A novel statistical model to determine the winner of drawn soccer matches

Apte Abhay

Abstract

A novel statistical model to determine the winner of drawn soccer matches is presented. An expression number based on three factors (shots, shots on goal and possession percentage) was used in a training set to calculate the expression numbers for all matches played by three La Liga teams in the 2018-2019 season. There were a range of expression numbers associated with the number of goals scored. The lowest and highest expression numbers for a specific number of goals for a particular team defined the lower and upper values of the overall capability of that team to score those number of goals associated with those expression numbers. For each team, an average expression number for a specific number of goals (0 through 5) was calculated. These goal-categorized expression numbers were plotted against the number of goals. The data was described with a linear model with a regression coefficient of > 83%. By assigning values of 0 and 1 for the lowest and highest expression number for each goal scored in a particular game, the real normalized weighted goal average ($nWGA_r$) for the team could be calculated using the expression number from that game. When the $nWGA_r$ was divided by the expression number, the resultant number, designated as an ideal weighted goal average (WGA_i), was obtained. The WGA_i represents the ratio of how well the team performed compared with its own playing statistics to how well it performed compared with an ideal statistic (an ideal expression number of 1.0). For drawn matches, an ideal weighted goal average (WGA_i) could be calculated for each team using their individual expression numbers for that game. The team with the greater WGA_i was determined to be the winner of the drawn match. This method represents a significant advantage over the traditional penalty shoot-out method of resolving drawn matches because it eliminates the luck-factor inherent in the penalty shoot-out approach. The model can also be used to assess whether play quality for each team was consistent with previous games. This model can offer coaches and players valuable information that can be leveraged toward more effective team development.

Apte Abhay, abhaysapte@vt.edu, abhaysapte@gmail.com, Colleyville Heritage High School, Colleyville, TX 76034. Current affiliation: Junior, Statistics and Economics, Virginia Tech, 1727 Aden lane Blacksburg, VA

Introduction

Tied matches are the bane of professional soccer. The penalty shoot-out method of deciding the winner of a drawn soccer match essentially relegates the 'beautiful game' to a gamble. Since the size of the net is too wide for the goalkeeper to cover against an unchallenged shot from 12 yards away, the penalty shoot-out essentially relies on luck rather than skill. It also places disproportionate emphasis on individual talent, which is at odds with the notion of soccer as a team sport.

Various alternatives to the penalty shoot-out (1-4) have been proposed. To my knowledge, the method conceived, developed and presented in this manuscript is the first to suggest the use of a statistical model to assign the winner of drawn soccer matches. In order to develop the statistical model, a 'training set' of match data points was obtained from the 2018-2019 season for three teams in the LaLiga league.

Model description and development

Barcelona, Real Madrid and *Espanyol* were the teams chosen to formulate the model. During the 2018-2019 season, *Barcelona* and *Real Madrid* were at the top of the league in points, while *Espanyol* was near the bottom. This allowed for evaluation of how the model performed across a range of low as well as high skill-level, while also evaluating its power to discern play quality and prediction between teams such as *Real Madrid* and *Barcelona* that had a lesser point differential in the league.

Table 1: Training set: Raw data for LaLiga matches 2018-2019 season

Game #	Barcelona				Real Madrid				Espanyol			
	Shots	Shots on target	PossessionPercent	Goals scored	Shots	Shots on target	Possession Percent	Goals scored	Shots	Shots on target	Possession Percent	Goals scored
1	13	5	66	1	14	7	54	2	12	1	52	1
2	12	8	76	2	28	8	59	0	21	2	61	2
3	7	6	59	3	14	5	26	2	6	0	45	0
4	15	7	72	3	23	6	69	2	13	6	36	2
5	15	7	64	2	16	9	64	4	17	2	58	2
6	10	2	67	0	22	8	71	3	15	5	61	2
7	17	8	60	4	11	4	66	3	7	1	49	0
8	9	4	51	1	19	7	57	1	13	2	68	1
9	12	6	69	3	19	5	61	2	15	6	54	3
10	16	9	44	4	17	3	49	0	5	2	35	1
11	13	3	70	2	15	6	62	4	13	3	50	0
12	15	9	68	4	22	7	53	2	4	2	30	0
13	21	10	64	2	16	8	71	3	9	3	51	1
14	10	2	74	0	15	5	68	1	10	2	45	2
15	11	4	50	2	13	7	56	2	6	2	56	2

Game #	Barcelona				Real Madrid				Espanyol			
16	16	6	77	2	13	6	71	1	19	5	50	2
17	23	12	62	1	11	4	57	3	9	2	48	1
18	8	1	65	0	12	4	67	0	10	4	46	1
19	8	3	61	2	9	3	51	0	11	5	50	3
20	11	2	72	0	24	11	60	3	11	4	47	2
21	8	4	75	2	6	2	46	1	10	4	37	2
22	17	10	65	5	9	2	47	0	7	2	45	0
23	8	1	74	0	17	4	48	0	5	1	65	0
24	8	4	57	2	17	11	43	3	15	3	67	0
25	10	6	62	2	22	4	66	1	7	2	41	2
26	15	8	56	4	17	5	56	2	17	3	56	1
27	9	4	56	3	25	11	65	3	10	4	36	1
28	18	9	62	5	7	4	45	1	16	2	59	0
29	10	7	61	4	15	4	55	2	11	3	62	0
30	11	6	77	2	14	3	56	0	18	3	51	0
31	13	5	66	1	17	10	62	4	14	3	48	1
32	25	17	66	5	14	4	63	0	13	4	46	1
33	9	5	47	2	22	11	62	5	10	1	50	1
34	9	2	52	0	22	7	63	0	6	2	54	1
35	8	4	69	4	10	5	59	4	12	5	53	2
36					18	10	53	3	6	2	39	0
37					12	6	64	2	16	4	38	2
38					18	8	61	2	9	3	59	0
39					19	6	64	1				
40					20	8	69	0				

Data obtained from https://www.whoscored.com for 2018-2019 LaLiga fixtures. Table 1 represents data from all the matches played by each team in the season, including matches played against other teams in the league.

A preliminary analysis of the data in Table 1 indicated that the number of goals scored depended on a relatively small number of critical variables. These included: the shots on target, the ratio of the shots on target to the shots, the possession percentage and the ratio of shots on target to the possession percentage. A shot is an attempt that is taken with the intent of scoring and is directed toward the goal. A shot on target (or on goal) is a shot that is on net. On the other hand, other variables such as the arial dual success, dribbles won, tackles, team ranking, home or away game, the number of shots from the 6 or the 18 yard box, the number of shots from outside the box, the number of crosses, through balls and short or long passes showed a weak correlation, if any, with the number of goals scored. An empirical mathematical relationship was formulated that allowed a linear model to describe the data for all the three teams (R^2 > 83%) for matches played in the 2018-2019 season.

$$expr = \left[\!\left[\frac{shotsongoal}{shots} \right]\!\right] + \left[\!\left[\frac{shotsongoal}{possessionpercent} \right]\!\right],$$

where expr = expression number Equation (1)

Table 2: Goal categorized expression numbers for *Real Madrid* for La Liga matches played in 2019

Goals scored	Number for individual match calculated using equation 1 and categorized according to the number of goals scored										Average	Range
0	0.421	0.236	0.393	0.392	0.265	0.319	0.268	0.349	0.429	0.515	0.359	0.236 to 0.515
1	0.491	0.407	0.546	0.377	0.242	0.410					0.412	0.242 to 0.546
2	0.629	0.549	0.348	0.345	0.450	0.663	0.383	0.339	0.593	0.409	0.471	0.339 to 0.629
3	0.476	0.424	0.613	0.434	0.642	0.609	0.744				0.563	0.424 to 0.744
4	0.703	0.497	0.749	0.585							0.634	0.497 to 0.749
5	0.677										0.677	Not applicable
Average and Range of the entire data set regardless of goals											0.472	0.236 to 0.749

Table 3: Goal categorized expression numbers for *Barcelona* for La Liga matches played in 2019

Goals scored	Number for individual match calculated using equation 1 and categorized according to the number of goals scored												Average	Range
0	0.230	0.227	0.140	0.210	0.139	0.261							0.201	0.139 to 0.261
1	0.460	0.523	0.715	0.460									0.540	0.460 to 0.715
2	0.772	0.576	0.274	0.632	0.444	0.453	0.424	0.563	0.570	0.697	0.623	0.662	0.558	0.274 to 0.772
3	0.516	0.587	0.564	0.959									0.657	0.516 to 0.959
4	0.558	0.815	0.676	0.732	0.767	0.604							0.692	0.558 to 0.815
5	0.937	0.742	0.645										0.775	0.645 to 0.937
Average and Range of the entire data set regardless of goals													0.558	0.139 to 0.959

Table 4: Goal categorized expression numbers for *Espanyol* Liga matches played in 2019

Goals scored	Number for individual match calculated using equation 1 and categorized according to the number of goals scored												Average	Range
0	0.00	0.163	0.291	0.567	0.330	0.215	0.245	0.159	0.321	0.225	0.385	0.385	0.274	0.00 to 0.567
1	0.102	0.183	0.457	0.392	0.263	0.487	0.230	0.510	0.277	0.395	0.120	0.370	0.316	0.102 to 0.510
2	0.128	0.628	0.152	0.415	0.244	0.369	0.363	0.449	0.508	0.334	0.511	0.355	0.371	0.128 to 0.628
3	0.511	0.555											0.533	0.511 to 0.555
Average and Range of the entire data set regardless of goals													0.332	0.00 to 0.628

Table 2 shows the results obtained by applying Equation 1 to games categorized by the number of goals scored for *Real Madrid*. For example, *Real Madrid* scored 0 goals in 10 of the matches, scored 1 goal in 6 matches, scored 2 goals in 10 matches, 3 goals in 7 of their matches, 4 goals in 4 of their matches and scored 5 goals in one match. For the training set, the expression in Equation 1 – designated as the 'expression number' (expr) - was used to calculate the expr for each game, using the data from Table 1. The last two columns in Table 2 are the average and range respectively for the goal categorized expression numbers. The average and range for the entire training set data was also calculated and is presented in the last row of the Table. Similarly calculated goal-categorized expression numbers are presented in Tables 3 and 4 for *Barcelona* and *Espanyol* respectively.

Figure 1: Correlation between the expression number and the number of goals scored for *Real Madrid* (from Table 2)

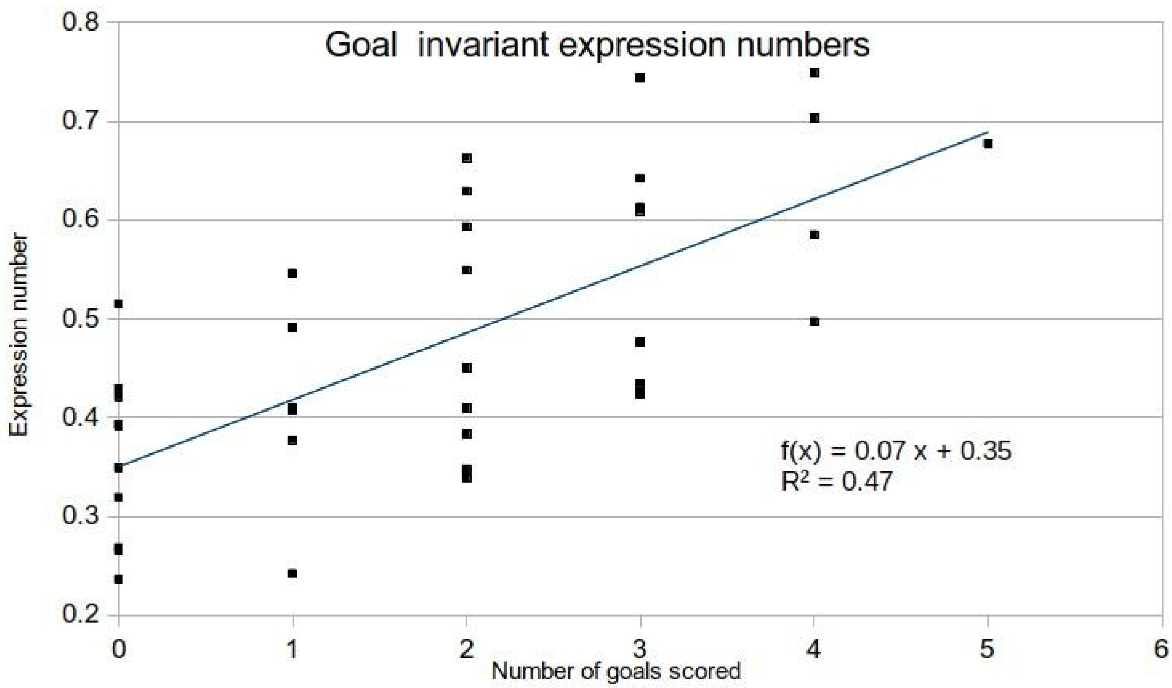


Figure 1 is a graphical representation of the data from Table 2. It can be seen that there is no relationship between the expression numbers calculated from the factors selected (shots, shots on target and percent possession) and the number of goals scored, with a regression coefficient < 0.5.

Due to the poor correlation between the number of goals scored and the expression numbers, the wide range of goals scored (between 0 and 5) and the failure of discriminating between the number of goals with a given expression number, it was evident that Equation 1, by itself, could not be used to predict the number of goals scored. In fact, the goal-invariant correlation was even poorer for *Barcelona* and *Espanyol* (Tables 3 and 4, data not shown). To resolve this conundrum, the graphical data in Figure 1 was considered to be discrete with respect to the number of goals scored. Using this strategy, it became clear that, for a fixed number of goals scored, there was an average and a range for the values calculated using Equation 1. By using this approach, the correlation coefficient for the average goal categorized expression number versus the number of goals was > 0.99 for *Real Madrid* (Figure 2). It is likely that with a larger data set (from more matches) these values could reasonably be expected to randomly lie between the minimum and maximum expression number for a fixed number of goals. The capability of scoring a particular number of goals was least with the least expression number for that goal and increased as the expression number for that particular goal increased. For example, using the data for *Real Madrid* (Table 2), minimum expression numbers of 0.242, 0.339, 0.424 and 0.497 were needed to score 1, 2, 3 or 4 goals respectively. However, it is not a certainty that a specific number of goals are guaranteed for a particular expression number, as can be seen from Row 1 in Table 2, where no goals were scored for expression numbers between 0.236 and 0.515. This underscores the necessity of using a weighted goal average (WGA) number to decide the winner of matches. The expression number was first 'capability normalized' using a linear function by calculating it as a number between 0 and 1 for the range of expression numbers for that particular number of goals. If the expression number was less than the least expression number for that number of goals, it was assigned a value of 0. If the expression number was greater than the largest expression number for that number of goals, it was assigned a value of 1. This 'capability normalized' expression number (exprn) was then multiplied by that particular number of goals. The summation of these multiplication products across the number of goals (6 for *Barcelona* and *Real Madrid*, and 4 for *Espanyol*) and division by the total number of goal categories, yielded a $nWGA_r$ for that team for that match. It should be pointed out that the multiplication product for zero goals would always be 0. This term was still included in the calculation of the $nWGA_r$ to be consistent with the definition of the WGA. Dividing the $nWGA_r$ by the expression number yielded an ideal weighted goal average (WGA_i). The winner of the drawn match was the one with the greater WGA_i. The physical significance of these arithmetic operations is described in further detail below.

Figure 2 (below) represents the goal categorized average expression number (from Table 2) plotted against the number of goals scored for that match for *Real Madrid*. The trend line is linear with a regression coefficient of > 0.99. This implies that the average goal categorized expression number is highly correlated with the number of goals, thereby justifying the choice of the selected variables (shots, shots on goal and possession) and their mathematical expression. The error bars for the average expression number represent the range of expression numbers for that number of goals (from Table 2).

Similarly, Figures 3 and 4 represent the goal categorized average expression numbers (from Tables 3 and 4) plotted against the number of goals for *Barcelona* and *Espanyol* respectively.

Figure 2: Training set data for *Real Madrid* (from Table 2)

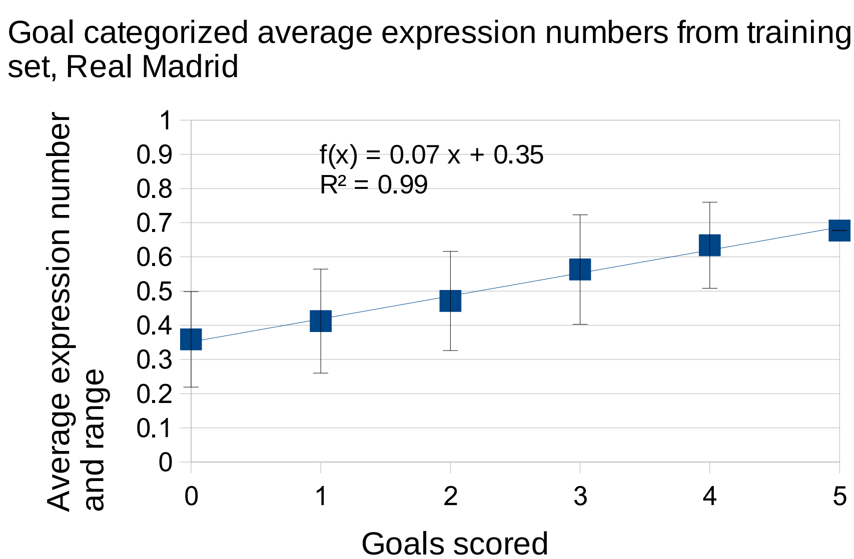


Figure 3: Training set data for *Barcelona* (From Table 3)

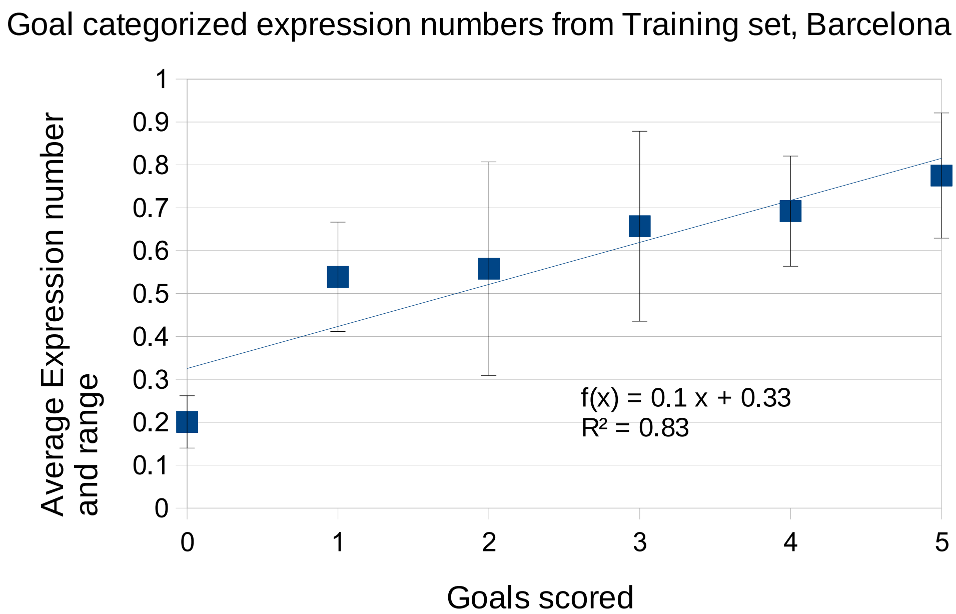
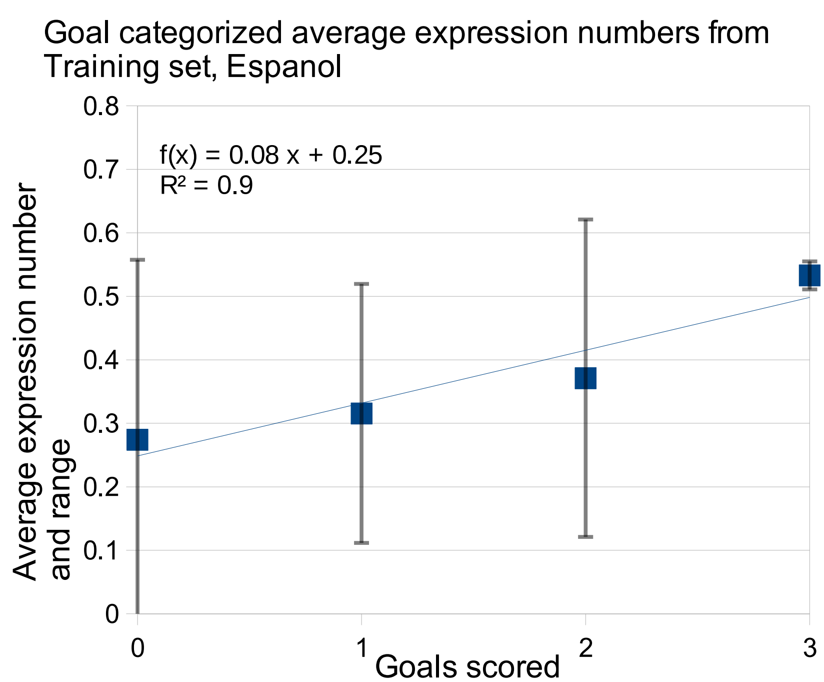


Figure 4: Training set data for *Espanyol* (from Table 4)



An example of the calculation of the WGA_i for a non tied match between Barcelona and Espanol (the first row of table 4) is presented. The expression number of Barcelona was calculated as 0.567 from equation 1

$$expr = \frac{9}{20} + \frac{9}{77} = 0.567$$

This expression number was capability normalized for each goal using the date from Table 3 and Figure 3. For example, for 2 goals, the least expression number of 0.274 was assigned a value of 0 and the greatest expression number of 0.772 was assigned a value of 1. A linear plot of the expression numbers on the Y-axis and the 0 to 1 interval on the X-axis yielded the equation y=0.498x + 0.274
The expression number, expr=0.567 was hence normalized to exprn=0.557. An algorithm was written in R to automate the calculations (Appendix 1). Using the same logic, the normalized expression numbers were calculated for all the goals scored. The exprn=1 for 0 goals scored because the expr of 0.567 is greater than the greatest expr (0.261) for 0 goals scored. The exprn=0 for 5 goals scored because the expr of 0.567 is lesser than the least expr (0.645) for 5 goals scored. The data are presented in Table 5.

Table 5: Calculation of $nWGA_r$ for an expr of 0.567 for Barcelona

Goals scored (g)	0	1	2	3	4	5
Normalized expression number (exprn)	1	0.605	0.557	0.767	0.704	0
(g) x (exprn)	0	0.605	1.114	2.301	2.816	0

$$nWGA_r = \frac{1}{n}\sum_{i=1}^{i=n} exprn_i \times g_i = 1.139$$

$$WGA_i = \frac{nWGA_r}{expr} = \frac{1.139}{0.567} = 2.009$$

This is the WGA_i that is presented for Barcelona in the first row of Table 6 below.

Application of the model to non-tied and tied matches

Table 6: Application of the model to non-tied matches, La Liga, 2018- 2019 season.

Team (Goals scored)	Shots	Shots on Goal	Possession	Expression Number	WGA_i	Opponent Team	Shots	Shots on Goal	Possession	Expression Number	WGA_i
Barcelona (5)	20	9	77%	0.567	2.01	Espanyol (0)	9	1	23%	0.155	0.93
Real Madrid (2)	23	8	64%	0.473	1.19	Espanyol (0)	8	3	36%	0.458	0.55
Real Madrid (0)	14	5	45%	0.468	1.20	Barcelona (3)	18	11	55%	0.811	2.69
Espanyol (1)	16	7	38%	0.622	1.96	Real Madrid (0)	12	5	62%	0.497	1.98
Espanyol (0)	11	2	40%	0.232	0.74	Real Madrid (2)	13	7	60%	0.655	1.65
Espanyol (1)	4	1	33%	0.280	0.67	Barcelona (4)	14	7	67%	0.604	1.92
Real Madrid (2)	22	14	42%	0.970	2.57	Barcelona (3)	16	9	58%	0.718	2.78
Espanyol (0)	6	3	35%	0.586	2.06	Barcelona (3)	15	9	65%	0.738	2.73
Espanyol (0)	14	3	53%	0.271	0.68	Real Madrid (6)	15	8	47%	0.704	3.03
Real Madrid (1)	19	5	65%	0.340	0.62	Espanyol (0)	10	3	35%	0.386	0.60
Espanyol (0)	5	8	31%	0.786	1.91	Real Madrid (1)	14	6	69%	0.516	1.94
Real Madrid (2)	13	5	44%	0.498	1.99	Barcelona (0)	9	4	56%	0.516	1.26

Out of 12, non-tied, matches played between the chosen teams in the 2018-2019 season, that were not part of the training set, the model accurately predicted the outcome of 11 matches using the expression number for those teams in that particular match. Furthermore, there was a significant correlation between the goal differential between the two teams and the difference between their WGA_i for the matches, as can be seen from figure 5. This means that the probability of the accuracy of the prediction, happening by random chance alone, is minimal. Hence, the model calculated WGA_i is underpinned by the true capability and talent of the team; not by chance or luck.

The model did not correctly predict the one match between Espanyol and Real Madrid (row # 4). The prediction failed by a narrow margin as can be seen from the WGA_i numbers for the two teams. In fact, assigning a possession percentage of 40% and 60% to Espanol and Real Madrid respectively enabled a correct prediction with WGA_i numbers of 1.987 and 1.979 for the two teams respectively. Since the possession was reported to a precision of ± 9%, inaccurate data reporting or entering can significantly alter the prediction results for matches that have a goal differential of 1. Since this does not seem to be applicable to all such matches (the model correctly predicts the outcome of two other matches with a goal differential of 1), it lends more credance to the occurance of an error in data acquisition or entry. Some other limitations of the model are described in the Limitations section below.

Figure 5: Prediction capability of WGA_i

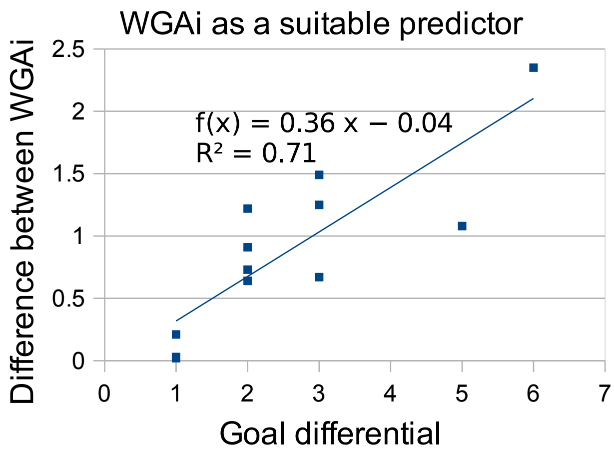


Table 7: Application of the model to drawn matches, La Liga, 2018- 2019 season.

Team (Goals scored)	Shots	Shots on Goal	Possession	Expression Number	WGA_i	Opponent Team	Shots	Shots on Goal	Possession	Expression Number	WGA_i
Barcelona (0)	9	2	52%	0.260	0.000	Real Madrid (0)	17	4	48%	0.318	0.197
Barcelona (2)	14	7	69%	0.601	1.929	Espanol (2)	7	3	31%	0.526	1.608
Espanol (1)	7	2	26%	0.360	0.601	Barcelona(1)	12	2	74%	0.194	0.000
Barcelona (2)	11	4	50%	0.444	0.372	Real Madrid (2)	17	5	56%	0.383	0.579
Barcelona (1)	11	3	55%	0.327	0.445	Real Madrid (1)	14	5	45%	0.468	1.198
Barcelona (2)	14	7	69%	0.601	1.929	Espanol (2)	7	3	31%	0.525	1.611
Espanol (0)	4	1	28%	0.286	0.666	Barcelona (0)	10	2	72%	0.228	0.000

Among the 7 drawn matches played between the three teams in 2018-2019 season (rows 1 and 4 are part of the training set), the model could predict the winner based on a greater WGA_i . For example, for the tied match in row 1 between Barcelona and Real Madrid, Real Madrid would be declared the winner based on its WGA_i of 0.197, as compared to a WGA_i of 0 for Barcelona. It must be emphasized that the model compares the differences in play quality between the opposing teams in that particular match based on two factors. The $nWGA_r$ predicts how well the team played in that match relative to the *same* team's average performance over the season. This, in turn, is derived from the training set data, i.e. Tables 2, 3 and 4 for Real Madrid, Barcelona and Espanol respectively. The physical significance of dividing the $nWGA_r$ by the expression number is that it normalizes the expression number to an 'ideal' expression number of 1.0. In other words, it calculates what the WGA would have been, had the expression number been equal to 1.0. It hence enables a comparison between different teams based on an ideal maximum expression number. The mathematical operation of multiplying a 'real' $nWGA_r$ by the 'ideal' inverse of the expression number may be considered analogous to the 'between-within mean variability' used in ANOVA.

Table 8: Interpreting expression numbers for the data in figures 2, 3 and 4.

Team	Slope	Intercept	Regression coefficient greater than
Espanol	0.083	0.249	0.89
Real Madrid	0.067	0.351	0.99
Barcelona	0.098	0.325	0.83

An examination of the data in Table 8 reveals that Real Madrid required the least increase in expression numbers for a specific increase in the number of goals scored, as evidenced from the (lowest) value of the slope, while Barcelona required the greatest. This means that Real Madrid required the least proportional increase in effort to obtain a given increase in goals scored. In other words, Real Madrid is more *efficient* at converting incremental effort into goals. This could mean a greater conversion of opportunities into goals, better finishing, more breakaways with greater speed, making the most of possession percentage, perhaps by pressing more and playing the ball more in the opponent's half with players who are physically fitter. In contrast, Barcelona required the greatest proportional increase in effort (~1.5 times that for Real Madrid) to obtain a given increase in goals scored. Therefore, everything else remaining the same, equal expression numbers for Barcelona and Real Madrid translated into a greater WGA_i for the latter.

Espanol has a 25% and a 30% lower intercept than either Barcelona or Real Madrid respectively. This significantly lower intercept translates into values of expression numbers that are not significantly different from Barcelona or Real Madrid even with significantly reduced possession percentage, shots on goal and shots. This implies that Espanol makes the most efficient use of the limited possession percentage it does get, in scoring goals. This also means that when it does attack, the probability of that attack resulting in a goal is greater for Espanol, than it is for Barcelona or Real Madrid. Converting possession into goals – execution – is Espanol's strength. However, the spread (range) of expression numbers for any particular goal is greater for Espanol, than for Real Madrid or Barcelona. In fact, the spread for Espanol is so large that there is not a significant difference between the largest expression numbers for 0 and 2 goals. This means that Espanol is inconsistent in its capability to score goals. This inconsistency could be due to the greater pressure to score given that the team has a lesser percent possession. The lack of goals beyond 3 is because Espanol – like Barcelona – too requires a larger proportional increase in effort to obtain a given increase in goals scored (slope = 0.083); an increase that does not materialize in goals due to a smaller possession percentage. It can be speculated that Espanol would play significantly better if it succeeded in keeping possession. In that regard, long passes, passes along the touch line, and pass accuracy may be targeted for improvement.

Limitations

Since the expression number is specific for a particular team as well as for a particular match, it follows that any significant changes in the team such as the roster, player(s) injured, traded or carded; a change in game tactics or strategy, which in turn arise from a change of coaching staff, philosophy or ownership can alter the expression number pattern usually associated with a particular team and goals, going forward. If the change(s) are significant and if chronological data before and after the change is captured and is part of the training set, such change(s) should manifest themselves as a change in the regression coefficient in figures 2, 3 and 4. Consequently, the goal categorized expression number average may not be as highly correlated with the number of goals scored.

The model assumes that the lowest expression number obtained from the training data set for a given team for a particular number of goals scored represents the lowest capability (0) of scoring those many goals, while the largest expression number obtained from the training data set for a given team for a particular number of goals scored represents the highest capability (1) of scoring those many goals. The model also assumes a linear increase of goal scoring capability with an increase in expression number for any particular number of goals.

The number of data points is < 30, the minimum acceptable for statistical evaluation. The number of data points was limited by the number of matches played between the three teams for the 2018-2019 season. Since the data-set was only extracted from the La Liga League for the 2018-2019 season, it may not be applicable to other leagues or during other seasons. The data used for formulating this model was obtained from the website **http://www.whoscored.com**. The source or veracity of the data uploaded to the site could neither be identified or confirmed.

Conclusions

A novel statistical model is proposed that assigns a winner to drawn soccer matches based on a statistical ideal Weighted Goal Average (WGA_i). The factors used in the model, i.e. the total shots, the shots on goal and the possession percentage are used to derive an expression number that is highly correlated (> 83%) to the number of goals scored in the training data set for three teams in the La Liga league for the 2018-2019 season. The model is sufficiently robust such that a > 91% prediction accuracy (11 out of 12 matches) is obtained for non-drawn matches - not part of the training set - with < 35 matches for each team used as the training set. The discriminatory power of the model ranges from the ability to correctly predict the outcome of matches played between the best teams to matches played between the best and worst team in the league. The model can also be used to determine if play quality was consistent with previous games, and serves as a play retrospection, skill-set and tactical development, and coaching tool. The model is a valuable tool to evaluate the match in terms of the training set statistical history, and may identify root-cause events, tactics, strategy or skill-set outliers that contribute to poor play quality and performance in specific matches.

References

1. Csato, L. 2020. “A comparison of penalty design shootouts in soccer.” https://arxiv.org/pdf/1806.01114.pdf

2. Farrell, T. 2008. http://www.theadgalternative.com/index.html#

3. Shannahan, C. 2015. https://sites.duke.edu/wcwp/2015/02/13/the-penalty-shootout-flaws-and-alternatives/

4. Tidley, W. 2012. https://bleacherreport.com/articles/1195909-10-alternatives-to-a-penalty-shootout#slide5

Appendix 1 : R code for calculating the ideal weighted goal average (WGA_i)

```
# This code calculates the ideal weighted goal average for the three LaLiga teams.
# Barcelona's raw data for the training set ----
shots <-
c(13,12,7,15,15,10,17,9,12,16,13,15,21,10,11,16,23,8,8,11,8,17,8,8,10,15,9,18,10,11,13,25,9,9,8)
shots
shotsOnTarget <- c(5,8,6,7,7,2,8,4,6,9,3,9,10,2,4,6,12,1,3,2,4,10,1,4,6,8,4,9,7,6,5,17,5,2,4)
positionPercent <-
c(66,76,59,72,64,67,60,51,69,44,70,68,64,74,50,77,62,65,61,72,75,65,74,57,62,56,56,62,61,77,66,66,4
7,52,69)
goals <- c(1,2,3,3,2,0,4,1,3,4,2,4,2,0,2,2,1,0,2,0,2,5,0,2,2,4,3,5,4,2,1,5,2,0,4)
expressionNumbers <- ((shotsOnTarget/shots)+(shotsOnTarget/positionPercent))
expressionNumbers
barcelonaSoccerData <- data.frame(shots = shots, shotsOnTarget = shotsOnTarget, percentPosition =
positionPercent, Goals = goals,
                                  ExpressionNumber = expressionNumbers )
barcelonaSoccerData
# Real Madrid raw data for the training set ----
rshots <-
c(14,28,14,23,16,22,11,19,19,17,15,22,16,15,13,13,11,12,9,24,6,9,17,17,22,17,25,7,15,14,17,14,22,22,
10,18,12,18,19,20)
rshotsOnTarget <- c(7,8,5,6,9,8,4,7,5,3,6,7,8,5,7,6,4,4,3,11,2,2,4,11,4,5,11,4,4,3,10,4,11,7,5,10,6,8,6,8)
rpositionPercent <-
c(54,59,26,69,64,71,66,57,61,49,62,53,71,68,56,71,57,67,51,60,46,47,48,43,66,56,65,45,55,56,62,63,6
2,63,59,53,64,61,64,69)
rgoals <- c(2,0,2,2,4,3,3,1,2,0,4,2,3,1,2,1,3,0,0,3,1,0,0,3,1,2,3,1,2,0,4,0,5,0,4,3,2,2,1,0)
rexpressionNumbers <- ((rshotsOnTarget/rshots)+(rshotsOnTarget/rpositionPercent))
realmadridSoccerData <- data.frame(shots = rshots, shotsOnTarget = rshotsOnTarget, percentPosition=
rpositionPercent, Goals = rgoals,
                                   ExpressionNumber = rexpressionNumbers)
realmadridSoccerData
# Espaniol raw data for the training set  ----
eshots <-
c(12,21,6,13,17,15,7,13,15,5,13,4,9,10,6,19,9,10,11,11,10,7,5,15,7,17,10,16,11,18,14,13,10,6,12,6,16,9
)
eshotsOnTarget <- c(1,2,0,6,2,5,1,2,6,2,3,2,3,2,2,5,2,4,5,4,4,2,1,3,2,3,4,2,3,3,3,4,1,2,5,2,4,3)
epositionPercent <-
c(52,61,45,36,58,61,49,68,54,35,50,30,51,45,56,50,48,46,50,47,37,45,65,67,41,56,36,59,62,51,48,46,5
0,54,53,39,38,59)
egoals <- c(1,2,0,2,2,2,0,1,3,1,0,0,1,2,2,2,1,1,3,2,2,0,0,0,2,1,1,0,0,0,1,1,1,1,2,0,2,0)
eexpressionNumbers <- ((eshotsOnTarget/eshots)+(eshotsOnTarget/epositionPercent))
espaniolSoccerData <- data.frame(shots = eshots, shotsOnTarget = eshotsOnTarget, percentPosition =
epositionPercent, Goals = egoals,
                                 ExpressionNumber = eexpressionNumbers)
espaniolSoccerData
```

```
# Variables beginning with the prefixes b,r, and e apply to Barcelona, Real Madrid, and Espaniol respectively.
# min and max refer to mimmimum and maximum expression numbers respectively, the number following the min or max represent the number of goals.
# Z is an input expression number for a particular game
# Prod represents the normalized expression number

#subset barca data via goals ----

b0g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 0,] #b0g and so on are barcelona's goals categorized with all other variables
b0g
prodb1g <- -1
prodb2g <- -1
prodb3g <- -1
prodb4g <- -1
prodb5g <- -1

b1g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 1,]
minb1g <- min(b1g[,"ExpressionNumber"])
maxb1g <- max(b1g[,"ExpressionNumber"])
b2g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 2,]
minb2g <- min(b2g[,"ExpressionNumber"])
maxb2g <- max(b2g[,"ExpressionNumber"])
b3g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 3,]
minb3g <- min(b3g[,"ExpressionNumber"])
maxb3g <- max(b3g[,"ExpressionNumber"])
b4g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 4,]
minb4g <- min(b4g[,"ExpressionNumber"])
maxb4g <- max(b4g[,"ExpressionNumber"])
b5g <- barcelonaSoccerData[barcelonaSoccerData$Goals == 5,]
minb5g <- min(b5g[,"ExpressionNumber"])
maxb5g <- max(b5g[,"ExpressionNumber"])
#average of expression numbers within 5 subsets----
byavg <-
c(mean(b0g[,"ExpressionNumber"]),mean(b1g[,"ExpressionNumber"]),mean(b2g[,"ExpressionNumbe
r"]),mean(b3g[,"ExpressionNumber"]),
  mean(b4g[,"ExpressionNumber"]),mean(b5g[,"ExpressionNumber"]))
byavg
bx <- (0:5)
bx
# linear plot with slope, intercept and correlation coefficient ----
plot(bx,byavg)
abline(lm(byavg ~ bx))
beq <- lm(formula = byavg ~ bx)
brsquared <- (cor(bx,byavg))^2
beq
brsquared
```

```
z <- .567

if(z<minb1g) {
  prodb1g <- 0}
if(z>maxb1g) {
  prodb1g <- 1}
if(z>minb1g && z<maxb1g) {
  prodb1g <- z*(maxb1g-minb1g)+minb1g
}

if(z<minb2g) {
  prodb2g <- 0}
if(z>maxb2g) {
  prodb2g <- 1}
if(z>minb2g && z<maxb2g) {
  prodb2g <- z*(maxb2g-minb2g)+minb2g
}

if(z<minb3g) {
  prodb3g <- 0}
if(z>maxb3g) {
  prodb3g <- 1}
if(z>minb3g && z<maxb3g) {
  prodb3g <- z*(maxb3g-minb3g)+minb3g
}

if(z<minb4g) {
  prodb4g <- 0}
if(z>maxb4g) {
  prodb4g <- 1}
if(z>minb4g && z<maxb4g) {
  prodb4g <- z*(maxb4g-minb4g)+minb4g
}

if(z<minb5g) {
  prodb5g <- 0}
if(z>maxb5g) {
  prodb5g <- 1}
if(z>minb5g && z<maxb5g) {
  prodb5g <- z*(maxb5g-minb5g)+minb5g
}


bwga <- (prodb1g+2*prodb2g+3*prodb3g+4*prodb4g+5*prodb5g)/6
bwga
bwgai <- bwga/z
bwgai
#subset Real Madrid data via goals ----
```

```
r0g <- realmadridSoccerData[realmadridSoccerData$Goals == 0,] #r0g and so on are real madrid's
goals categorized with all other variables
r0g

prodr1g <- -1
prodr2g <- -1
prodr3g <- -1
prodr4g <- -1
prodr5g <- -1

r1g <- realmadridSoccerData[realmadridSoccerData$Goals == 1,]
minr1g <- min(r1g[,"ExpressionNumber"])
maxr1g <- max(r1g[,"ExpressionNumber"])
r2g <- realmadridSoccerData[realmadridSoccerData$Goals == 2,]
minr2g <- min(r2g[,"ExpressionNumber"])
maxr2g <- max(r2g[,"ExpressionNumber"])
r3g <- realmadridSoccerData[realmadridSoccerData$Goals == 3,]
minr3g <- min(r3g[,"ExpressionNumber"])
maxr3g <- max(r3g[,"ExpressionNumber"])
r4g <- realmadridSoccerData[realmadridSoccerData$Goals == 4,]
minr4g <- min(r4g[,"ExpressionNumber"])
maxr4g <- max(r4g[,"ExpressionNumber"])
r5g <- realmadridSoccerData[realmadridSoccerData$Goals == 5,]
minr5g <- min(r5g[,"ExpressionNumber"])
maxr5g <- max(r5g[,"ExpressionNumber"])
#average of expression numbers within 5 subsets----
rmyavg <-
c(mean(r0g[,"ExpressionNumber"]),mean(r1g[,"ExpressionNumber"]),mean(r2g[,"ExpressionNumber"
]),mean(r3g[,"ExpressionNumber"]),
        mean(r4g[,"ExpressionNumber"]),mean(r5g[,"ExpressionNumber"]))
rmyavg
rmx <- (0:5)
rmx
# linear plot with slope, intercept and correlation coefficient ----
plot(rmx,rmyavg)
abline(lm(rmyavg ~ rmx))
rmeq <- lm(formula = rmyavg ~ rmx)
rmrsquared <- (cor(rmx,rmyavg))^2
rmeq
rmrsquared


z <- .473

if(z<minr1g) {
  prodr1g <- 0}
if(z>maxr1g) {
  prodr1g <- 1}
```

```
if(z>minr1g && z<maxr1g) {
  prodr1g <- z*(maxr1g-minr1g)+minr1g
}

if(z<minr2g) {
  prodr2g <- 0}
if(z>maxr2g) {
  prodr2g <- 1}
if(z>minr2g && z<maxr2g) {
  prodr2g <- z*(maxr2g-minr2g)+minr2g
}

if(z<minr3g) {
  prodr3g <- 0}
if(z>maxr3g) {
  prodr3g <- 1}
if(z>minr3g && z<maxr3g) {
  prodr3g <- z*(maxr3g-minr3g)+minr3g
}

if(z<minr4g) {
  prodr4g <- 0}
if(z>maxr4g) {
  prodr4g <- 1}
if(z>minr4g && z<maxr4g) {
  prodr4g <- z*(maxr4g-minr4g)+minr4g
}

if(z<minr5g) {
  prodr5g <- 0}
if(z>maxr5g) {
  prodr5g <- 1}
if(z>minr5g && z<maxr5g) {
  prodr5g <- z*(maxr5g-minr5g)+minr5g
}


rwga <- (prodr1g+2*prodr2g+3*prodr3g+4*prodr4g+5*prodr5g)/6
rwga
rwgai <- rwga/z
rwgai
#subset Espanol data via goals ----

e0g <- espaniolSoccerData[espaniolSoccerData$Goals == 0,] #e0g and so on are real madrid's goals
categorized with all other variables
e0g

prode1g <- -1
prode2g <- -1
```

```
prode3g <- -1

e1g <- espaniolSoccerData[espaniolSoccerData$Goals == 1,]
mine1g <- min(e1g[,"ExpressionNumber"])
maxe1g <- max(e1g[,"ExpressionNumber"])
e2g <- espaniolSoccerData[espaniolSoccerData$Goals == 2,]
mine2g <- min(e2g[,"ExpressionNumber"])
maxe2g <- max(e2g[,"ExpressionNumber"])
e3g <- espaniolSoccerData[espaniolSoccerData$Goals == 3,]
mine3g <- min(e3g[,"ExpressionNumber"])
maxe3g <- max(e3g[,"ExpressionNumber"])

#average of expression numbers within 5 subsets----
eyavg <-
c(mean(e0g[,"ExpressionNumber"]),mean(e1g[,"ExpressionNumber"]),mean(e2g[,"ExpressionNumber
"]),mean(e3g[,"ExpressionNumber"]))

eyavg
ex <- (0:3)
ex
# linear plot with slope, intercept and correlation coefficient ----
plot(ex,eyavg)
abline(lm(eyavg ~ ex))
eeq <- lm(formula = eyavg ~ ex)
ersquared <- (cor(ex,eyavg))^2
eeq
ersquared
e1g[,"ExpressionNumber"]
x <- 1:12
min(e1g)
max(e1g)
abline(lm(e1g[,"ExpressionNumber"] ~ x))
plot(x,e1g[,"ExpressionNumber"])


z <- .622

if(z<mine1g) {
  prode1g <- 0}
if(z>maxe1g) {
  prode1g <- 1}
if(z>mine1g && z<maxe1g) {
  prode1g <- z*(maxe1g-mine1g)+mine1g
}

if(z<mine2g) {
  prode2g <- 0}
if(z>maxe2g) {
  prode2g <- 1}
```

```
if(z>mine2g && z<maxe2g) {
  prode2g <- z*(maxe2g-mine2g)+mine2g
}

if(z<mine3g) {
  prode3g <- 0}
if(z>maxe3g) {
  prode3g <- 1}
if(z>mine3g && z<maxe3g) {
  prode3g <- z*(maxe3g-mine3g)+mine3g
}


ewga <- (prode1g+2*prode2g+3*prode3g)/4
ewga
ewgai <- ewga/z
ewgai
```