Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (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, Piazza S. Francesco, 19 - 55100 Lucca, Italy
Giorgio Gnecco[2],
Laboratory for the Analysis of CompleX Economics Systems (AXES) and Game Science Research Center, IMT School for Advanced Studies, Piazza S. Francesco, 19 - 55100 Lucca, Italy
Rodolfo Metulini[3],
Department of Economics (DSE), University of Bergamo, Via dei Caniana, 2 - 24127 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 its 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 (2022) 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 (2022), 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 (2022) 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 (2022) 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 (2022) 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 (2022).

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 $i = 1, \ldots, n$ 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 $T \in \mathscr{T}$ as $T_1, \ldots, T_{|T|}$. 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 $T^{(2)} \in \mathscr{T}$, $(T^{(1)}, T^{(2)})$ 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 $T^{(2)}$).

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

In the above, $\mathscr{T}$ 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, $\upsilon : \mathscr{T} \to \mathbb{R}$ (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 (2022) 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.

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

[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 Beldsoe	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
Whesley Matthews	2,564,753	0.0398	0.0376	2019/20	MIL	3,623,000	Yes
Donte Di Vincenzo	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 Di Vincenzo	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	30,800,000	Yes
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 Bogdamovic	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 Clarxson	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	1,628,573	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 Clarxson	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.

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 normalized

WGS 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

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$^{\cdot}$
	(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: · 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: · 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).

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

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.

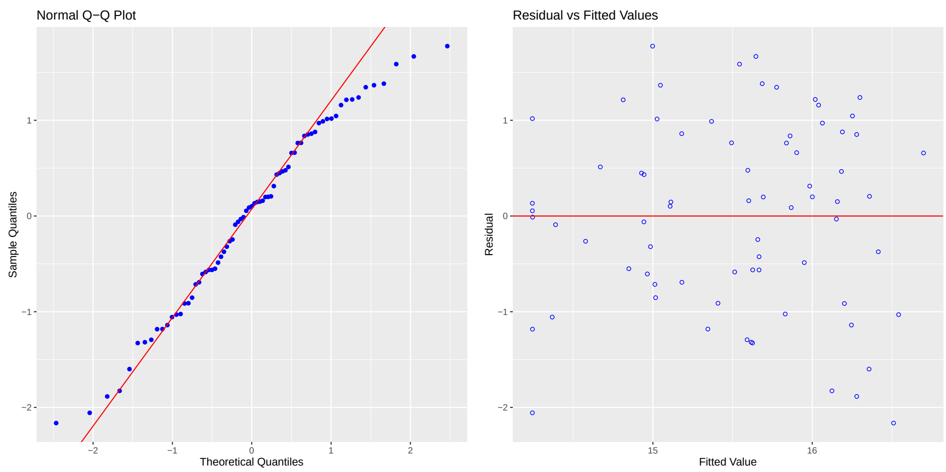

Figure 1: Q-Q plot for normality of residuals (left). Plot of the residuals versus fitted values (right). First regression of Table 4.

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.

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 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 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 ($\hat{salary}_t$), 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%[6]. 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 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

[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.

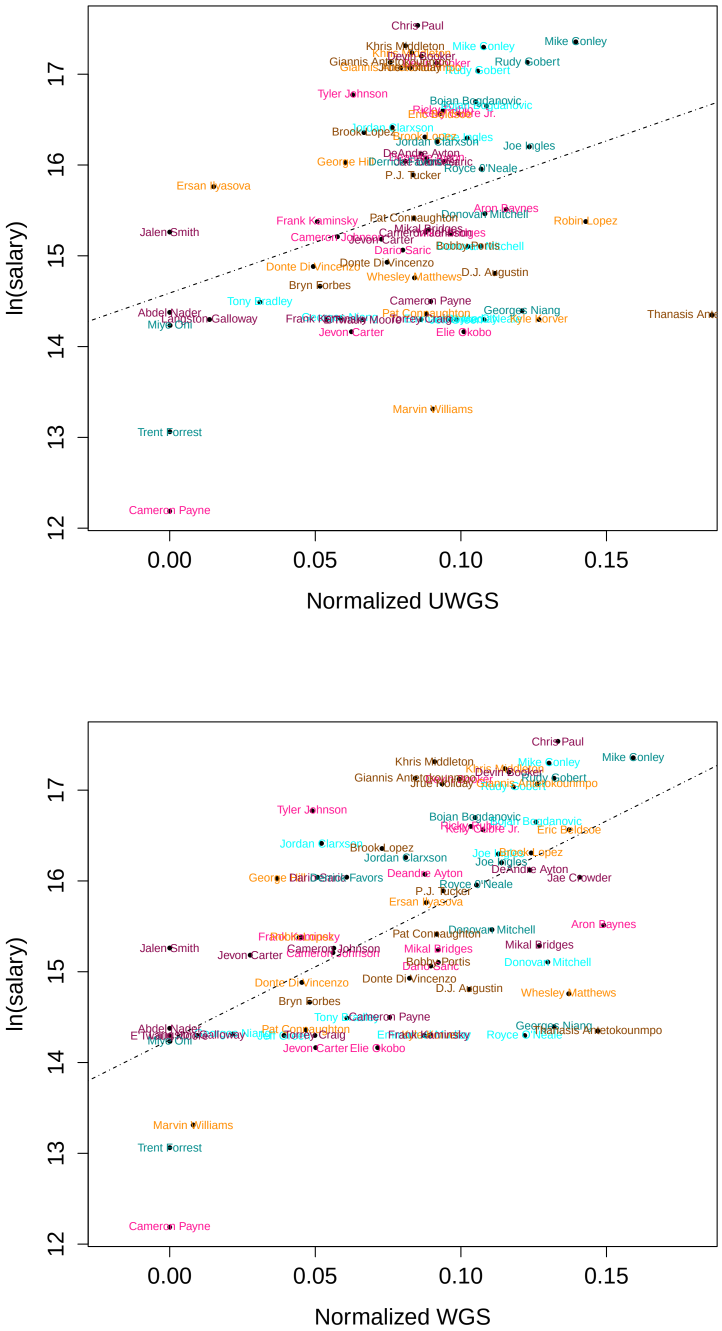

Figure 2: Scatter plot of the natural logarithm of salary (*ln(salary)*, on the Y-axis) and, on the X- axis: (top chart) the *normalized UWGS value*, with regression lines according to the regression in the second column of Table 2; (bottom chart) the *normalized WGS value*, with regression lines according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

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.

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 ago 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 abovementioned 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 (2022).

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 (2022). 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 $\mathscr{L}_i$ 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 $k = 1, 2$:

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

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 presence in the quintet). Under the stated assumptions, the estimate (8) is unbiased, and its variance is inversely proportional to $|\mathscr{L}_i|$. 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. (2022). Measuring players' importance in basketball using the generalized Shapley value. In *Annals of Operations Research*. doi: 10.1007/s10479-022-04653-z.

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., Szczepański, 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.