An Advanced Real Plus Minus for Soccer

A More Accurate Measure of Player Value in the World of Soccer.

Abstract

Real Plus Minus (RPM) is an advanced statistic used in major team sports that was created in order to assess a player's true value to their team. While standard box plus minus statistics factor only the score differential while a player is on the court/field, RPM accounts for the level of supporting talent and competition. The idea was first popularized in basketball but has not been utilized extensively for the sport of soccer. For this paper, we wanted to not only expand this popular metric to soccer but incorporate various statistics in addition to goals that highlight a player's added contributions to their team. This advanced version of RPM serves as a holistic measure of the impact a player has on a game. To calculate the metric, we employed a ridge regression using data from the 2019-20 Premier League season. By weighting each game statistic, we created an overall offensive and defensive advanced RPM for each player. While this metric would typically be used to identify transfer targets that are considered valuable, we believe it can be applied on a game level allowing managers to construct optimal lineups based on their game plan for the opposing team as well. Our methodology can be replicated for any soccer league across the world, and the advanced RPM concept is applicable to different sports.

Data

To create our real plus-minus metric, we used the fcscrapR package [1](Yurko) to get ESPN game commentary for every match during the 2019-2020 Premier League season. The publicly available github repository had a function to scrape game ID's which were used to loop through all of the matches. We then utilized the rvest package to scrape the lineups of each game using the same game ID's in order to obtain all of the starters and substitutes.

To separate when a player is on and off the field for a given match, we created a new stint anytime there was a substitution, or a player was sent off with a red card or a second yellow. A stint is defined by the 22 total players on the field, and the average game may have 6 unique stints if there are five substitutions at separate points in the game. Using the play-by-play commentary, we recorded how many times each event happened during each stint, while also noting the players involved in each stint and the amount of time played per stint. Each line of the data frame includes every player that played in the Premier League during the season, with a 1 under the names of the players who played during that specific stint for the home team, a -1 under the players who played for the away team, and a 0 for everyone who was not playing in that stint. This combination of 1s, 0s, and -1s essentially created a "lineup matrix" that allowed us to know which players were on the field at any given time for any given match. Each line also includes the start and end time of each stint, as well as the number of different statistics that occurred during the stint. The length of the stint allows us to create per 90 statistics to improve upon accuracy. The events we recorded for each stint include:

a. Goals and Goals Allowed
b. Shots on Target and Shots on Target Allowed
c. Free Kicks won in the Attacking Half and Free Kicks given up in the Attacking Half
d. Red Cards for Opponents and Red Cards
e. Yellow Cards for Opponents and Yellow Cards

[1] https://github.com/ryurko/fcscrapR

f. Corners and Corners Allowed
g. Counter Attacks and Counter Attacks Allowed
h. Key Passes and Key Passes Allowed
i. Penalties Won and Penalties Allowed

Here is an example of what the stint data would look like for one game:

Game ID	Stint ID	Home Team	Away Team	Goals Scored	Aaron Connolly	Aaron Cresswell	Aaron Lennon	Angelo Ogbonna	Conor Hourihane
541465	541465_1	West Ham United	Aston Villa	0	0	0	0	1	-1
541465	541465_2	West Ham United	Aston Villa	0	0	0	0	1	-1
541465	541465_3	West Ham United	Aston Villa	0	0	0	0	1	-1
541465	541465_4	West Ham United	Aston Villa	1	0	0	0	1	0
541465	541465_5	West Ham United	Aston Villa	0	0	0	0	1	0
541465	541465_6	West Ham United	Aston Villa	0	0	0	0	1	0

Figure 1: Example Stint Data for One Game

There were a few discrepancies in the ESPN data which we had to manually fix within the R data frames. The first issue was with player names, as the lineup and commentary scrapes produced different iterations of some names (ex: Jeff vs Jeffrey). We also had to fix occasional game statistics that were missing for a select few matches.

After we calculated the RPM, we used publicly available FBref data to obtain the following attributes for each player as they would increase the functionality of the statistic:

a. Position
b. Club
c. Age
d. Games Played
e. Minutes Played

Methodology

After collecting our data, we used a Ridge Regression to calculate our Advanced RPM. Ridge regression uses a type of shrinkage estimator called a ridge estimator. Shrinkage estimators theoretically produce new estimators that are shrunk closer to the "true" population parameters. In theory, the ridge will add bias to estimates to depict their true population values. To explain this in terms of soccer, the ridge regression will add "bias" to players who are playing against better players, hence their contributions during that game will be even more noteworthy. Essentially, this is how we account for the talent on the field allowing for the Advanced RPM metric to be accurate.

For each statistic, we ran a ridge regression with the event being the dependent variable and the lineup matrix being the independent variables. The coefficient value (x) from each ridge regression for each player implies that when that player is on the pitch, the team will have x more amount of that event than they usually would. For example, our ShotsOnTargert Ridge, predicts the amount of additional shots on target a player's team will have when he is on the field. Kevin De Bruyne's Shots on Target coefficient was 0.30951. This implies that when he is on the field, his team is predicted to have 0.30951 more Shots on Target than they normally would. Below is an example of a ridge regression's output:

```
Call:
linearRidge(formula = shot_on_target ~ . - 1, data = Per90OffShotsOnTarget)

          `Aaron Connolly`           `Aaron Cresswell`                `Aaron Lennon`                  `Aaron Mooy`
               0.0143494179               0.0001486909               0.0483644713              -0.1520229479
          `Aaron Ramsdale`           `Aaron Wan-Bissaka`          `Abdoulaye Doucouré`                 `Adam Idah`
              -0.1818949066               0.0490923946              -0.0645410391              -0.2123526297
           `Adam Lallana`                `Adam Masina`                 `Adam Smith`               `Adam Webster`
              -0.0067615796              -0.0282124082              -0.2093192067              -0.0547777436
           `Adama Traoré`                      Adrián            `Adrian Mariappa`          `Ahmed El Mohamady`
              -0.0669363993               0.0409819588              -0.0501983643              -0.1078359393
```

Figure 2: Example Output of a Linear Ridge

The results of each ridge regression were then used as part of a weighted sum to create our final Advanced RPM value. We weigh each event based on their "goal value" which is how likely that event is to result in a goal scored or allowed for a team. For example, goals scored and goals allowed would have a weight of 1 and -1 respectively. Below are the formulas used to calculate the weighted sum:

$$AdvancedAttRPM = \alpha_1 GoalsCoeff + \alpha_2 SoTCoeff + \alpha_3 FKCoeff + \alpha_4 RCCoeff + \alpha_5 YCCoeff + \alpha_6 CornerCoeff + \alpha_7 CounterAttacksCoeff + \alpha_8 KeyPassesCoeff + \alpha_9 PenaltiesCoeff + \alpha_{10} ShotAttCoeff$$

$$AdvancedDefRPM = \beta_1 GoalsCoeff + \beta_2 SoTCoeff + \beta_3 FKCoeff + \beta_4 RCCoeff + \beta_5 YCCoeff + \beta_6 CornerCoeff + \beta_7 CounterAttacksCoeff + \beta_8 KeyPassesCoeff + \beta_9 PenaltiesCoeff + \beta_{10} ShotAttCoeff$$

The α_x represents the weight for each event with $\beta_x = -\alpha_x$. The β_x values are the negative of α_x as we want the defending RPM to be scaled in a manner where the higher a player's defending RPM, the better their defending capability. For example, the GoalsCoeff in the advanced DefRPM formula represents the amount of goals a player allows when he is on the field, hence a negative coefficient implies that player is better. Therefore, multiplying it with a negative value will allow us to scale it so that a greater DefRPM value implies you are better on defense. The weights were developed using previous studies that help in understanding the relationship of each event on the possibility of scoring a goal. For example, penalties are weighted as 0.76 as the xG value of a penalty is 0.76. Below is a graph showing the weights (α_x):

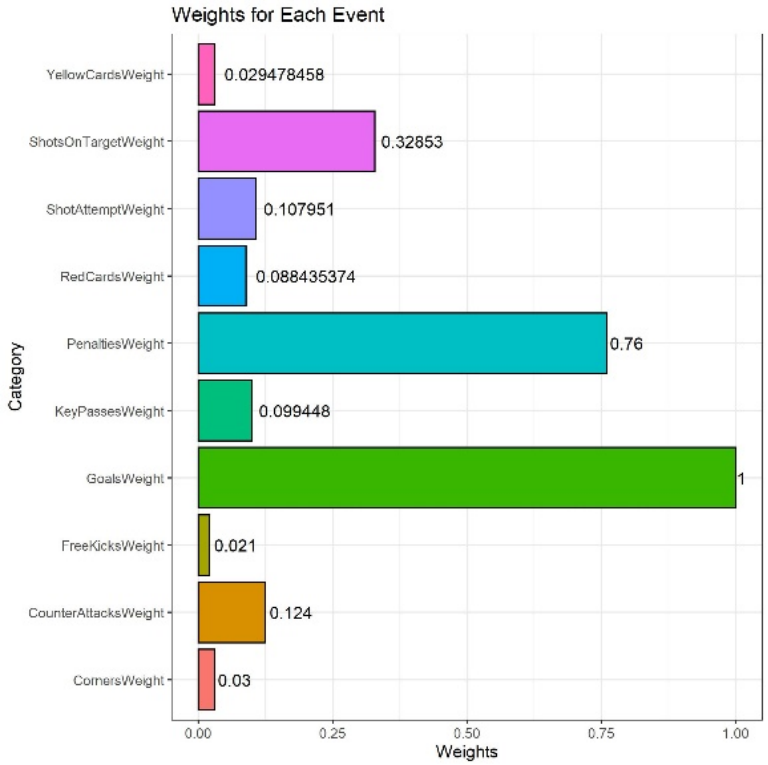

Figure 3: Weights of Each Event

The overall RPM of a player is simply calculated by adding the Advanced Attacking RPM and Advanced Defending RPM as seen by this formula:

$$Advanced\ Real\ Plus\ Minus = AttackingRPM + DefendingRPM$$

Here is an example of the RPM calculation for Kevin De Bruyne:

$$AdvancedAttRPM = (0.125) + (0.108 * 1.1034) + (0.4297 * 0.03) + (0.0175 * 0.0295) + (-0.025 * 0.088) + (0.056 * 0.021) + (-0.00008 * 0.76) + (0.013 * 0.124) + (0.310 * 0.329) + (0.064 * 0.099)$$

$$AdvancedDefRPM = (-1 * -0.102) + (-0.712 * -0.108) + (-0.06 * -0.03) + (-0.058 * -0.0295) + (0.0008 * -0.088) + (-0.113 * -0.021) + (-0.002 * -0.76) + (-0.002 * -0.124) + (-0.363 * -0.329) + (-0.0116 * -0.099)$$

$$AttRPM = 0.366,\ DefRPM = 0.307,\ Real\ Plus\ Minus = 0.673, (0.366 + 0.307)$$

*The values used above have been rounded up for the sake of understanding and hence are not 100% accurate.

Findings and Results:

As mentioned earlier, our results were both conclusive and arguably similar to that of the eye test of an avid Premier League fan but also reveal a few hidden gems in the Premier League. In this section we have visualizations that validate our findings, highlight the output of our research, and how it can be applied in the future.

In the first three visualizations below, we utilized a minutes played requirement to ensure that the results had only players that significantly contributed to their teams during the 2019-20 season. Although it may have cut out some youngsters or role players, it provides a more accurate sample of who the best players in the categories are as players with less than 1140 minutes may have a skewed RPM statistic due to a small sample size. We selected 1140 minutes as the requirement as it represents 33% of the total 3420 minutes of a 38 game Premier League season for an individual player. Since many young prospects were not included in the first four, we will look at all players under the age of 21 later in the paper. Similar to before, we placed a minutes played requirement, but in this case, it was only at least 342 minutes, which represents 10% or more. We chose a smaller requirement because as with most teams in the Premier League, it is often hard to find minutes for younger players due to the competitiveness of the league. It is not ideal to throw many of the underdeveloped and inexperienced players into the heat of the battle. Therefore, we believe a requirement of 342 minutes is effective as many U21 players do not play as much as veterans. In terms of the output, a scatter plot is used to identify the key performers. When reading it, players who are further right and higher up, think northeast in direction, are those who are considered the most effective players as these are players with a high RPM over a large sample. On the other hand, those who are the furthest right and furthest down, think southeast in direction, are the least effective players as these are players with a low RPM over a large sample.

Before we look at the visualizations it is extremely important to understand that Soccer, unlike Basketball or Hockey, do not require every player to be heavily involved on the both sides of the field, attacking and defending. While we would expect every player in Basketball or Hockey to be good on offense and defense, in soccer we do not expect a striker to excel when defending. Therefore, we believe for attacking players, their attacking RPM is a far better measure of their value than the holistic RPM. The same can be said for defenders and their defending RPM and midfielders and their holistic RPM. However, the

attacking RPM of defensive players can still be used to indicate which defenders are more involved in corners or in pushing up and vice versa for the defending RPMs of attacking players to a certain degree.

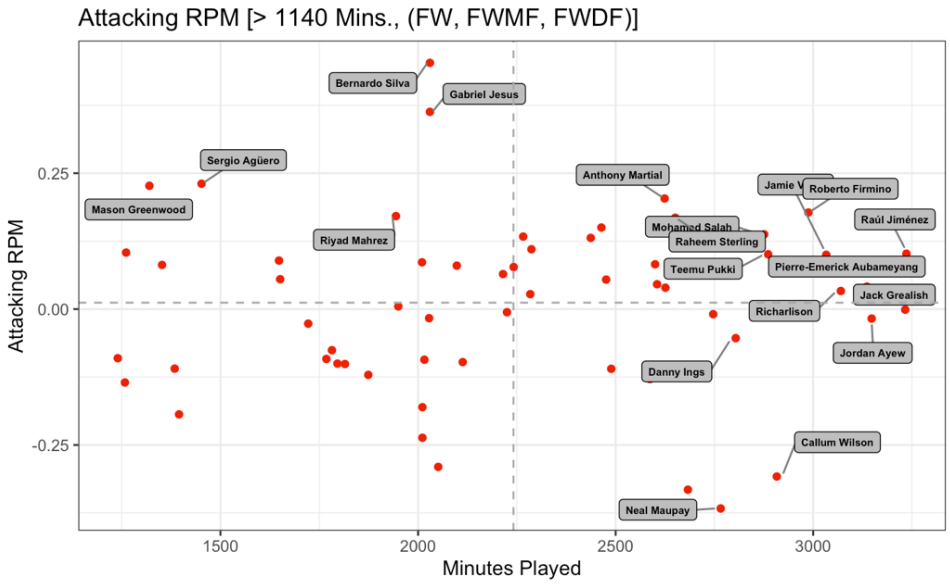

Figure 4: Attacking RPM of Attacking Players (> 1140 Mins.)

The first position group we looked at were the attacking players of the Premier League, or those who were labeled as FW (forward), FWMF (forward-midfielder), or FWDF (forward-defender). For this visualization, the attacking RPM statistic was used as it is most effective at revealing the most valuable attackers. According to the plot, we can identify Bernardo Silva, Gabriel Jesus, Anthony Martial, and Roberto Firmino as the most effective attackers. These results are important as it brings validity to the reasoning and output behind our RPM figure as many consider these players to be some of the elite attackers in the Premier League. While these results are not the end-all to an argument of who is better, it can provide a better outlet for analysts and fans to identify other players who provide the most value.

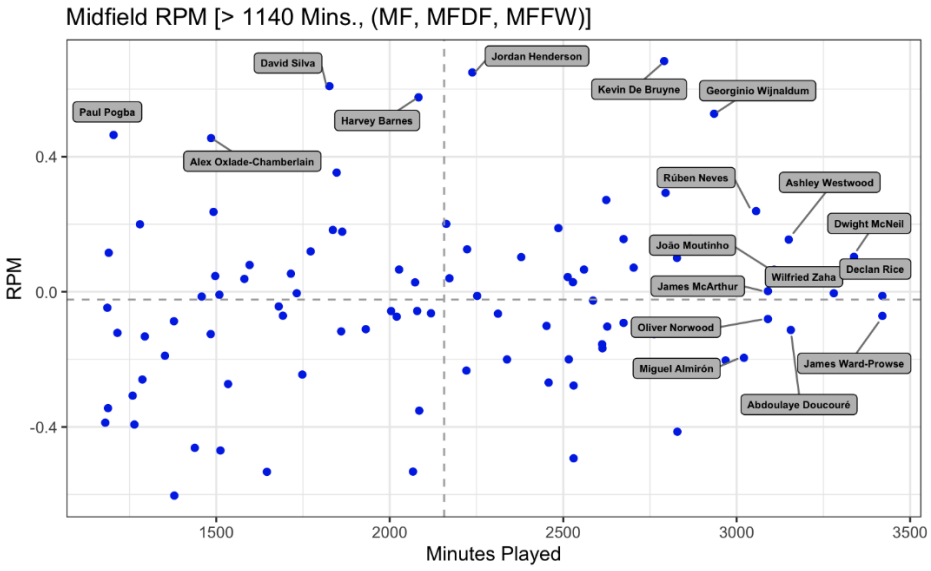

Figure 5: RPM of Midfield Players

Next, we looked at the midfield players of the Premier League. These positions are labeled MF (midfielder), MFDF (midfield-defender), and MFFW (midfield-forward). Instead of using attacking RPM as before, total RPM (attacking + defending) was used as midfield players are usually required to contribute both offensively and defensively during the span of the match. By using the same process as before, we can identify Kevin De Bruyne, Jordan Henderson, and Georgino Wijnaldum as the most effective midfielders. Once again, these results are promising as De Bruyne is widely regarded as the best midfielder in the Premier League and arguably in the world.

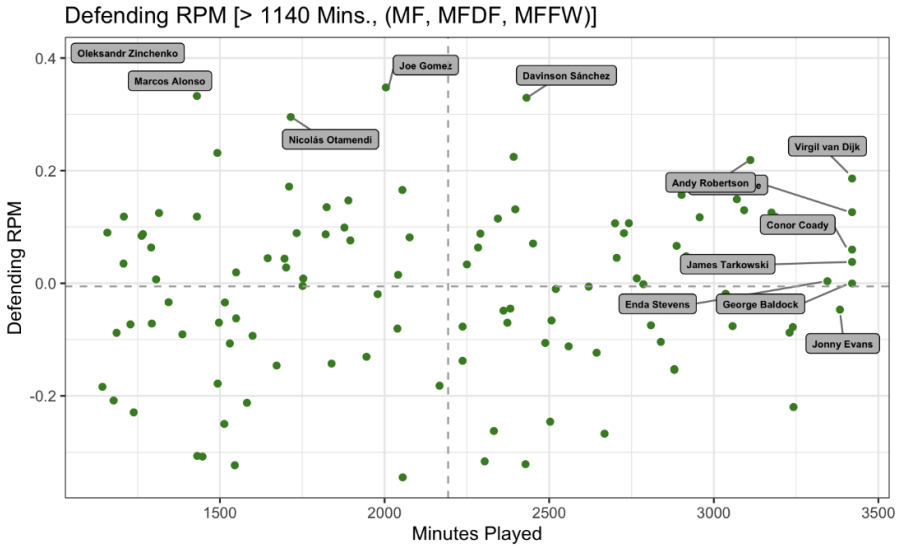

Figure 6: Defensive RPM of Defensive Players

Our next analysis was that of the defensive Premier League players. Positions were labeled as DF (defender), DFMF (defensive-midfielder), and DFFW (defensive-forward). In this case, defensive RPM was used as it most effectively values the players used in defensive positions. Andy Robertson, Virgil Van Dijk, and Davinson Sanchez were amongst the most effective according to their defensive RPM output.

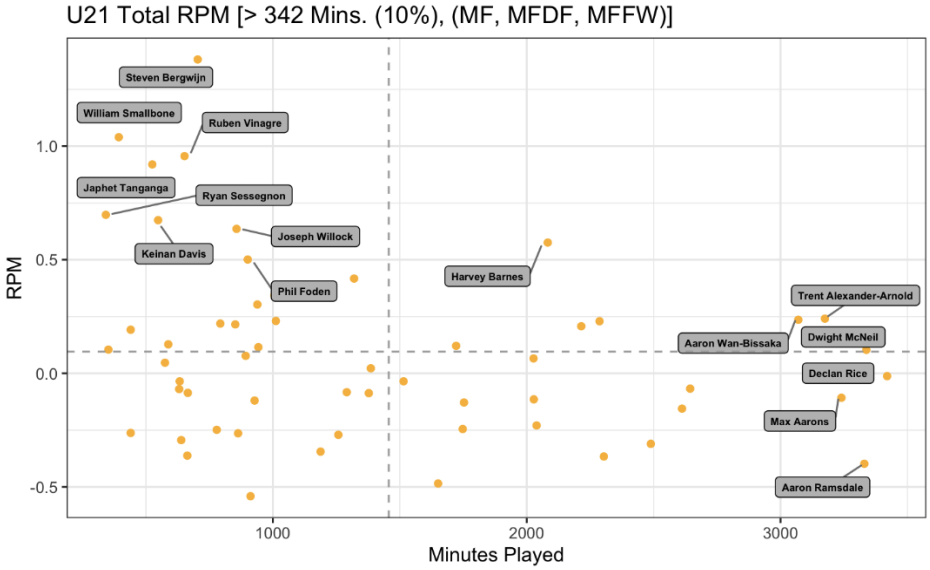

Figure 7: RPM of U21 Players

Next, as touched on before, a look at the top U21 performers during the 2019-20 Premier League season. Harvey Barnes, Trent Alexander-Arnold, and Aaron Wan-Bassaka were amongst the best youngsters in the league. This output gives us a great look at which young prospects are breaking through, especially those with a high RPM and minutes played, as it means they are performing consistently well across a large sample. It is also important to note that many of these players have not played as much as the players in the earlier visualizations, so one must keep in mind that a slight sample size bias may either improperly inflate or deflate their RPM statistic.

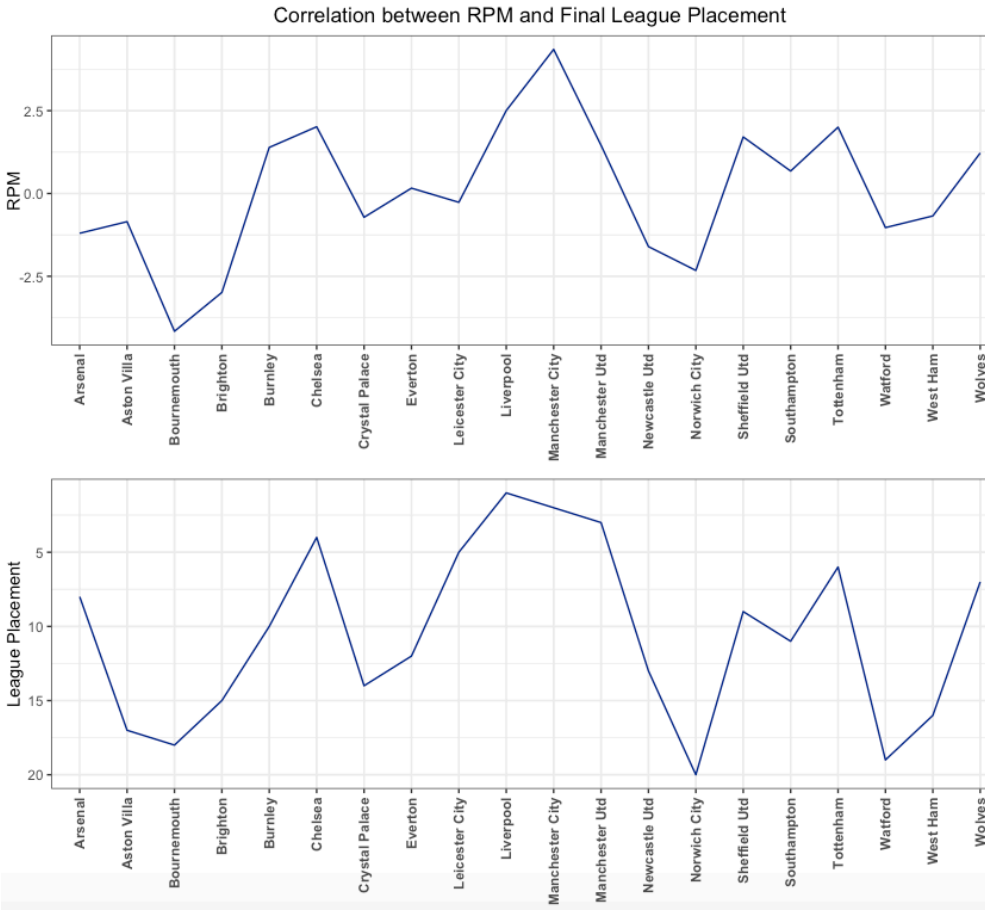

Figure 8: Correlation Between League Finish & Total RPM of Clubs

We expect a high league finish (1st, 2nd, 3rd), to correlate with the highest team RPM values. As shown above, there is indeed a relationship between the two. It is most notably shown between Liverpool and Manchester City, where the RPM line graph spikes and plateaus simultaneously with their 1st and 2nd league finishes. Subsequently, some of the worst clubs in terms of RPM, Bournemouth and Norwich City, finished in the bottom three of the premier league. This confirmation of a relationship during the 2019/20 season is extremely important as it strongly verifies the validity of our RPM statistics which accurately depicts which clubs and players are performing the best. Only players who met the minutes played minimum were included in the team totals as to not skew the results.

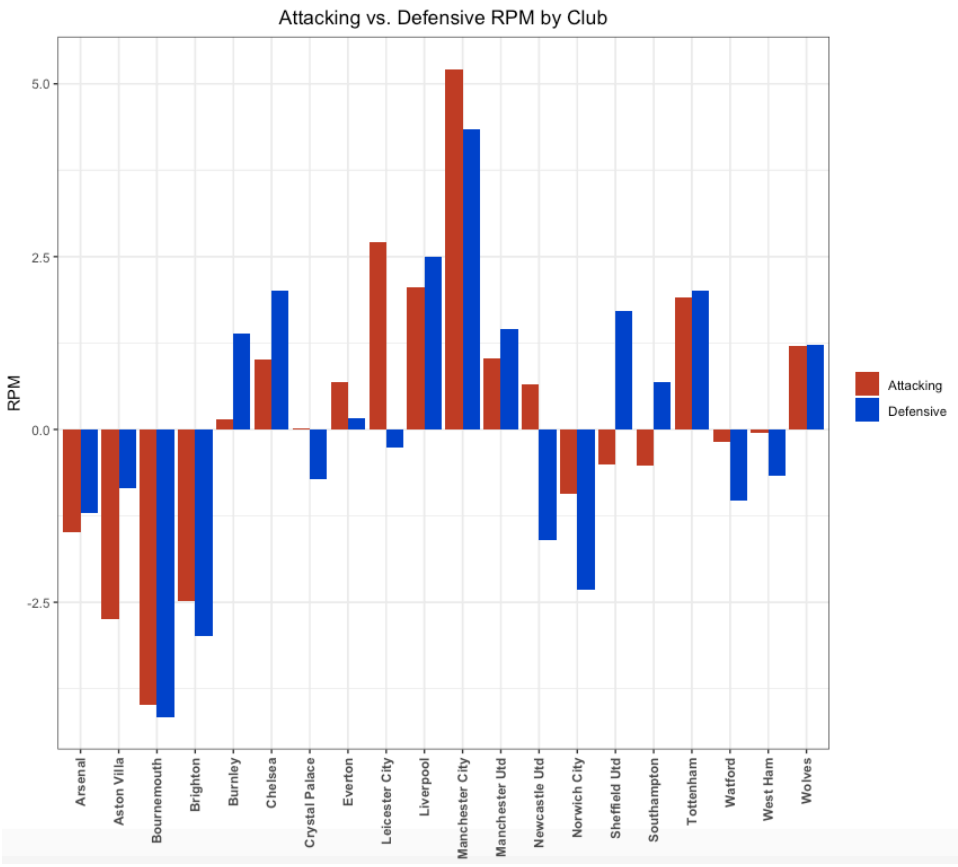

Figure 9: Difference Between Attacking & Defending RPM of Club

In a further breakdown of how the clubs performed individually, we take a look at the attacking and defending RPM of each, instead of total RPM. Manchester City were surprisingly elite defensively, and their league best attack was not enough to surpass Liverpool in clinching the title. On the other hand, all three relegated teams (Bournemouth, Watford, and Norwich City.) were by far some of the worst defensively, which undoubtedly can be attributed to their plummets to the Championship.

Application

Our Advanced RPM statistic brings a variety of analysis and application options to the table that have not been capitalized upon before. A principal use of Advanced RPM can be within talent scouting and acquisition, as it offers a new variable that integrates several important and often undiscussed measures of player performance. It is not surprising to see the Advanced RPM statistic identify some of the star players that most fans are familiar with; it is surprising (and extremely useful) when Advanced RPM highlights players that are not recognized by some of the more basic stats like goals, assists, shots, etc. This can allow managers and coaches to scout younger or overshadowed players that are not getting the recognition they deserve. This application can be measured with respect to salary or market value, and thus be used to see which players are under or overperforming. This can aid managers in determining lengths of player contracts as well as whether or not to sign new players.

An additional key application of the Advanced RPM statistic concerns lineup analysis. With this statistic, coaches will be able to determine which players are playing best with one another, and thus determine what lineups will be most ideal in any given situation. As more and more substitutions are being allowed, Advanced RPM proves more valuable in game-time decisions. Finally, our Advanced RPM statistic is not

limited to just the EPL; it's utility can be applied to soccer leagues and teams around the world. It is often incredibly difficult to decide which players deserve the call up to represent their country, so Advanced RPM can assist national team managers in making these unnecessarily stressful decisions.

Possible Improvements

The weights we applied to each metric could have been worked on and fine-tuned to be slightly more accurate, as most were taken from previous research. A possible improvement that may be possible with better data would be incorporating some more advanced metrics such as expected goals. Furthermore, the difference in play style between goalkeepers and other positions may limit the metric's possible goalkeeper analysis.

A previous edition of this research for the 2017/2018 Premier League season included manual data collection, whereas this new project involved automating the commentary and lineup scrapes so that the regressions can be reproduced either on a yearly basis or with completely different leagues.

Conclusion

Overall, our RPM stat produced very promising results. As stated above, our RPM rankings generally followed what an avid and knowledgeable Premier League fan would traditionally think with players such as Virgil van Dijk and Kevin De Bruyne near the top of their respective player groups. Moreover, we strongly believe the comparison between a club's RPM and their final position on the table strongly validates this metric. We believe our RPM stat is a valuable metric that can be used by leagues and teams everywhere for various purposes, and hopefully further research will be done to increase the accuracy of the metric even more with the inclusion of more advanced statistic variables.

Bibliography

2017-2018 Premier League Player Stats. (n.d.). Retrieved July 02, 2020, from https://fbref.com/en/comps/9/1631/stats/2017-2018-Premier-League-Stats

Caley, M. (2014, June 17). What is a corner kick worth in soccer? Retrieved July 02, 2020, from https://www.washingtonpost.com/news/fancy-stats/wp/2014/06/17/what-is-a-corner-kick-worth-in-soccer/

Hobbs, J., Power, P., Sha, L., Ruiz, H., & Lucey, P. (2018). MIT Sloan Sports Analytics *Conference* [Scholarly project]. Retrieved 2020, from http://www.sloansportsconference.com/wp-content/uploads/2018/02/2008.pdf

How many goals are scored from free kicks? (n.d.). Retrieved July 02, 2020, from https://www.bettingwell.com/sports-betting-guide/football-bettors-guide/how-many-goals-are-scored-free-kicks

Moritz, S. (2020, March 17). Ridge.pdf. Retrieved July 02, 2020, from https://cran.r-project.org/web/packages/ridge/ridge.pdf

Sakellaris, D. (2018, June 19). The Correlation Analysis of Scored Goals and Red Cards. Retrieved July 02, 2020, from https://statathlon.com/the-correlation-analysis-of-scored-goals-and-red-cards/

Suchman, A. (2014, April 19). Explaining ESPN's Real Plus-Minus. Retrieved July 02, 2020, from https://cornerthreehoops.wordpress.com/2014/04/17/explaining-espns-real-plus-minus/