Using Random Walks to Simulate Premier League Seasons

Andrew Nober, John Rho

May 8, 2023

1 Introduction

The English Premier League (EPL) operates on a simple premise. The 20 teams in the League play each other twice, once at home and once away, in a double round robin format. Teams play all 19 other teams at least once before playing any other team twice. At the end of the season, the team with the most points wins the league championship. While the EPL has a few teams that consistently place high in the final table, it still varies from year to year. It is difficult to predict exactly how many points a team will finish with at the end of the season.

Points are awarded based on the outcome of games between two teams. If a team wins, they earn 3 points to their season total. A tie earns a team 1 point. A loss earns a team nothing. There are 38 matchweeks in a season (a matchweek is one round of games). Ties between teams with the same number of points at the end of the season are broken by goal difference, which is the number of goals a team scored during the season minus the goals they conceded. A team's final points total at the end of the season is given by the formula:

$$P = 3W + T + 0L \text{ where } W + T + L = 38$$

A team can earn a maximum of 114 points (38 games x 3 points per win), a minimum of 0 points (38 losses), and almost any number of points in between.

At the end of the 2017-2018 season (competition typically runs August to May), Manchester City was crowned champions with 100 points, the most any team had ever earned. The next year, the competition was extremely close. Manchester City won again with 98 points and second-place Liverpool FC earned 97 points. These teams were clearly exceptional, showing some of the most dominant performances in the history of the league. **However, how "exceptional" were these teams statistically? Can their performance be explained as simply the outcome of a random process, or are they dependent on more complicated factors like the team's recent success or inherent strength?**

Most soccer leagues compete over an entire year with the winner of the league's championship being decided by who has the most points at the end of the year. In England, teams in the Premier League play 38 games throughout the season (one home and one away game against each of the other 19 teams) and earn 3 points for a win, 1 point for a draw, and 0 points for a loss. Fans often speculate about how "good" a team is in the absolute, since players change from season to season. The point system, however, gives us a quantitative tool that can help us determine how "good" a team is. Hence, we hope to use random walks to help explain these real-world distributions.

2 Modeling the Premier League

2.1 Overview

We will run three different types of random walks, each averaged over several trials. The outcomes will be compared to the last 30 seasons of the Premier League, which is a soccer league in England. These are the proposed forms and their motivations:

- Pure Random Walk: this is our control, simulating if every team was the same
- Sliding Window Random Walk: this factors for the team's recent performance
- Team-Strength Random Walk: this factors for pre-existing differences in team strength/talent

For each match, there are three possible outcomes: team Home wins (A: +3, H: +0), the Away team wins (A: +0, H: +3), or both teams tie (A: +1, H: +1). We model over 38 matches for each team in a double round-robin format (each team plays every other team once at home and away). There are 20 teams overall.

2.2 Constant p win-probability function (pure random walk)

In this scenario, the outcome probability function P is a constant

$$<\textit{(home win prob)}, \textit{(away win prob)}, \textit{(draw prob)}> = <\frac{1}{3}, \frac{1}{3}, \frac{1}{3}>.$$

With this model, every single game is model as an independent event, and previous performance has no impact on the current game. For this reason, teams have a constant probability of $\frac{1}{3}$ to win, lose, or tie a game. In terms of the model, this is the case where $R = 0$, meaning that a better performance over the previous 3 games has **no** impact on the current game. This model is the simplest case, and it models the Premier League as a true random walk.

2.3 Sliding window p win-probability function

This model was a modification of the pure random walk. In this case, we adjusted the probability of winning based on the performance of the last three games. We likened this simulation to having "hot streaks," in which teams tend to go on winning/losing streaks. Note that we let the game run as a pure random walk for the first 18 matches for each team, and then we activated the sliding window. This was because the sliding window effect would create near-linear relationships since teams that randomly initialize to start at a certain strength would remain as such. Moreover, we implemented a level of noise so that the model would not be fixated at a certain level of performance, which we likened to a hot streak being broken in a game.

Hence, we define the following relationships to define the home/away win and draw probabilities:

$$\begin{aligned}
\textit{home advantage} &= 0 \\
\textit{home record} &= \frac{\sum_i (\textit{last three home})}{\textit{len(last three home)} * 3} \\
\textit{away record} &= \frac{\sum_i (\textit{last three away})}{\textit{len(last three away)} * 3} \\
\textit{record difference} &= \textit{(home record)} - \textit{(away record)} \\
\textit{home win prob} &= \frac{1}{1 + \exp\left(-(\textit{record difference}) + (\textit{home advantage})\right)} \\
\textit{away win prob} &= \frac{1}{1 + \exp\left((\textit{record difference}) + (\textit{home advantage})\right)} \\
\textit{draw prob} &= 1 - \textit{(home win prob)} - \textit{(away win prob)}
\end{aligned}$$

We modeled the probabilities by scaling them using a sigmoid, which ensured that the values were between 0 and 1. Note that record difference can have a minimum value of 0 and a maximum value of 9. Teams can earn anywhere from 0 to 9 points in 3 matches, excluding 8 points, which a team cannot earn in 3 matches. Our model preserves the idea that it is not how good a team is playing in the absolute but how strong their hot streak is compared to their opponent. Note, to avoid linearity in the simulations, we also added random noise to the win probabilities as follows:

$$\begin{aligned}
\textit{noise} &= \mathcal{N}(0, 0.6) \\
\textit{home win prob} &= \textit{(home win prob)} + (\textit{noise})/2 \\
\textit{away win prob} &= \textit{(away win prob)} + (\textit{noise})/2 \\
\textit{draw prob} &= \textit{(draw prob)} - \textit{(noise)}
\end{aligned}$$

Furthermore, a *home advantage* variable is defined as 0.5, indicating there is not an advantage. This is simply the baseline for the sigmoid input.

2.4 Dependent p win-probability using goal-scoring distribution and team strength estimates

Our third model involved a more complex approach. The objective was to determine if a factor such as inherent team strength was a better determinant of team performance. We compared weighted distributions of the number of goals scored for each team, which would give a more comprehensive simulation of a team's firepower. These distributions were determined by preloaded team strengths, which mirrored the distribution of the Premier League.

Importantly, we calculated preloaded strength of the teams. In the real world, there are only a select few teams that are considered "juggernauts" that are much stronger than their competition, largely due to inherently more talented rosters. Hence, to represent this, we used the Gamma distribution, which has positive skewness (showing that there are a few outlier teams that are very good). We implemented these changes to produce a model that is more accurate to the real-world data.

Hence, we develop the following relationships:

$$
\begin{gathered}
\textit{team strengths}_{\textit{team}} = \sqrt{2} \times \Gamma(\textit{shape}, \textit{scale}) \\
\textit{home advantage} = 0.5 \\
\textit{strength diff} = ((\textit{home strength}) \times (\textit{home record})) - ((\textit{away strength}) \times (\textit{away record})) \\
\textit{noise} = \mathcal{N}(0, 0.5) \\
\textit{strength diff} = (\textit{strength diff}) + (\textit{noise}) \\
\textit{home\_win\_prob} = \frac{1}{1 + \exp\left(-(\textit{strength diff}) + (\textit{home advantage})\right)} \\
\textit{away\_win\_prob} = \frac{1}{1 + \exp\left((\textit{strength diff}) + (\textit{home advantage})\right)} \\
\textit{draw\_prob} = 1 - (\textit{home win prob}) - (\textit{away win prob})
\end{gathered}
$$

Note that we initialized *shape* and *scale* to be 20 and 0.05, respectively (these were determined with experimentation as good values to increase right-skewedness). Like above, we also scale the probabilities using a sigmoid in order to guarantee values between 0 and 1.

3 Codebase

The full codebase can be found at the following Google Colab notebook. The codebase is thoroughly documented. Some cells may not run independently without running prior cells. To ensure that all the code runs as intended, make sure to run the notebook from start to finish to generate all intended outputs. Exact outputs may very based on differences in random seed or local optima.

4 Results and Analysis

To simulate 30 seasons of play (with 38 matches played by 20 teams each), we created a for-loop that generated a list of all 20 teams' final score for all 30 seasons (giving 600 values in total). We did this for each of the five scenarios, including the real-world data. Then, we plotted the distributions to overlap each other and calculated mean, median, standard deviation, and kurtosis for each of the simulations' distributions.

Our objective is to compare each method with the actual data. We will use the calculated statistics to guide our analyses of the best "explanations" behind the outcomes we observe.

4.1 Overall Distribution Statistics

Table 1: Points Distribution

Data Source	Mean	Median	Std	Kurtosis
Real-World	51.137	48.0	14.867	0.8036
(1) Pure Random	50.643	51.0	7.754	0.0937
(2) Sliding Window	52.202	50.0	19.679	-1.000
(3) Team Strength	52.785	52.0	12.861	0.0689

To match the real-world data, the simulations above was calculated over 30 seasons each. Note that by nature of the random walk, we may see discrepancies each time we run the program.

Note that all of the simulations model the centrality (mean/median) and spread (std) relatively accurately. The pure random, expectedly given its symmetry of win probabilities, does not model the spread as well, but interestingly, its kurtosis values are closer to the real-world data. The Team Strength model tends to do the best overall, with the closest spread to the real-world data as well as the second-closest kurtosis.

The spread calculates how spread-out the team success is. Moreover, kurtosis models the presence of extreme values, particularly teams that are incredibly bad or exceptionally good. It is interesting to note that the sliding window approach has negative kurtosis. Hence we can interpret the data as suggesting that the performance of extremely good (or bad) teams are not dictated as much by hot streaks. Rather, inherent team strength seems to be a better measure of explaining why these "exceptional" teams perform the way they do.

4.2 Real-World Data

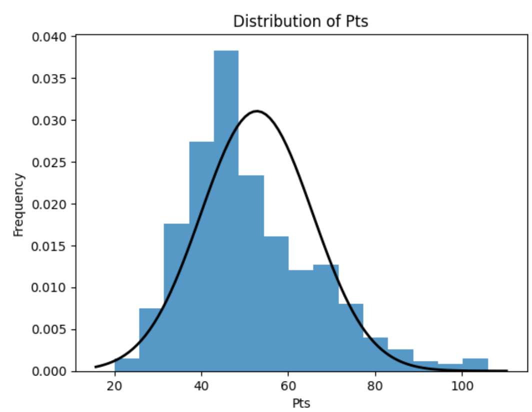

Figure 1: Distribution of Pts for Real-World Data (30 Seasons, 20 teams each with 38 Matches)

Note that the distribution of points is skewed right. This is because there are only a few teams that are "juggernauts," which can be attributed to different factors such as inherent team strength, hot winning streaks, among others. However, the right-skwedness of the distribution indicates that these "juggernauts" are rare or constitute a small portion of the teams. It also suggests that the best "juggernaut" teams are significantly better than the average team. Interestingly, the same characteristics have different incidences among worse teams. Inherent team weakness and cold losing streaks are more consistent than their successful counterparts.

4.3 Method 1: Pure Random Walk

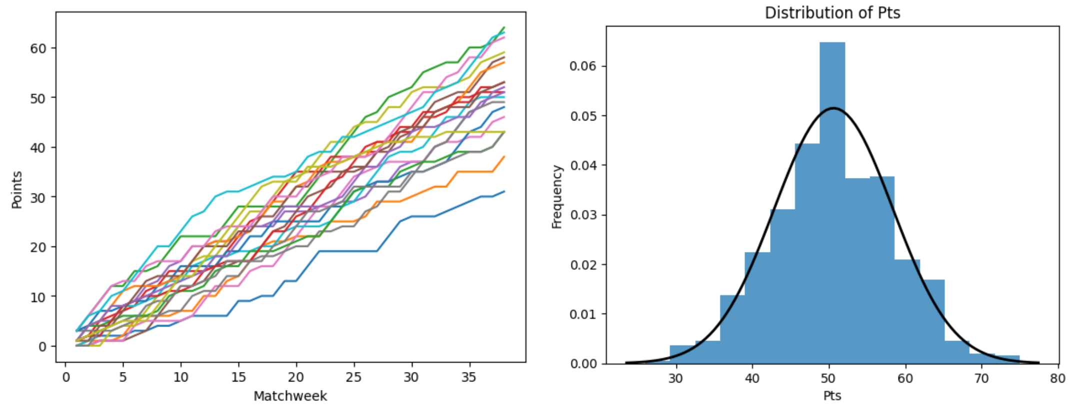

Figure 2: **Left**: Single season simulation || **Right**: (30 Seasons, 20 teams each with 38 Matches)

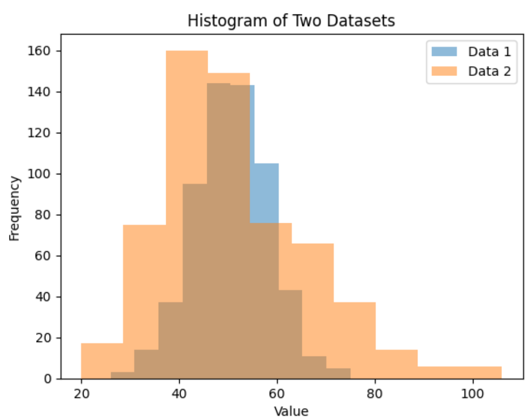

Figure 3: Comparison of Pure Random Walk Simulation (Data 1) to Real-World Distribution (Data 2) (10 bins)

The pure simple random walk method produces a symmetrical and bell-shaped distribution of final points. The mean final points (analogous to the mean final position of a walker) of teams theoretically falls at the season expected points value. For the simplest random walk model, the expected value of any one match is

$$\text{Points} = \sum_{i=1}^{38} 3 \times [\text{Team win probability}] + 1 \times [\text{Tie probability}] + 0 \times [\text{Lose probability}]$$

Since the teams have a constant win, tie, and lose probability of $\frac{1}{3}$, the expected value of the entire season should be 49.4 (or 50 because teams cannot attain fractional points). The simulated center of the distribution falls at 50.643, meaning the model closely approximates the mean performance of a team. However, this model shows that the influence of teams sliding window and relative strength must be high because it does a poor job at modeling the best and worst teams. Our model's standard deviation and kurtosis show this problem. The real world data has a standard deviation nearly double that of the simple random walk model, indicating that team quality has a high variance. The simple random walk model especially under-performs at explaining the very best teams and the real-world data's lengthy right tail.

4.4 Method 2: Sliding Window-Weighted Random Walk

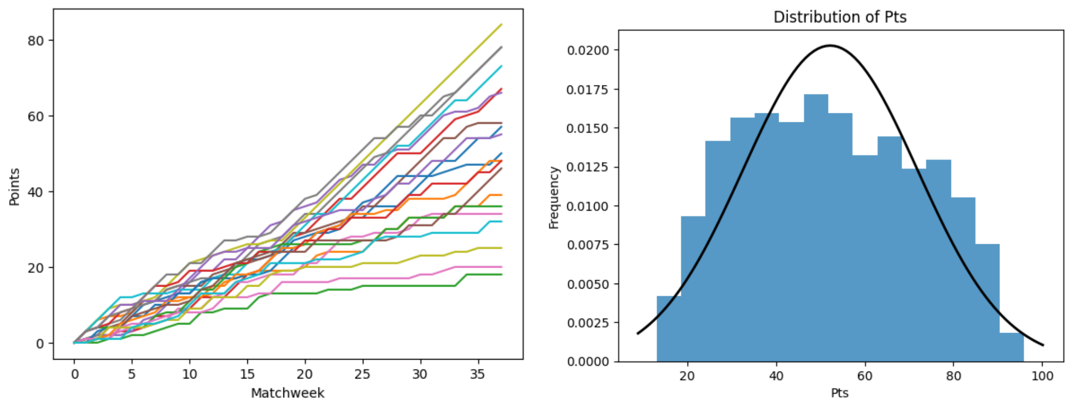

Figure 4: **Left**: Single season simulation || **Right**: Distribution of Pts for Sliding Window Simulation (30 Seasons, 20 teams each with 38 Matches)

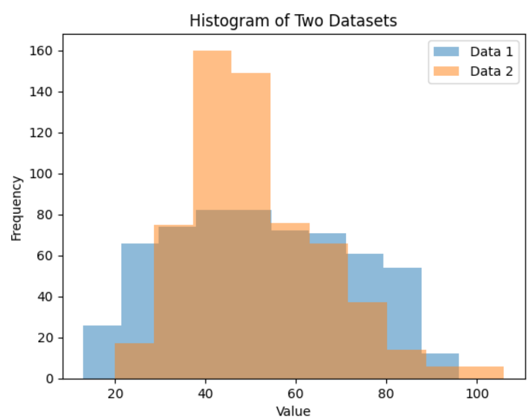

Figure 5: Comparison of Sliding Window Simulation (Data 1) to Real-World Distribution (Data 2) (10 bins)

As shown in the data, the representation was able to capture comparable values of centrality (mean/median approximately 50). Similarly, the spread (measured by std deviation) was also comparable, with the real world having 14.867 whereas the sliding window simulation had 19.679. Hence, the simulation is able to capture the major features of the real-world distribution, particularly compared to a pure random walk. Hence, we can see that the influence of "hot streaks" allows us to model more closely to the real-world data, suggesting that hot streaks (window size 3) do indeed impact teams' success in the real world.

Note, however, that the simulation did not capture the right-skewness of the real-world distribution. Thus, the teams, as the single season simulation suggests, tend to have final pointages that are vastly spread apart. We can see this in the histogram as well, with its more uniform shape compared to the real-world distribution. Note that we programmed the sliding window to take into effect at Matchweek 18, which then tends to continue pre-existing trends (with some noise) until the final Matchweek.

Note that the sliding window, in its naive form without added noise, is actually a relatively poor representation. The season simulations become linear based on their initialization of performance because the teams continue to perform based on precedent. This is not really representative of real-world data. Hence, we introduce random noise to the simulation to overcome this problem and produce the results above.

4.5 Method 3: Team Strength-Weighted Random Walk

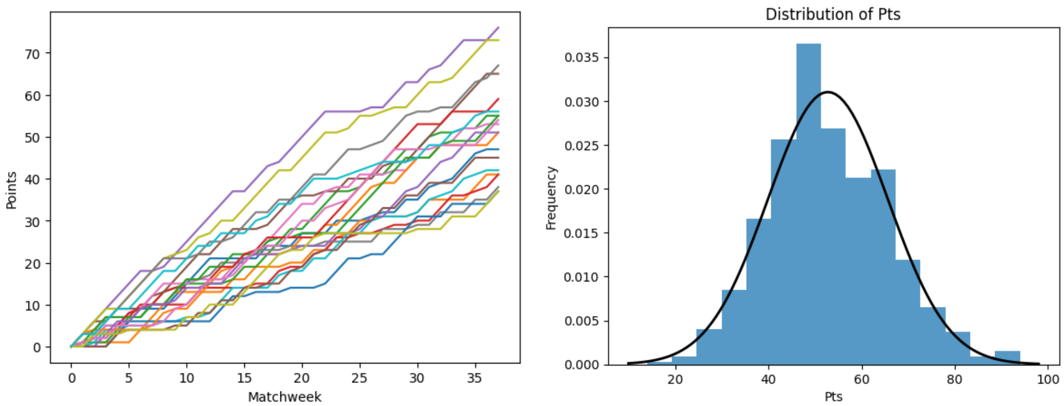

Figure 6: **Left**: Single season simulation || **Right**: Distribution of Pts for Team Strength Simulation (30 Seasons, 20 teams each with 38 Matches)

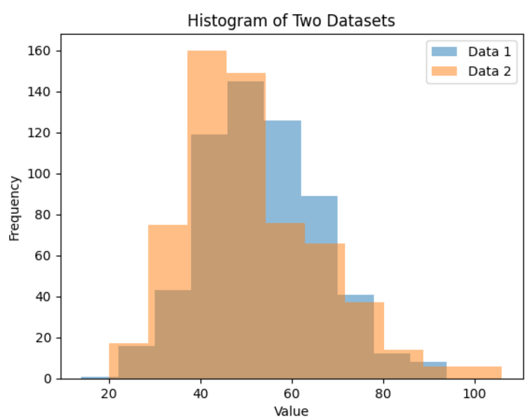

Figure 7: Comparison of Team Strength Simulation (Data 1) to Real-World Distribution (Data 2) (10 bins)

This model built upon the pure random walk by introducing predefined team-strength values that influenced the win probability. Note that compared to the sliding window method, the data has a lower spread that more closely models the real-world data. However, the graph is not as right-skew as the real-world distribution, suggesting that the Gamma distribution we selected, though in the right direction of skewing the data to the right, still underestimates how exceptional the top performing teams are. This suggests that the top-end teams in the real-world distribution are leagues above the average team; moreover, the real-world distribution has many more ”stinkers” of teams that are exceptionally bad.

5 Project Conclusions

5.1 Overview:

The data suggest that it is possible to use a random walk to approximately simulate the performance of a Premier League season. Moreover, by weighting the random walk probabilities using specific features (like sliding windows or team strength), we also modeled the influence of these factors compared to the real world distribution. Specifically, **inherent differences in team strength**, rather than hot streaks, appear to model the exceptionalism of "juggernaut" teams in the real-world data. This was modeled in comparison to a pure random walk.

Moreover, it is important to note that 30 seasons is a relatively small sample size. Even within our own testing, there was variance with the exact simulations (given the random nature). Combining other soccer league data (and normalizing the points system) could help us get a better picture of these random walk-esque relationships. Similarly, we could also run more 30-season simulations and average these outcomes. In terms of how this contributes to our understanding of what phenomenon are behind the performance of extremely good teams, our models demonstrate that the Premier League is a league of unequal teams. The EPL does not start each season at the beginning of a simple random walk. For the average team, the simple random walk model may be what they experience, but the worst and best teams experience a much more complicated season. Modeling teams performance based on their short run success additionally does not accommodate for the variance and distribution of the actual Premier League. Our third model suggests that a team's inherent quality, determined by the skill of its players, the depth of its squad, and their manager, can often outweigh a worse team's hot streak.

5.2 Future Directions:

In the future, possible improvements include using a form of regression (or a more complex model, like a neural network trained on the game data). This could allow us to better determine the impact of specific factors like team strength, roster construction, etc. compared to a pure random walk. Moreover, we could also diversify our data source to include other soccer leagues, or we can perform the analysis for other sports as well (American football, basketball, etc.) to compare the different impacts of these simulated factors. There is also the interesting question of whether the Premier League models similarly to other leagues that operate in the same format. The English Championship (a second tier league similar to AAA for baseball), Spanish La Liga, Italian Serie A, and German Bundesliga are all major leagues that use the same format (some have a different number of teams). While the Premier League depends on the inherent strength of teams, this could be due to the particular financial and historical structures of the League. Other Leagues may approximate other models better.

References

[1] Guillermo Abramson, Nicolas E Boccara, and Vincent D Blondel. Dynamics of tournaments: the soccer case. a random walk approach modeling soccer leagues. *Journal of Statistical Physics*, 157(5):977–989, 2014.

[2] Evan Gower. English premier league standings. https://www.kaggle.com/evangrower/english-premier-league-standings, 2021. Accessed: May 7, 2023.

[3] National Institute of Standards and Technology. Exploratory Data Analysis (EDA) – 1.5 Graphical Methods for Describing Data, 2003.