Analyzing the Influence of Player Tracking Statistics on Winning Basketball Teams

Igor Stančin* and Alan Jović*
* University of Zagreb Faculty of Electrical Engineering and Computing / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
stancin.igor@gmail.com, alan.jovic@fer.hr

***Abstract* - Basketball player tracking and hustle statistics became available since 2013-2014 season from National Basketball Association (NBA), USA. These statistics provided us with more detailed information about the played games. In this paper, we analyze statistically significant differences in these recent statistical categories between winning and losing teams. The main goal is to identify the most significant differences and thus obtain new insight about what it usually may take to be a winner. The analysis is done on three different scales: marking a winner in each game as a winning team, marking teams with 50 or more wins at the end of the season as a winning team, and marking teams with 50 or more wins in a season, but considering only their winning games, as a winning team. The results of the analysis reveal a few categories that are significantly different between the winning and the losing teams, such as: the number of uncontested shots made, the number of assists and secondary assists, and the number of defensive rebound chances. Based on these results, we propose the effective passing ratio, a novel statistical category, which also demonstrates large differences between winning and losing teams.**

Keywords* - *player tracking statistics, hustle statistics, statistical analysis, correlation, winning team, basketball

I. INTRODUCTION

Basketball is a fast and dynamic game with many different kinds of events that are happening on the court very frequently. Every possession of the ball starts with a defensive rebound, steal or inbound pass, ends with shot, free throw or a turnover [1]. During ball possession, many different sequences of events can happen, including empty sequence, because right after the start of a possession the end of the possession may come. Although a single attack in basketball is limited to 24 seconds, a possession of the ball can last a minute or more. When a team misses a shot and grabs offensive rebound (ball retrieved after a missed field goal or free throw), a new possession is not started, only a new play is started. Therefore, a possession of the ball is not limited to 24 seconds [1]. Most of the events are happening around the ball, but some events like fouls or screens can happen away from the ball, anywhere on the court. Collecting complete statistics about all events in a game is therefore a challenging task.

National Basketball Association (NBA) is, incontestably, the best basketball league in the world, both on and off the court. On the court, we find many of the best players in the world. They prove their abilities every four year at the Olympic games, by winning a gold medal in the majority of cases. Off the court, there are specialized personnel who try to improve each segment of the game, e.g. doctors, psychologists, basketball experts and scouts. One of the most interesting facts is that the teams in the league, and the league itself, have expert teams for data analysis that provide them with valuable information about every segment of the game. There are a few teams who make all of their decisions exclusively based on advanced analysis of statistics. Few years ago, this part of organization became even more important, because NBA invested into a computer vision system that collects position of every player and the ball, 25 frames per second [2]. This system made data collection easier and more detailed. Many new statistical categories started to be measured and many of the existing categories were improved.

The goal of this paper is to analyze some of these new statistical categories and their relationship between the winning and the losing basketball teams. Categories that will be analyzed are grouped in the box score (structured summary of all statistics and results of each game) player tracking statistics and box score hustle statistics. Many interesting statistical categories are included in these groups. The aim is to get answers to questions like: “Is more passes always better?”, “Is more running better?”, and “Which one is more important, contested or uncontested shots?”. Statistics that measure defensive events were traditionally neglected compared to offensive statistics [3]. Statistics that are going to be analyzed in our paper contain several defensive types of statistics and we are going to identify the important ones. We intend to find some interesting correlations between these statistical categories, and based on the correlations, try to identify a winning formula, if there is one.

II. SIMILAR PAPERS

Many papers were written based on this new technology of collecting data. Most of the papers rely on positions of each player and the ball in each moment and then build interesting research around it. Franks et al. [3] used a combination of player tracking data and statistics to create a new model for evaluation of defensive plays. Lucey et al. [5] analyzed the information about position of all players 3 seconds before each 3-point shot was taken and tried to determine which features lead to an open 3-point shot.

Goldsberry [6] used the data to rank the best shooters in the NBA using "Court Vision", and, more recently, used it to rank individual defenders [7]. Wiens et al. [8] took a close look at situations in which teams should go for an offensive rebound.

These papers are interesting and they rely on position of each player in each moment, not on pure play-by-play statistics. There are far less papers that analyze only the new statistical categories, without information about players' positions. For example, Sampaio et al. [2] analyze the difference between all-star and non all-star players (each year fans and coaches pick the best individuals from each conference and then these two conference teams play against each other in an all-star game), based on pure play-by-play player tracking statistics. Before that, in 2003, they analyzed the statistics from Portuguese basketball league to determine the difference between winning and losing teams in different types of games [9].

III. DATA SET

The data that is going to be analyzed is freely available from the official web page of NBA [4]. We built a web scraper that got us all the needed information. Four games were removed from the data set, because there were no data recorded for player tracking statistics in these four games. Altogether, the data set consists of 1226 games from the 2016/2017 regular season. We analyze only team statistics, not individually for each player. Although statistics are measured per player, team level statistics are also calculated and freely available. We assume that team level statistics are calculated by aggregating player level statistics for categories for which this makes sense, while categories with percentages and averages were then calculated from the aggregated values. Below are explanations of every statistical category that is in our data set.

Player tracking statistics:

- DIST - Distance run (in miles)
- SPD - Average speed
- TCHS - Touches with the ball
- PASS - Passes
- AST - Assists
- SAST - A player is awarded a secondary assist if they passed the ball to a player who recorded an assist within 1 second and without dribbling
- DFGM - Field goals made by the opponent while the player or team was defending the rim
- DFGA - Field goals attempted by the opponent while the player or team was defending the rim
- DFG% - Percentage of field goal attempts the opponent makes while the player or team was defending the rim
- ORBC - Offensive rebound chances
- DRBC - Defensive rebound chances
- RBC - Rebound chances
- FG% - Field goal percentage
- CFGM - Contested field goals made
- CFGA - Contested field goals attempted
- CFG% - Contested field goal percentage
- UFGM - Uncontested field goals made
- UFGA - Uncontested field goals attempted
- UFG% - Uncontested field goal percentage
- FTAST - A player is awarded a free throw assist if they passed the ball to a player who drew a shooting foul within one dribble of receiving the pass

Hustle statistics:

- Screen Assists - The number of times an offensive player or team sets a screen for a teammate that directly leads to a made field goal by that teammate
- Deflections - The number of times a defensive player or team gets his hand on the ball on a non-shot attempt
- Loose Balls Recovered - The number of times a player or team gains sole possession of a live ball that is not in the control of either team
- Charges Drawn - The number of times a defensive player or team draws a charge (offensive foul)
- Contested 2PT Shots - The number of times a defensive player or team closes out and raises a hand to contest a 2 point shot prior to its release
- Contested 3PT Shots - The number of times a defensive player or team closes out and raises a hand to contest a 3 point shot prior to its release
- Contested Shots - The number of times a defensive player or team closes out and raises a hand to contest a shot prior to its release

Explanations of the categories are available on the official NBA stats API [4]. There are a few glossaries for all the categories, with minor difference between them (e.g. SAST in box score for player tracking glossary represents secondary assists, while in main help/glossary section it represents screen assists). To avoid any confusion, in this paper we will use the glossary as written above.

IV. EXPERIMENTS DESIGN

Since we are analyzing the relationship of the statistical categories between the winning and the losing teams, we will split the data set into several different separations. The first separation will compare statistics for all winning teams in each game and statistics for all losing teams in each

Table 1. Most significant results from first separation of data set.

Statistical category	Mean		Standard deviation		α	p-value
	Winning team	Losing team	Winning team	Losing team		
FG%	0.48	0.43	0.049	0.046	3.7E-4	1.79E-114
UFG%	0.46	0.40	0.074	0.073	3.7E-4	3.24E-67
AST	24.27	20.97	5.26	4.65	3.7E-4	2.72E-54
UFGM	18.52	16.04	4.06	3.88	3.7E-4	1.31E-49
CFG%	0.50	0.46	0.073	0.07	3.7E-4	2.10E-49
DRBC	59.51	53.99	9.04	9.61	3.7E-4	4.94E-48
SAST	5.90	4.89	2.79	2.33	3.7E-4	3.83E-19
CFGM	22.44	20.98	4.43	4.31	3.7E-4	2.02E-15
SCREEN ASSISTS	10.49	9.42	4.09	3.81	3.7E-4	1.99E-11
FTAST	2.24	1.98	1.55	1.43	3.7E-4	3.52E-05

Table 2. Most significant results from second separation of data set.

Statistical category	Mean		Standard deviation		α	p-value
	Winning team	Losing team	Winning team	Losing team		
SCREEN ASSISTS	10.89	9.17	4.08	3.60	3.7E-4	6.25E-17
FG%	0.47	0.45	0.05	0.05	3.7E-4	1.99E-14
UFG%	0.45	0.42	0.08	0.07	3.7E-4	5.11E-13
UFGM	18.36	17.04	4.33	3.93	3.7E-4	1.46E-09
AST	23.61	21.94	5.88	4.67	3.7E-4	2.28E-09
SAST	5.89	5.09	2.97	2.42	3.7E-4	3.89E-07
CFG%	0.49	0.47	0.07	0.07	3.7E-4	2.99E-05
DIST	16.59	16.92	0.72	0.73	3.7E-4	1.27E-22
TCHS	417.89	430.12	35.02	35.08	3.7E-4	4.29E-10
PASS	296.76	307.55	33.35	31.49	3.7E-4	7.51E-09

Table 3. Most significant results from third separation of data set.

Statistical category	Mean		Standard deviation		α	p-value
	Winning team	Losing team	Winning team	Losing team		
FG%	0.49	0.44	0.05	0.05	3.7E-4	6.66E-56
UFG%	0.48	0.41	0.08	0.07	3.7E-4	8.43E-38
AST	25.13	21.12	5.79	4.49	3.7E-4	2.52E-30
UFGM	19.47	16.43	4.13	3.84	3.7E-4	1.44E-30
CFG%	0.50	0.46	0.07	0.07	3.7E-4	9.08E-23
SCREEN ASSISTS	11.22	8.93	4.18	3.51	3.7E-4	8.98E-19
DRBC	58.79	53.81	8.43	9.42	3.7E-4	3.40E-19
SAST	6.42	4.89	3.08	2.37	3.7E-4	2.92E-16
DIST	16.63	16.94	0.74	0.74	3.7E-4	2.16E-14
DFG%	0.52	0.57	0.12	0.12	3.7E-4	1.31E-11
PASS	298.97	308.82	33.12	32.17	3.7E-4	2.70E-05

game. The second separation will be done by marking teams with 50 or more wins at the end of a season as a winning team, and on the other side, marking teams with 35 or less wins in a season as a losing team. This gives us groups that are comparable, because of a similar number of entities in each group. The third separation is the same as the second one, but this time, we are considering only the winning games for teams with 50 or more wins and only losing games for teams with 35 wins or less.

For every separation we make, we calculate the mean and standard deviation of each category. For establishing whether the differences between the distributions in each separation are relevant, we use the two-tailed Mann-Whitney U Test. This test is appropriate for establishing difference in means between two categories, regardless of the categories' data distribution and regardless of the direction of categories' relationship. In this way, we establish significance of each statistical category. Significance level alpha is set to $\alpha_0 = 0.01$, but since we are repeating our test for 26 statistical categories (plus one novel category, see section V.B), in order to establish relevance of each category, we will use Bonferroni correction to try to avoid false positive results. Bonferroni correction formula is

$$\alpha = \frac{\alpha_0}{m} \quad (1)$$

Where *m* is 27, so our new alpha is 3.7E-4.

Also, for every separation, correlations between categories are calculated and we try to identify some interesting rules between two features, or in a group of features.

V. Results and Discussion

In Tables 1–3, we show the most significant differences in statistical categories between winning and losing teams.

A. Interpretation of results

We can see that the field goal percentage (FG%) has a large significance (low p-value) in all three tables. This is expected, because the main goal of basketball is to score a basket, so the teams that cannot do it effectively may have smaller chances of winning. Field goal percentage is not a new statistical category, so it is not of our main interest, but it is good to establish its importance, so that we can compare it with the importance of percentages of other categories, such as uncontested and contested shots.

If we think about contested and uncontested shots before conducting the analysis, we can make two claims. First, we claim that everybody will make their uncontested shots with approximately the same percentage and that the difference between winning and losing teams would be a better percentage of contested shots. Second, we claim that better teams will have more uncontested shot attempts, which would lead them to gain easy points by making those uncontested shots and eventually to winning the game. However, if we take a look at the statistics, we can see that in the Tables 1–3, there is no uncontested shot attempts (UFGA). This means that there is no significant difference between winning and losing teams in this category, so the second claim is disproven. The results for uncontested field goal percentage (UFG%) in Tables 1 and 2 also refute the first claim.

It can be observed that uncontested field goal percentage shows larger difference between winning and losing teams than contested field goal percentage in all three tables. This could be confusing at first, but if we think about the game and the shots, we can say that most of the uncontested shots are shots for 3-points or shots of players who have weaker offensive skills, because defenses are leaving them free for shot deliberately, while contested shots attempts are mostly near the rim. This may be the most probable reason for this confusing difference in percentages.

Uncontested field goals made (UFGM) are significant in all three tables as well. The difference between winning and losing teams are around 2.5 per game. Contested field

Table 4. Correlation for all three separations.

Categories that correlate		Correlations					
		First separation		Second separation		Third separation	
		Winning	Losing	Winning	Losing	Winning	Losing
FG%	UFG%	0.62	0.59	0.70	0.65	0.67	0.61
FG%	CFG%	0.71	0.69	0.72	0.70	0.68	0.65
FG%	UFGM	-	-	0.51	-	0.44	-
FG%	AST	0.47	-	0.53	0.49	0.45	0.40
SAST	AST	0.55	0.48	0.61	0.49	0.60	0.50
UFGM	AST	0.47	-	0.54	-	0.49	0.43
UFGM	UFG%	0.68	0.70	0.70	0.68	0.65	0.67
CFGM	CFG%	0.59	0.65	0.60	0.62	0.58	0.62
ORBC	CFGA	0.47	-	-	0.48	0.42	0.47
PASS	SAST	0.36	-	0.37	-	0.37	-
PASS	TCHS	0.98	0.97	0.98	0.97	0.98	0.98
DIST	TCHS	0.48	0.55	0.50	0.57	0.51	0.59

goals made (CFGM) are significant only in the first separation of data, while in the other two cases, the significance level is negligible. With all this information, we can conclude that the uncontested shots are more important than contested ones for quantifying the difference, but this opens a new question: is it better to have a few uncontested field goals made more, but with a slightly lower percentage, or the other way around?

Screen assists (SCREEN ASSISTS) have significant differences in all three separations. Since Table 2 and Table 3 represent the differences between the best and the worst teams at the end of a regular season, this difference implies that season-round better teams are setting better screens. Two things can explain why this could be a case. The first one is that better teams have better and stronger players, so it is hard for a defender to get around the screen. The second one is that better teams have better offensive actions, so they easily create points with screens.

The other result that implies that winning teams have better players is distance covered (DIST). The difference is significant in Tables 2 and 3, but this time, winning teams have a lower average, which means that they run less than the losing teams. A possible explanation for this could be that their players have a better intuition for where at the court they should be, so they manage to do things with less effort in running.

Defensive rebound (DRBC) chances are significant in Table 1. There is also a high significance for defensive rebound chances in the third separation (Table 3). If an opponent team has a lower field goal percentage, then that leads directly to a higher number of defensive rebound chances. Significance of defensive rebound chances is just another confirmation for the importance of field goal percentage.

Additionally, we tried to find some connection between player quality and any of these statistical categories or winning teams by adding into the data set the average Player Efficiency Rating (PER) [10] of 4 players that played the most minutes in game. We choose PER as the currently best measure for quality of players. Unfortunately, there was no clear connection or correlation between any statistical category and the PER average.

Table 4 contains the largest and most interesting correlations between categories. Values that are missing were not written because they are too small for consideration (below 0.35). We were hoping to find the "winning formula" through correlations, but unfortunately, there is no such a thing, because the data set is too complex, having too many factors involved. Still, some moderately strong correlations exists, and based on them (for winning teams), we might say that more running leads to more touches and passes, and that leads to more assists and secondary assists. Finally, considering all that, we are getting better shots and our field goal percentages are larger, which leads to winning. However, this conclusion would be somewhat far-fetched, because the correlations are too small for this kind of conclusion and we can not be sure that this kind of chaining of correlations is appropriate, because some other factors that we did not take into consideration might be involved. Actually, significance levels of differences between those two distributions shown in Tables 1–3 are a confirmation that this conclusion would be wrong. So, we can say that we could not find a clear winning formula, at least not using correlation.

B. Suggestion of a novel statistical category

Assists have a high level of significance in Tables 1–3. If we consider only Tables 1 and 3, which contain only the individual winning games, we can see that the difference in average for assists is around 4 assists per game. Since assists in many cases create easy points, this represents a big difference between winning and losing the game.

A player is awarded a secondary assist (SAST) if he passed the ball to a player who recorded an assist within 1 second and without dribbling [4]. From the definition of SAST, we can say that there is a high correlation between AST and SAST and that means that a high number of secondary assists should also be a significant difference between winning and losing teams. Confirmation for this conclusion is shown in Tables 1–3.

Passes (PASS) also have significant differences in Tables 2 and 3. What is surprising is that winning teams have less passes then the losing teams. Every year, basketball is more and more played as a team sport (in terms of offensive actions), while isolations (offensive action where teammates back away to draw their defenders as far from the ball as possible and the ball-handler tries to beat a defender one-on-one) are less frequent. The first thing that could explain the fact that winning teams have less passes is that it is not enough to run and pass the ball, what losing teams often do, there still needs to be some clear plan on what is the goal of passing. The second explanation could be that winning teams have better individuals, who can attract double-team defense on them, and then just pass to an open player who then has a much easier job.

Since assists and secondary assists are passes, and also free throw assists (FTAST) have significant difference in

Table 5. Statistics for EPR.

Values for EPR in:	Mean		Standard deviation		Alpha	p-value
	Winning team	Losing team	Winning team	Losing team		
First separation	0.109173	0.09325	0.023968	0.021157	3.7E-4	3.38E-63
Second separation	0.106989	0.094896	0.027418	0.020208	3.7E-4	4.72E-20
Third separation	0.113273	0.090872	0.027674	0.019107	3.7E-4	2.08E-42

the first separation, the formula for the effective passing ratio (EPR) automatically imposes itself:

$$EPR = \frac{AST + SAST + FTAST}{PASS} \quad (2)$$

It is a simple formula that gives us a percentage of passes that are any kind of assists. Screen assists are not involved in this formula, because screen assists are not passing assists. If a passing assist is involved in the play with a screen assist, it would be added through AST. For example, if a player gets an open position by his teammate screen, he gets the ball and makes a field goal, that assist is already added within AST and adding it one more time within the screen assist would not be correct.

Since each factor in the formula is significant in at least one of first three tables, EPR is also significant. Significance level for EPR is shown in Table 5. We can observe that differences in all three categories are highly significant. Based on that, we can conclude that EPR is a very important difference between winning and losing teams. Although statistically significant differences in no way imply causality, it may be beneficial that teams become aware of the importance of EPR and try to adjust their game in order to improve it and thus potentially increase their chances of winning.

In order to demonstrate the significance of EPR, we build a logistic regression model based only on EPR of a single team in a single game and try to predict whether that team won or lost the game. We used 10-fold cross-validation for the purpose on the whole dataset and we compared the results with random model for each separation. Random model gave us around 50% accuracy, as expected. Logistic regression model based on EPR gave us the accuracy of 63.3% in first separation, 62.8% in second separation, and 65.7% in the third separation. Thus, the model achieves more than 20% improvement compared to random case, which confirms our supposition that EPR makes an important difference between winning and losing teams. Due to space limitation, future work with EPR would include a detailed comparison with simple assists, such as identification of teams that are good in one category but bad in the other one, observation of EPR on the player level and similar.

VI. CONCLUSION

The analysis has shown some significant differences between winning and losing teams. The most important differences are: 1) field goal percentage, especially uncontested (UFG%), but also contested (CFG%); 2) uncontested field goal made (UFGM) are perhaps even more important than uncontested field goal percentage and it is certain that the uncontested shots are more important than the contested ones; 3) better teams have lower distance covered; 4) better teams have less passes and touches with the ball, despite the current high popularity of motion attacks (type of offensive actions where all the players and the ball are constantly moving); 5) better teams have more defensive rebound chances, which can be linked to poor opponent field goal percentage; and 6) better teams have more assists and secondary assists, despite the fact that they have less passes.

Correlations between categories exists, but they are too small to make any kind of "winning formula". Based on the analysis, in order to become a winning team, the most prudent approach would be to concentrate on the categories that were mentioned in the paper and try to make some adaptations in the team in order to improve these categories. For instance, creating offensive plays with more efficient running, more effective passes with goal of creating uncontested shots.

Effective passing is a very important difference between winning and losing teams. For that purposes, we introduced the effective passing ratio (EPR) measure. We can safely conclude that this measure is one of the most important differences between winning and losing team. The current results suggest that teams start taking the measure of quality of their passes as a reliable indicator of their future success in the game.

For future work, we plan to analyze the data from several seasons more deeply using various data mining methods in order to find additional, previously undiscovered models of winning teams' success.

REFERENCES

[1] J. Kubatko, D. Oliver, K. Pelton, and D. T. Rosenbaum, "A starting point for analyzing basketball statistics," Journal of Quantitative Analysis in Sports 3(3), p. 1, 2007; doi:10.2202/1559-0410.1070

[2] J. Sampaio, T. McGarry, J. Calleja-Gonzalez, S. Jimenez Saiz, X. S. I. del Alcazar, and M. Balciunas, "Exploring Game Performance in the National Basketball Association Using Player Tracking Data," PLOS ONE 10(7): e0132894, 2015; doi:10.1371/journal.pone. 0132894

[3] A. Franks, A. Miller, L. Bornn, and K. Goldsberry, "Counterpoints: Advanced Defensive Metrics for NBA Basketball." MIT Sloan, 9th Annual Sports Analytics Conference (SSAC15), Boston, MA, USA, Feb. 27-28, 2015, pp. 1–8, 2015.

[4] National Basketball Association. "Official NBA stats API," https://stats.nba.com/

[5] P. Lucey, A. Bialkowski, P. Carr, Y. Yue, and I. Matthews, "How to Get an Open Shot: Analyzing Team Movement in Basketball using Tracking Data," MIT Sloan, 8th Annual Sports Analytics Conference (SSAC14), Boston MA, USA, Feb. 28 – Mar. 1, 2014, pp. 1–8, 2014.

[6] K. Goldsberry, "CourtVision: New Visual and Spatial Analytics for the NBA", MIT Sloan, 6th Annual Sports Analytics Conference (SSAC12), Boston MA, USA, Mar. 2-3, 2012, pp. 1–7, 2012.

[7] K. Goldsberry and E. Weiss, "The Dwight Effect: A New Ensemble of Interior Defense Analytics for the NBA", MIT Sloan, 7th Annual Sports Analytics Conference (SSAC13), Boston, MA, USA, Mar. 1-2, 2013, pp. 1–11, 2013.

[8] J. Wiens, G. Balakrishnan, J. Brooks, and J. Guttag, "To Crash or Not To Crash: A Quantitative Look at the Relationship Between Offensive Rebounding and Transition Defense in the NBA," MIT Sloan, 7th Annual Sports Analytics Conference (SSAC13), Boston, MA, USA, Mar. 1-2, 2013, pp. 1–7, 2013.

[9] J. Sampaio and M. Janeira, "Statistical analyses of basketball team performance: understanding teams' wins and losses according to a different index of ball possessions," International Journal of Performance Analysis in Sport 3(1), pp. 40–49, 2003.

[10] J. Hollinger "Pro Basketball Forecast," Potomac Books Inc., 2005.