CSCanada
Progress in Applied Mathematics
Vol. 7, No. 2, 2014, pp. 1-8
DOI: 10.3968/4886

ISSN 1925-251X [Print]
ISSN 1925-2528 [Online]
www.cscanada.net
www.cscanada.org

Statistical Analysis for the First Bundesliga in the Current Soccer Season

Holger Broich[a], Joachim Mester[b], [c], Florian Seifriz[b], [c], and Zengyuan YUE[b], [c],*

[a]Bayer 04 Leverkusen Fussball GmbH, BayArena, Bismarckstr. 122 – 124, 51373 Leverkusen, Germany.
[b]Institute of Training Science and Sport Informatics, German Sport University Cologne, Carl-Diem-Weg 6, 50933 Cologne, Germany.
[c]The German Research Centre of Elite Sport.
*Corresponding author.

Address: Institute of Training Science and Sport Informatics, German Sport University Cologne, Carl-Diem-Weg 6, 50933 Cologne, Germany. E-mail: z.yue@dshs-koeln.de
Received: October 14, 2013/accepted: January 15, 2014/Published online: April 26, 2014

Abstract: Statistical analysis for the 153 matches of First Bundesliga, i.e. the first national soccer league in Germany, in the first 17 "playing days" (August 10, 2013 to January 29, 2014) of the current soccer season was made. Various team parameters were compared between the winning and losing teams in the 118 non-drawing matches. The results support the conclusions of our earlier analysis (Yue, Broich, & Mester, 2014) that *the quality of shots, represented by the goal efficiency, defined by the number of goals divided by the number of shots, is more important than the quantity of shots for winning a soccer game*. This conclusion is also confirmed by the correlation analysis based on all the 153 matches: The correlation between the number of goals and the goal efficiency is found to be much stronger than the correlation between the number of goals and the number of shots. The team parameters of the second to the fourth importance are the number of shots, the number of passes and the number of ball contacts respectively. In contrast, the distance coverage is found to be statistically not important for winning a game.
Key words: Soccer matches; Satistical analysis; Goal efficiency; Coaching; Bundesliga

Holger, B., Joachim, M., Florian, S., & Yue, Z. Y. (2014). Statistical Analysis for the First Bundesliga in the Current Soccer Season. *Progress in Applied Mathematics, 7*(2), 1-8. Available from http://www.cscanada.net/index.php/pam/article/view/10.3968/4886 **DOI:** 10.3968/4886

1. INTRODUCTION

Compared to the world-wide popularity of soccer games, the performance analysis for soccer game is always very limited because of the complexity of the game. The motions of players

are not deterministic like in a mechanical system, but always with some random factors. Therefore, statistical analysis has been the only approach used in this field up to now. Such statistical analysis can be traced to the pioneering work of Reep & Benjamin[1] based on a large amount of matches between 1953 and 1968. Their results favoured "direct play", i.e. shorter passing sequence, rather than "possession play", i.e. longer passing sequence. The results of further researches in this field (Bate[2], Hughes, Robertson, & Nicholson[3], Hughes & Franks[4]) were controversial regarding whether "direct play" (also called "direct attack") or "possession play" (also called "elaborate attack") would be more effective. Different data from different samples of matches, as well as different statistical methods, used in these researches may contribute to the controversy. Even based on the same data, different statistical methods may lead to different conclusions (Lago-Ballesteros, Lago-Penas, & Rey[5]). In addition to different samples of matches and different statistical methods, there could be a more basic reason for the controversy, as we pointed out in our earlier paper (Yue, Broich, & Mester[6]), that "direct play" or "possession play" each has advantages and disadvantages, and some other factors must be involved in determining which one would be more effective. Thus, we cannot simply draw a general conclusion. We therefore suggested a two-steps strategy to answer the question what kind of tactics would be favourable for winning a soccer game (Yue, Broich, & Mester[6]). The first step would be to find out which factors are more important for winning a soccer game by directly analysing a large number of matches instead of a large number of team possessions as many researchers did before, and the second step would be to find out what kind of tactics, not just the length of passing sequence, would be favourable for raising the most important factor obtained by the first step so as to raise the chance of winning. The major finding of our earlier paper (Yue, Broich, & Mester[6]) was that the quality of shots, represented by the goal efficiency, defined by the number of goals divided by the number of shots, is more important than the quantity of shots for winning a soccer game. However, our earlier analysis (Yue, Broich, & Mester[6]) was only based on the 50 non-drawing matches where the difference of the numbers of goals between the winning and the losing teams was not smaller than 2 among the 126 matches of the First Bundesliga during the period August 5 to Novenber 27, 2011. Thus the question remains: If all the non-drawing matches were included, could we still draw the same conclusion? The purpose of the present analysis is to answer this question by including all the 118 non-drawing matches during the 17 "playing days" (August 10, 2013 to January 29, 2014) of the current Soccer Season of First Bundesliga. The methods and the results will be shown in the next two sections respectively. Some discussions will then be made, and conclusions will be eventually drawn.

2. METHOD

2.1 Data Source and the Conversion of The Data to Convenient Format

Although the data of player's statistics of all the matches in the current Soccer Season of First Bundesliga could be found in the official website (http://www.bundesliga.de/de/) , it would be tremendous amount of manual work to copy all the numbers by hand. Therefore, we used an alternative approach. We started from the PDF files of the data for all the team parameters and all the 153 matches of the 17 "playing days" (August 10, 2013 to January 29, 2014) of the current Soccer Season of First Bundesliga, and converted these files to EXCEL files by a special program. We can then compare these team parameters between the winning and the losing teams to find out which team parameters are more important for the result of the match. Some simple descriptive statistics could be obtained directly from these EXCEL

files, while more professional analyses, e.g. *T*-tests and correlation analysis were carried out by professional statistical package, e.g. STATISTICA.

2.2 Samples of Matche

Unlike the sample used in our earlier paper (Yue, Broich, & Mester[6]) where only the matches with a goal difference not smaller than 2 were included, the present analysis take all the non-drawing matches into account. Namely we only require

$$\Delta G=G_A-G_B \geq 1 \quad , \tag{1}$$

where G means the number of goals, and the subscripts A and B stand for the winning and the losing teams respectively. Among the 9×17=153 matches in the first 17 "playing days" of the current soccer season of First Bundesliga, there were 118 matches which meet condition (1).

2.3 Notations

In the present analysis, we use the same notations as in our earlier paper (Yue, Broich, & Mester[6]) (except *OO*, see the explanation below):

A: the *winning team*;

B: the *losing team*;

G: the *number of goals* of the team in the match;

E: the *goal efficiency* defined by the number of goals divided by the number of shots of the team in the match;

S: the *number of shots* of the team in the match;

P: the *number of passes* of the team in the match;

C: the *number of ball contacts* of the team in the match;

D: the *distance coverage* of the team in the match;

R: the *number of sprints* of the team in the match;

OO: the percentage of winning *"one and one"* of the team in the match;

(Note, in our earlier paper (Yue, Broich, & Mester[6]) *OO* meant the *number of "one and one"* of the team in the match.)

M: *mean* over all the matches;

SD: *standard deviation* over all the matches;

M_d: *mean of differences* between the winning and the losing teams over all the matches;

SD_d: *standard deviation of differences* between the winning and the losing teams over all the matches;

p: *significance level* of the difference by paired *T*-test. The difference would be regarded as significant if $p < 0.05$;

r: *Pearson's correlation coefficient*, which would be regarded as significant if p (for r)< 0.05; *q*: *relative size of the difference*, defined by

$$q = (M_A - M_B)/[\frac{1}{2}(M_A + M_B)] \quad , \tag{2}$$

i.e. the difference of the means between the winning and the losing teams divided by the mean of the means.

From the definition of *goal efficiency E*, we have

$$G=S\times E \quad , \tag{3}$$

S and *E* can be considered as the quantity and the quality of the shots respectively. Equation (3) itself does not tell which one of *S* and *E* is more important for winning a game. Our earlier analysis (Yue, Broich, & Mester[6]) based on the reduced sample with $\Delta G \geq 2$ pointed out that *E* is more important than *S*. Now we will examine whether this is still true when we use the full sample of non-drawing matches with $\Delta G \geq 1$.

As in our earlier paper (Yue, Broich, & Mester[6]), *p* and *q* will be used to judge the importance of each of the team parameters. The value of *p* tells how true the difference of the team parameter between the two teams is. The exact mathematical meaning of *p* is the probability of making *Type I error*, i.e. claiming a difference when it does not exist. Usually, a difference could be asserted as significant if $p<0.05$. Under this condition, we would have more than 95% confidence to believe that the difference found in the samples is a true effect rather than just by chance. A significant difference does not necessarily mean that the difference is big or important particularly for the case of big samples. For very big samples, such as the one we use in this paper (N=118), small difference could become very significant. Therefore, we need another parameter *q* to characterize the relative size or importance of the difference. *For two team parameters, if $p_1<p_2$ and $q_1>q_2$, parameter 1 would be obviously more important than parameter 2; otherwise, if p_1 and p_2 are both smaller than 0.05, or very close to each other, the parameter with larger q would be regarded as more important.* We will use this criterion to determine the order of importance for all the team parameters, and in particular, to judge which team parameter is the most notable one, for the result of the game.

3. RESULT ANALYSES

The results of the statistical analysis over all the 118 non-drawing matches of First Bundesliga during the 17 "playing days" (August 10, 2013 to January 29, 2014) are summarized in Table 1.

Table 1
Statistics Over the 118 Non-Drawing Matches of the First Bundesliga (August 10, 2013 to January 29, 2014)

Parameter	Teams	*M*	*SD*	M_d	SD_d	*p*	*q* (%)	*r*
E	A	0.1866	0.0904	0.1272	0.0998	0.000000*	103.4	0.32*
	B	0.0594	0.0800					
S	A	15.08	5.2412	3.14	8.37	0.000082*	23.28	-0.42*
	B	11.93	4.6680					
P	A	440.22	139.68	58.93	214.34	0.003435*	14.35	-0.57*
	B	381.29	101.18					
C	A	643.16	130.72	60.64	213.62	0.002553*	9.89	-0.67*
	B	582.53	102.75					
D (km)	A	118.04	3.78	0.98	3.45	0.002620*	0.83	0.68*
	B	117.06	4.63					
R	A	207.74	25.40	1.64	24.89	0.4768	0.79	0.57*
	B	206.10	28.10					
OO(%)	A	51.18	4.29	2.35	8.58	0.003516*	4.71	-1.0*
	B	48.82	4.29					

Note. A and B mean the winning and the losing teams, respectively. The values of *p* with stars mean significant differences ($p<0.05$). The values of *r* with stars mean significant correlations (*p* for *r*, which is not shown in this table, is smaller than 0.05).

The results shown in Table 1 are similar to those obtained in our earlier paper (Yue, Broich, & Mester[6]): The goal efficiency *E* is still by far the most important team parameter

to determine the result of the match because the winning teams' mean goal efficiency is significantly much higher than the losing team (q =103.4%, $p < 0.000001$). The team parameter of the second importance is the number of shots S (q=23.28%, p=0.000082). The team parameters of the third and the fourth importance are the number of passes P (q = 14.35%, p = 0.003435) and the number of ball contacts C (q = 9.89%, p = 0.002553) respectively. The remaining three team parameters (D, R and OO) are not as important as the first four because the relative sizes are all smaller than 5%, in which only for OO (the percentage of winning the "one and one") we have q=4.71%, while for D (the distance coverage) and R (the number of sprints), we both have q<1%. Note, the significance of the small difference for D (mean values 118.04 for winning teams vs. 117.06 for losing teams) is caused by the large sample (N=118). However, the small relative size (q=0.83%) indicates that D is not an important parameter for winning a soccer game.

The last column of Table 1 shows the correlation coefficients between the winning and the losing teams for various team parameters. The negative correlations for S, P and C and the positive correlations for D and R can all be well understood as we explained in our earlier paper (Yue, Broich, & Mester[6]): S, P and C are competitive parameters in the sense that the winning team's better performance related to these parameters would tend to suppress the performance of the losing team. For example, when the winning team has more shots, the losing team has to spend more time for defending. This would tend to reduce the number of shots of the losing team. Similarly, the more passes and ball contacts the winning team has, the fewer passes and ball contacts the losing team would tend to have. This explains the negative correlations of S, P and C between the winning and the losing teams. The positive correlations for D and R can be understood as follows. If the players of one team run faster back and forth, covering bigger distances in the field, the players of the other team would have to follow and run faster, covering bigger distances in the field too. The same is true for the number of sprints R. The correlation coefficient r=-1 for OO is simply caused by the definition of OO in this paper: It is now the percentage of winning the "one and one", rather than the number of winning "one and one". Therefore we have OO_A+OO_B=1, and we must have r=-1.

As an alternative way of comparing the importance of different team parameters, we have also calculated the correlation coefficients between G and each of E, S, and D over the entire 153 matches, shown in Figures 1, 2, and 3, respectively.

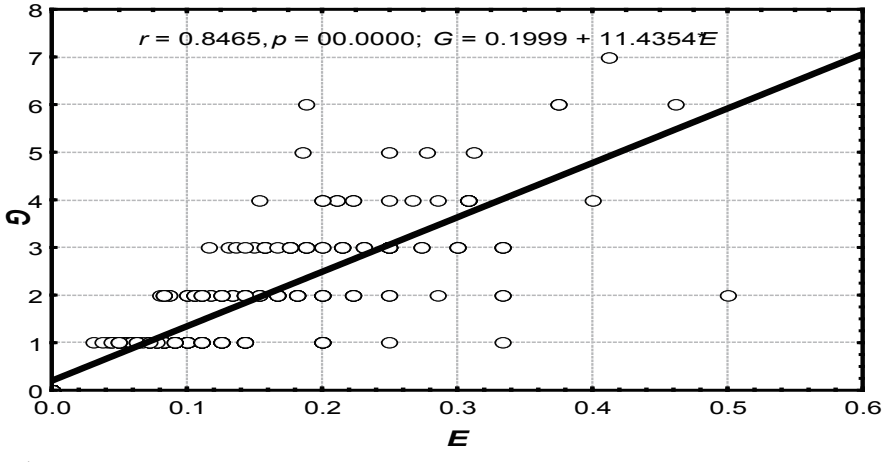

Figure 1
Correlation Between the Number of Goals (G) and the Goal Efficiency (E)

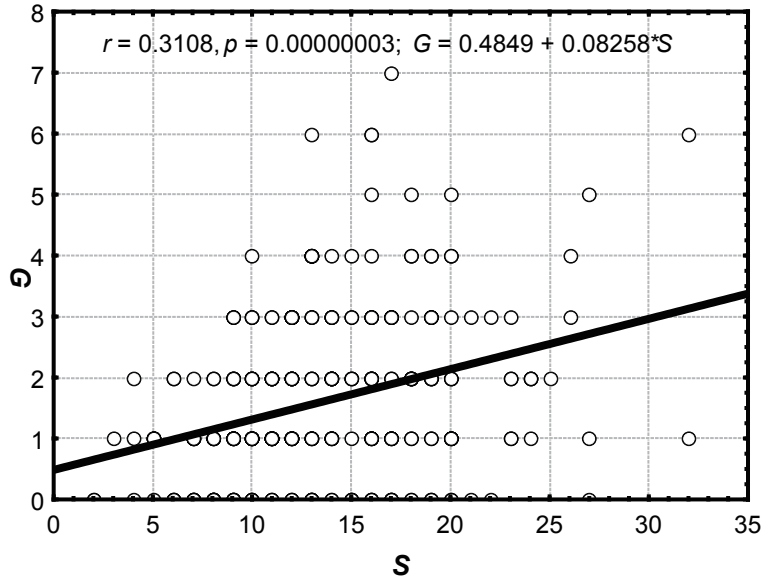

Figure 2
Correlation Between the Number of Goals (G) and the Number of Shots (S)

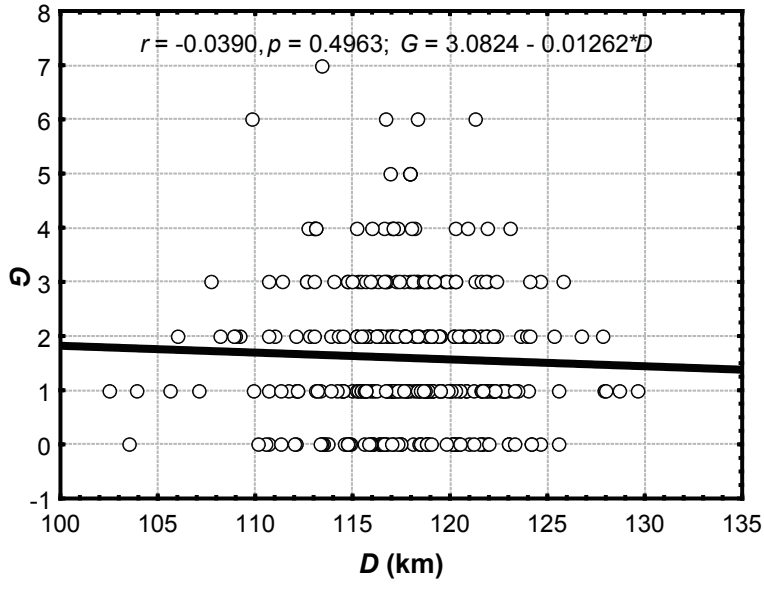

Figure 3
Correlation Between the Number of Goals (G) and the Distance Coverage (D)

From Figure 1 and Figure 2 we see that the number of goals *G* has stronger correlation with *E* (r=0.8465) than with *S* (r=0.3108). Again we see that, compared to the quantity of shooting (*S*), the quality of shooting (*E*) has a more important influence to the number of

goals. Fig. 3 shows a very weak (even negative) and very insignificant correlation between the number of goals G and the distance coverage D (r=-0.0390, p=0.4693), showing that the distance coverage is indeed not statistically important for winning a game.

4. DISCUSSIONS

The present analysis and our earlier paper (Yue, Broich, & Mester[6]) both confirm that the goal efficiency E is by far the most important team parameter for winning a game. Thus, the important question is how to raise the goal efficiency. Although some results in the current literature, e.g. Hughes & Franks[4], show that "direct play" has higher goal efficiency than "possession play", the length of passing sequence is certainly not the only factor, and perhaps not the major factor, to influence the goal efficiency. In order to compare the possible influences of various factors to the goal efficiency, we calculated the correlation coefficients between the goal efficiency E and each of various team parameters, e.g. *SH_G* (shots on the goal), *SH_inner* (shots in the penalty box), *SH_out* (shots out of the penalty box), *SH_head* (shots with head), *Corner*, as well as *S, P, C, D, R, OO* defined in the method section. We found that the goal efficiency E has the strongest and most significant correlation with *SH_G* (r=0.2847, p=0.0000004) compared with other team parameters. This suggests that, in order to raise the goal efficiency E, it is important in order to create more shots on goal from close range under favourable shooting conditions. The creation of favourable shooting conditions depends upon the velocities of the players, the ability of getting rid of the defenders, the coordination among the forwards, the pattern of the collective behaviours of all forwards, and so on. Although some analyses on the playing styles (Pollard, Reep, & Hartley[7]), on the classification of kicked shots and headed shots (Pollard & Reep[8]), and on the techniques of attacking and shooting in general (Hughes[9]) are available in the literature, it is still an open question what kind of tactics, not just the length of passing sequences, would be favourable for the goal efficiency. The overall team parameters alone may not be sufficient to respond to answer this question. More detailed data for the detailed motions of all the players and the ball, as we used in our former studies (Yue et al. [10-12]) but unfortunately only for a single match, would be very helpful, if available for many matches.

5. CONCLUSION

By taking all the 118 non-drawing matches ($\Delta G \geq 1$) during the 17 "playing days" (August 10, 2013-January 29, 2014) of the current Soccer Season of First Bundesliga into account, the present statistical analysis further confirms our earlier conclusion (Yue, Broich, & Mester[6]), which was based only on 50 matches with $\Delta G \geq 2$, that the *goal efficiency*, defined by the number of goals divided by the number of shots, is by far the most important team parameter for the result of the match. This means that *the quality of shots is more important than the quantity of shots for winning a soccer game*. The second important team parameter is the *number of shots*. The team parameters of the third and the fourth importance are the *number of passes* and the *number of ball contacts* respectively. Although the present result favours "direct play", which has higher goal efficiency according to the statistical analysis in the literature, than "possession play", the shooting condition, including the location and the situation of the shooting, and the associated

"last few passes" before the shooting may be more influential to the goal efficiency than the length of the passing sequence alone. Much further analysis on the influence of various tactics to the goal efficiency, based on more detailed data on the detailed motions of all the players and the ball for a large number of matches, remains to be carried out. One additional conclusion is that the distance coverage is statistically not significant for winning a soccer game.

REFERENCES

[1] Reep, C., & Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society, A*, *131*, 581-585.

[2] Bate, R. (1988). Football chance: Tactics and strategy. In T. Reilly, A. Lees, K. Davids, & W. J. Murphy (Eds.), *Science and Football* (pp. 293-301). London: E & FN Spon.

[3] Hughes, M., Robertson, K., & Nicholson, A. (1988). Comparison of patterns of play of successful and unsuccessful teams in the 1986 world cup for soccer. In T. Reilly, A. Lees, K. Davids & W. J. Murphy (Eds.), *Science and football* (pp. 363-367). London: E & FN Spon.

[4] Hughes, M., & Franks, I. (2005). Analysis of passing sequences, shots and goals in soccer. *Journal of Sports Sciences*, *23*, 509-514.

[5] Lago-Ballesteros, J., Lago-Penas, C., & Rey, E. (2012). The effect of playing tactics and situational variables on achieving score-box possessions in a professional soccer team. *Journal of Sports Sciences*, *30*, 1455-1461.

[6] Yue, Z., Broich, H., & Mester, J. (2014). Statistical analysis for the soccer matches of first Bundesliga. *International Journal for Sport Science and Coaching*, *9*(3), 553-560.

[7] Pollard, R., Reep, C. & Hartley S. (1988). The quantitative comparison of playing styles in soccer. In T. Reilly, A. Lees, K. Davids, & W. Murphy (Eds.), *Science and football* (pp. 309-315). London: E & FN Spon.

[8] Pollard, R., & Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *The Statistician*, *46*, 541-550.

[9] Hughes, C. (1980). *The football association coaching book of soccer tactics and skills*. London: Queen Anne Press.

[10] Yue, Z., Broich, H., Seifriz, F., & Mester, J. (2008a). Mathematical analysis of a soccer game. Part I: Individual and collective behaviours. *Studies in Applied Mathematics*, *121*, 223-243.

[11] Yue, Z., Broich, H., Seifriz, F., & Mester, J. (2008b). Mathematical analysis of a soccer game. Part II: Energy, spectral and correlation analyses. *Studies in Applied Mathematics*, *121*, 245-261.

[12] Yue, Z., Broich, H., Seifriz, F., & Mester, J. (2011). Kinetic energy analysis for soccer players and soccer matches. *Progress in Applied Mathematics*, *1*, 98-105.