Beyond the Boxscore: Applying Team and Individual Performance Evaluation Metrics to Canadian Basketball

Abstract

The increased usage of statistical tools and methodologies in basketball has led to significant improvements in player and team performance evaluation. This quantitative analysis has produced many statistical metrics that provide teams with comparative advantages over their opponents. These metrics are readily available in many American and International basketball leagues but are particularly underutilized at many levels of basketball in Canada, even though similar raw data is collected through mediums such as boxscores and play-by-play data. This paper applies some of these boxscore-based metrics to the Canadian Elite Basketball League and University Sports Basketball, with an accompanying website [1] that makes these statistics publicly accessible. Following a brief history of the advancements of analytics in basketball and background on the boxscore and its uses, the Four Factors of Basketball will be used to derive expected win percentage models for teams over the course of a season. The introduction of these models will be followed by discussions on their statistical background and characteristics, and case studies will illustrate unique findings resulting from their construction. Advanced individual player metrics such as Box Plus Minus, Player Efficiency Rating, Value over Replacement Player, and Points per Possession will also be calculated to improve our notion of player value within these leagues.

1 Background

Basketball statistics began with tabulating points scored, personal fouls, and free throws and dividing these aggregate totals by the number of games played to produce *per-game* statistics. These statistics are easy to compute, widely understood by the general public, and still used by sports broadcasters and fans alike. However, they hold very little value to analysts when describing player performance because they fail to consider playing time, making them poor indicators of efficiency and productivity. Introducing basic *percentages* to measure shooting proficiency resolved some differences in usage rates but gave way to untrustworthy sample sizes and did not properly account for the lower-percentage but increased value of the 3-point shot. As new metrics were introduced to try and combat these deficiencies, ambiguity arose about their context and employment and little headway in basketball analytics was made until the late 1990s. It was then that *regression-based* statistics were introduced to compare the predictive power and standard error of models and per-game statistics were succeeded by the more stable and informative *per-minute* statistics. *Possession-based* metrics introduced a new wave of analytics that provided a strong correlation with wins on a team level and could be broken down for individual players as well [2]. Since teams alternate possessions in a game, playing basketball was abstracted to an optimization problem of how to maximize a team's per-possession efficiency [3]. Current leaders in advanced NBA analytics involve *plus-minus* statistics [4], such as Box Plus-Minus (BPM) [5], Estimated Plus-Minus (EPM), and Daily Plus-Minus (DPM). Other metrics with more exotic acronyms have experienced increased popularity in recent years, including LEBRON, a player role and luck-adjusted metric from BBall Index [6], and RAPTOR, the Robust Algorithm (using) Player Tracking (and) On/Off Ratings developed by Nate Silver of FiveThirtyEight [7].

1.1 Data Collection and Structure

Boxscores are the most accessible source of detailed game analysis at all levels of basketball. They are compact tables that contain the summarized raw counting statistics of players, which are then aggregated to comprise team performance, giving people both a cursory and in-depth review of a game's happenings with greater detail than just viewing the final score. However, boxscores do not tell the whole story, and they have been succeeded by more informative data

sources such as *play-by-play* and *player tracking*. Play-by-play data refers to a game report that gives the sequential listing of events that determine a game's outcome, chronicling the score and players on the court at the time of any event occurrence. They contain useful information for predicting the outcome of a particular matchup and simulating the progression of events that can lead to such an outcome [8]. The introduction of spatio-temporal data via player tracking and wearable technologies in the 2010s has revolutionized basketball analytics, as cameras are now able to determine the locations of all the players on the court, as well as the (x,y,z) coordinates of the ball [9] with incredible accuracy and precision. This has given researchers the ability to determine the relationship between shot selection and opponent location, as well as provide improved analysis of individual defensive performance [10] - something historically undervalued by boxscore and play-by-play-based statistics. Bayesian regression models and Markov Chains have been used to model player career projections and team evaluations, leading to the creation of new metrics [11].

The leagues under consideration are University Sports (**U SPORTS**), the governing body and leader of university sport in Canada [12], and the Canadian Elite Basketball League (**CEBL**), the top domestic league in Canada. U SPORTS player and team data is taken from 'Coach's View' Season Summary Statistics that are publicly available for each basketball team by PrestoSports [13]. Information about player position, which is incorporated in individual performance evaluation metrics, is taken from usportshoops [14]. Statistics from the 2015-2016 through 2019-2020 seasons were used as the training set for created models, and the 2021-2022 season was used as the test set for predictions (the 2020-2021 season was cancelled due to the covid-19 pandemic). The CEBL played its inaugural season in 2019 and is comprised of talent from leagues such as the NBA G-League, USPORTS, the NCAA, and multiple high-level international leagues. Season summary data and roster information for the CEBL is taken from RealGM [15]. Since neither league uses player tracking and play-by-play data can be difficult to parse efficiently, the season summaries were taken exclusively from game boxscores. The data was collected using web-scraping in the R [16] programming language, and scripts were written to calculate advanced statistics from these basic statistics and output them in easy-to-read Microsoft Excel tables.

1.2 Team Evaluation Using the Four Factors of Basketball

The Four Factors of Basketball were introduced by statistician and basketball analytics pioneer Dean Oliver in his 2003 book *Basketball on Paper* [17]. They are based on the four main events that affect an offensive possession in basketball - a field goal attempt, a turnover, an offensive rebound, and a foul. While other more obscure events can affect or terminate a possession, the four factors capture the overwhelming majority of possession-ending occurrences, making models that utilize them relatively simple and incredibly informative. Since these factors can be expressed in terms of percentages they are comparable across seasons and even eras of basketball with some slight modifications and estimations.

- **Effective Field Goal Percentage** (eFG%) calculates how efficiently a team scores, adjusting for the extra value gained from a 3-point shot versus a 2-point shot. $eFG\% = \frac{FG + 0.5(3P)}{FGA}$

- **Turnover Percentage** (TOV%) measures the percentage of a team's possessions that end in a turnover. Since possessions are not officially tracked in U SPORTS boxscores, estimates with good empirical accuracy are used in their place: $POSS \approx FGA + 0.44 \times FTA - OFF + TO$ [3]. $TOV\% = \frac{TO}{POSS}$

- **Offensive Rebound Percentage** (ORB%) is defined as the percentage of offensive rebounds that a team obtains out of total available rebounds from missed field goal attempts and free throws. $ORB\% = \frac{OFF}{TOT}$

- **Free Throw Factor** (FTF%), the percentage of "free "points a team gets per field goal attempt. It underscores the conventional wisdom that the ability to get to the line does not have much bearing if a team is not proficient at making free throws. $FTF\% = \frac{FT}{FGA}$

The rest of the paper is organized in the following manner. The Four Factors of Basketball will be used in Section 2 on U SPORTS team data to model a team's expected win percentage, recalculating the coefficients to better understand each factor's impact on the men's and women's games. Section 3 introduces techniques to quantify individual player performance in the CEBL by extending aggregate methods and reports the results, which are summarized in Section 4.

2 Expected Win Percentage

The expected wins model modifies baseball's "Expected Runs" metric [18] by regressing a team's wins on the four factors, a useful application of which is the ability to determine the "importance" of each factor. Once the factors have been ranked, a strategic framework can be devised within a team's operations to properly allocate time to developing skills that better align with the significance of the factor in question. By adhering to a regimented training model based on the four factors and their relationship to winning, teams can manipulate their factor scores to overcome talent constraints and optimize their expected win percentage.

2.1 Methodology

A multiple linear regression is a common empirical research tool that attempts to explain a measurable outcome (the dependent variable, Y_i) using well-understood determinants (independent variables), while accounting for random error ϵ. Independent variables are represented as the set $X = \{X_m, m \in 0, 1, 2, \ldots M\}$, where $X_0 \equiv 1$ to provide the traditional intercept (constant) term. The general form of a multiple regression model is then defined as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M + \epsilon$$

such that

$$E(Y|X_1, \ldots, X_M) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M$$

Multiple regression model assumptions [19] must hold so that the parameters $\beta_0, \beta_1, \ldots, \beta_M$ of the model can be estimated using *Ordinary Least Squares Estimators*. Ordinary least squares estimators minimize the sum of squared errors for each parameter, represented as

$$S(\beta_m) = \sum_{i=1}^{N}(y_i - E(y_i))^2 = \sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_M x_{iM})^2$$

With these assumptions satisfied, this process will result in the estimators $\hat{\beta_m}$ being the best linear unbiased estimators of the parameters. If the error term of the model is normally distributed, then the least-squares estimators and the dependent variable y_i for each observation will also be normally distributed. The coefficient of determination, more commonly referred to as a model's R^2 value, will be used in the following discussions to demonstrate how well the four factors predict team success in a season. *Adjusted R^2 ($\bar{R}^2$)* is an alternative measure of goodness-of-fit that accounts for some deficiencies present within the basic R^2 construct and penalizes the inclusion of variables that do not add value to a model's predictive ability.

2.2 Model Specification

Because U SPORTS teams in different conferences play a different number of regular season games, expected winning percentage is used as the response variable in place of expected wins so that the model is properly generalized across the league. To use the four factors as explanatory variables to model a team's expected win percentage, their offensive and defensive factor performances are split into two categories. The offensive factors are denoted by their familiar name, while the factors of their opposition are prepended with "opp", for eight total variables. Since the actual win percentage is a value ranging from 0 to 100 inclusive and is linear in nature, there is no guard against expected values outside the realm of possibility, such as values over 100% or negative percentages. Due to its proprietary nature it is unclear whether Oliver's model included an intercept term or not, and the terms' effects are worth considering. The inclusion of a constant term often has an inconsequential interpretation, but excluding it could have significant ramifications on a model's R^2 and normality assumptions. Therefore, models with and without the constant term will be generated in the following manner and compared during analysis:

$$
\begin{aligned}
WinPecentage = {} & \beta_0 + \beta_1 eFG\% + \beta_2 TOV\% + \beta_3 ORB\% + \beta_4 FTF\% \\
& + \beta_5 oppeFG\% + \beta_6 oppTOV\% + \beta_7 oppORB\% + \beta_8 oppFTF\%
\end{aligned}
\tag{1}
$$

$$
\begin{aligned}
WinPecentage = {} & \beta_1 eFG\% + \beta_2 TOV\% + \beta_3 ORB\% + \beta_4 FTF\% + \beta_5 oppeFG\% \\
& + \beta_6 oppTOV\% + \beta_7 oppORB\% + \beta_8 oppFTF\%
\end{aligned}
\tag{2}
$$

2.3 Analysis

Table 1 report the output of the regressions run from equations 1 (M_1 and W_1) and 2 (M_2 and W_2), courtesy of the stargazer package from R [20]. The signs on each coefficient align with a basic understanding of the game of basketball: shooting, rebounding, and free throws will have a positive impact on a team's success, while turnovers will have a negative effect. The opposite intuitively holds for the opposition and is reflected by the coefficients on the "opp"variables. Each coefficient is statistically significant at the 1% level except for oppFTF in M_2, which has a p-value of 0.012.

Table 1: U SPORTS Win Percentages

|	*Dependent variable:*			
	Win Percentage			
	M_1	M_2	W_1	W_2
eFG	2.333***	2.855***	2.242***	2.779***
	(0.214)	(0.161)	(0.201)	(0.164)
TOV	−1.962***	−1.734***	−1.761***	−1.554***
	(0.260)	(0.258)	(0.186)	(0.187)
ORB	0.695***	0.936***	0.678***	0.813***
	(0.167)	(0.157)	(0.140)	(0.142)
FTF	0.562***	0.702***	0.659***	0.718***
	(0.185)	(0.185)	(0.177)	(0.183)
oppeFG	−2.726***	−2.231***	−2.294***	−1.793***
	(0.224)	(0.181)	(0.219)	(0.193)
oppTOV	2.062***	2.022***	1.795***	1.867***
	(0.227)	(0.232)	(0.211)	(0.218)
oppORB	−0.970***	−0.596***	−1.038***	−0.806***
	(0.234)	(0.214)	(0.169)	(0.166)
oppFTF	−0.544***	−0.479**	−0.734***	−0.621***
	(0.185)	(0.189)	(0.184)	(0.189)
Constant	0.747***		0.643***	
	(0.209)		(0.149)	
N	236	236	236	236
R^2	0.872	0.975	0.904	0.977
$\bar{R}^2$	0.868	0.974	0.901	0.976
RSE	0.087 (df = 227)	0.090 (df = 228)	0.084 (df = 227)	0.087 (df = 228)

Note: *$p<0.1$; **$p<0.05$; ***$p<0.01$

In M_1, the adjusted R^2 of 0.868 means that the explanatory variables account for nearly 87% of the variation in the response variable (win percentage) of the model. This is extremely high and comparable with results from NBA datasets [21][22], providing strong evidence in defence of the factors' relationship to winning. The minuscule difference between the original and adjusted goodness-of-fit values demonstrates how little penalty each added variable receives. Since the factors and dependent variable can be expressed with percentages, they have a very straightforward interpretation. For example, a 1 percent change in eFG% is expected to change a team's win percentage by 2.333 percent, and a team that reduces its TOV% by 1 percent is expected to enjoy a 1.962 percent increase in win percentage. M_2 omits the constant term and consequently produces some interesting results. To start, the coefficient of determination in M_2 experiences a substantial increase to 0.974, nearly 11 points greater than that of M_1. Another

Model	R^2	RMSE	RSE
M_1	0.876	0.093	0.174
M_2	0.870	0.096	0.179
W_1	0.875	0.103	0.166
W_2	0.868	0.106	0.172

Table 2: Comparing Predictive Abilities of Models

observation from comparing the two models is the defensive emphasis of M_1 compared to the more offense-heavy focus given to the coefficients of M_2. The values of eFG% and oppeFG% are seemingly flipped between the models, and rebounding also experiences a similar switch in coefficients on both sides of the ball, from 0.695 to 0.936 offensively and -0.97 to -0.596 on defense. Some coefficients remain relatively unchanged, such as oppTOV (2.062 vs 2.022) and oppFTF (-0.544 vs -0.479). The intercept term β_0 is incorporated to ensure that the mean of the residuals will be zero and protect the model against bias, part of the model assumptions referenced in Section 2.1. Even though the histograms of the residuals of M_1 and M_2 have slightly different shapes as show in Figure 2.3, those of M_2 are still very normal in their distribution, with a mean of 0.000555. Normality tests such as that of Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling all report p-values greater than 0.05 (0.084, 0.85, and 0.33, respectively), so any uncertainties about M_2's legitimacy can be quickly put to rest.
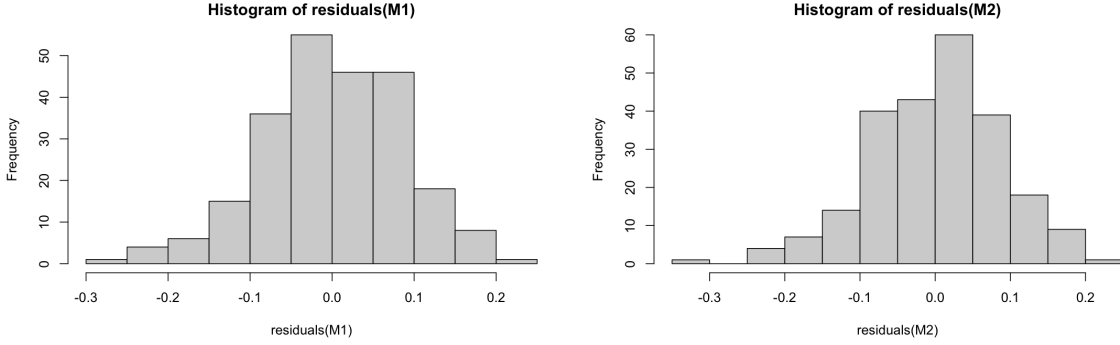
The women's regression output W_1 was similar to the men's model, with intuitively correct coefficients and an excellent R^2 value of 0.904. W_2 exhibits a comparable offensive emphasis to M_2 when estimating the factor coefficients and experiences a large increase in R^2 with almost zero penalization from the dependent variables. The ability to explain nearly 98% of the model's variation is substantial and further proves that the four factors are transferable across multiple levels of basketball. Apart from eFG%, which displays a considerable difference in offense and defense between W_1 and W_2, the other factor coefficients are relatively unchanged. It is interesting to note that even though the average eFG% in U SPORTS basketball is vastly different between men and women (48% to 41.3%), the coefficient on these terms for both models is similar.

2.4 Predictive Ability

The models used the games played during the 2021-22 season to test their predictive abilities and were compared in Table 2 using their R^2, root-mean-square error (RMSE), and residual square error (RSE). Even though the models without intercepts had superior R^2 values to those with them in the regression output with the training set, they were very similar in predicting win percentage on the test data set. The men's and women's models produced nearly identical results in explaining the variation of win percentage in the test set, but the RSEs of the women's models were lower than M_1 and M_2, with W_1 being the only model to dip under 17%. Figure 1 compares the expected win percentage from M_1 (horizontal axis) and actual win percentage (vertical axis) for the 48 men's basketball teams during the 2021-2022 season. Within this figure lie various applications of the four factors, two of which are explored in the following case studies.
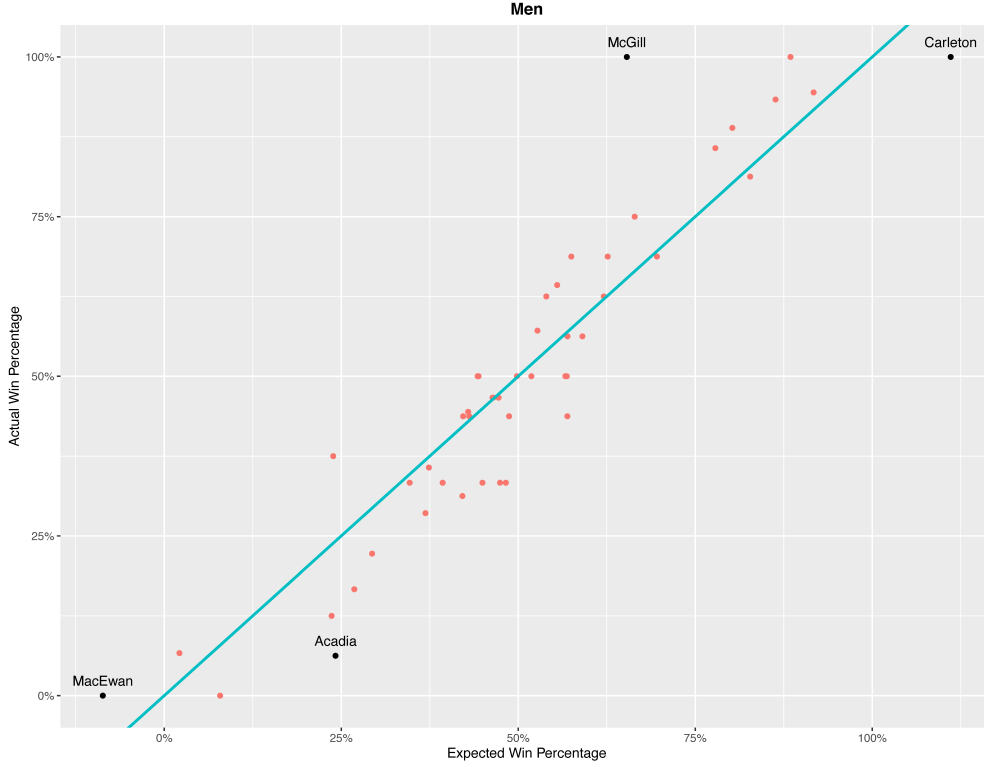
Figure 1: Comparing Expected and Actual Win Percentage for the 2021 - 2022 Season

eFG	TOV	ORB	FTF	oppeFG	oppTOV	oppORB	oppFTF
40.4	23.0	22.3	16.6	56.8	19.2	34.8	14.9

Table 3: Macewan Griffins 2021-2022 Regular Season Statline

2.5 Case Study: Measuring Statistical Accomplishment

One application of the expected win percentage model is using the predicted values as measures of statistical dominance and inferiority. Here, the "best" and "worst" statistical teams are defined as those with the highest and lowest expected win percentages, the extreme values from the x-axis shown in Figure 1. Coming as little surprise to anyone familiar with U SPORTS men's basketball over the past decade, the best statistical team during the 2021-2022 season were the eventual national champion *Carleton Ravens*. They were the cream of the crop amongst a deep pool of talented teams, with an expected win percentage nearly 20 points greater than the next best school, crusing to a 14-0 regular season. Statistically, Carleton boasted the 2nd best eFG%, 2nd best ORB%, and the 2nd best opponent ORB% in U SPORTS, at 53.7%, 38.2%, and 20.9%, respectively. However, their defensive performance is what warranted their infeasible expected winning percentage of 111%. Their 37.8% opponent eFG was the lowest recorded in U SPORTS men's basketball, not only in the season of interest but across the entire 6-year data set. Only four other teams, including the 2017-2018 Ravens, had recorded a season allowing an opponent eFG under 40% since the 2015-2016 season. When examining statistical inferiority among men's U SPORTS teams, one does not have to look any further than the Grant Macewan University Griffins. The Griffins failed to win a game during the regular season and their prohibitively poor performance resulted in them achieving an expected win percentage of -8%, the only predicted value below zero among men's results. Shown in Table 3, their 40.4% eFG% was the lowest in U SPORTS in the 2021-2022 season, and the second-worst output across the entire 6-year data set. Defensively, Macewan's opponent eFG% of 56.9% was the highest surrendered in the six years under consideration, nearly 1% worse than the next closest team, the winless 2017-2018 Trinity Western Spartans. The Griffins were among the bottom-five teams in nearly every other statistical category during 2021-2022, including the 3rd-worst TOV%, 4th-worst ORB% and FTF%, and 2nd-worst opponent ORB%. An interesting thing to note is that Macewan had the 3rd-best opponent FTF% in the country, and this detail demonstrates why FTF% is the least valuable of the four factors - excellence in this category cannot overcome poor results among the other factors.

2.6 Case Study: Assessing Achievement Using Residuals

Another byproduct of the predicted model construction in Figure 1 is that the difference between expected and actual win percentage gives the residuals, visually described as the vertical distance between the 45° line and the point of interest. Taking the maximum and minimum values of these residuals yields a unique interpretation - teams with the greatest under/over-achievement. Across the men's teams, Acadia University had the dubious distinction of U SPORTS' most underachieving team. The lowly Axemen finished dead last in their conference with a 6.25 win percentage (1 win, 15 losses) but had an expected win percentage nearly 18 points higher, at 24% (4 wins and 12 losses). Acadia had solid offensive production with a 49.6% eFG% that was good for 16th in the country, and coupled this shooting proficiency with an above-average 18.1% TOV%, 20th in U SPORTS. This set the precedence for perceived success but is where the positive reviews come to a crashing halt. The Axemen were ranked 41st in the country in opponent eFG%, 44th in opponent FTF%, and 46th in ORB% at a paltry 21.4%. However, there is more than meets the eye about Acadia's undisciplined defense, a detail that provides clarity to Acadia's win percentage discrepancy. The Axemen received 13 disqualifications during the regular season, the third highest total in U SPORTS. This statistic is not shown by the four factors and occurs when a player is ejected from a game for committing five personal fouls. The early exits of players who were key offensive contributors cost Acadia on multiple occasions late in games, with five of their losses coming by 6 points or less. On the opposite end of the achievement spectrum, the McGill Redbirds were an expected 8-win 4-loss team that took advantage of a weak conference schedule en route to a perfect 12-0 record, winning the RSEQ division. The Redbirds' 50% difference between their win percentage and that of the second place team (Concordia, 6-6) was the largest such gap in any division in U SPORTS. Additionally, their division had the smallest divisional parity index as per win percentage standard deviation ratios [23], an indicator of the overall closeness of talent within that division. McGill achieved success with an elite defensive scheme, ranking 5th overall in opponent FTF%, 6th in opponent TOV%, and 11th in opponent eFG%. This staunch defense won the Redbirds multiple "close" regular season games - 9 of their 12 victories came by 6 points or less. However, their modest 49.5% eFG% and below-average 21.9% TOV% proved to be insurmountable against the stronger out-of-conference opponents that McGill faced in the USPORTS Final 8. In the quarterfinal, they were overwhelmed by the statistically superior and eventual U SPORTS bronze medallists Alberta Golden Bears and then fell to the Canada West Champion Victoria Vikes in the consolation semifinal. This marked a disappointing end to McGill's once hopeful season, summarized by a lack of offensive prowess that was hidden behind the relative weaknesses of their regular season opponents.

3 Individual Player Evaluation

A question that arises following the effective performance of advanced analytical tools in evaluating team performance is whether these methods can be extended to serve as proxies for quantifying an individual player's talent. Such proxies provide great value to coaching staff for scouting reports and playing time allocation, and to fans and other media members in making informed year end league award nominations. Individual player metrics generally fall into two categories, the first being statistics that measure a player's value in terms of their contributions to their team's success in a specific area, a top-down approach. This category takes possession-related statistics such as the four factors and points per possession (PPP) and disaggregates them so that player statlines can be used to calculate individual PPP, eFG%, TOV%, and FTF%. This technique used in paid applications such as Synergy Sports Technology [24] and provides insight into identifying the relative strengths (and weaknesses) of each player to properly contextualize their valuation. Additionally, play-by-play and spatial data make it possible to determine a player's ORB%, which cannot be computed using only boxscore numbers. The other branch of individual player metrics are called "catch-all" statistics, a bottom-up approach that aims to summarize observable contributions of a player and conglomerate them into one number to determine their value. Catch-all metrics are extremely useful for overall comparisons of players but are a hotly contested subject area in NBA circles due to the relative secrecy of their formulation. However, some of these calculations have been publicized, and three of them; Player Efficiency Rating (PER) [25], Box-Plus Minus (BPM) [5], and Value over Replacement Player (VORP) [5], will be discussed in greater depth.

3.1 Catch-All Statistics

PER was created by former ESPN columnist John Hollinger and was the first all-in-one statistic to attempt to quantify a player's value. It takes a players' boxscore contributions: field goals, free throws, 3's, assists, rebounds, blocks and steals, turnovers, missed shots, and fouls, adding the positive contributions and penalizing the negative ones using detailed formulas [26]. It adjusts for per-minute productivity and team pace, as well as the league averages in all the statistical categories, and is normalized such that a PER of 15 is considered league average. It is biased towards offensive output since these totals are better recorded in boxscores so players with an unseen defensive impact will be considered underrated overall, but as the first of its kind PER revolutionized player evaluations during the 2010s. BPM is the successor to PER and has replaced it in most NBA front offices. Created by Daniel Myers, it is much more comprehensive and better-rounded metric than its predecessor (though still a bit offensively biased). BPM estimates a player's contribution in points above league average per 100 possessions played. It is produced using similar boxscore-based statistics as PER, but also includes a player's position, offensive role, and their teams' overall performance in the even more comprehensive calculations [27]. BPM is set such that league average is 0.0 (0 points above or below average). For example, a BPM of 5.0 means that a player's team is 5 points per 100 possessions better with them on the floor than with average production from another player. VORP converts BPM, which does not take playing time into account, into an estimate of a player's overall contribution to their team. It is measured against what a theoretical bench/replacement player would provide (someone with a BPM of -2.0 by definition), by including the number of possessions that a player has featured in and the fraction of their teams' games that they have played. VORP removes any biases in pace and is considered by many as one of the better measures of actual value contributed to a team [28].

	Name	Team	BPM	VORP	eFG	TOV	FTF	PPP	PER
1	Tony Carr	SR	5.2	6.1	50.4	15.70	36.9	1.39	21.30
2	Khalil Ahmad	NIA	4.9	6.0	48.7	12.90	39.4	1.66	22.80
3	Shane Gibson	FVB	5.7	6.0	57.8	12.00	12.8	1.49	20.70
4	Christian Vital	HHB	9.1	5.7	54.8	11.00	14.8	1.60	24.40
5	Jordan Baker	ES	5.0	5.3	52.8	16.10	16.5	0.88	18.50
6	Alex Campbell	FVB	3.9	5.2	56.0	9.30	20.2	0.97	16.00
7	Isiah Osborne	MON	6.6	5.2	53.7	4.50	11.3	1.09	17.60
8	Deng Adel	OB	4.5	4.9	53.6	19.00	30.3	1.32	18.30
9	Caleb Agada	HHB	4.7	4.4	46.8	17.80	12.1	1.23	18.90
10	Isiaha Mike	SSS	3.8	4.2	58.7	12.10	13.1	1.29	18.60

Table 4: Advanced Player Metrics for the Top 10 CEBL Players During the 2022 Season, per VORP

3.2 Application to the CEBL

Table 3.1 displays the advanced statistics of the top ten CEBL players, as per VORP. Due to the relative infancy of the CEBL, these calculations are the first publicly available set of their kind [1]. There are many interesting discussions that arise from viewing this table, and one example is the power of VORP over BPM and PER in terms of strict player valuation. Take Hamilton's Christian Vital, who led the CEBL in BPM by a considerable margin and was second in both PER and PPP but was credited as the fourth on the VORP total and wasn't even shortlisted for the league's most valuable player honors. This is because Vital only appeared in 13 out his team's 20 regular season games, and this through no fault of his own - part of his time away was spent playing for the Toronto Raptors during the 2022 NBA Summer League. BPM and PER, which are rate statistics, correctly identify Vital as a prolific producer when on the court, and it comes as no surprise that upon returning he led his team to the CEBL championship and was named finals MVP. However, since his overall impact over the full course of the season was not felt to his team as much as the players ranked ahead of him (Carr, 18 games played, Gibson 19, Ahmad 20), VORP accurately definition his value in terms of overall contribution to his team's success.

When comparing the players from Table 3.1 to their league award considerations, a couple of significant omissions can be found that demonstrate the incompleteness of valuations based on traditional counting statistics. Montreal's Isiah Osborne, who was left off of both Sixth Man of the Year and Top Canadian ballots, was second only to the aforementioned Christian Vital in BPM and led the CEBL with a TOV% of 4.5%, nearly half that of the next closest player (E.J. Onu, 8.5%). This mistake-free and efficient style of basketball is highly valued by

statistical metrics but unfortunately caused Osborne to be overlooked by more prolific point accumulators. An above average interior scorer, Osborne averaged a "pedestrian "10.3 points per game while playing for the league's worst team, and these factors led to him not being properly credited for the impact he carried. Similarly, Edmonton's Jordan Baker was left off of the Top Canadian ballot even though he led all Canadian-born players in VORP and was top-5 among Canadians in both BPM and PER. Baker is the CEBL's all-time leader in rebounds and assists and had an impressive 2022 campaign, acting as a major focal point of his team's offensive schemes even though he often wasn't the main scoring option and recorded a PPP total reflective of that. His undervaluation comes partially due to his unseen defensive impact and high creation output, which, though not seen by basic counting statistics and some of the offensive top-down statistics, were recognized accordingly by advanced statistics.

4   Conclusion

Advanced statistical methods utilizing regressions and Bayesian-based constructs have been used for nearly 30 years in the NBA to enhance the quantitative understanding of the game of basketball. This paper investigated possession-based metrics and their applications within U SPORTS and the CEBL to assess their efficacy when addressing questions of how to evaluate team and individual performance. It has been demonstrated that these analytical instruments maintain excellent functionality when applied to data from such leagues, with models utilizing the four factors alone explaining up to 98% of a team's variation in win percentage for U SPORTS men's and women's teams. Expected win percentage predictions perform extremely well, with low error rates and R^2 values near or above 87% for the 2021-2022 U SPORTS regular season. Established individual player metrics such as Box Plus Minus, Value Over Replacement Player, and Player Efficiency Rating were applied to CEBL players to illustrate oversights in traditional player valuations and provide observers with a better ranking system. The introduction of a publicly accessible repository and website to access this data allows both fans and decision makers to utilize them to maximize and understand their team's success. Advances in these areas and the inclusion of more descriptive data types such as spatial and play-by-play will ensure that Canadian Basketball leagues continue to produce high-quality basketball as it grows.

4.1   Future Works

There are limitations to the individual player evaluation applications from Section 3 due to lack of information or available data surrounding their original regression and plus-minus methods. In PER's case, a low-level description of the coefficients that compute the league value of an offensive possession and the importance of each player's contribution is not publicly available. Therefore, while it is effective in general for intra-league player comparisons, its values become increasingly approximate the farther away the style of play of the league is from that of the NBA. Box Plus Minus tolerates different league contexts much better by comparing players to their own teammates before adjusting based on individual statistics but struggles to properly adjust coefficients in leagues that have widely different PPP or possession-ending event type distributions. For example, differing offensive rebounding distributions and large eFG% discrepancies that are apparent between U SPORTS men and women should result in a different coefficient weights given to each contribution. Therefore, attaining the additional information and data necessary to recreate BPM, PER, and similar catch-all metrics from the ground up will improve our notion of player value within U SPORTS and the CEBL.

4.2   Supplementary Material

The data and code necessary to reproduce the analyses performed in this paper are available from GitHub at `https://github.com/awosoga/CMSAC_2022`.

References

[1] David Awosoga. Canadian basketball analytics. `https://canadianbasketballanalytics.carrd.co`. Accessed: 2022-08-12.

[2] Jeremias Engelmann. Possession-based player performance analysis in basketball (adjusted +/- and related concepts). 2016.

[3] Justin Kubatko, Dean Oliver, Kevin Pelton, and Dan Rosenbaum. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3:1–1, 02 2007.

[4] Austin Clemens. Nylon calculus 101: Plus-minus and adjusted plus-minus. `https://fansided.com/2014/09/25/glossary-plus-minus-adjusted-plus-minus/`. Accessed: 2022-02-26.

[5] Daniel Myers. About box plus/minus (bpm). `https://www.basketball-reference.com/about/bpm2.html`. Accessed: 2022-02-21.

[6] Krishna Narsu and Tim/Cranjis McBasketball. Lebron introduction. `https://www.bball-index.com/lebron-introduction/`. Accessed: 2022-03-01.

[7] Nate Silver. Introducing raptor, our new metric for the modern nba. `https://fivethirtyeight.com/features/introducing-raptor-our-new-metric-for-the-modern-nba/`. Accessed: 2022-03-01.

[8] Petar Vračar, Erik Štrumbelj, and Igor Kononenko. Modeling basketball play-by-play data. *Expert Systems with Applications*, 44:58–66, 2016.

[9] Zachary Terner and Alexander Franks. Modeling player and team performance in basketball, 2020.

[10] Alexander Franks, Andrew Miller, Luke Bornn, and Kirk Goldsberry. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9, 05 2014.

[11] Edgar Santos Fernandez, Paul Wu, and Kerrie Mengersen. Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4):289–312, 2019.

[12] U SPORTS. U sports. `https://usports.ca/en`. Accessed: 2022-02-21.

[13] PrestoSports. Prestosports | all-in-one sports technology platform. `https://www.prestosports.com/landing/index`. Accessed: 2022-02-21.

[14] Martin Timmerman. U sports hoops, university baskbetball in canada. `https://usportshoops.ca`. Accessed: 2022-02-21.

[15] RealGM. Canadian elite basketball league. `https://basketball.realgm.com/international/league/128/Canadian-Elite-Basketball-League/`. Accessed: 2022-08-01.

[16] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022.

[17] Dean Oliver. *Basketball on paper : rules and tools for performance analysis / Dean Oliver.* Brassey's, Inc., Washington, D.C, 1st ed. edition, 2004.

[18] T.M. Tango, M.G. Lichtman, and A.E. Dolphin. *The Book: Playing the Percentages in Baseball.* Potomac Books, 2007.

[19] R. Carter Hill, William E Griffiths, and G. C. (Guay C.) Lim. *Principles of econometrics / R. Carter Hill, William E. Griffiths, Guay C. Lim.* Wiley, Hoboken, fifth edition, 2018.

[20] Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.3, 2022.

[21] Tarek Al Baghal. Are the four factors indicators of one factor? an application of structural equation modeling methodology to nba data in prediction of winning percentage. *Journal of Quantitative Analysis in Sports*, 8, 01 2012.

[22] Konstantinos Kotzias. The four factors of basketball as a measure of success. `https://statathlon.com/four-factors-basketball-success/`. Accessed: 2022-05-09.

[23] Duane Rockerbie. *The Economics of Professional Sports 2017.* 11 2017.

[24] Synergy Sports Technology. Synergy sports. `https://synergysports.com`. Accessed: 2022-04-18.

[25] J. Hollinger. *Pro Basketball Forecast.* PRO BASKETBALL PROSPECTUS. Potomac Books, Incorporated, 2005.

[26] Basketball Reference. Calculating per. `https://www.basketball-reference.com/about/per.html`. Accessed: 2022-08-12.

[27] Daniel Myers. Box plus/minus 2.0 example calculations. `https://docs.google.com/spreadsheets/d/1PhD9eo3IqzpQo21-yVJPQzYjpXl_h-ZonIKqGEKBqwY/edit#gid=307166562`. Accessed: 2022-08-15.

[28] Bryan Kalbrosky. What is the best advanced statistic for basketball? nba executives weigh in. `https://hoopshype.com/lists/advanced-stats-nba-real-plus-minus-rapm-win-shares-analytics/`. Accessed: 2022-03-01.