Streaks of wins and losses

Andrew Chan Alex Church Dan Bouchard James Templeman Yiyoung Kim

November 6, 2015

Abstract

Professional sport provides big business opportunities. Wins and losses of a team or player in major sporting events affects not only players, coaches, team owners, sponsors, but also sports betting companies, local economies, and sometimes even the global economy! Therefore, successful predictions of the outcome would benefit a huge amount of stakeholders. This report attempts to predict the outcome of a sporting event by breaking down data from the barclays premier league into three hypothesis

1. What extent does previous scores impact future results and can we use the average goals scored by each team to predict the future scores
2. What extent does previous results impact on the results of latter games
3. To what extent does the total cost of a squad affect the performance of a team can we use this to simulate the likely win/draw/loss streaks of a season

The first hypothesis looks into whether either scores or results can be accurately predicted throughout a whole season 2014-15, using the data from the 2012-13 and 2013-14 seasons. Analysis of the number of goals that each team scored and conceded both home and away provided an understanding of the results. Predictions were then made of how many goals an individual team would most probably score against another and therefore gave the most likely score in a given match.

The second hypothesis looks into whether last games' results impact on latter results in a statistical viewpoint. The last games' results are collected for 16 years. Both of predictions and probabilities are based on the results ratio over whole 16 years of results.

The third hypothesis questions whether we can predict the outcome of a match based on the squads cost. By making a team's own 'Elo rating' based solely on the squad cost the chances of outcomes can be created and simulated which correlate strongly with the actual win/loss/draw streaks of the premier league.

Contents

1 Introduction

It is very difficult to try and attempt to guess the outcome of a sporting match, one of the reasons betting is so popular. However confining ourselves to team sports, it is obvious that strong teams are likely to win and weak teams are not. Nevertheless, even a strong team may not be constantly winning. The performance of a team would be subject to perpetual waxes and wanes over a season or different seasons, influenced by moods, managing emotions and handling stress - all part of human behaviour. The aim of this report is to find patterns in sports records and exploit them for successful predictions. There are many aspects within sport to access when trying to predict the outcome of a game or match, things like the weather, player behaviours, spectators, individual ability, luck. The factors affecting results and win or loss streaks in football include: recent form, comparing average and team goals, team strength and league position. The report was separated into three hypotheses which look into a range of these factors both short and long term to see if it is possible to predict both individual matches and the win and loss streaks of teams.

2 Literature Review

Everyday millions of people try to predict the outcome of a sporting event whether that is two children arguing over what the score will be in their football match or people betting on the results in the world cup. There has been lots of research done around this subject looking into hundreds of different variables they may affect a game especially with regards the barclays premier league. Multiple big corporations have statistics on the outcome of games and the details involved including players. However no corporation tries to predict the outcome of a game, even though a TV pundit may hazard a guess now and again. The studies that we have looked at do not directly look into the outcome of barclays premier league football matches howbeit the way research is collected in their respective reports as well as the way they have come to their conclusions about games and players is of interest, so they will not be reviewed in depth however will be referred to when appropriate.

When considering the variables that affect a sporting event, we had certain areas of particular interest. ‘Understanding baseball team standings and streaks’ [5] was an important article to analyse as it gave insight to the effects of previous results on baseball teams. The report conducted by C.Sire and S.Redner uses the Bradley-terry model and the rank independence of the average wins and losses in major league baseball to form predictions on the outcome of games. This proved useful in two of our hypothesis. The second hypothesis studies the effects of win/loss/draw streaks and the third uses a team ranking system to make its predictions (even though the rating is not based on the results of previous games). The report found that a uniformly distributed rank of teams from a value of 0.44 up to 1 gave the best match for the win and loss streak records. It also mentioned how seasons were not quite long enough to accurately predict individual win and loss streaks but if the model was simulated for seasons with more games, the win/loss data converges to the prediction of the equation. Even though this report was not relevant to football or the barclays premier league, it gave relevance to a team sport and how history relates to their current form.

Further to this, the impact of individual players performances are discussed in the review “twenty years of hot hand research review and critique” [6]. The report looks into whether win streaks are due to player confidence and that feeling of being ‘on a roll’ or if it is just chance such as a streak of many heads in coin tosses. It seems to come to a conclusion that streaks tend to be chance, so the likelihood of scoring is largely independent of players prior performance and streaks represent random accidents that are unlikely to continue. It did also say that stronger teams tend to have more frequent win streaks but the times that these streaks occur is still random and hard to predict. Applying this to teams in football and the third hypothesis, we can simulate the win/draw/loss results of season using the strength of the teams as the total squad cost with the current streaks not having an effect on the next result - so the streaks are random as the article concludes. This means that if a team has won 3 times in a row it is still equally likely to win as before.

3 Methods

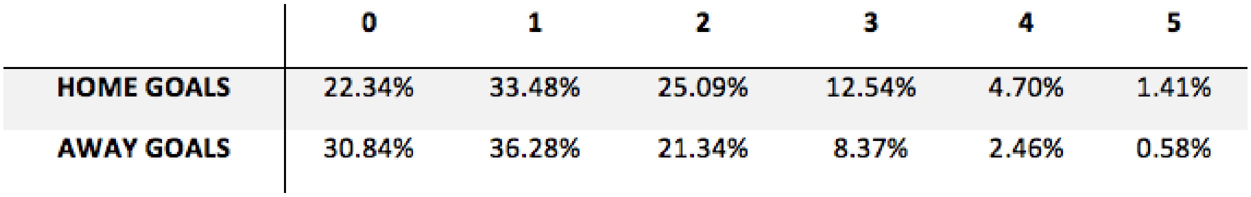
3.1 Hypothesis One

Data collection:
Data was collected from the 2012-13 and 2013-14 English Premier league seasons in an attempt to predict scores in the 2014-15 season. The initial aim was to look into whether there was any pattern in how many goals were scored in individual matches and if the total goals could be predicted. As the data was analysed it became clear that the home team tended to score more goals than the away team on a consistent basis. Looking into the individual goals scored both home and away for every match and taking the average gave 1.5 for the the mean number of goals scored at home and 1.18 for the mean scored away from home. It can be seen from this that the most likely amount of goals scored from each team would be either 1 or 2 but the best way to show this would be to use the poisson distribution. The poisson distribution takes in a mean value and calculates the probability of individual discrete value of goals (which is what is required in this case). So it seemed the best probability distribution because information on the amount of goals in a set period of time (one match) was available from a large data set and individual probabilities could be easily calculated. Binomial distribution could not be used because the exact probability of goal was not known, the event of a goal happening or not happening was not independent and the amount of opportunities were not the same for every match. The poisson distribution equation is given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{1}$$

Where x = number of goals
λ = mean

So the individual probability of home goals and away goals using the above mean values gives the below distribution:



	0	1	2	3	4	5
HOME GOALS	22.34%	33.48%	25.09%	12.54%	4.70%	1.41%
AWAY GOALS	30.84%	36.28%	21.34%	8.37%	2.46%	0.58%

Figure 1

The poisson distribution seems to be a good fit to the discrete values of how many goals a team scores because it is very likely that one team will score 2 goals or less and less likely more than this. However, this distribution is not unique to specific matches and cannot be applied to all matches so there needed to be a way to produce lambda values (mean, μ) that are similar in value to the current home and away goals distribution but a better fit to the actual matches. The poisson distribution is still a good approach but needs to be modified slightly to achieve more reliable predictions.

A recent article (Ratcliffe, 2014) gave a way of using the poisson distribution to predict the outcome in sporting matches (also specifically the premier league) and gave a method in making the lambda values specific to each individual match. The idea was to assign home and away, attack and defence coefficients so each team was given 4 coefficients and the value of these coefficients was based on recent data, the data being used in this case is the 2012-13 and 2013-14 seasons. The equations for average home and away goals were given as follows:

Team x -home team

Team y -away team

$$\begin{aligned}\text{Team x goals} &= \text{average home goals} \times \text{home attack coefficient} \\ &\qquad \times \text{away defence coefficient}\end{aligned} \tag{2}$$

$$\begin{aligned}\text{Team y goals} &= \text{average away goals} \times \text{away attack coefficient} \\ &\qquad \times \text{home defence coefficient}\end{aligned} \tag{3}$$

So if two equally average teams played each other all the coefficients would equal 1 and the score would have a distribution as above. So each of the coefficients had to be relative to each other and the way to do this is to calculate the attack coefficients per team as:

$$\text{Home attack coefficient} = \frac{\text{average goals scored at home}}{\text{league average home goals scored}} \tag{4}$$

$$\text{Away attack coefficient} = \frac{\text{average goals scored away}}{\text{league average away goals scored}} \tag{5}$$

The defence coefficients were calculate in a similar way but the amount home goals conceded is the same as the amount of away goals scored, so the defence coefficients were:

$$\text{Home defence coefficient} = \frac{\text{average goals conceded at home}}{\text{league average away goals scored}} \tag{6}$$

$$\text{Away defence coefficient} = \frac{\text{average goals conceded away}}{\text{league average home goals scored}} \tag{7}$$

Then using these coefficients calculated for each team and equations “2” and “3” the lambda values (mean home and away goals) can be calculated for each match and their value varies slightly from the overall average home or away goals making them specific to each match. From the coefficients it can be seen that strong teams have high attack coefficients and low defence coefficients. To run this model for a whole season of fixtures, a python program file was used that outputted all the lambda values for each match along with the score prediction. The python program inputted all of this information into an excel spreadsheet and this spreadsheet could be used to compare actual and predicted results. One problem was that there was no data on Leicester and Burnley so those matches had to be ignored as we could not attempt to predict those matches.

3.1.1 Results

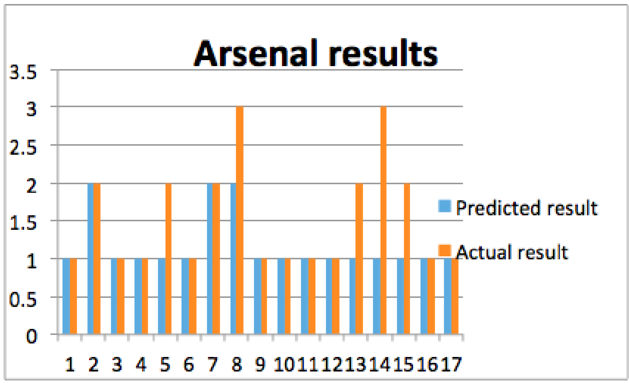


Figure 2: A figure

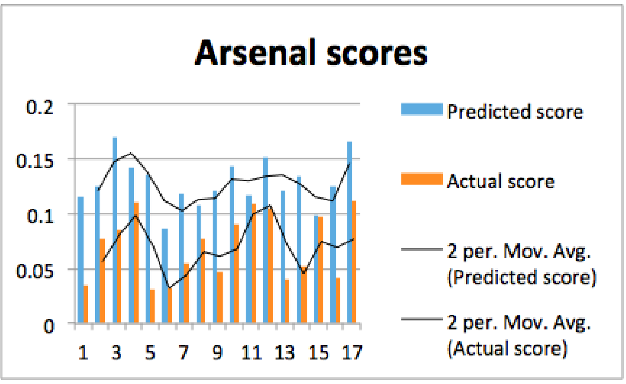


Figure 3: Another figure

To compare the actual and predicted results, for every game: a home win was assigned a value 1, a draw was assigned a value 2 and an away win was assigned a value 3. Figure 2 looks at comparing the actual and predicted results for all of Arsenals home matches in the 2014-15 season. Figure 2 shows a good number of predictions with 12 out of 17 being correct, 70.6%.

To show a graph of results, the probability of both the predicted scores and the actual scores was plotted and a successful prediction was where the two bars joined up. For figure 3 none of the bars join up so there were no successful predictions of the Arsenal scores, however the scores follow a similar trend with the trend lines. This shows that even though no successful predictions were made, the predictions made were a good estimate because the probabilities of the actual result were generally not far off. Furthermore, it also shows when an unlikely scoreline occurred with the height of the actual score bar being really low.

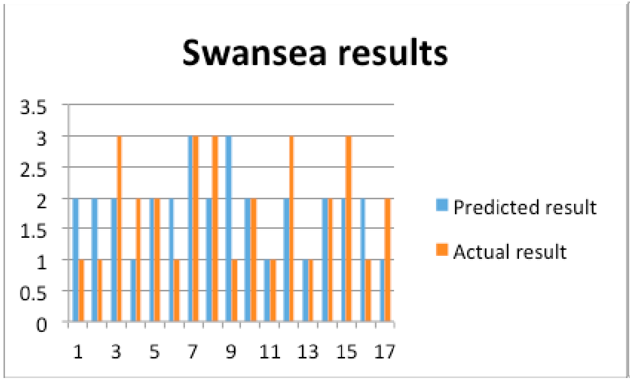


Figure 4

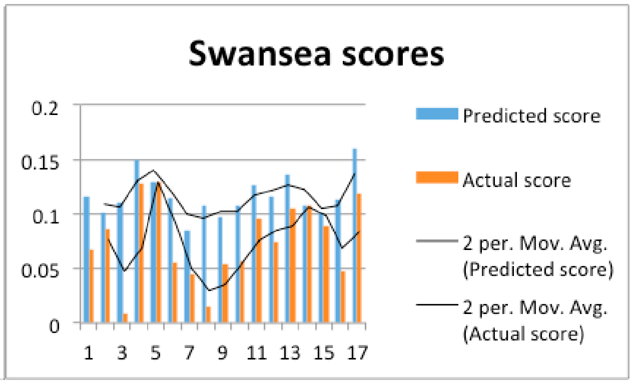


Figure 5

Figure 4 for Swansea's home games are not as good and only 6/17 results were correctly predicted, 35.3%. Figure 5 is also less predictable as the probabilities of the actual scores seem to vary a lot and do not follow the same sort of trend as the predicted scores. However, two scores were accurately predicted which does not occur with Arsenal. Overall, 48.04% of results and 12.74% of scores were predicted correctly.

3.1.2 Discussion

Generally all the graphs follow similar patterns to the 4 above with the matches of the top 5 or 6 teams in the league being better predicted but the mid-table teams or very weak teams were less predictable and very abstract results tended to occur. “This observation fits with the general principle that outliers become progressively rarer in a highly competitive environment” [5]. The amount of correctly predicted results and scores was very low which was probably because of the vast amount of factors that could have been included. One important factor that could have been included to improve the predictions was to try and include recent form and possibly even head-to-head results. A way that this could have been done would have been to possibly make recent form coefficients for each team in a similar way to before (comparing recent amount of goals and average amount of goals) and include them in equations 2 and 3. This would have needed to be compared to the actual results again but the python program would have had to be significantly altered to include dates to take in the most recent results and with the time remaining this was not possible. A head-to-head coefficient could have been computed in a similar way but many other factors such as manager sackings, player injuries, derby games and relegation battles are less quantitative and harder to include.

3.2 Hypothesis Two

To what extent do previous games' results impact future games from a statistical viewpoint. Probabilities found are determined based on data collected from the 1999 to 2014 premier league seasons. The collected data consists of each team's 38 results per season. It is divided into two methods to predict matches, one is calculating the probability based only on the previous game and the other depends on the results from the last six games.

Method 1: Predicting the result depending on only one previous game
Probabilities are determined based on the result ratio depending on last game's result:

1. Count each case when the last game's result is Win/Draw/Loss.
2. Count each case when the next game is Win/Draw/Loss depending on the last game's result.

$$\text{Probability} = \frac{\text{The total cases of the result of upcoming game depending on the last game's result}}{\text{Total cases of the result of the last game}} \tag{8}$$

Based on the uploaded data on excel, each case is counted by connecting between a Java program and excel file.

Total cases two consecutive results of games:
Total cases of two consecutive results of games $= 20 \times 16 \times 37 = 11840$
Where:
$20 =$ the number of teams in one season
$16 =$ the number of years collected
$37 =$ As the last game does not have the next game, the number of cases per team is 37

In addition, the latter games' results are also counted depending on the last game result by a java program. In Figure 6, there are each counted cases in bracket.

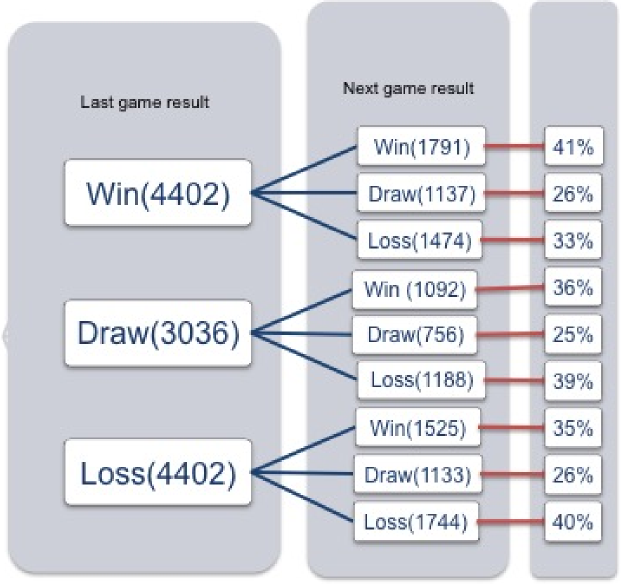


Figure 6: Figure caption

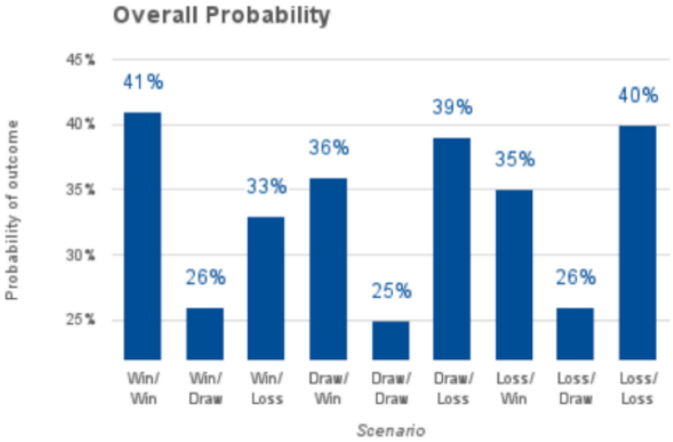


Figure 7: Figure caption

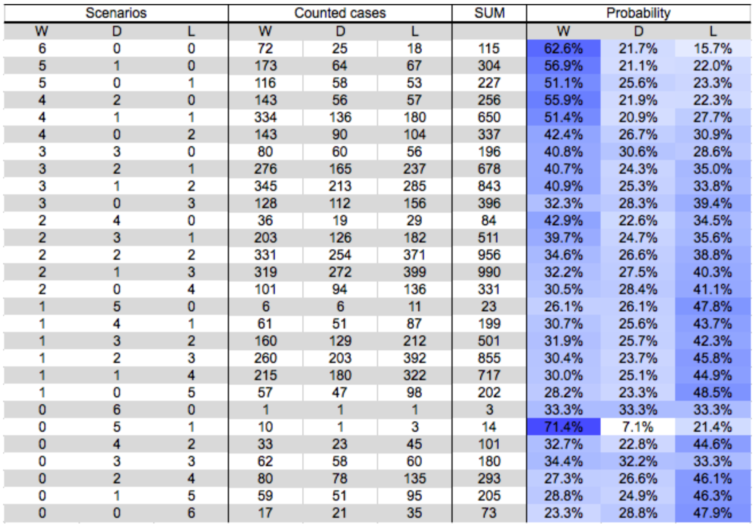
Figure 7 consists of the probabilities arranged as a bar chart, this shows the probability of the next games result depending on the previous, for example there is a 41% chance of winning the next game when the previous game was won.

It also demonstrates that the probability of drawing is about 25 percent, regardless of a result of the last game. It clearly shows that the win/loss streaks get a higher probability than other cases. So if you win you are more likely to win again and if you lose you are more likely to lose again (both of these probabilities are around 40%) Drawing the last game results in a higher probability of losing the next game opposed to winning.

Method 2: Predicting the next game's result depending on recent six games

There are many draws in football compared to other sports so the draw cannot be ignored, this means there

are 28 different permutations of the number of wins, draws, and losses in 6 matches (as shown in figure 8). Total cases six consecutive results of games Total cases of six consecutive results of games $= 20 \times 16 \times 32 = 10240$ Where:
$20 =$ The number of teams in one season
$16 =$ The number of years collected
$32 =$ The number of games per team (The first six games are not counted).
It is enough large to look into a probability based on the total cases. Other java program is made to count each cases. It works when a scenarios what I want to get is inputted, counts the seventh game results.



Scenarios			Counted cases			SUM	Probability		
W	D	L	W	D	L		W	D	L
6	0	0	72	25	18	115	62.6%	21.7%	15.7%
5	1	0	173	64	67	304	56.9%	21.1%	22.0%
5	0	1	116	58	53	227	51.1%	25.6%	23.3%
4	2	0	143	56	57	256	55.9%	21.9%	22.3%
4	1	1	334	136	180	650	51.4%	20.9%	27.7%
4	0	2	143	90	104	337	42.4%	26.7%	30.9%
3	3	0	80	60	56	196	40.8%	30.6%	28.6%
3	2	1	276	165	237	678	40.7%	24.3%	35.0%
3	1	2	345	213	285	843	40.9%	25.3%	33.8%
3	0	3	128	112	156	396	32.3%	28.3%	39.4%
2	4	0	36	19	29	84	42.9%	22.6%	34.5%
2	3	1	203	126	182	511	39.7%	24.7%	35.6%
2	2	2	331	254	371	956	34.6%	26.6%	38.8%
2	1	3	319	272	399	990	32.2%	27.5%	40.3%
2	0	4	101	94	136	331	30.5%	28.4%	41.1%
1	5	0	6	6	11	23	26.1%	26.1%	47.8%
1	4	1	61	51	87	199	30.7%	25.6%	43.7%
1	3	2	160	129	212	501	31.9%	25.7%	42.3%
1	2	3	260	203	392	855	30.4%	23.7%	45.8%
1	1	4	215	180	322	717	30.0%	25.1%	44.9%
1	0	5	57	47	98	202	28.2%	23.3%	48.5%
0	6	0	1	1	1	3	33.3%	33.3%	33.3%
0	5	1	10	1	3	14	71.4%	7.1%	21.4%
0	4	2	33	23	45	101	32.7%	22.8%	44.6%
0	3	3	62	58	60	180	34.4%	32.2%	33.3%
0	2	4	80	78	135	293	27.3%	26.6%	46.1%
0	1	5	59	51	95	205	28.8%	24.9%	46.3%
0	0	6	17	21	35	73	23.3%	28.8%	47.9%

Figure 8: Figure caption

In Figure 8 the left column, scenarios, is the number of wins, draws, and losses. There are 28 different possibilities, the middle column is the number of cases of the 7th result for each scenario. 'Sum' is a sum of total cases of game results. The final column, probability, is the probability of getting the 7th result.

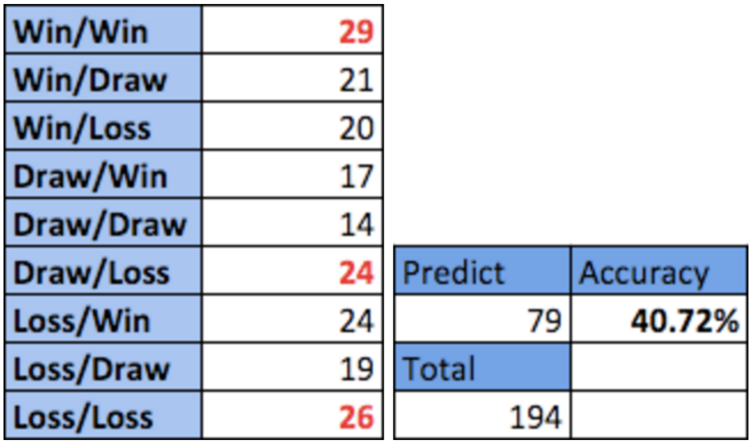
$$Probability = \frac{\text{the number of counted 7th game results}}{\text{the sum of results depending on each scenario}} \quad (9)$$

In Figure 8 the cells shaded with a darker blue have a higher probability.

Figure 8 demonstrates recent game results tend to impact on later games, because some scenarios have quite high percentage for particular result. In order to predict a result, for example, it is likely to predict a win when a team has five draws and one loss in their last six games. Events are likely to occur when probability is not only absolute like above example but also ambiguous to predict. It is sometimes hard to predict result between two teams, which are at a similar level. If so, this percentage can help to determine a winning team. In this principle, it is worth to compare between probabilities of two teams' next result.

Simple simulation on 2015/2016 season

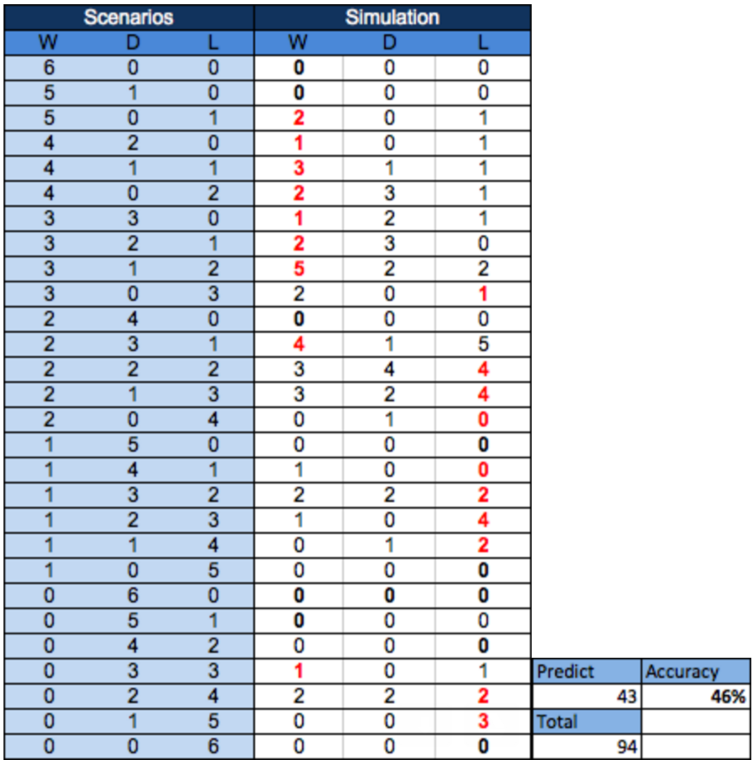
The simple simulation is to count games, the predicted result is the highest probability of two methods. Each result in Figure 9 have the number of counted games based on scenarios in 2015/16 premier league season. There are only 10 or 11 games per team (so far), so the total number of games is smaller than other seasons. It is a simple simulation and accuracy is not reliable, because the real simulation has to compare the two teams' last results. However, those two simulation can give a rough accuracy and counted games of further season, not counted games before. In Figure 9 & 10, red marks are the counted games predicted depend on the highest calculated probability. Accuracy is calculated by the number of predicted games over total cases.



Win/Win	29
Win/Draw	21
Win/Loss	20
Draw/Win	17
Draw/Draw	14
Draw/Loss	24
Loss/Win	24
Loss/Draw	19
Loss/Loss	26

Predict	Accuracy
79	40.72%
Total	
194	

Figure 9: Method 1 Results



Scenarios			Simulation		
W	D	L	W	D	L
6	0	0	0	0	0
5	1	0	0	0	0
5	0	1	2	0	1
4	2	0	1	0	1
4	1	1	3	1	1
4	0	2	2	3	1
3	3	0	1	2	1
3	2	1	2	3	0
3	1	2	5	2	2
3	0	3	2	0	1
2	4	0	0	0	0
2	3	1	4	1	5
2	2	2	3	4	4
2	1	3	3	2	4
2	0	4	0	1	0
1	5	0	0	0	0
1	4	1	1	0	0
1	3	2	2	2	2
1	2	3	1	0	4
1	1	4	0	1	2
1	0	5	0	0	0
0	6	0	0	0	0
0	5	1	0	0	0
0	4	2	0	0	0
0	3	3	1	0	1
0	2	4	2	2	2
0	1	5	0	0	3
0	0	6	0	0	0

Predict	Accuracy
43	46%
Total	
94	

Figure 10: Method 2 Results

3.3 Hypothesis Three

Data Collection

Data was collected on the longest streaks achieved by a team in a given tournament, the sample space chosen was the range from the 2004-05 tournament to the 2014-15 tournament. In order to avoid problems with relegations and promotions of teams their end of season rank was used as their identification, i.e. team of rank 1 finished in first position for that season. This data was then averaged over the 10 year sample range in order to find the likely magnitudes of each type of streak occurring and the general trends of these streaks throughout the team ranks. These results were then used as a comparison for the results collected from our model.

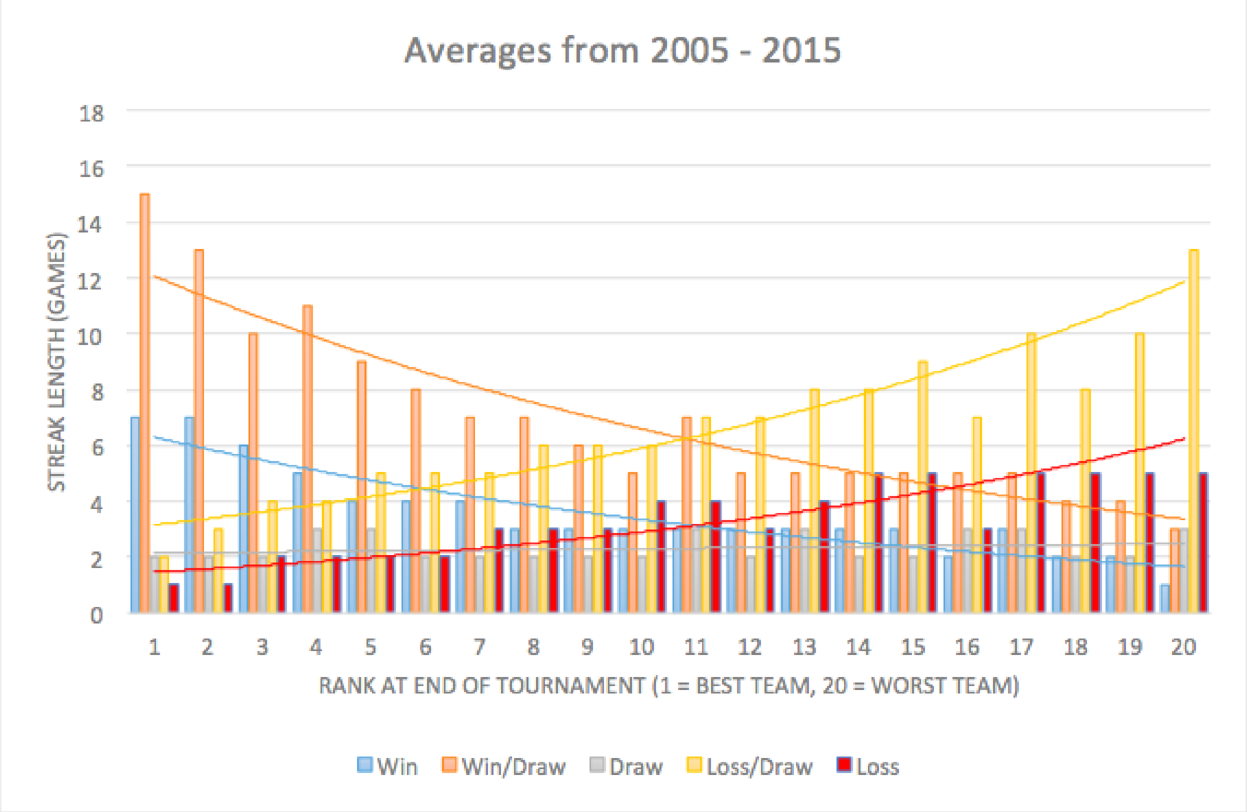


Figure 11: Figure caption.

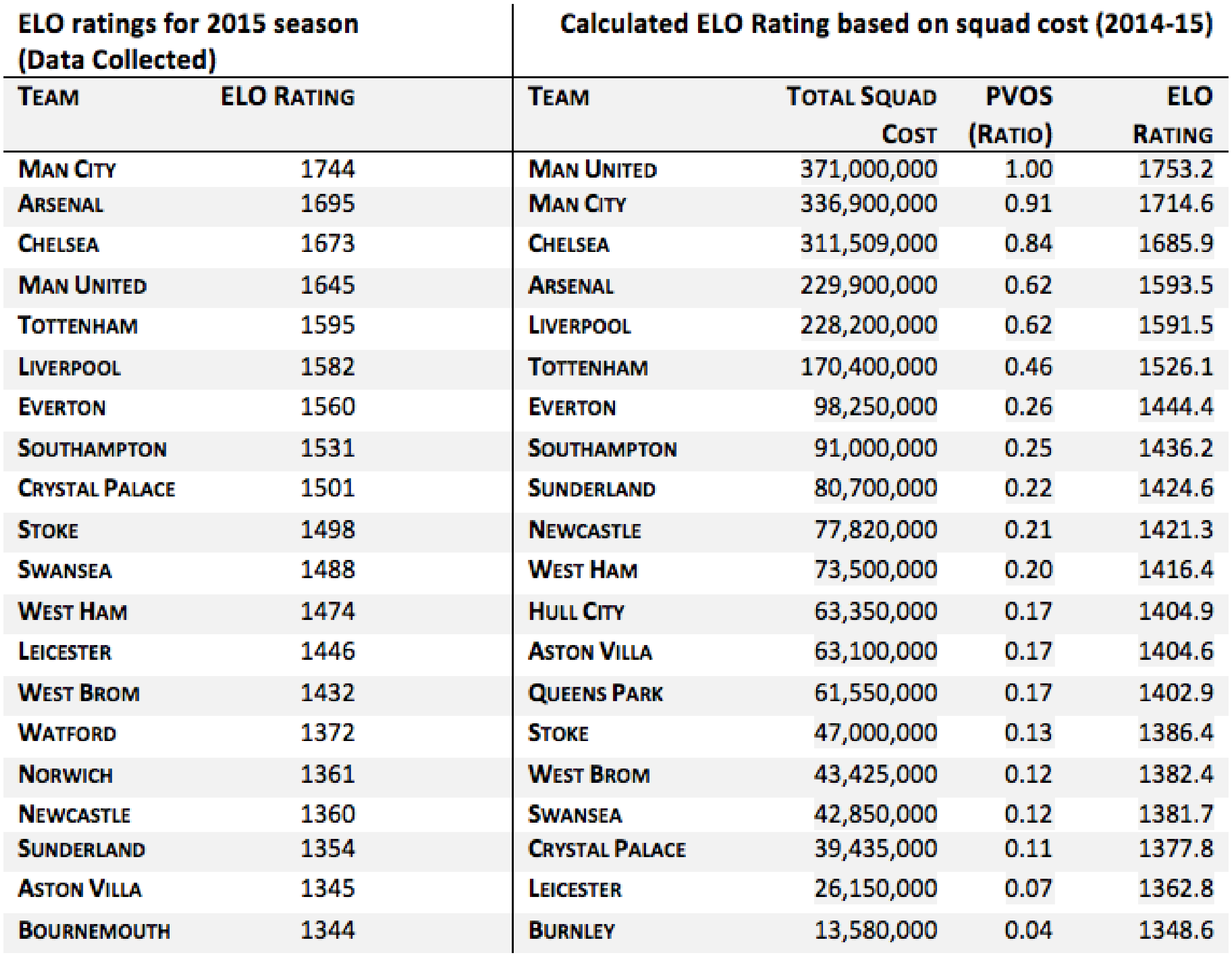
Figure 11 demonstrates the trends and magnitudes of the streaks from the data collected, some features to note here would be:

- Draw streaks stay at steady average throughout the ranks, at around 2
- The graph is almost symmetric however the Loss and Loss/Draw streak lengths for the lower ranked teams tend to be slightly lower in magnitude than the respective Win and Win/Draw streak lengths of the higher ranked teams.
- As expected the win/draw and loss/draw streaks are the highest as these are more likely to occur than any single result. The higher ranked teams always have higher win streaks than the lower ranked teams similarly they have lower loss streaks lower loss streaks, however as this is only averaged over 10 tournaments it is still possible for trends to be broken and anomalies to occur.

Squad Cost and Ratings Approach:
Rating were calculated based on the total cost of the team, the range of where these ratings should be based was decided by the current Elo Ratings of all teams in the premier league. We created our ratings by taking the range between the top Elo Rating team and the lowest and then took a ratio for their squad costs with the most expensive team being one. From that we then used Formula 10 to give our teams our own Elo Ratings.

$$\text{Calculated Elo Rating} = \Bigg(\bigg(\Big(\text{Ratio} - 0.04\Big) * 100\bigg) * 4.2\Bigg) + 1350 \tag{10}$$

the 0.04 is to take into account the lowest ratio value as well as the 1350 being the lowest Elo Rating. 4.2 is the same difference per Elo Rating to 0.01 of the ratio. In Figure 12 both the PVOS ratio and Elo Rating rounded which is why they do not reflect exact results.



ELO ratings for 2015 season (Data Collected)		Calculated ELO Rating based on squad cost (2014-15)			
Team	ELO Rating	Team	Total Squad Cost	PVOS (Ratio)	ELO Rating
Man City	1744	Man United	371,000,000	1.00	1753.2
Arsenal	1695	Man City	336,900,000	0.91	1714.6
Chelsea	1673	Chelsea	311,509,000	0.84	1685.9
Man United	1645	Arsenal	229,900,000	0.62	1593.5
Tottenham	1595	Liverpool	228,200,000	0.62	1591.5
Liverpool	1582	Tottenham	170,400,000	0.46	1526.1
Everton	1560	Everton	98,250,000	0.26	1444.4
Southampton	1531	Southampton	91,000,000	0.25	1436.2
Crystal Palace	1501	Sunderland	80,700,000	0.22	1424.6
Stoke	1498	Newcastle	77,820,000	0.21	1421.3
Swansea	1488	West Ham	73,500,000	0.20	1416.4
West Ham	1474	Hull City	63,350,000	0.17	1404.9
Leicester	1446	Aston Villa	63,100,000	0.17	1404.6
West Brom	1432	Queens Park	61,550,000	0.17	1402.9
Watford	1372	Stoke	47,000,000	0.13	1386.4
Norwich	1361	West Brom	43,425,000	0.12	1382.4
Newcastle	1360	Swansea	42,850,000	0.12	1381.7
Sunderland	1354	Crystal Palace	39,435,000	0.11	1377.8
Aston Villa	1345	Leicester	26,150,000	0.07	1362.8
Bournemouth	1344	Burnley	13,580,000	0.04	1348.6

Figure 12: Figure caption

How ratings were used to calculate win, loss and draw percentages

$$\text{Expected Outcome(Team X)} : E_x = \frac{100}{1 + 10\left(\dfrac{RankY - RankX}{400}\right)} \tag{11}$$

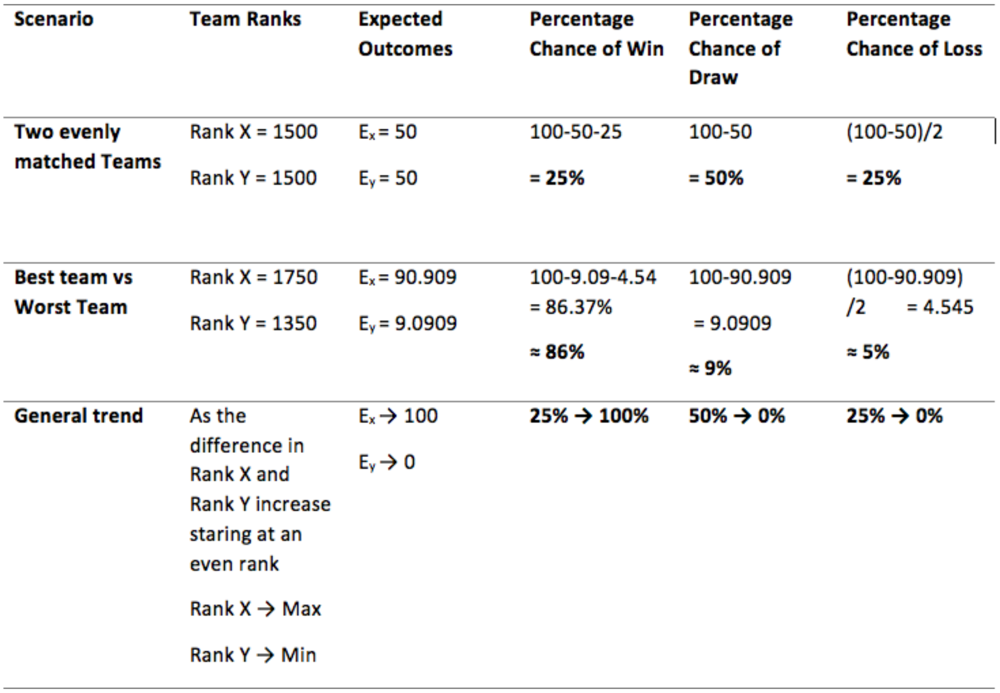
$$\text{Similarly Expected Outcome(Team Y)} : E_y = \frac{100}{1 + 10\left(\dfrac{RankX - RankY}{400}\right)} \tag{12}$$

Expected outcomes of matches were calculated using a known formula given for Elo Ratings, this formula gives a value between 0 and 100 based on the difference between the rank of the two teams. The smaller the difference in ranks the closer to 0 power of 10 becomes and hence the closer to 50 the expected outcome becomes. The value 400 in these equations is chosen because it is approximately the max difference between the rank of the top and bottom teams, adjusting this value affects the rate at which the expected outcome reaches the minimum and maximum values i.e. lowering this value results in a greater change of the expected outcome for the same difference in the ranks of the two teams. The results were then separated further into percentage chances of winning, losing and drawing using the equations below.

$$\text{Percentage Chance of Draw} : P_{X_D} = 100 - E_x = E_y \tag{13}$$

$$\text{Percentage Chance of Loss} : P_{X_L} = \frac{100 - E_x}{2} \tag{14}$$

$$\text{Percentage Chance of Win} : P_{X_W} = 100 - \left(P_{X_D} + P_{X_L}\right) \tag{15}$$



Scenario	Team Ranks	Expected Outcomes	Percentage Chance of Win	Percentage Chance of Draw	Percentage Chance of Loss
Two evenly matched Teams	Rank X = 1500 Rank Y = 1500	E_x = 50 E_y = 50	100-50-25 **= 25%**	100-50 **= 50%**	(100-50)/2 **= 25%**
Best team vs Worst Team	Rank X = 1750 Rank Y = 1350	E_x = 90.909 E_y = 9.0909	100-9.09-4.54 = 86.37% **≈ 86%**	100-90.909 = 9.0909 **≈ 9%**	(100-90.909) /2 = 4.545 **≈ 5%**
General trend	As the difference in Rank X and Rank Y increase staring at an even rank Rank X → Max Rank Y → Min	E_x → 100 E_y → 0	**25% → 100%**	**50% → 0%**	**25% → 0%**

Figure 13: Significant outcome examples

Results of Squad Cost and Ratings Approach

This was then ran through a simulation of a premier league tournament in which each team plays every other team twice, although home advantage was neglected in this case. The results given from this model are as shown below in Figure 14.

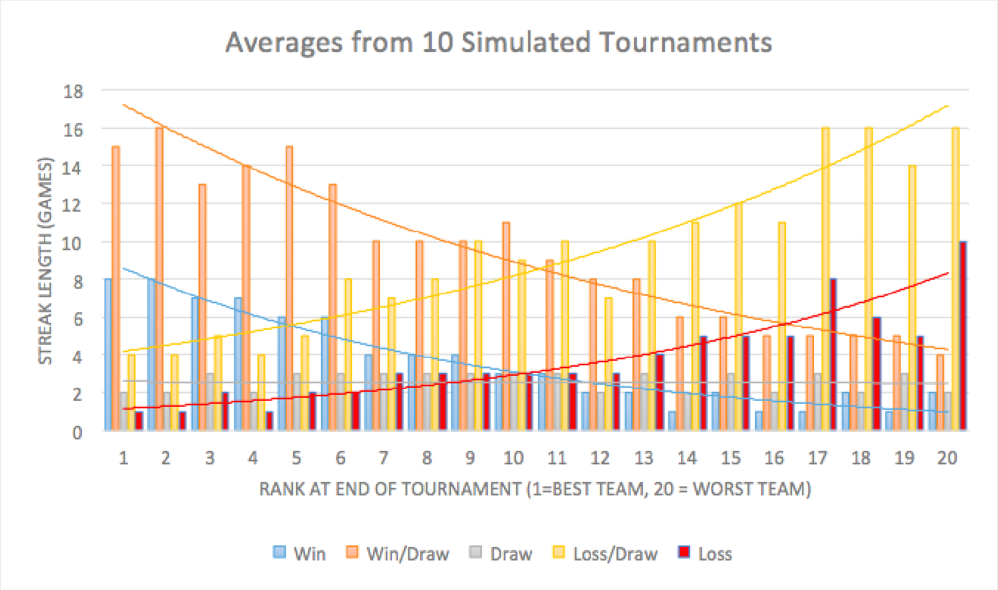
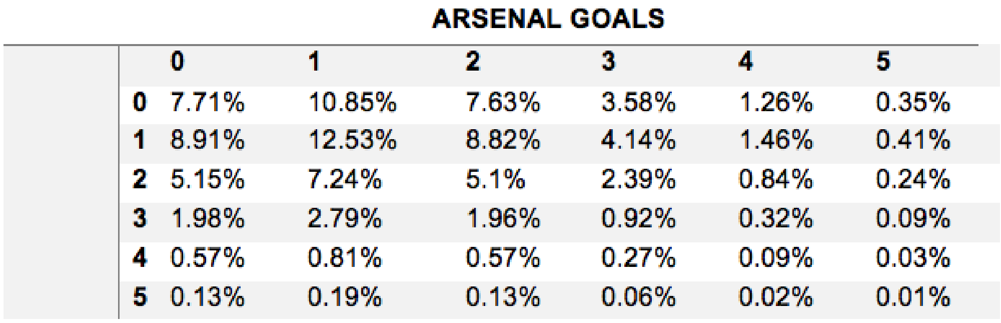


Figure 14

This demonstrates that the results from this model match the trends found in the data we have collected, although the streak lengths tend to be slightly higher, especially for the mid ranked teams. This could lead to the conclusion that in this model longer streaks occur more naturally, this could be due to factors such as pressure or fatigue getting to individual players during a win streak or a loss streak. However because our model takes into account a lot of assumptions it is difficult to draw any concrete conclusions from these results. One of the reasons this approach may not give the most accurate results is that often teams invest in younger players, this would mean that although a teams price is inflated for a particular season, you may not see the return of that investment until several seasons later. Including other factors in this case could also lead to more accurate results, for example including a home advantage or continually adjusting the ratings of teams based on their current streaks.

Probability data approach

In order to improve the model we used the probability research data found in section 3.1. The win, loss and draw probabilities for each team against every other team was calculated using the predictions from hypothesis 1. Using the lambda values for each team, the probability of each score (up to five goals where probabilities become negligible) can be calculated as follows: Arsenal v Chelsea



ARSENAL GOALS

	0	1	2	3	4	5
0	7.71%	10.85%	7.63%	3.58%	1.26%	0.35%
1	8.91%	12.53%	8.82%	4.14%	1.46%	0.41%
2	5.15%	7.24%	5.1%	2.39%	0.84%	0.24%
3	1.98%	2.79%	1.96%	0.92%	0.32%	0.09%
4	0.57%	0.81%	0.57%	0.27%	0.09%	0.03%
5	0.13%	0.19%	0.13%	0.06%	0.02%	0.01%

Figure 15: Figure caption

So the probability of a home win in this case is the sum of the upper triangle of probabilities if you discount the diagonal, the probability of an away win is the sum of the lower triangle of probabilities and the probability of a draw is the sum of the diagonal entries. However, the total sum does not quite add up to 100% instead it is 99.55% and for some games it was 98% or lower, so by dividing the win, draw, loss probabilities by the total table probability scaled the probabilities nicely up to total 100%.

There was no data on Burnley or Leicester (they were newly promoted) so after looking at the final table Figure 16 of the 2014-15 season the probability values for Burnley and Leicester was taken as the midpoint of the two teams either side of them in the final table as a good approximation.

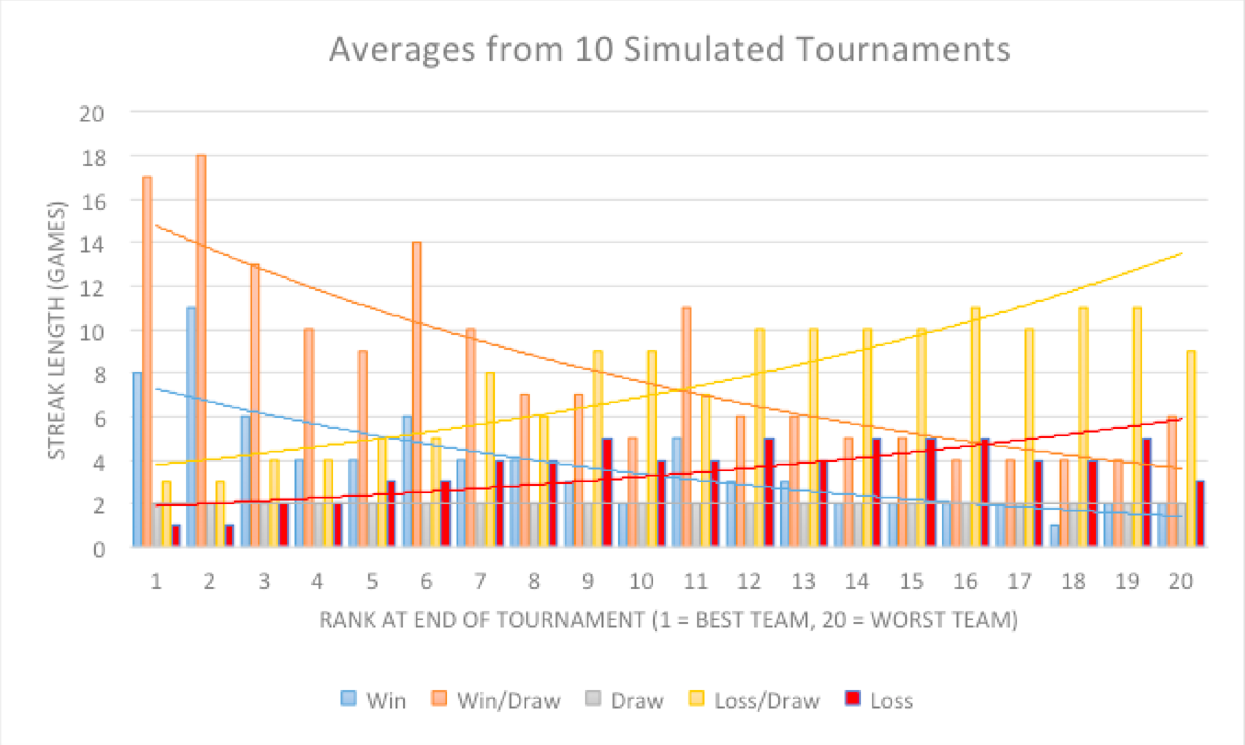


Figure 16: Results of probability data approach

The results of this approach match pretty closely to the data collected and in most scenarios match closer than the first approach with squad cost and ratings. For example, the data still follows the same trends as however the magnitudes of streaks tend to match closer to the data collected in this model, this is evidenced best in the mid ranked teams. This model also tends to favour the win streaks of teams, this is evidenced by the win streaks being slightly higher and the loss streaks being slightly lower than the data we collected. Another area where this model seems to lack is in the loss and loss/draw streaks, where they tend to reach their max by the mid ranked teams, in the data collected it keeps rising into the lower ranked teams.

Also with this model we decided to count the anomalies that occur, the greatest win streak achieved in the premier league is a streak of 14 achieved by arsenal in 2003-04, during this win streak they also went on a unbeaten streak (Win/Draw streak) of 49 games however this stretched over two seasons. In order to determine the likelihood of these streaks in our model we counted the number of times a win streak of length 14 and a win/draw streak of length 37 (max of one tournament) occurred.

Over 10,000 tournaments

- The number of times a team went on a win streak greater than 14 : 3564
- The number of times a team went on a win draw streak greater than 37 : 189

This leaves the likelihood of a team going on a 14 game or higher win streak at 3564/10000 or approximately approximately 7 times every 22 years. In the data we have collected this has occurred once in the 22 tournaments since the premier league has started in 1992. Similarly an unbeaten streak has only occurred once in 22 tournaments however in this model the likelihood is much smaller at only 189/10000 or approximately once every once every 50 years.

4 Discussion/ Conclusions

From the first hypothesis we found that predicting outcomes of games was far more accurate than predicting the individual score which allowed us to further focus the models on outcome (win/loss/draw) streaks as opposed to score-streaks. The percentage of results accurately predicted was quite low and this was probably due to the high number of factors that were involved but could not be included in the time as well as the unpredictability of football. The second hypothesis showed that looking into the results of the previous six games gave a more accurate prediction than just observing the previous one result and comparing the predictions from this to the 2015/16 season gave a successful prediction percentage similar to hypothesis 1. The third hypothesis found that when simulating a premier league tournament using the probability data found in section 3.1 a more accurate result is given than when compared to a tournament based upon squad cost, however there are still differences to the real world data.

Overall we conclude that the current streak that a team is on in the premier league will have an effect on the outcome of an upcoming game, however due to the myriad of other factors that can influence a game it is difficult to tell the extent of which this has. We would predict that although there is an effect, this will most likely not be the defining factor of a game result and could in fact be negligible in comparison with the other factors. The reason we have come to this prediction is that when creating a model for the outcome of a game without the influence of streaks the results still match relatively closely to the data collected. Due to the fact that the results of the models created also tend to have longer streaks than in the data collected we also came to the conclusion that the effect that a win streak is having on the team may not always be positive. For example although the team morale may be boosted during a win streak it will likely be at a trade off with other factors, this could include injury or fatigue on players after winning several games in a row. This correlates with the previous prediction that the streaks effect will probably not be the defining factor in the outcome of a game as the fatigue or injury will overpower this.

The article 'Twenty years of "hot hand" research' also reaches a similar conclusion in regard to the importance of the streaks and whether predictions can be made about these streaks. The article concludes that streaks of wins and losses are random and occur just like streaks of heads or tails. So the article is saying that streaks are independent of the probability of a win, loss or draw. We concluded that by looking at streaks you cannot predict the outcome of the next game accurately, so although we looked at it from two different viewpoints, both came to the same conclusion that the two events (streaks and next outcome) are independent. Better teams will tend to have more frequent streaks of wins but the ability to predict when the streak starts or finishes (the individual outcomes) does not seem to follow a pattern, seems to be random.

Looking back at the article 'Understanding baseball team standings and streaks' we reached different conclusions for the the distribution of team strengths for producing win and loss streaks. The article concluded the team strengths had a uniform distribution but our 'Elo Ratings' were not uniform for the 20 teams. Additional research into this and whether the actual win/loss streaks and the produced win/loss streaks match up might produce different results.

In order to understand whether the results from this research would be applicable to other sports, factors where the premier league may differ to other sports or tournaments need to be considered. One factor would be that football is a team based sports although most of the results seem applicable to individual player based sports, this will however affect the influence of some of the assumptions, for example an injury to a player will have a less significant impact in a team based sport. Another factor would be the way in which the tournament is performed, It would be likely that a knockout competition will have a very different set of results when compared with a competition in which each team plays every other team twice, which is the case of the premier league. For this reason the conclusions of this report should be considered only applicable to the premier league although it is likely that there are other sports tournaments that this also could be considered for. A broader study of many different sports and tournaments would need to be performed in order to confirm this.

References

[1] Ratcliffe, J.R. 2014. *Poisson Distribution: Predict a soccer betting winner*. 12 August. [Online]. [Accessed 12 October 2015]. Available from: `http://www.pinnaclesports.com/en/betting-articles/soccer/how-to-calculate-poisson-distribution`

[2] *Premier league Elo Ratings*. [Online]. [Accessed 28 October 2015]. Available from: `http://sinceawin.com/data/elo/league/div/e0`

[3] *Premier league squad costs- transfer league*. 2015-2016. [Online]. [Accessed 26 October 2015]. Available from: `http://www.transferleague.co.uk`

[4] *Progress results streaks over 16 seasons*. [Online]. [Accessed 28 October 2015]. Available from: `http://www.whoscored.com/Regions/252/Tournaments/2/England-Premier-League`

[5] Sire, C and Redner. 2009. *Understanding baseball team standings and streaks*. The European Physics Journal B. 67,473-481.

[6] Bar-Elia, M & Avugosa, S & Raab, M. 2006. *Twenty years of "hot hand" research: Review and critique. Psychology of sport and exercise*. 7, 525-553.

5 Appendix