b_2x^2 + c_2x + d_2, & if\ x \in (x_2, x_3] \\ \qquad\qquad\qquad\qquad \dots & \\ a_nx^3 + b_nx^2 + c_nx + d_n, & if\ x \in (x_n, x_{n+1}] \end{cases}$$

With properly chosen coefficients, $a_i, b_i, c_i\ and\ d_i$ for the polynomials, the resulting function traverses the points smoothly. For determining the coefficients, several equations are formulated which all together compose a uniquely solvable system of equations, such as the natural spline, the not-a-knot spline, the periodic spline and the quadratic spline. In our case, we employed a natural spline boundary condition, in order to minimize possible extrapolations. It should be noted that the natural spline is defined as setting the second derivative of the first and the last polynomial equal to zero in the interpolation function's boundary points.

3.1.3. STINE INTERPOLATION

A method sharing the good elements of the different method is also proposed (Stineman, 1980). In Stineman interpolation, missing values are replaced by piecewise rational interpolation and by default the time index associated with each unique object is used for interpolation. According to Stineman, the interpolation procedure has the following properties: If values of the coordinates of the specified points change monotonically, and the slopes of the line segments joining the points change monotonically, then the interpolating curve and its slope will change monotonically. If the slopes of the line segments joining the specified points change monotonically, then the slopes of the interpolating curve will change monotonically. Finally, suppose that the first and second conditions are satisfied by a set of points, but a small change in the ordinate or slope at one of the points will result the previous mentioned conditions being no longer satisfied. Then making this small change in the ordinate or slope at a point will cause no more than a small change in the interpolating curve.

Let x_i, y_i denotes the rectangular coordinates of the i^{th} point on curve and y_i' the slope of the curve at i^{th} point. Given x such that $x_i \leq x \leq x_{i+1}$, the procedure for calculating y, which denotes the corresponding interpolated value, is defined using a slope between two points by $s_i = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}$. Next $\Delta y_i = y_i + y_i'(x - x_i) - y_0$, where Δy_i denotes the vertical distance from the point (x, y_0) to a line through (x_i, y_i) with slope y_i'. Similarly, $\Delta y_{i+1} = y_{i+1} + y_{i+1}'(x - x_{i+1}) - y_0$, denotes the vertical distance from the point (x, y_0) to a line through (x_{i+1}, y_{i+1}) with slope y_{i+1}'. Therefore, according to Stineman, y can be calculated as follows:

If $\Delta y_{i'} \Delta y_{i+1} > 0$, then Δy and Δy_{i+1} have the same sign and the equation is the following:

$$y = y_0 + \frac{\Delta y_i \Delta y_{i+1}}{\Delta y_i + \Delta y_{i+1}} \tag{1}$$

If $\Delta y_{i'} \Delta y_{i+1} < 0$, an inflection point exists between x_i and x_{i+1} and the equation is derived accordingly from Stineman (1980):

$$y = y_0 + \frac{\Delta y_i \Delta y_{i+1}(x - x_i + x - x_{i+1})}{(\Delta y_i + \Delta y_{i+1})(x_{i+1} - x_i)} \tag{2}$$

In this work, we tried also to observe the predictive power of some robust models beyond the interpolation algorithms' ability. The most suitable algorithms for such problems when the predicted output is a continuous numerical value as in our case, are mainly supervised learning algorithms which are based on regression type techniques. We wanted also to observe the effectiveness of models based on non-linear regression, in order to see in that way, which ones responds better in such cases. Specifically, three machine learning regression algorithms have been employed, where each one of the algorithms estimated separately both (x, y) coordinates of each player again. It should be mentioned that both (x, y) coordinates, did not show some correlation between them. This means that each variable was affected by different factors and variables and for that reason we chose to proceed with the estimation of the missing positions using only univariate models.

3.1.4. RANDOM FOREST

The first approach under scope was a random forest regression algorithm. As first proposed, random forests build multiple decision trees and merge their predictions together to get a more accurate and stable prediction rather than relying on individual decision trees (Ho, 1995). The samples of the training observations, are drawn with replacement, known as bootstrapping, which means that some samples will be used multiple times in a single tree and a prediction is recorded for each sample. After all of the predictions have been assessed, the ensemble prediction is calculated by averaging the predictions of the above trees producing the final estimations.

For the implementation of the specific model, we kept only the initial actual values of each entity and omitted any extra values, which in fact did not offer something to our analysis. This step was necessary, in order to precisely estimate the prediction accuracies of our model. The most correlated variables for both (y, x) coordinates, were the "Frame" and "Timestamp" variables. In order to properly define the actual samples for both training and testing sets, we deemed appropriately to use as a training set each time, all the initial observable (k) values of the response variable (either y, x each time) and as a testing set all the missing values of the response variable, that were induced due to the censoring effect from the camera. Therefore, each time the splitting proportions between the training and testing sets, varied according to the missing proportions of each player. Afterwards, a decision tree was generated associated to these (k) data points, by defining also, the correct number of trees. For a new data point, we made each one of the defined trees, predict the value of y and x as the case may be, for the data point in question and assigned the new data point to the average across all the predicted values of the response variable each time.

3.1.5. EXTREME GRADIENT BOOSTING

The second non-linear regression algorithm, was a stochastic gradient boosting algorithm, or otherwise called as extreme gradient boosting algorithm. Each object treated as a unique entity again and the positional data (x, y coordinates) were separately calculated, by omitting all the extra values that were filled in the beginning and keep only the initial actual values. The most related variables for both response variables ((x, y) coordinates), were again the frame and timestamp of each player. For the training part, we used all the initial (k) observable values of the response variable and as a testing set all the missing information. A fraction of the training observations sampled, in order in each iteration a new tree generated from each subsample and any errors from the previous trees to be corrected. The training procedure that we followed, proceeded iteratively adding new trees for the prediction of the residuals of the errors of prior trees and then combined with previous trees to make the final prediction for our response variable. Finally, extra tuning parameters calculated at each iteration, in order to define the best selected model and at the end. RMSE was used to select the optimal model using the smallest value.

3.1.6. K-NEAREST NEIGHBORS REGRESSION

The third non-linear regression algorithm that we used, was the nearest-neighbor regression. The idea is to identify from the training set observations with regressors similar to the one we want to predict. Similarity is based on some distance, Euclidean in the standardized variables in our case. Then a weighted average of the k-nearest observations is obtained as the prediction. In the training phase of the algorithm, only the features are stored, by defining appropriately each time the optimal k value, which determines the number of neighbors we look at when we assign a value to any new observation. Regarding the modelling phase, the calculation of the most related variables, the preprocessing splitting procedures and the omission of the "extra" values have been commonly implemented as the previous two non-linear algorithms. For each object the size of the neighborhood was calculated by using a heuristically optimal number k based on the RMSE and a cross-validation technique in order to select the value of k that minimizes the mean squared error. The Euclidean distance computed in order to find the corresponding k number of neighbors each time and the distance was increased by ordering the labeled examples. An inverse distance weighted average was also estimated with a number of k-nearest multivariate neighbors. Finally, for the evaluation purposes, the last optimal k RMSE value was selected in the last iteration of the whole mentioned procedure, using the smallest possible value.

3.1.7. TIME SERIES METHODS

On the grounds that our data are in fact time series data, we wanted also to observe whether a time series model, could also respond to the present problem and how well compared to the other executed methods. As a first step, we create twenty-nine different time series objects, in order to observe possible various patterns of each object (22 starting lineup players, 6 substitute players and the ball). Each object as it is reasonable was composed of a different number of observations, due to the fact that each player was observed at different time instances from the camera. We deemed appropriate also to observe the different characteristics of each one of the different objects, by examining their trend, seasonality and random components for each object's coordinates. A specific trend and seasonality did not exist for the majority of objects, comprising mainly of random noise, as the observations were either decreasing or increasing depending on the position, speed and acceleration of the object. In order to stabilize variance, we used the technique of the kernel smoothing, by appropriately defining previously the optimal corresponding bandwidth. We also extracted the logarithmic values for both (x, y) coordinates in order to observe possible differences and we differenced as well for the integration part the logarithmic values to turn the data into its stationary form.

As the coordinates of the objects were inextricably linked also with future values, we reduced the sample size to achieve stationarity of each object. For the autocorrelation and partial autocorrelation part, the majority of players presented a strong positive autocorrelation, decreasing slowly over time and thus indicating that each object could not achieve absolute stationarity. This can be explained due to the structural format of the data. After all the preliminary checks, we defined an autoregressive integrated moving average (ARIMA) model for every object on the pitch, by including a necessary differencing step to eliminate the non-stationarity. Regarding the technical implementation of each ARIMA model, we employed for the training part of the model the previous ten observations of the y or x coordinate as the case may be, to be fitted in the model and predict accordingly the following ten observations of every single object. The constant term defined as the average period-to-period change in Y. The number of orders changed every time a new object was selected, by including also a linear drift term each time a model was running. The frequency parameter was equal to the number of tenths of seconds observed at each training subsample. If a missing value was observed, the model imputed the specific value with the corresponding prediction and continued the whole process, by dividing the training part into small subsamples of ten corresponding observations, up until a player appeared on the pitch again. Adding covariates is also an option but this creates estimation problems and improved very little the results.

Finally note that the three interpolation methods are applied to the position data only. The three supervised methods are using covariate information while the time series approach is based on solely the past observation of the relevant variable. Also, we emphasize that we have used other methods also but for saving space we do not present their results since they were inferior.

3.2 METRICS USED

In order to compare the different methods, we employed several metrics and criteria. Each assessment criterion was computed separately for each object. The first metric used, was the root mean square error (RMSE). In our case, let y_i denote the i^{th} observation of the actual coordinate of a player, $\hat{y}_i$ the imputed value and n the number of missing values of the specific coordinate. Then, the RMSE is given by:

$$RMSE(\hat{y}, y) = \sqrt{\tfrac{1}{n}\sum(y_i - \hat{y}_i)^2}\ . \tag{3}$$

Another measure is the mean absolute percentage error (MAPE), which is a measure of prediction accuracy. MAPE is commonly used as a loss function for regression problems, due to its very intuitive interpretation in terms of relative error. MAPE is given by:

$$MAPE(\hat{y}, y) = 100\frac{\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|}{n}\ . \tag{4}$$

We have also employed the typical Pearson correlation between the imputed and the original values (PMCC). We also use Cohen's h, which is a measure of distance between two proportions of probabilities, where for $h = 0.2$ implies a small difference, $h = 0.5$ a medium difference and $h = 0.8$ implies a large difference (Cohen, 1960). Cohen's h metric is defined as:

$$h(\hat{y}, y) = \frac{M_y - M_{\hat{y}}}{SD_{y\hat{y}}}, \qquad SD_{y\hat{y}} = \sqrt{\frac{SD_y^2 + SD_{\hat{y}}^2}{2}} \tag{5}$$

where M_y and $M_{\hat{y}}$ are the sample means of y and $\hat{y}$ respectively and SD_y and $SD_{\hat{y}}$ the respective standard deviations.

Finally, as an extra evaluation criterion for accessing the accuracy between the actual and the interpolated observations, we compared also the average distances per minute among all players. The actual and imputed average distances were calculated by the Euclidean distance as:

$$d_{avg(actual)} = \sqrt{\left(y_i - y_j\right)^2 + \left(x_i - x_j\right)^2},$$

$$\hat{d}_{avg(imputed)} = \sqrt{\left(\hat{y}_i - \hat{y}_j\right)^2 + \left(\hat{x}_i - \hat{x}_j\right)^2} \quad . \tag{6}$$

In order also to properly define the accuracy effect among the actual and interpolated distances, we calculated the RMSE and MAPE between the distances. In our case let x_i, y_i denote the i^{th}observation of the actual x, y coordinates of a player, and x_j, y_j the j^{th} observation of the actual x, y coordinates of another player. In addition, let $\hat{x}_i, \hat{y}_i$ denote the interpolated values of the missing coordinates happened in the same time instance of a player and $\hat{x}_j, \hat{y}_j$ the interpolated coordinates of another player as well. The corresponding error metrics after the calculation of the average distances per minute among the players, were the following:

$$RMSE\left(\hat{d}_{imputed}, d_{actual}\right) = \sqrt{\frac{\sum_{t=1}^{n}\left(\hat{d}_{imputed} - d_{actual}\right)^2}{n}} \tag{7}$$

$$MAPE\left(\hat{d}_{imputed}, d_{actual}\right) = 100\frac{\sum_{t=1}^{n}\left|\frac{\hat{d}_{imputed} - d_{actual}}{d_{actual}}\right|}{n} \tag{8}$$

where $\hat{d}_{imputed}$ denotes the average distance per minute between each object after the necessary interpolation procedures, d_{actual} the actual average distances per minute among each object and n denotes the complete data points.

4. RESULTS

4.1 COMPARING THE DIFFERENT METHODS

After discussing the proposed methodology, we can present at this point all of our findings and make the necessary comparisons between the evaluation metric results of each separate algorithm. Our experiment was as follows: For each player we used the observed data, i.e., all the time points that their position was available. Then we randomly selected a number of points to be considered as missing and we applied the different approaches to recover this together with the positions that they were missing in the data set. The reported metrics are based on the comparison of the actual available methods that we tried to predict and the ones derived from each methodology. For each player the proportion of data that we hide for the experiment was proportional to the observed missing proportion we had. Looking the results from Table 1, regarding the interpolation algorithms, all of them presented satisfactory results, with the best performing algorithms for all metrics, to be the Stine interpolation, followed by the linear interpolation. For all the evaluation metrics, the densities concerning the errors between the actual and the imputed values are smaller than the other methods. Both samples concerning the actual and imputed values, seem to present great similarities and very small differences (value < 0.2). It is worth mentioning, that all algorithms perform better, in lower rates of missingness but in higher rates the results were also very satisfactory.

According to the three mentioned supervised methods, namely random forest and extreme gradient boosting algorithms, presented mediocre results, with big variations concerning the differences between the actual and the predicted values. Random forest behaved better on cases with lower missing rates, while the extreme gradient boosting exactly the opposite. On the other hand, k-NN regressor presented very good results and quite close as the corresponding results derived from the Stine and linear interpolation algorithms. This seems logical as the k-NN algorithm assumes the similarity and as an extension the prediction of the most common neighbors of a missing observation. The similarities between distances of the actual and the predicted values were very high, indicating that the specific model has a fairly high predictive efficiency as well. Regarding the evaluation results of the time series forecasting implementation, fluctuated at the same levels as the evaluation results of the random forest and the extreme gradient boosting algorithms. It handled more efficiently, players with lower missingness rates and with a higher speed and acceleration. Finally, the similarities between the distances of the actual and predicted values were lower in comparison with Stine, linear and k-NN models.

Table 1: Evaluation metrics derived from the different imputation algorithms

	RMSE		MAPE		Pearson Correlation		Cohen	
	y	x	y	x	y	x	y	x
Interpolation								
Linear	0.207	0.165	0.297	0.310	0.999	0.999	0.172	0.167
Spline	0.247	0.279	0.271	0.318	0.999	0.999	0.180	0.153
Stine	0.198	0.158	0.259	0.287	0.999	0.999	0.172	0.166
Supervised Methods								
k-NN regression	0.500	0.547	1.069	1.184	0.995	0.999	0.172	0.165
Random forest regression	1.230	1.883	3.179	4.028	0.992	0.997	0.169	0.157
Extreme gradient boosting regression	2.513	3.216	9.257	8.521	0.978	0.991	0.162	0.165
Time Series								
ARIMA	2.062	2.421	7.089	7.885	0.953	0.991	0.213	0.170

Recapitulating, the best algorithms were the interpolation algorithms using the Stine interpolation and the k-nearest-neighbor regressor, as both models indicate the lowest RMSE and MAPE values and with the smallest dispersion in their predictions between the actual and the predicted values. Afterwards, the ARIMA and the random forest algorithms seem to fluctuate at the same levels, while the extreme gradient boosting algorithm seems to present a higher variability concerning both evaluation metrics and as a result it cannot be considered as a suitable approach (Table 1). Regression type models fail to account for the temporal effect and thus they provide worse results. K-NN regression is the only one that can consider such effects by selecting points close to the time points under examination.

Figure 3 presents the average distances per minute among the players comparing the actual values and the interpolated ones. We plot only selected methods to save space. The Stine interpolation and k-NN regression derived the best results.

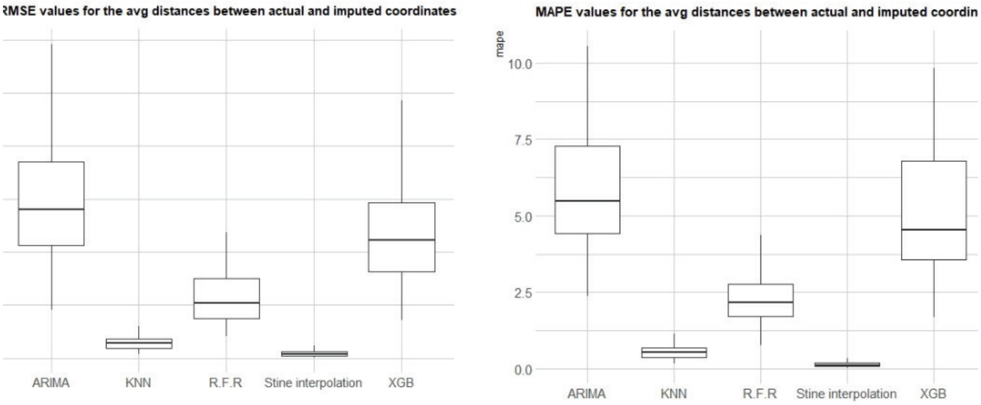


Figure 3: RMSE and MAPE results for the average distances per minute among players derived from each method

In conclusion, the most efficient algorithm is considered to be the interpolation algorithm using the Stine approach, as the predicted values of both x, y coordinates are very close to the actual ones. Therefore, it is proposed to be used when the censoring effect induced from the camera is present both for higher and for lower missing rates of objects. However, it must be stressed that all algorithms present satisfactory results as the values of the RMSE metric yielded an average of 0.5 - 2.4 (except from the XGB algorithm), indicating a margin error of approximately 0.5 ~ 2.4 meters prediction of a player's actual position. An example of the positional values of a single player for a time interval of 10 minutes after the interpolation procedures using the Stine method, can be observed below (Figure 4). The upper two (2) line plots represent the actual position values in both of his x, y axes on the pitch, while the tow (2) line plots below constitute the corresponding values after the interpolation.

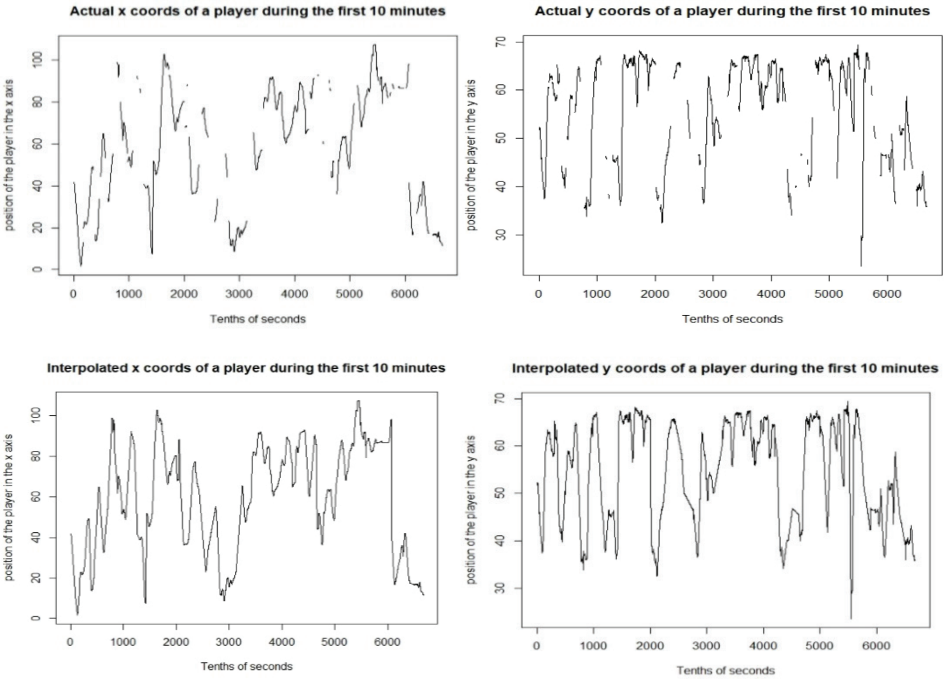


Figure 4: Positional values of a player for the first 10 minutes of the game after the interpolation procedures

4.2 APPLICATION OF INTERPOLATED DATA

In order to have also a more concrete view about the usefulness of this approach, we proceeded to some comparisons between the usage of the full information generated using the interpolated coordinates in relation with the actual data as they have been recorded from the broadcast camera. We provide some examples based on typical metrics based on tracking data.

4.2.1. HEATMAPS

A representative example can be considered for cases like players heatmaps showing the appearance of the players in different parts of the pitch and occasions like the players' soccer paths during a certain period of time (Figure 5). In the left-hand side one can see the heatmap for Ronaldo (Juventus) based on the actual available data (top) and the same for when the missing data have been imputed (interpolated). One can see a very close agreement between the two plots. At the right part of the figure one can see the observed paths of Dybala for the same match (top) and the paths when missing data are interpolated. Both cases show a large agreement.

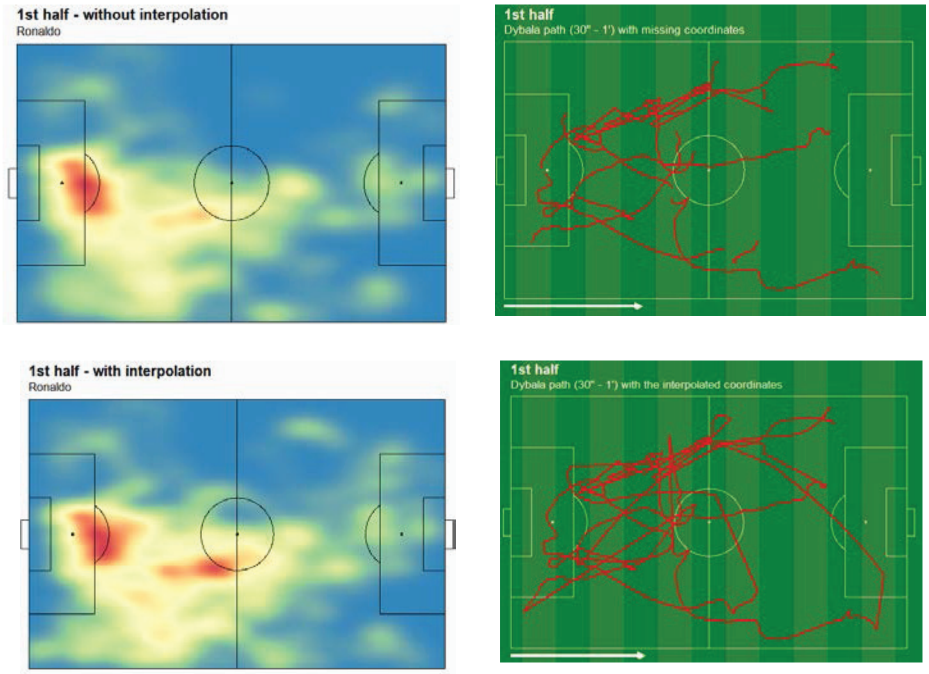
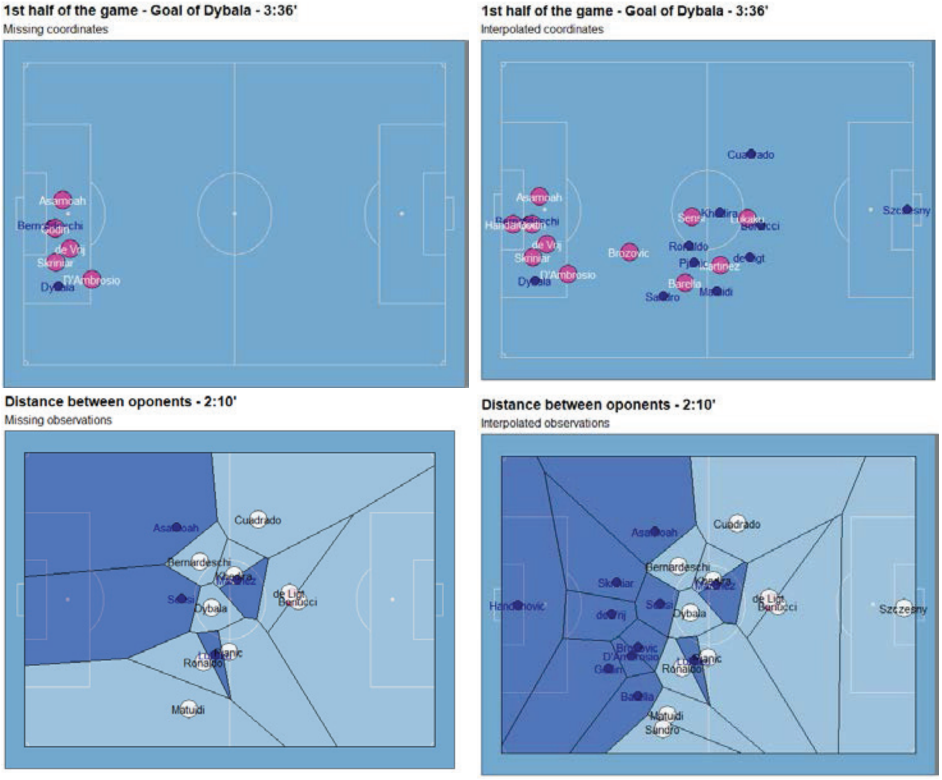


Figure 5: Football heatmaps showing the effectiveness of a player in different parts of the pitch and player's soccer path during a certain time instance

4.2.2. VORONOI DIAGRAMS

In mathematics, a Voronoi diagram is a partitioning of a plane into regions based on distance to points in a specific subset of the plane. These objects, which are mainly called as seeds of the plane, are just finitely many points in the plane. For each seed of the plane, there is an appropriate region, called a Voronoi cell, consisting of all points of the plane closer to that seed than to any other. The Voronoi diagram of a set of points is dual to its Delaunay triangulation. In football, these seeds represent the different positions of the players and the plane is referring to the instance of the field. A Voronoi diagram actually indicates the space that each player controls in the sense that if the ball is in this field then this player is closer to it. Hence missing data prohibit the creation of such plots. Imputed positions of the players can help better represent such situations. In Figure 6 one can see such a plot for a particular time stamp from the match considered. In Figure 6, one can see at the left part the Voronoi diagram from the available data and the right part based on the data after the imputation.


1st half of the game - Goal of Dybala - 3:36'
Missing coordinates

1st half of the game - Goal of Dybala - 3:36'
Interpolated coordinates

Distance between oponents - 2:10'
Missing observations

Distance between oponents - 2:10'
Interpolated observations

Figure 6: Players movement in real time

4.2.3. AVERAGE FORMATION LINE

A valuable match-level metric also, that is frequently monitored after post-match analyses, concerns the average formation line (AFL) of the teams. This metric corresponds to the average position on the long axis from the own goal of a team to the opponent's goal of the other team including all team events with the ball. By using this indicator, teams can better understand time intervals where their team put more pressure on the opposing team and instances when the team was in a more defending role. Figure 7 presents the AFL of the two teams. The left plot uses only the available players and hence estimates badly the AFL since some players are missing. The right panel is based on interpolated data using the Stine approach. Regarding the differences generated, it is clear that by using the initial data, the home team during the 1st half presents a more attacking formation system, while in the 2nd plot does not seem to be so much forward to the opponent's goal. On the other hand, as far as it concerns the 2nd half, the home team based on the interpolated data seems to have a more attacking behavior, while regarding the plot with the actual data it reflects a more conservative formation, as it seems to be playing quite further back.

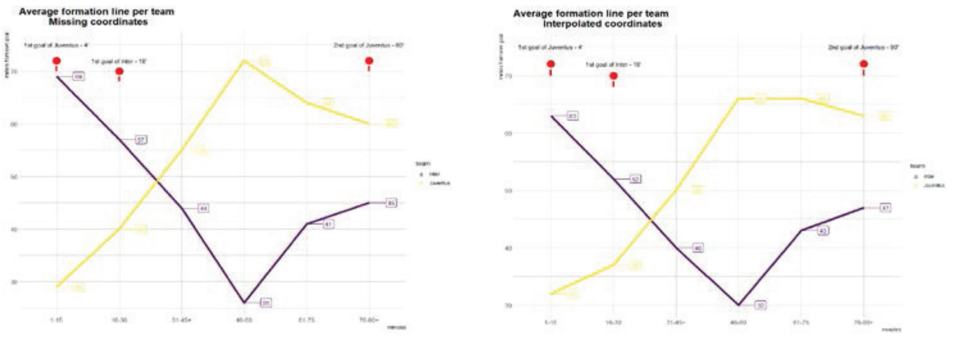


Figure 7: Average formation line per team using both actual and interpolated data

4.2.4. TEAM COMPACTNESS

Team compactness is another space-related concept that highlights the distances between one team's players (Santos and Penas, 2019). The idea is that players from one team keeping the biggest possible distance between themselves maintaining links between each other that will keep them in control of the space they occupy and action inside their occupied structure. Seven key metrics were employed for both positional defensive and attacking plays where among others are included the average last defender's and first pressing player's lines, the average left and right lines, the average teams' depth and width and the defensive/attacking squares. By observing the defending and attacking squares of each approach (Figure 8), it is clear that for both positional plays, the home team has a more compact system when the interpolated coordinates are employed. By using the complete dataset also, the home team presents a more stable shape and secure space-control system, with narrowly defensive and attacking plays and a bigger concentration of the players in the inside zones of the pitch.

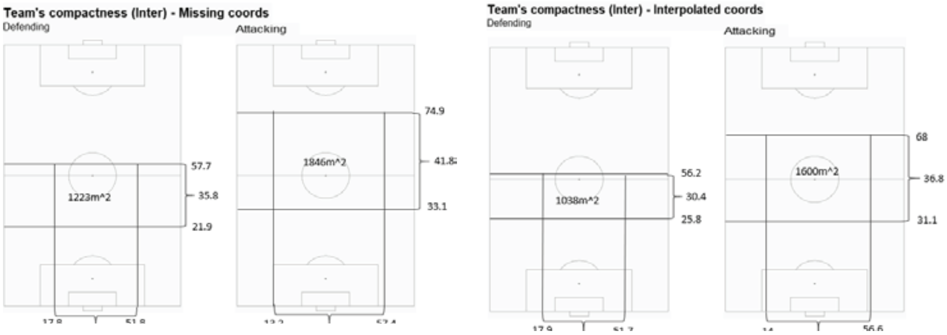


Figure 8: Home team's compactness using both actual positional before and after the interpolation procedure

5. CONCLUDING REMARKS AND FURTHER RESEARCH

In this paper we examined the case when tracking data based on broadcasting were used. Contrary to tracking data from a multicamera system, this is a cheap alternative to producing less data. The scope of this paper was to examine whether interpolation methods can retrieve the missing information. The findings are very promising in the sense that the use of existing interpolation methods provided very satisfactory results, while several metrics used in practice calculated from the interpolated data was shown to have quite good performance in comparison with full data cases. Of course, this is rather a first step towards this research problem.

Some limitations of our approaches are also present. To start with, we have used only one match as a proof-of-concept case so a more detailed examination would have been much more insightful. Also note that we base our approach in existing interpolation methods which perhaps do not take in full account the particular data at hand. So, an interesting problem is whether better and more suitable for the problem interpolation is possible. Also, the missingness mechanism needs further investigation. Here implicitly we assume missing at random mechanism but one may argue that players further away from the ball are more probably missing, so the mechanism can be somewhat informative. This deserves some more investigation. Finally, note that it would have been interesting to measure the effect of the interpolation to other metrics used in football based on tracking data. In this paper while we demonstrated the potential with some of them, it is perhaps the case that for other metrics more sophisticated interpolation may be needed.

To sum up, we believe that since broadcast-based tracking data are simple to obtain and very cheap their usage will be of increasing interest in the near future and thus we expect increasing interest and perhaps improved methodologies to obtain and analyze them.

ACKNOWLEDGMENTS

The authors would like to thank two anonymous referees for detailed comments that helped us improving the manuscript.

REFERENCES

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. In *The American Statistician*, 46(3), 175–185.

Bialkowski A., Lucey P., Carr P. and Matthews I. (2016). Discovering team structures in soccer from spatiotemporal data. In *IEEE Transactions on Knowledge and Data Engineering* 28(10), 2596-2605.

Cohen, J. (1960). A coefficient of agreement for nominal scales. In *Educational and Psychological Measurement,* 20, 37–46.

Diquigiovanni J. and Scarpa B. (2018). Analysis of association football playing styles: An innovative method to cluster networks. In *Statistical Modelling* 19(1) 28–54.

Dvorak J., Junge A., Chomiak J., Graf-Baumann T., Peterson L., Rösch D. and Hodgson R. (2000). Risk factor analysis for injuries in football players. Possibilities for a prevention program. In *American Journal of Sports Medicine*. 2000;28(5 Suppl): S69-74.

Grothendieck G. and Zeileis A. (2005). zoo: S3 infrastructure for regular and irregular time series. In *Journal of Statistical Software* 14(i06).

Horton M., Gudmundsson J., Chawla S. and Estephan J. (2014). Automated classification of passing in football. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp 319-330, Springer, Cham.

Ho, T.K. (1995). Random decision forests, In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278-282 vol.1.

Karlis D. and Ntzoufras I. (2003). Analysis of sports data by using bivariate Poisson models. In *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3), 381-393.

Link D., Lang S. and Seidenschwarz P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. In *PLoS One* 11(12): e0168768.

Lu, W. L., Ting, J. A., Little, J. J. and Murphy, K. P. (2013). Learning to track and identify players from broadcast sports videos. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 35(7), 1704-1716.

Mohr M., Krustrup P. and Bangsbo J. (2003). Match performance of high-standard soccer players with special reference to development of fatigue. In *Journal of Sports Sciences* 21(7):519-28.

Moritz S. and Beielstein T.B. (2017). imputeTS: Time series missing value imputation in R. In *The R Journal* 9, (1), 207.

Mortensen, J. (2020). *Statistical Methods for Tracking Data in Sports.* Doctoral dissertation, Science: Department of Statistics and Actuarial Science, Simon Fraser University, Canada.

Power P., Ruiz H., Wei X. and Lucey P. (2017). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In *The 23rd ACM SIGKDD International Conference* :1605-1613.

Rampinini E., Coutts A. J., Castagna C. Sassi E. and Impellizzeri F. M. (2007). Variation in top level soccer match performance. In *International Journal of Sports Medicine,* 28(12), 1018-1024.

Rein R. and Memmert D. (2016). Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. In *SpringerPlus* 5(1) 1-13.

Santos P. M. and Penas C. L. (2019). Defensive positioning on the pitch in relation with situational variables of a professional football team during regaining possession. In *Human Movement* 20(2):50-56.

Shaw L. and Glickman M. (2019). Dynamic analysis of team strategy in professional football. In *Barca Sports Analytics Summit 13.*

Stineman, R. W. (1980). A consistently well-behaved method of interpolation. In *Creative Computing,* 6(7), 54-57.

https://www.skillcorner.com Internet Document

https://uppsala.instructure.com/courses/28112 Internet Document

https://medium.com/skillcorner/a-new-world-of-performance-insight-from-video-tracking-technology-f0d7c0deb767 Internet Document

https://sportlogiq.com/en/ Internet Document

doi.org/10.26398/IJAS.0035-011

GENDER COMPARISON OF IN-MATCH PSYCHOLOGICAL TRAITS OF TENNIS PLAYERS: DYNAMIC NETWORK ANALYSIS

Antonina Milekhina[1]
Higher School of Economics, Moscow, Russia - ORCID 0000-0002-4188-4557

Kristijan Breznik
International School for Social and Business Studies, Celje, Slovenia - ORCID 0000-0002-8136-9592

Marialuisa Restaino
University of Salerno, Fisciano, Italy - ORCID 0000-0002-1150-8278

Abstract This study aims to investigate the existence of professional tennis players' psychological traits. For this purpose datasets on tennis matches of professional male (ATP) and female (WTA) tennis players were collected. Dynamical network analysis was applied with the RSiena program. Results revealed differences in in-match psychological characteristics' influence on the ability to build positive head-to-head for male and female players. Furthermore, a revealed tendency of head-to-head networks to cyclic structures is discussed. The study represents one of the first attempts to explore sport-related content with tools of dynamical network analysis.

Keywords: *Tennis, Gender comparison, Match analysis, Dynamic network analysis, RSiena.*

1. INTRODUCTION

Data analytics is growing its popularity due to the rapid development and increasing availability of big data. The possibility of large amounts of information is creating more and more opportunities for organizations and individuals for making use of and processing them.

Over the years, various approaches have been adopted for analyzing both the results of sports matches and competitions and the performance of teams and individual athletes in numerous sports.

Among other sports, tennis stands out for some reasons. First of all, tennis is a mostly individual sport, which means that the results could be attributed to the

[1] Corresponding author: Antonina Milekhina, antonina.milekhina@yandex.ru

player as opposed to team sports, where team success should take into account the efforts of each team member. The second feature of tennis is the fact that a tennis match is played till the last point is won. Therefore, there is no time restriction for a match and comebacks are not unusual in tennis matches. Thus, psychological momentum plays important role in a champion's mind setup.

Third, a tennis player's career is usually relatively long and can span over fifteen-twenty years. Also, there are cases when athletes achieved a peak in their career towards their end. Fourth, tournaments of different calibers are played almost every week. So we have weekly ranking updates and the opportunity to follow one's career changes in tiny pieces for a long period of time. Ranking updates provide information about players' progress within the calendar year and in the latest 12 months.

This provides us with a good opportunity to study career trajectory and explore mechanisms of its development, also taking into consideration that ranking and head-to-head information only can be misleading due to the incompleteness of its information.

This study aims to investigate the existence of professional tennis players' psychological traits. And if this trait exists, gender comparison would be additionally explored.

The rest of the paper is structured in the following order. In Section 2, we provide a comprehensive literature review of the topic. After the literature review, a description of networks and models with additional information is provided. Data collection and a brief descriptive data overview are reported in Section 4. The main results of analyzed networks and network models with discussion are presented in Section 5, followed by conclusions.

2. LITERATURE REVIEW

There is a lot of literature on investigating tennis matches and players' results, and exploring the differences and similarities between male and female players by looking at the various aspects of a match.

Short-term player's results have been largely examined on the level of point, match, and tournament (Klaassen and Magnus, 2001; Knottenbelt et al., 2012; O'Donoghue and Brown, 2009; Ovaska and Sumell, 2014; Serwe and Frings, 2006). In particular, Klaassen and Magnus (2001) focused on predicting the winner of a point in a match. They showed that the probability of winning a point is not random and win of the previous point has a positive effect on winning the current. Then, O'Donoghue and Brown (2009) argued that the sequence of points that occur in a tennis match does not differ from random, leading to a discussion of psychological momentum and its impact on one's game. In fact, as outlined in the research of psychological momentum (PM) by Iso-Ahola and Dotson (2014), if initial success leads a competitor to perceive *Self* as the superior performer (e.g., higher competence) and concurrently *Opponent* as the inferior performer, positive PM is likely to occur and success to ensue. On the other hand, if the self-opponent perceptions are reversed following the initial performance, negative PM is experienced and downward spiraling performance is likely to follow. Finally, when the perceptions are closely matched, neither competitor achieves PM (Iso-Ahola and Dotson, 2014).

For analyzing and predicting the outcome of the match, different regression models and machine-learning algorithms have been adopted (Candila and Palazzo, 2020; Jayal et al., 2018; Lisi and Zanella, 2017; Sipko and Knottenbelt, 2015). In Candila and Palazzo (2020), the artificial neural networks (ANNs) were implemented to forecast the probability of winning in tennis matches, while in Lisi and Zanella (2017) the logistic regression was estimated for predicting the winner.

An interesting field of research deals with examining long-term career trajectories of tennis players by using rankings (Kovalchik et al., 2017; Li et al., 2018; Lisi and Zanella, 2017). In Kovalchik et al. (2017), it was found a strong association between the shape of the ranking trajectory and the highest career ranking earned. Then, for evaluating ranking trajectory some important marks in a player's career were analyzed (e.g., age of first being ranked, age of achieving top-100, et cetera) (Li et al., 2018).

Moreover, different approaches have been proposed for evaluating and analyzing the ranking of players, such as paired comparison of players (Baker and McHale, 2014; Kovalchik, 2020; McHale and Morton, 2011), incomplete pairwise comparison (Bozóki et al., 2016; Szàdoczki et al., 2022), sorting algorithms (Spanias and Knottenbelt, 2013). Then, a broad discussion is also on whether the current ranking rules are fair and represent the true level of the player and whether ranking could be a good predictor for match results (Del Corral and Prieto-Rodriguez, 2010; Dingle et al., 2012). Further, probit regression has been implemented by Del Corral and Prieto-Rodriguez (2010) for the prediction of outcomes in tennis matches based on the ranking difference between two players.

Finally, some studies have proposed to apply tools of social network analysis for establishing players' hierarchy and creating alternative ranking systems, e.g. subgraph-based ranking (Aparício et al., 2016; Maquirriain, 2014; Spanias and Knottenbelt, 2013), link analysis of hubs and authorities (HITS algorithm, Hyperlink-Induced Topic Search), simple PageRank and PageRank with teleportation (Michieli, 2018), time-dependent PageRank algorithm (Breznik, 2015; London et al., 2014; Radicchi, 2011) network-based dynamical ranking system (Motegi and Masuda, 2012). Moreover, the weighted networks are also used for predicting the probability of winning (Arcagni et al., 2022). These approaches were used to compare players' greatness, both concurrently and all-time (Radicchi, 2011).

An important recent issue in sports is to analyze whether there are differences and similarities between men and women (Blanca Torres, 2019; Cross, 2014; Cross and Pollard, 2009; Fernández-García et al., 2019; Hizan et al., 2011, 2015). The perspective of differences in various aspects of the game of tennis, e.g. the number of aces, double faults, unforced errors, winners, tiebreak sets, games per set, and points per game, was analyzed by Cross (2014). Then, Blanca Torres (2019) explored the differences in the duration of the match, the number of sets, and the duration of sets between junior and absolute categories and regarding the surface in both genders. It turns out that in the absolute category, there were no differences in the duration of the match, nor in the duration and number of sets. Meanwhile, significant differences between genders were observed in the performance indicators between winners and losers in London 2012 Olympic games tennis tournament (Fernández-García et al., 2019). Break points won was the most relevant prediction variable to the result of the match in the female category. In the male category, it was joined by first-serve points won and first-serve return points won. Then, still speaking of gender, players showed different serve patterns as men served not surprisingly faster, with higher success, and placed their serves more frequently to the external areas of the service boxes. Women's serve was directed more to the body of their opponent.

Another relevant topic when analyzing the difference between males and females lies in investigating patterns of hemispheric specialization for cognitive function and those patterns may be related to handedness (Lake and Bryden, 1976). In Vogel et al. (2003) and Rilea et al. (2004) it was proven that the difference in spatial ability scores between left-handed and right-handed persons has been higher among males compared to females. Meanwhile, Breznik (2013) asserted that the advantage of left-handed professional tennis players over their right-handed opponents is higher in males compared to females. Then, the quality of players and matches is inversely proportional to the advantage of left-handers against their right-handed counterparts.

Physical abilities are without a doubt crucial for a tennis player. However, physical abilities alone cannot explain the success or failure of some players. If a player's success were related only to physical abilities, predicting would be very much straightforward (faster, higher, et cetera will always win). We are aware of cases when players with smaller heights achieve great success against higher players or players with higher speed of serve. And thus we might assume that psychological traits contribute to a player's success as well, which was already stated by Morgan (1985) in his mental health model.

In order to achieve success, a player should be mentally tough as summarized by Connaughton et al. (2008). Mental toughness is defined by Jones (2002) as "having the natural or developed psychological edge that enables you to (1) generally, cope better than your opponents with the many demands (competition, training, lifestyle) that sport places on a performer, and (2) specifically, be more consistent and better than your opponents in remaining determined, focused, confident, and in control under pressure."

Moreover, Jones (2002) ranks twelve mental toughness attributes. Part of them is lifestyle characteristics, which are out of focus for our research. Another part, however, could be attributed to the in-match mindset:

1. Having an unshakable self-belief in your ability to achieve your competition goals.

2. Bouncing back from performance setbacks as a result of increased determination to succeed.

3. Thriving under the pressure of competition.

4. Accepting that competition anxiety is inevitable and knowing that you can cope with it.

5. Not being adversely affected by others' good and bad performances.

6. Remaining fully-focused on the task at hand in the face of competition-specific distractions.

7. Pushing back the boundaries of physical and emotional pain, while still maintaining technique and effort under distress (in training and competition).

Though Jones (2002) describes this as "perception of the attributes of the ideal mentally tough performer", we might suggest that these characteristics can be expressed and inspected through match results. Unshakable self-belief (1) can represent the ability of a player to win against higher-ranked players. Thriving under pressure (3) and remaining fully-focused (6) correspond to the ability to win tough sets. Not being adversely affected by others' performance (5) can also be expressed by not losing easy sets.

Building together a network of tennis matches and the psychological traits of a player, we may suggest the following scheme to conceptualize our research: both the psychological traits of a player and his/her existing embeddedness into the current network influence result of a given match, which in turn influences the overall performance of one player against another one (one's head-to-head), which builds a directed link in the network. This affects network structure and leads to network evolution, which in turn changes the player's position in the network and thus the player's chances for success.

Based on the theoretical framework, we will focus on two main hypotheses with some sub-hypotheses below. Each of them will be tested separately for male and female competition.

Hypothesis 1: Relation of head-to-head match records between professional tennis players will show the tendency to transitivity.

This hypothesis can be postulated in the following way: if the head-to-head results of the first player to the second one are positive and the head-to-head of the second player against the third one is also positive, we to the same extent predict positive head-to-head result from the first player against the third one.

Hypothesis 2: The probability of creating the positive/negative head-to-head match record of one player against another player depends on his/her psychological traits and the psychological traits of his/her opponents.

Hypothesis 2a: The ability to thrive under pressure will enhance one's chance to build a positive head-to-head against other players.

Hypothesis 2b: Being prone to lose easy sets and matches will decrease one's chances against other opponents (to build outgoing ties in the network).

Hypothesis 2c: Ability to win against higher-ranked opponents will enhance one's head-to-head against other opponents.

3. NETWORK(S) AND MODEL DESCRIPTION

Social network analysis deals with a structural approach to interactions among social actors (Freeman, 2004). The building blocks of each social network are nodes and links (directed or undirected). Nodes are representing actors which are in our case tennis players. A relation is presented with a directed link (or an arc) and denotes the current match record between two tennis players in a network.

A match record constitutes a dyadic relationship being a result of previous interactions between two individuals with both of them contributing to it. In this sense, it's impossible to attribute the current head-to-head record to the action of one of the players as compared to e.g. runners' results. On the other hand, the match record is not the result of cooperation between two players (as in team sports). A dyadic relationship of two actors within one match record has competition in nature with the winner taking it all. This establishes the clear hierarchical structure of who is a better player in the dyad. Each further match between two actors contributes to dyadic relationship development, which can preserve current head-to-head or dissolve (by balancing it out). Putting such a relationship between several actors together we build a network of head-to-heads and observe its development from year to year.

We would expect that outline of such a network will follow the outline of a network built by a preferential attachment mechanism, which assumes that actors with a higher degree (or outdegree in our case) are easier to get another outgoing link (Newman, 2001). Indeed a player with higher outdegree is a player, who has better head-to-head records, also having a positive head-to-head with many players presumes winning many matches and thus potentially having a higher ranking. A player with a higher ranking is expected to be a better player and wins against lower-ranked players. However, from time to time an unexpected link occurs when a lower-ranked player (so-called underdog) wins.

In such a network, the direction of the link is determined not only by the individual characteristics of both players but also by the whole structure of the network relationship, in which both of them are embedded. Network embeddedness was applied by Granovetter (1985) to market societies and described as "attempts at purposive action embedded in concrete, ongoing systems of social relations", but this mechanism is applicable in sport as well since we assume that wish to win is a purposive action. The network structure of two players can explain the direction of a potential link between them (which is who will win more matches between them and thus will have a positive head-to-head). As each new match potentially establishes a new link, the network of match records (and the network of head-to-heads as its derivative) is in constant evolution.

Due to the way tournaments are organized, tennis matches are very suitable for network modeling. The network is constructed so that the players represent the nodes, and the characteristics of their mutual matches describe the links between them. The first work that used network modeling in the field of tennis is represented by Situngkir (2007), where the network generation was introduced and explained, and then some analysis were performed on single Grand Slams tournaments matches only.

In order to analyze the evolution of the network of match records and dependencies within such evolution, the stochastic actor-oriented model (SAOM) suggested by Tom Snijders was applied (Snijders, 2001). SAOM's purpose is to explain the evolution of a network as a function of the structural effects of the network itself (e.g. reciprocity, transitivity, et cetera), actor variables (constant and varying), and dyadic variables (constant and varying). With help of SAOM, longitudinal networks for several snapshots of observations can be evaluated. SAOM are "actor-oriented", which means that they model change from the perspective of the actors, both focal actors (or egos) and their neighboring actors (or alters) (Snijders, 1996). Actors are active subjects with their goals constrained by the current network structure (Snijders, 1996). SAOM are continuous-time Markov chain models, allowing also to simulate the evolution of a network with given effects. The evolution of a network is regarded as the number of probabilistic sequential ministeps. A ministep implies an actor's action of creating, maintaining, or terminating outgoing tie. The transition between two waves of network observation consists of all ministeps done (Ripley et al., 2011). To explain changes in a network and in an actor's behavior, different actor covariates and network effects can be used.

We propose the following network modeling setup: choose a time period (called a wave) for all steps (e.g. 3-year periods), then choose a starting point (e.g. 1 January 2001). A wave is built on three (accumulated) networks. The first network represents a head-to-head record at the end of the first year based on all match records for this year, the second network represents a head-to-head record at the end of the second year based on all match records for the first two years et cetera. Such an approach helps to gradually accumulate information about each period with no big jumps in the number and list of network participants. Participants from all networks are present in the SIENA model starting from the first one as per RSiena requirement. We can model those players who were not present as structural zeros, even if this is not obligatory. We can think of these players as present on tour and in this sense, they also had the opportunity to build a link in earlier networks but either willingly or due to some circumstances haven't done it. A link in a such network represents a head-to-head record of two players. The link will go from the player with more wins to a player with fewer wins in their match history for a particular period. If there is no match history between two players or their head-to-head is even, then there is no link. Such representation will give us information about who is a better player in the course of the time period and not only in a particular match.

In such a model setup, there is the possibility of both tie emerging (a player wins a match and a positive head-to-head emerges which is represented by outgoing tie) and tie dissolving (a player used to have a positive head-to-head, but loses a match and head-to-head balances out which is represented by tie dissolving). Thus we will evaluate network effects for the possibility of tie emerging, maintaining, and dissolving (function evaluation). Network evaluation function for player *i* is given by

$$f_i^{net}(x) = \sum_k \beta_k^{net} s_{ik}^{net} x \tag{1}$$

where β_k^{net} are parameters and $s_{ik}^{net} x$ are effects (Ripley et al., 2011). The equation (1) can be understood in a similar way as for instance multiple regression model equation. For each player, we can estimate his or her score with the linear combination of chosen effects. Varying effects used, we can produce several different models. The list of specific models used in our research with associated effects is presented below. For each model, associated coefficients β_k^{net} will be calculated by model estimation.

Due to limitations of calculating power because of the high number of actors in networks, we were forced to put covariate and network effects into separate models. We evaluate them for both ATP and WTA tours in separate runs. In order to build a model we introduce a notion of a wave. The wave is a time period included into one run of the model within which several shapshots of the network are taken and the development between each snapshot and the beginning of the wave is observed. This helps us to take into account both previous network setup and development for each year within one wave. We propose the following network modeling setup: (1) choose the length, starting point, number of snapshots per wave, number of waves per model, and lag between current and next wave starting points, (2) run the same SIENA model for all waves, (3) choose significant effects for most waves. Overall we estimated five models for each tour applied to ten waves of networks, each wave in turn consists of 3 consecutive years.

To test Hypothesis 1, we propose the first model which includes two network effects: transitivity and 3-node cycle effects. The transitivity effect (or transitivity closure) is a hierarchical effect and can be described as follows: an outgoing tie from node A to node B and from node B to node C increases the chance of an outgoing tie from node A to node C. Such effect presumes a clear understanding of who is the best player in this triad. A cycle effect is the number of 3-node cycles a player is involved in. In other words, a cycle will emerge if player A has a positive head-to-head over player B, player B has a positive head-to-head over player C and player C has a positive head-to-head over player A. Such effect presumes no hierarchy in the triad. Moreover, it's impossible to single out the best player in the triad, which means that on-paper underdogs can win as well. Unfortunately, we were able to achieve convergence for ATP only for 6 waves out of 10 (1-5 waves and 8). Regarding the postulated Hypothesis 1 we expect in the first model tensity towards transitivity but not to 3-node cycles.

Hypothesis 2 with its sub-hypotheses was tested with the following four models (models two to five). The second model includes ego effects with creation function for all covariate effects included. With help of this model, we evaluated whether any included effects increase the chance of creating an outgoing tie (building positive head-to-head) for a player. Unfortunately, we were not able to achieve convergence for ATP data for this model, so we only discuss results for WTA data.

The third model includes alter effects with a creation function for all covariate effects. Hence we assess whether any of included effects increases the chance of incoming tie (or negative head-to-head) for ego.

The fourth model includes ego effects with endowment function for all covariate effects, thus we examine the chance that the tie will dissolve (or head-to-head is balanced out or reversed to the benefit of ego's opponents).

The fifth model includes alter effects with endowment function for all covariate effects. In other words, this model evaluates the chance that the incoming ties will dissolve (or head-to-head is balanced out or reversed to the benefit of ego).

It should be indicated that each year is unique and due to circumstances that happened (e.g. injuries of players, other outer circumstances), head-to-head network evolution for some years may diverge significantly from others, so a longer observation period is of benefit. However 10-year period provided us with some understanding of which effects to look at. For the estimation of SAOM, a computer program called SIENA (Simulation Investigation for Empirical Network Analysis) and its R package RSiena were applied. SAOM models computed within these programs are commonly called SIENA models. The manual for RSiena contains a lot of necessary information for SIENA modeling in R (Ripley et al., 2011).

4. COLLECTED DATA OVERVIEW

The data used for the study include information on tennis matches played among male (ATP) and female (WTA) professional tennis players. The data were freely available online at Kaggle datasets repository (Sackmann, 2022 and Hakeem, 2022), for ATP and WTA respectively and they were analyzed in other researches such as Candila and Palazzo (2020); Gollub (2021); Khder and Fujo (2022); Yue et al. (2022).

Results for the men's ATP tour are recorded from 2001 to 2012, while for the women's WTA tour results and historical betting odds are available from 2007 to 2019. Among the competitions, we excluded Federation Cup in ATP and Davis Cup in WTA.

The datasets contain information about the following: match circumstances (e.g. tournament name, level of tournament, draw size, date, type of surface, round et cetera), match results (final score, length in minutes, number of aces, first serves in, break points saved et cetera) and players' information (name, age, hand, seeding, ranking et cetera).

From the raw data provided in the initial dataset several characteristics of matches were calculated for further analysis. For exploratory analysis, we compared each player's characteristics for their respective year on tour. For social network modeling we took these characteristics for a calendar year.

These characteristics, described briefly in Appendix A, are:

- number of tough sets won against all number of tough sets played (*tsw_tsp*).
- number of easy sets lost against all number of easy sets played (*esl_esp*),
- number of matches won against higher-ranked opponents out of all matches against higher-ranked opponents (*wahr)*,
- number of matches lost against lower-ranked opponents out of all matches against lower-ranked opponents *(lalr)*.

Before the modeling, an exploratory analysis was conducted. In total there are 32870 and 32053 matches in datasets, respectively for ATP and WTA. Then, 1640 men unique players and 1528 unique women players were involved. Among them, 463 men players and 570 women players have ever won a match.

In Table 1 the descriptive statistics of demographic characteristics of players in ATP and WTA are shown. Among players, the average player's height is 185.3 cm and 173.9 cm respectively for ATP and WTA winners, while for losers the average height is 184.9 cm for men and 173.2 for women. Both winners' and losers' age in ATP and WTA have approximately the same range.

The differences in mean for age and height between men and females, and between losers and winners have been tested, and the null hypothesis is always

Table 1: Descriptive statistics of demographic characteristics of players in ATP and WTA

ATP				
Variable	Winner/Loser	mean	s.d.	range=max - min
Age	winner	25.51	3.64	23.9
	loser	25.72	3.63	29.3
Height	winner	185.3	6.52	40
	loser	184.89	6.61	40
WTA				
Age	winner	24.64	4.04	30.8
	loser	24.43	4.26	32.8
Height	winner	173.87	6.64	37
	loser	173.17	6.63	37

rejected (all p-values are less than the smallest $\alpha = 0.001$). Thus, the age and height between winners and losers and with respect to gender are statistically significant.

In Figures 1-4, a comparison between the distribution of age and height in ATP and WTA for winners and losers is shown.

Looking at the age of winner and loser, we can see that in WTA there are a larger number of outliers, meaning that women play tennis for a longer time than men. Moreover, for winners both distributions of ATP and WTA are positively skewed and have a similar variability, while for losers in ATP the distribution of age seems to be negatively skewed, and for WTA it seems almost symmetric (Figures 1 and 2).

Some differences can be observed in winners and losers' height for ATP and WTA (Figures 3 and 4). In fact, for both winners and losers the distribution of men's height moves towards higher values, while women's height has a distribution around lower values. Moreover, for winners it seems there is a symmetric behavior, while for losers a negative skewness in ATP is opposed to a positive skewness in WTA.

Furthermore, the match length, measured in minutes, is longer in ATP, which is due to the fact that at some tournaments men are playing 5-set matches, and the variability is higher in WTA (Figure 5).

Among players, there were significantly more right-handed players than left-

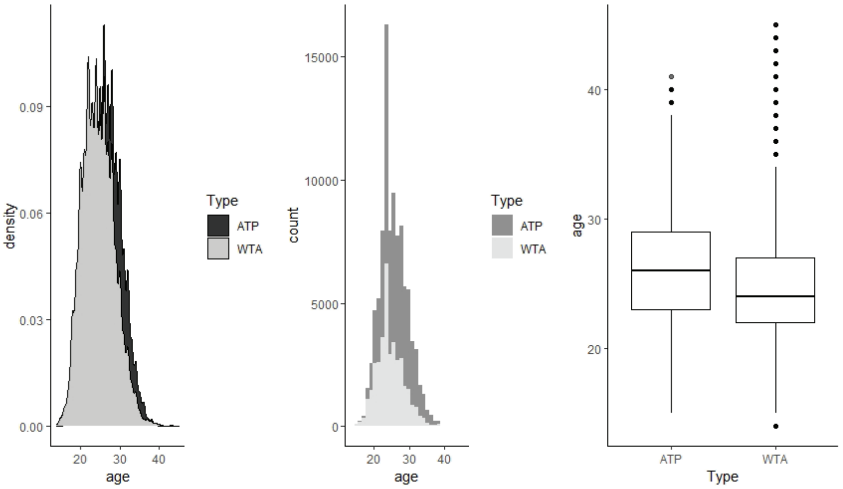


Figure 1: Histogram, density plot and boxplot of age for winners in ATP and WTA

handed ones for both winners and losers (right-hand winner: 87.97% for men and 89.93% for women; right-hand loser 86.25% for men and 88.36% for women) (Figures 6a and 6b).

5. NETWORK ANALYTIC RESULTS AND DISCUSSION

For our analysis we chose three years as the length of study period for a wave with three snapshots taken at the end of each year and a lag of a year between the starting point of two waves. E.g. a wave can start on 1st January 2001 and end on 31st December 2003 with three snapshots taken on 31st December 2001, 31 December 2002, and 31 December 2003. The first network represents head-to-head record at the end of the first year based on all match records for this year, the second network represents head-to-head record at the end of the second year based on all match records for the first two years et cetera. Summary of RSiena models was extracted over 10 waves for each tour. In Table 2 we indicate the number of times each effect was significant out of ten waves and its sign (if significant). Please refer to Appendix A for name of effect deciphering and details. Ego or Alter column indicates if the effect was evaluated from ego or alter perspective. Function type stands for one of the network evaluation functions (tie emerging, maintaining, dissolving). The example RSiena output for years 2008-2010 on ATP tour for the model with alter effects and creation function is in Table 3. The Table presents estimates (log-probability ratios for the estimated parameter to provide a

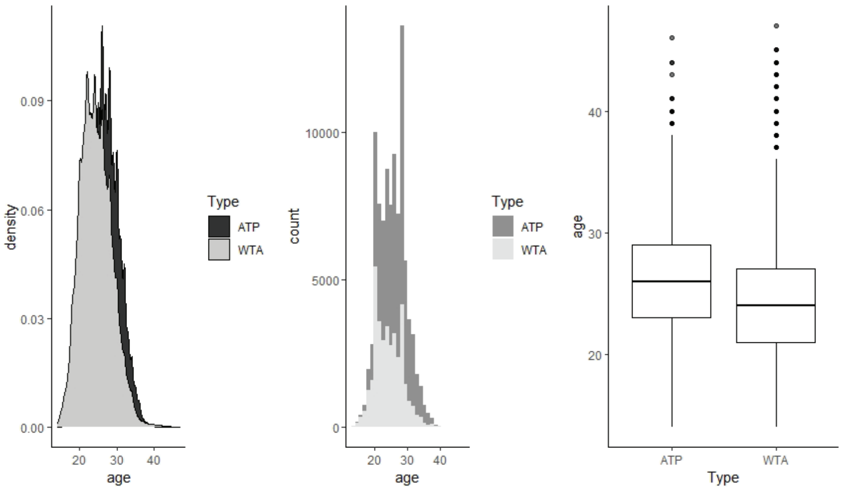


Figure 2: Histogram, density plot and boxplot of age for losers in ATP and WTA

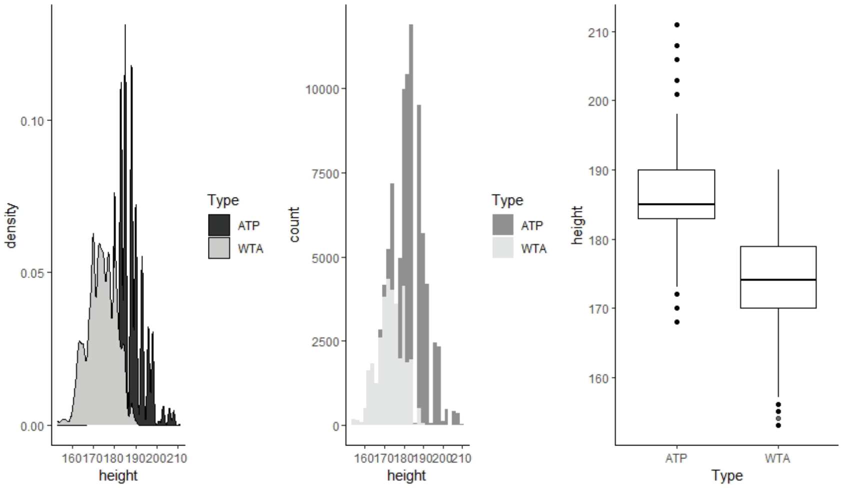


Figure 3: Histogram, density plot and boxplot of height for winners in ATP and WTA

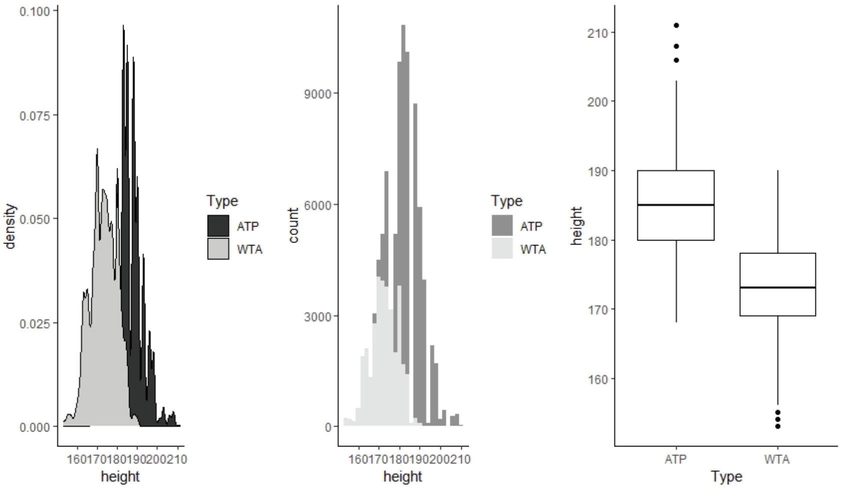


Figure 4: Histogram, density plot and boxplot of height for losers in ATP and WTA

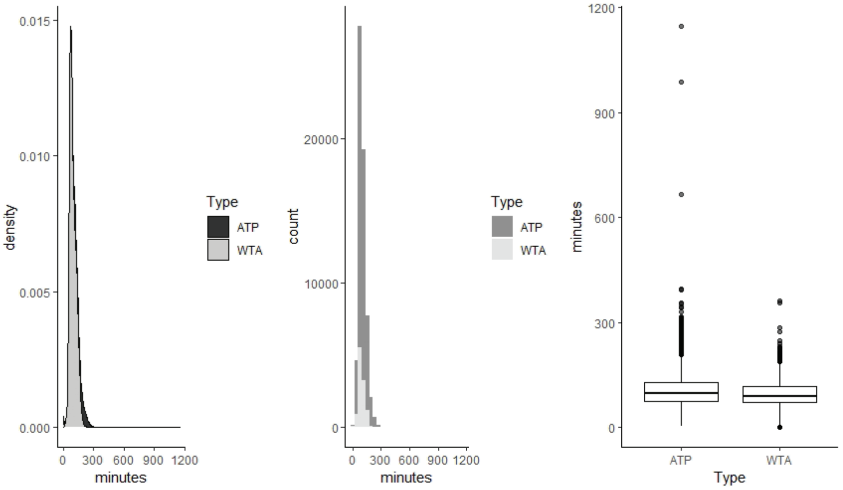
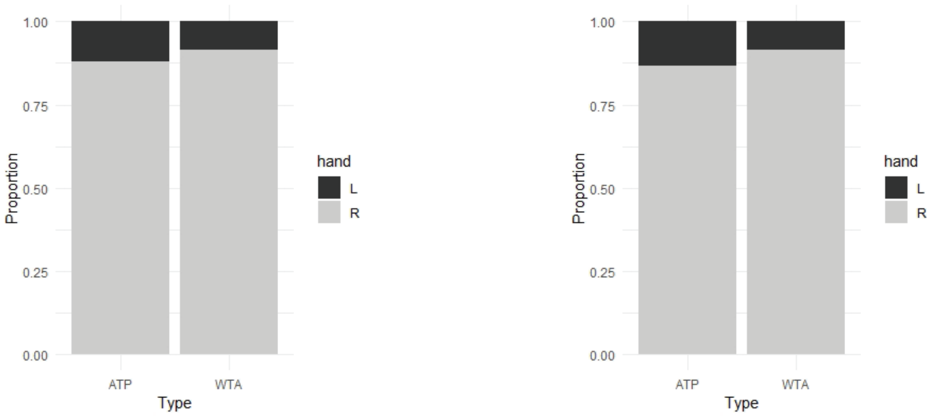


Figure 5: Histogram, density plot and boxplot of match length in minutes in ATP and WTA



(a) Winner hands **(b) Loser hand**

Figure 6: Plot of players hands in ATP and WTA

change to the network) and standard errors for network and behavior dynamics, convergence t-ratios (or t-statistics for deviations from targets, which is an indicator of model convergence) and overall maximum convergence ratio (which is another indicator of convergence for the whole model). In Figure 7 we provided an example of the network presentation of wave 10 of male competition (ATP). It can be observed that the network is very dense, and the calculated average degree is slightly over 26. On average, players are linked with 26 opponents, i.e. players established on average 26 head-to-head relations with their opponents between 2010 and 2012 (both years included). The highest out-degree in this network (i.e. the highest number of positive head-to-head), is obtained by David Ferrer by achieving a positive score against 108 opponents. He also faced a relatively high in-degree number (i.e. negative head-to-head). His score was negative against 22 opponents. The "big three", Federer, Nadal and Djokovic, are following closely. Federer with the second highest outgoing degree (out-degree: 92; in-degree: 8), Nadal fourth (86;8), and Djokovic sixth (83;9).

A summary of RSiena models was extracted over 10 waves for each tour. For simplicity, in Table 2 we indicate the number of times each effect was significant out of ten waves and its sign (if significant). The RSiena output for years 2008-2010 on ATP tour for the model with alter effects and creation function is in Table 3.

For the WTA network effect of weights by respective head-to-head transitivity is either non-significant or of neglectable magnitude (less than 0.1). This

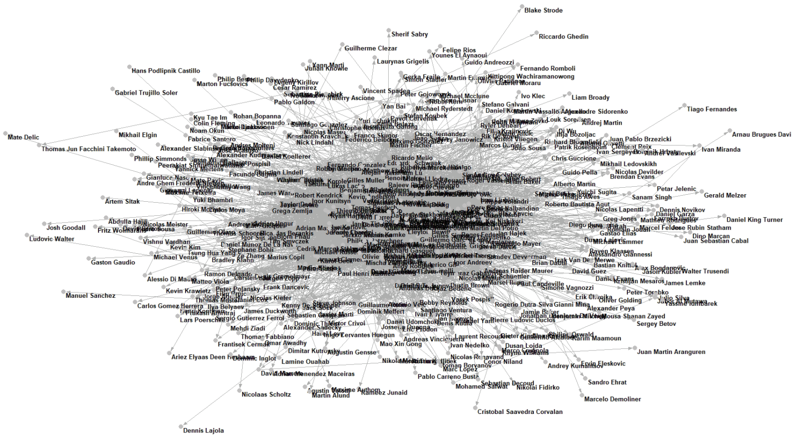


Figure 7: Network presentation of the wave 10

Table 2: Summary of models

Name of effect	Ego or Alter	Function type	ATP	WTA
Transitive closure	-	Evaluation	1 (out of 6, +)	4 (1 +/ 1-)
3-actor cycle	-	Evaluation	5 (out of 6, +)	10 (+)
tsw_tsp	Ego	Creation	N/A	10 (+)
tsw_tsp	Alter	Creation	10 (+)	10 (+)
tsw_tsp	Ego	Endowment	10 (-)	10 (+)
tsw_tsp	Alter	Endowment	8 (-)	10 (+)
esl_esp	Ego	Creation	N/A	8 (-)
esl_esp	Alter	Creation	10 (-)	9 (+)
esl_esp	Ego	Endowment	10 (+)	4 (+)
esl_esp	Alter	Endowment	10 (+)	8 (+)
wahr	Ego	Creation	N/A	4 (-)
wahr	Alter	Creation	6 (+)	0
wahr	Ego	Endowment	6 (-)	0
wahr	Alter	Endowment	3 (-)	0
lalr	Ego	Creation	N/A	8 (4 +/ 4 -)
lalr	Alter	Creation	8 (-)	6 (+)
lalr	Ego	Endowment	7 (+)	0
lalr	Alter	Endowment	7 (+)	0

means that the chance of player A to win against player C and build a positive head-to-head if player A has a positive head-to-head against player B and player B has a positive head-to-head against player C, is close to guessing; thus we should reject Hypothesis 1 for the WTA.

For the ATP the same effect was significant only in one wave with four waves not converging to produce results. However, the magnitude of the effect was very high (170.6665), which indicates a high transitivity tendency of the network. Since in other years, the transitivity effect was not significant, we should reject Hypotheses 1 for the ATP as well.

Additional evidence against Hypothesis 1 is a significant tendency of the networks to build a 3-node cycle. Though not strong for the WTA tour and rather strong for the ATP tour. Since the cycle is not a hierarchical structure, it is impossible to rank players within the cycle and cast doubts on the possibility of "fairer" ranking attempts (Aparício et al., 2016; Dingle et al., 2012; Motegi and Masuda, 2012; Spanias and Knottenbelt, 2013) as well as great predicting power of rankings (Boulier and Stekler, 1999; Del Corral and Prieto-Rodriguez, 2010).

Hypothesis 2 is partially supported by different strengths and parameters for ATP and WTA. Winning tough sets which we regard as the ability to thrive under pressure is a significant effect for both the ATP and the WTA players. A positive parameter implies the tendency that players with higher winning tough sets percentage increase their positive head-to-head towards other players more rapidly than average, which is the case for the WTA tour. Also, the magnitude of this parameter is high (between 7.2622 and 48.2832 for different waves). This result is in line with the indication that both men and women perform worse under the pressure of advanced stages of tournaments with women being more prone to unforced errors at crucial junctures of matches (Paserman, 2007).

Controversial to this result, a high percentage of winning tough sets also contributes to higher incoming ties (or negative head-to-head for a player) which we observe for both ATP and WTA players. However magnitude of the effect is small ($0.3842 - 1.3247$ for ATP and $0.2013 - 1.4888$ for WTA). A potential explanation would be the strong possibility to build positive head-to-head due to the ability to thrive on pressure and the weak possibility to build negative head-to-head due to letting the set to be tough because of neglectance. Thus Hypothesis 2a is partially confirmed for ATP and WTA players.

Interestingly, winning tough sets decreases the chance for ATP players to lose head-to-head advantage (dissolve positive head-to-head and get negative). On the opposite, for the WTA players such chance increases, that corresponds with higher confidence ability for male athletes (Nicholls et al., 2019). Our results correspond with the finding that if a male player wins a tough tie-break, he has a 60% chance of winning the next set (Page and Coates, 2017). However, in Page and Coates (2017) no effect was found for women.

The ability to win against higher-ranked opponents turned out to be irrelevant for WTA players. In some waves, it even decreases the chance to build positive head-to-head, however in less than 50% waves. Winning against higher-ranked players slightly increases the chance to get negative head-to-head for male players and decreases the chance to averse existing negative head-to-head. However it also greatly decreases the chance of averse positive head-to-head as well. Thus, we have to reject Hypothesis 2C.

Being prone to lose against lower-ranked players controversially increases the chance to build positive head-to-head for female players in some waves. However, this could be explained by the fact that the chance of losing against lower-ranked players is generally higher for higher-ranked players with them being still able to build many positive head-to-heads despite of losing a lot against lower-ranked players. For other waves, this effect decreases the chance to build a positive head-to-head, which was anticipated. For female players losing against lower-ranked players is irrelevant for existing head-to-heads, while for male players losing against lower-ranked players will slightly decrease the chance of getting negative head-to-head (magnitude of the parameter between -0.2982 and -0.9993), strongly increase the chance of averting positive head-to-head (magnitude of the parameter between 5.3690 and 51.6434) and less strongly increase the chance of averting negative head-to-head (magnitude of the parameter between (2.4922 and 11.9606).

Losing easy sets percentage is crucial for male players. On one hand, a high percentage of easy sets lost counterintuitively decreases the chances of negative head-to-head for the player. On the other hand, a high percentage of easy sets lost increases the chance of reversing head-to-head from positive to balanced or zero, which is expected. It also increases the chance of reversing head-to-head from negative to balanced or positive. For female players losing easy sets percentage decrease the chance of building positive head-to-head and increase the chance of building negative head-to-head, which is different to ATP. The influence of losing easy sets percentage on reversing head-to-head is the same as for ATP players though significant in fewer waves. This is in line with findings by De Paola and Scoppa (2017) that women are more discouraged when facing the pressure of falling behind in sets and receiving negative feedback.

Table 3: RSiena output for years 2008-2010 on ATP tour for model with alter effects and creation function

Name of effect	estimates	SE	conv t-ratios
Network Dynamics			
1. rate constant (period 1)	4.2885***	(0.1117)	0.0645
2. rate constant (period 2)	3.9953***	(0.0864)	0.0379
3. creat tsw_tsp alter	1.3247***	(0.0770)	0.0748
4. creat esl_esp alter	-1.1097***	(0.0700)	0.0072
5. creat wahr alter	0.2467	(0.3205)	0.0400
6. creat lalr alter	-0.2982	(0.1562)	-0.1074
Behavior Dynamics			
7. rate ranking behaviour (period 1)	0.8083***	(0.0889)	-0.0540
8. rate ranking behaviour (period 2)	1.0115***	(0.1331)	-0.0314
Overall maximum convergence ratio:	0.1610		
Total of 949 iteration steps.			
*** - significant at the 0.001 level.			

6. CONCLUSION

Our study concentrates on the in-match psychological traits of players and the network effect of head-to-head networks. For both ATP and WTA tours in-match psychological traits of players are crucial for their ability to get and keep positive records against other players. For female athletes, the most crucial statistic is winning tough sets, while for male players − not losing easy sets. Also ranking matters more for male players while winning against higher-ranked players or losing against lower-ranked players do not provide much information on a female's potential to get a positive record against other players. Furthermore, a tendency of head-to-head networks to cyclic structures is revealed.

There exist some practical implications. On the one hand, players should understand that psychological traits can be sometimes at least as important as physical condition. They will be able to build a tennis match strategy taking into account the psychological characteristics of their opponents. Tennis coaches should be aware of differences in psychological traits between females and males that were identified in this study. Moreover, those coaches who will be able to quantify explored properties in psychological traits will be able to advise their players to make better decisions on the court. On the other hand, results should be interesting also for sports fans and journalists who cover tennis sport very closely.

This work has some limitations. First of all, we were limited by the dataset. Recently, a lot more information regarding tennis matches is being collected and therefore more options to develop hypotheses with related characteristics will be tested in the future. Secondly, working with RSiena on a relatively large network requires a lot of computer power. Consequently, we were to a great extent restricted by the number of applicable model effects. Thirdly, in the current version of RSiena there is no option to work with weighted networks.

To the authors' best knowledge, this study represents the first attempt to use RSiena in the field of sports and we believe that this study can be a good starting point for further work. Besides including other model effects, additional players' characteristics can be used. It would be interesting to compare differences in network dynamics between ATP and WTA regarding the surface (clay, grass, hard court) or tournament type (Grand Slams versus others). Considering the popularity of sports in general and the availability of data, there are plenty of opportunities to carry out similar research in other sports. In particular, individual sports with match-related characteristics and many matches played among players such as table tennis, badminton, snooker and similar can benefit a lot from this kind of study.

References

Aparício, D., Ribeiro, P. and Silva, F. (2016). A subgraph-based ranking system for professional tennis players. In Cherifi, H., Gonçalves, B., Menezes, R. and Sinatra, R., editors, *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, pages 159–171. Springer International Publishing.

Arcagni, A., Candila, V. and Grassi, R. (2022). A new model for predicting the winner in tennis based on the eigenvector centrality. In *Annals of Operations Research*. https://doi.org/0.1007/s10479-022-04594-7.

Baker, R. and McHale, I. (2014). A dynamic paired comparisons model: Who is the greatest tennis player? In *European Journal of Operational Research*. 236(2): 677–684.

Blanca Torres, J. (2019). Influence of the category and gender in temporary variables in the individual elite tennis. In *Journal of Sport and Health Research*. 11(1): 69–78.

Boulier, B. and Stekler, H. (1999). Are sports seedings good predictors? An evaluation. In *International Journal of Forecasting*. 15(1): 83–91.

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

Breznik, K. (2013). On the gender effects of handedness in professional tennis. In *Journal of Sports Science and Medicine*. 12: 346 - 353.

Breznik, K. (2015). Revealing the best doubles teams and players in tennis history. In *International Journal of Performance Analysis in Sport*. 15(3): 1213–1226.

Candila, V. and Palazzo, L. (2020). Neural networks and betting strategies. In *Risks*, 8(3):68.

Connaughton, D., Hanton, S., Jones, G., Wadey, R., et al. (2008). Mental toughness research: Key issues in this area. In *International Journal of Sport Psychology*, 39(3): 192–204.

Cross, R. (2014). Men’s tennis vs. women’s tennis. In *ITF Coaching and Sport Science Review*. 62(22): 3-5.

Cross, R. and Pollard, G. (2009). Grand slam men’s singles tennis 1991–2009 serve speeds and other related data. In *ITF Coaching and Sport Science Review*. 49: 8–10.

De Paola, M. and Scoppa, V. (2017). Gender differences in reaction to psychological pressure: evidence from tennis players. In *European Journal of Work and Organizational Psychology*. 26(3): 444–456.

Del Corral, J. and Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for grand slam tennis matches? In *International Journal of Forecasting*. 26(3): 551–563.

Dingle, N., Knottenbelt, W. and Spanias, D. (2012). On the (page) ranking of professional tennis players. In Tribastone, M. Gilmore, S., editors, *Computer Performance Engineering. EPEW UKPEW 2012 2012. Lecture Notes in Computer Science, vol 7587.* Springer. 237–247.

Fernández-García, A., Blanca-Torres, J., Nikolaidis, P. and Torres-Luque, G. (2019). Differences in competition statistics between winners and losers in male and female tennis players in Olympic games. In *German Journal of Exercise and Sport Research*. 49(3): 313–318.

Freeman, L. (2004). The development of social network analysis. In *A Study in the Sociology of Science*, 1: 687.

Gollub, J. (2021). Forecasting serve performance in professional tennis matches. *In Journal of Sports Analytics*. 7(7): 223–233.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. In *American Journal of Sociology*. 91(3): 481–510.

Hakeem, H. (2022). Atp and Wta tennis results and betting odds data. https://www.kaggle.com/datasets/hakeem/atp-and-wta-tennis-data. Last access: 16 February 2022.

Hizan, H., Whipp, P. and Reid, M. (2011). Comparison of serve and serve return statistics of high performance male and female tennis players from different age-groups. In *International Journal of Performance Analysis in Sport*. 11(2): 365–375.

Hizan, H., Whipp, P. and Reid, M. (2015). Gender differences in the spatial distributions of the tennis serve. In *International Journal of Sports Science and Coaching*. 10(1): 87–96.

Iso-Ahola, S. E. and Dotson, C. O. (2014). Psychological momentum: Why success breeds success. In *Review of General Psychology*. 18(1): 19–33.

Jayal, A., McRobert, A., Oatley, G. and O'Donoghue, P. (2018). In *Sports Analytics: Analysis, Visualisation and Decision Making in Sports Performance*. Routledge.

Jones, G. (2002). What is this thing called mental toughness? An investigation of elite sport performers. In *Journal of Applied Sport Psychology*. 14(3): 205–218.

Khder, M. and Fujo, S. (2022). Applying machine learning-supervised learning techniques for tennis players dataset analysis. In *International Journal of Advances in Soft Computing & Its Applications*. 14(3): 189–214.

Klaassen, F. and Magnus, J. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. In *Journal of the American Statistical Association*. 96(454): 500–509.

Knottenbelt, W., Spanias, D. and Madurska, A. M. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. In *Computers & Mathematics with Applications*. 64(12): 3820–3827.

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. *In International Journal of Forecasting*. 36(4): 1329–1341.

Kovalchik, S., Bane, M. and Reid, M. (2017). Getting to the top: An analysis of 25 years of career rankings trajectories for professional women's tennis. In *Journal of Sports Sciences*. 35(19): 1904–1910.

Lake, D. and Bryden, M. (1976). Handedness and sex differences in hemispheric asymmetry. In *Brain and Language*. 3(2): 266-282.

Li, P., De Bosscher, V. and Weissensteiner, J. R. (2018). The journey to elite success: A thirty-year longitudinal study of the career trajectories of top professional tennis players. In *International Journal of Performance Analysis in Sport*. 18(6): 961–972.

Lisi, F. and Zanella, G. (2017). Tennis betting: Can statistics beat bookmakers? *In Electronic Journal of Applied Statistical Analysis*. 10(3): 790–808.

London, A., Németh, J. and Németh, T. (2014). Time-dependent network algorithm for ranking in sports. In *Acta Cybernetica*. 21(3): 495–506.

Maquirriain, J. (2014). Analysis of tennis champions' career: How did top-ranked players perform the previous years? In *SpringerPlus*. 3(1): 504.

McHale, I. and Morton, A. (2011). A Bradley–Terry type model for forecasting tennis match results. In *International Journal of Forecasting*. 27(2): 619–630.

Michieli, U. (2018). Complex network analysis of men single ATP tennis matches. *arXiv paper*. https://doi.org/10.48550/arXiv.1804.08138.

Morgan, W. (1985). Selected psychological factors limiting performance: A mental health model. In Clarke, D.H. and Eckert, H.M.E., editors, *Limits of Human Performance*. Champaign, IL: Human Kinetics. 70–80.

Motegi, S. and Masuda, N. (2012). A network-based dynamical ranking system for competitive sports. *Scientific Reports*. 2: 904.

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. In *Physical Review E*. 64(2): 025102.

Nicholls, A., Polman, R., Levy, A. and Backhoused, S. (2019). Mental toughness in sport: Achievement level, gender, age, experience, and sport type differences. In *Personality and Individual Differences*. 47(1): 73–75.

O'Donoghue, P. and Brown, E. (2009). Sequences of service points and the misperception of momentum in elite tennis. In *International Journal of Performance Analysis in Sport*. 9(1): 113–127.

Ovaska, T. and Sumell, A. (2014). Who has the advantage? An economic exploration of winning in men's professional tennis. In *The American Economist*. 59(1): 34–51.

Page, L. and Coates, J. (2017). Winner and loser effects in human competitions. Evidence from equally matched tennis players. In *Evolution and Human Behavior*. 38(4): 530–535.

Paserman, M. (2007). Gender differences in performance in competitive environments: Evidence from professional tennis players. In *IZA Discussion Paper*. 2834: 1–58.

Radicchi, F. (2011). Who is the best player ever? A complex network analysis of the history of professional tennis. In *PloS One*. 6(2): 1-7.

Rilea, S., Roskos-Ewoldsen, B. and Boles, D. (2004). Sex differences in spatial ability: A lateralization of function approach. *Brain and Cognition*. 56(3): 332-343.

Ripley, R. M., Snijders, T., Boda, Z., Vörös, A. and Preciado, P. (2011). *Manual for RSiena.* University of Oxford, Department of Statistics, Nuffield College, 1.

Sackmann, J. (2022). Atp matches. https://github.com/JeffSuchman/tennis_atp.git.

Serwe, S. and Frings, C. (2006). Who will win Wimbledon? The recognition heuristic in predicting sports events. In *Journal of Behavioral Decision Making*. 19(4): 321–332.

Sipko, M. and Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches. *MEng Computing-Final Year Project - Imperial College London*. https://www.doc.ic.ac.uk/teaching/distinguished-projects/2015/m.sipko.pdf.

Situngkir, H. (2007). Small world network of athletes: Graph representation of the world professional tennis player. Available at SSRN: https://ssrn.com/abstract=1001917 or http://dx.doi.org/10.2139/ssrn.1001917.

Snijders, T. (1996). Stochastic actor-oriented models for network change. In *Journal of Mathematical Sociology*. 21(1-2): 149–172.

Snijders, T. (2001). The statistical evaluation of social network dynamics. In *Sociological Methodology*. 31(1): 361–395.

Spanias, A. and Knottenbelt, B. (2013). Tennis player ranking using quantitative models. Manuscript. http://www. doc. ic. ac. uk/wjk/publications/spanias-knottenbelt-mis-2013.pdf.

Szàdoczki, Z., Bozòki, S., Juhàsz, P., Kadenko, S. and Tsyganok, V. (2022). Incomplete pairwise comparison matrices based on graphs with average degree approximately. In *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04819-9.

Vogel, J., Bowers, C. and Vogel, D. (2003). Cerebral lateralization of spatial abilities: A meta-analysis. In *Brain and Cognition*. 52(2): 197-204.

Yue, J., Chou, E., Hsieh, M.-H. and Hsiao, L.-C. (2022). A study of forecasting tennis matches via the Glicko model. In *PLoS One*. 14(1): e0266838.

A. Appendix: A description of variables used in SIENA model

As described in Section 1, we considered some characteristics for estimating SIENA models. Here is their brief description.

- *number of tough sets won against all number of tough sets played - tsw_tsp*

  We transfer from match to set statistics because it's the lowest available level of analysis as a unit. Moreover, a match can consist of a different number of sets. Also, a match can consist of sets of different toughness. Averaging match toughness will result in losing variability information.

  A set should be considered tough if a loser has won at least 5 games in the set. We should note that if both players won 6 games each, in most tournaments a tie-breaker is played, which requires maximum concentration.

- *number of easy sets lost against all number of easy sets played - esl_esp*

  A set should be considered easy if the loser has won no more than 2 games in the set. As opposed to tough sets played statistics, the easy-sets statistic should be rather regarded as a ratio of how more often a player is on the winning side of easy sets than on the losing. The more a player is losing easy sets, the less he will win.

- *number of matches won against higher-ranked opponents out of all matches against higher-ranked opponents - wahr*

  Usually, a higher-ranked opponent is considered a favorite to win the match. The fact that a higher-ranked player has lost a match can be explained by one of the following:

  - A higher-ranked opponent performed badly in the match.
  - A lower-ranked opponent performed outstanding in the match.

  Both players could be unaware of the abilities of the opponent due to the fact that they have never played before.

  When calculating these characteristics we've used raw ranking and not ranking cohorts in RSIENA model. In this respect, we should note that players of approximately the same strength might have close ranking and if one's ranking is higher, it doesn't provide much useful information. Still, we consider winning against the higher-ranked opponent as a strong will to win.

- *number of matches lost against lower-ranked opponents out of all matches against lower-ranked opponents - lalr*

  This is the opposite of winning against higher-ranked opponents.

  We should mention that the higher is the person in the ranking, the less opportunity he has to play against higher-ranked opponents and the more opportunity he has to play (and lose) against lower-ranked ones, which can potentially affect statistics.