Analysis of the winning probability and the scoring actions in the American professional soccer championship

Análisis de la probabilidad de ganar y de las acciones que conducen al gol en la liga americana de fútbol profesional

Albert Altarriba-Bartés[1,2]; M. Luz Calle[3]; Antonio Susín[4]; Bruno Gonçalves[5]; Marc Vives[6]; Jaime Sampaio[5], & Javier Peña[1,2]

1. Sport and Physical Activity Studies Center (CEEAF), Universitat de Vic-Universitat Central de Catalunya, Vic. Barcelona. Spain
2. Sport Performance Analysis Research Group (SPARG), Universitat de Vic-Universitat Central de Catalunya, Vic. Barcelona. Spain
3. Biosciences Department, Universitat de Vic-Universitat Cental de Catalunya, Vic. Barcelona. Spain.
4. Math Department, UPC-BarcelonaTech, Barcelona. Spain
5. Research Centre for Sport Sciences, Health and Human Development (CIDESD-FCT), Universidade de Trás-os-Montes e Alto Douro (UTAD). Portugal
6. Sports Performance Department. Institut Nacional d'Educació Física de Catalunya (INEFC), Universitat de Barcelona (UB), Barcelona. Spain

Abstract

This study aimed to assess the effect of the scoring moment on the conditional probability of winning or losing a professional soccer match, as well as identified the most influential variables contributing to victory in Major League Soccer (MLS), the men's professional league in the United States. Data from 680 matches played in the 2015 and 2016 regular seasons were analysed, by dividing the matches into fifteen-minute periods. Additionally, the influence of playing home or away on the match outcome and the type of technical-tactical actions that lead to a goal was also analysed. The temporal analysis revealed that scoring first increased the probability of winning the match significantly and showed dependency on the time in which the goal was scored. The two principal components of the principal component analysis (PCA) were counterattacks (PC1) and crosses (PC2). These were the most critical variables during open play to determine how MLS teams scored goals. Nevertheless, scoring first, playing as a home team always gave a better chance to win the game than scoring first and playing away (0.72 vs. 0.32 probability). As the match approached the end, winning or losing was even more determinant and less reversible (0.85 vs. 0.72 for the home and away team respectively when they were ahead on the score in the minute 75 or later). These findings can contribute to a better understanding of the performance indicators in professional soccer, helping coaches to determine the right strategies and improving the tactical patterns to succeed in competition.

Key words: goals; principal component analysis; coaching; team sports.

Resumen

Este estudio tiene como objetivo evaluar el efecto del momento de anotar un gol en la probabilidad condicional de ganar o perder un partido de fútbol profesional, así como identificar las variables más influyentes que contribuyen a la victoria en la Major League Soccer (MLS), la liga masculina profesional de los Estados Unidos. Se analizaron los datos de los 680 partidos jugados durante las temporadas regulares de 2015 y 2016, dividiendo los partidos en períodos de quince minutos. Además, también se analizó la influencia de jugar en casa o fuera en el resultado final del partido y el tipo de acciones técnico-tácticas que conducen a lograr marcar un gol. El análisis temporal reveló que conseguir marcar un gol antes que el rival aumentaba significativamente la probabilidad de ganar el partido y mostró dependencia del período de tiempo en el que se anotó éste. Los dos componentes principales del análisis de componentes principales (PCA) fueron los contraataques (PC1) y los centros (PC2). Estas fueron las variables más determinantes durante el juego abierto para determinar cómo anotan los goles los equipos de la MLS. Sin embargo, anotar primero, jugando como local siempre obtuvo mayores posibilidades de ganar el partido que anotar primero jugando como visitante (probabilidad 0,72 vs 0,32). A medida que el partido se acercaba al final, ganar o perder era aún más determinante y menos reversible (0,85 frente a 0,72 para el equipo local y visitante respectivamente, cuando se tenía ventaja en el marcador en el minuto 75 o posterior). Estos hallazgos pueden contribuir a una mejor comprensión de los indicadores de rendimiento en el fútbol profesional, ayudando a los entrenadores a determinar las estrategias correctas y mejorando los patrones tácticos para tener éxito en la competición.

Palabras clave: goles; análisis de componentes principales; entrenamiento; deportes de equipo.

Correspondence/correspondencia: Albert Altarriba-Bartés
Systems Biology Department, Universitat de Vic-Universitat Cental de Catalunya, Vic. Spain
Email: albert.altarriba@uvic.cat

Introduction

To score a goal is the primary offensive aim of a soccer team. The difficulty of scoring when playing at the highest level is what gives this sport its marked character of uncertainty and complexity. Due to this importance, it has been attempted in the past to define the factors that lead to the achievement of a goal in elite sport.

In 2010, Lago-Peñas; Lago-Ballesteros; Dellal, and Gomez (2010) conducted a study about the match-related statistics that differentiated the winning teams in the Spanish Soccer league from the ones that drew or lost the matches. They identified the total shots, shots on goal, crosses, crosses against, ball possession and venue as the significant key variables, with the statistics related to the efficacy of shots on goal and the generation of scoring opportunities the differential factor between teams. Similarly, Castellano; Casamichana, and Lago (2012) analysed 177 matches played during 2002, 2006 and 2010 World Cup tournaments. In their findings, total shots, shots on target and ball possession were the variables related to the offensive actions that best set winning teams apart from those that drew and lost. Again, the effectiveness of attacking actions seemed to be the primary explanatory factor of a victory in soccer. Moura; Martins, and Cunha (2014) using a principal component analysis (PCA) and a Silhouette Coefficient calculation deduced the number of shots, number of shots on target, playing time with ball possession and percentage of ball possession as the most critical discriminating factors in the match outcome. Their PCA analysis revealed that the first two principal components (shots and shots on target) retained more than 64% of the total variance, contributing significantly to the final score. Recent research by Liu; Gómez; Gonçalves, and Sampaio (2016) assessing the Spanish male first division and, including data from 320 close matches from the 2012-2013 season found a moderate positive effect from the variable shots on target (3.4 extra wins per 10 matches; 99% confidence limits ±1.0), and a small positive effect from the number of total shots performed by the team (1.7 extra wins; ±1.0). The research mentioned above suggests that better values in the highlighted performance factors are essential to winning a soccer match. Nonetheless, consistency in the match actions seems to play also an important role. Shafizadeh; Taylor, and Peñas (2013) suggest that teams with better position at the end of a tournament, show a higher degree of achievement in all performance indicators (goal-related, offensive-related and defensive-related) than their opponents. The differences not only were observed in individual matches but across the overall men UEFA European Championship 2012 tournament.

The importance of some other indicators on the outcome of a soccer match is yet to be clearly elucidated. Saavedra; Aguilar; Marques; Tobío, and Romero (2013) attempted to calculate home advantage in different Union of European Football Associations (UEFA) professional leagues. The authors showed that home advantage was significant in 32 of the 52 analysed UEFA Championships and that its value was 55.6% during the first decade of the 21st Century. Home-advantage describes the benefit that the home team has over the visiting team and is commonly calculated by dividing points won at home by total points won, home and away (Gómez and Pollard, 2014).

Collet (2012) explored the impact of different performance indicators in the English Premier League, the French "Ligue 1", the German "Bundesliga", the Italian "Serie A" and the Spanish "La Liga", and observed the relationship between ball retention, possession time, and team success in domestic leagues. In the study, the teams that held the ball for longer achieved a significantly higher number of points. However, when the highest quality teams, were omitted or faced each other, the importance of possession was less influential. When team quality was included in the analysis as an independent variable, the impact of higher possession became negative. This effect was stronger in the German and French championships than in the English and Spanish leagues. In the Italian championship and the Champions League, the effect of possession time was negligible on the match outcome. In the end, teams maintaining a high percentage of ball possession develop different strategies to achieve success when compared to low percentage possession teams. Ball possession, however, is influenced by situational variables (such as score and match-status) and not only by the playing style (Bradley, Lago-Peñas, Rey, and Sampaio, 2014).

To date, several studies have examined the temporal distribution of the goals scored during a soccer match. Mitrotasios and Armatas (2014) assessing 31 matches from the European Football Championship played in Poland and Ukraine in 2012, found that 57.9% of the goals in the tournament were scored during the second half and most of them (21.1%) in the last 15-minute period. The majority were scored during open plays (72.4%) with the remaining 27.6%, from set plays, such as corners and free kicks. Leite (2013) studying the same European tournament also observed that the team scoring first managed to win the match in 70.9% of the matches. In the same study, when considering the attack strategy used to score a goal, an organised attack was the most successful instrument (44.1% of the cases). Research from other major European Leagues also shows a higher goal-scoring frequency in the second half of the English, French, Italian and Spanish professional championships (55.1%) with, the last 15 minutes of the matches presenting the highest scoring rate again (Alberti, Iaia, Arcelli, Cavaggioni, and Rampinini, 2013). Firdaus; Ali; Katis; Patsika, and Kellis (2015), analysing the Greek soccer tournament also highlighted a significant increase in the frequency of goal scoring during the second half and at the extra time of each half of the matches and pointed at kicking (with 2/3 of the total scored goals) and the header (18.5% of the scored goals) as the most relevant technical skills used. The researchers found a higher percentage of goals scored (+21%) when teams were playing in their home pitch.

Different data analysis techniques have been used in the past to try to establish a deeper understanding of soccer, complementing practical knowledge. However, not much attention has been given to the temporality of the winning probabilities, in a way that can increase the understanding of how scoring has different effects across game time. Our study aimed to assess the effect of the scoring moment on the conditional probability of winning or losing a soccer match and the influence of home-field advantage, as well as identifies the most influential variables in the victory of the teams of the American Major League Soccer (MLS).

Methods

Sample

A total of 1893 goals were observed and coded from 680 official matches played over the 2015-2016 regular seasons of the MLS. This study meets the ethical standards in sports and exercise science research (Harris and Atkinson, 2015). The information and data used were obtained from the Wyscout platform (Wyscout Spa, Italy).

Procedure

We produced our dataset consisting of the key moment of the goal (non-tracking variables) and the situational variables (Table 1). In order to classify every goal in each of the different non-tracking variables, two experienced analysts (professional MLS analysts with more than five years of experience in soccer performance analysis) were trained to identify the absolute key moment in the play before a goal was scored. Both analysts encoded all goals separately, using Hudl Sportscode (Agile Sports Technologies Inc, USA), and compared the individual outcome of each of them. Comparisons were made between observers, and the goal indicators with the minimum Cohen's kappa value for all variables exceeded 0.90, demonstrating high inter-rater reliability (O'Donoghue, 2010).

Variables

The information collected from the MLS games included 12 variables selected as non-tracking variables in the analyses. Following Leite (2013), these 12 variables were grouped into three different attack modes employed to score. Additionally, the analysed variables included three situational variables (Table 1).

Table 1. Type of variables, description and modes of attack following Leite (2013) criteria.

Type of variable	Name (Variable)	Description	Modes of attack
Non-tracking	Crosses (Crosses)	Number of goals scored from open-play crosses (excluding goals from corner or free-kick).	Organized attack (OA)
	Combinations (Possession...Combination)	Number of goals scored after an attack characterized by short, quick passes and movement between two or three players.	
	Possession and penetrating passes (Possession...Penetrating.Pass)	Number of goals scored after an attack characterized by the team in possession playing passes between and behind opponent's lines.	
	Direct plays (Direct)	Number of goals scored after an attack characterized by long passes played into a target forward or behind the defensive line for an attacker to run onto.	
	Individual plays (Individual)	Number of goals scored after an attack relying upon the individual skill of a player in 1v1 situations.	
	Quick restarts (Quick.Restarts)	Number of goals scored after the action to quickly take a direct or indirect free kick, thereby not allowing the defense to recover or set up, reorganize.	
	Own goals (Own.Goals)	Number of goals scored when a player from the defending team accidentally puts the ball into their own goal.	
	Corners kicks (Corner)	Number of goals scored from corner kick plays.	Stopped balls (SB)
	Free kicks (Free. Kick)	Number of goals scored from free kick plays.	
	Throw-ins (Throw.Ins)	Number of goals scored from throw-ins.	
	Penalty kicks (Penalty.Kicks)	Number of goals scored from penalty kicks.	
	Counter-attacks (Counters)	Number of goals scored after a quick transition from defense into attack, exploiting the space behind the team that lost the ball.	Counterattacks (CA)
Situational	Match outcome	The outcome of the match according to the local team: win, lose, draw.	-
	Number of goals	Total number of goals scored and conceded by the local team.	
	Moment of the goal	Minute that each goal was scored or conceeded by the local team.	

Statistical analysis

To assess the importance of playing home or away, and the effect of scoring before the opposing team, initial and conditional probabilities of winning the match for the local and visitor teams were estimated. The initial probability (at time 0) that the local team would win the match was estimated as the relative frequency of matches won by the local team in our dataset (number of matches won by a local team divided by the total number of matches). Similarly, the initial probability that the visitor team will win the match was also estimated (number of matches won by a visitor team divided by the total number of matches). These two probabilities do not add a total of 1 because some matches ended in a draw. For times "t" equal to 15, 30, 45, 60 or 75 minutes after the start of the match, it was calculated the conditional probability that the local team will win the match given that the local team was winning at a time "t". This conditional probability was computed as the relative frequency of matches won by the local team among those that the local team was winning at time "t". For instance, in our database, the local team was winning at minute 15th in 111 matches, and among these, 80 ended with its victory. Thus, the conditional probability that the local team wins the match conditioned to the fact that the local team is winning at minute 15th is equal to 80/111=0.72. On the other hand, the visitor team was winning at minute 15th in 90 matches and among these 43 ended with the victory of the local team. In this case, the conditional probability that the local team wins the match conditioned to the fact that the visitor team is winning at minute 15th is equal to 43/90=0.47. Similarly, the other conditional probabilities were also estimated.

We applied three multivariate exploratory techniques to analyse the information available on the 12 performance variables of every team: Principal Component Analysis (PCA), hierarchical clustering, and heatmap.

PCA is a dimension-reduction approach that transforms some correlated variables into a smaller number of uncorrelated variables called principal components. Each principal component is a linear combination of the initial variables. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability. PCA provides a low-dimensional representation of the data that contains as much as possible of the variation, and thus, it is a useful tool for data visualization and exploration of possible data patterns. (Jackson, 1991, Federolf, Reid, Gilgien, Haugen and Smith, 2012, Jolliffe, 2002, Jolliffe, 2016). We used a PCA to explore similarities between the different teams according to their performance.

Hierarchical clustering is a tree-like visual representation of the observations, called a dendrogram, that allows viewing the data organized into groups or clusters according to their degree of similarity. The height in the dendrogram represents the distance between the teams: the smaller the height, the more similar the teams are. (Hastie, Tibshirani and Friedman, 2009, Johnson, 1967, Li, Chiusano and D'Elia, 2010, Prabu, Sudhaghar, Viswajith, Venkata and Srikaanth, 2019).

A heatmap is a useful graphical representation of a dataset where colours represent the different values. A heatmap plot with hierarchical clustering of both the teams and the performance variables was also performed. (Wilkinson and Friendly, 2009).

All statistical analyses were performed with the statistical package R (The R Foundation for Statistical Computing, Vienna, Austria), version 3.5.0.

Results

Data analysis (Table 2) showed that a total of 1893 goals were accounted for during 2015 and 2016 seasons: 820 (43.3%) in the 1st half (246 between 1st-15th min, 263 between 16th-30th min, 311 between 31st-45th min) and 1073 (56.7%) in the 2nd half (325 between 46th-60th min, 309 between 61st-75th min, 439 between 76th-90th min). A total of 822 (43.4%) were scored from an Organized Attack (OA), 573 (30.3%) from Stopped Balls (SB) and 498 (26,3%) from Counter Attacks (CA).

Table 2. Goals for time and period of the game during 2015 and 2016 seasons.

Season	Time	Minutes	Number of goals (%)	Total goals (%)
2015	1st half	1st-15th	129 (13.8)	406 (43.3)
		16th-30th	122 (13.0)	
		31st-45th	155 (16.5)	
	2nd half	46th-60th	152 (16.3)	531 (56.7)
		61st-75th	166 (17.7)	
		76th-90th	213 (22.7)	
2016	1st half	1st-15th	117 (12.3)	414 (43.3)
		16th-30th	141 (14.7)	
		31st-45th	156 (16.3)	
	2nd half	46th-60th	173 (18.1)	542 (56.7)
		61st-75th	143 (15.0)	
		76th-90th	226 (23.6)	

Table 3. Way of attack that teams scored goals during 2015 and 2016 seasons.

Season	OA (%)	CA (%)	SB (%)	Total goals
2015	395 (48.1)	261 (52.4)	281 (49.0)	937
2016	427 (51.9)	237 (47.6)	292 (51.0)	956
Total	822	498	573	1893

Figure 1 depicts the initial and conditional probabilities of winning. Before the starting of the match, and only for the home team, the probability of winning the match is almost double (0.53 home team vs. 0.27 away team). After fifteen minutes of the first blow of the whistle, if the home team was already winning, it had a probability of 0.72 to get the three points at the end of the match. However, the probability dropped to 0.48 if the away team was the one who was

winning at this point of the match. In the same way, if the home team was losing at the fifteenth minute of the match, the probability of getting the victory at the end of the ninety minutes was 0.32, but if it was the away team the one who was losing after fifteen minutes, the probability to finally win the match was only 0.1. The same tendency was observed at the 30th, 45th, 60th, and 75th minute. As the match was approaching the end, winning or losing was more determinant and less reversible for the rival. Scoring first, at any point in the game, playing as a home team gives more chances to finally win the game than scoring first playing away (see circular references in figure 1).
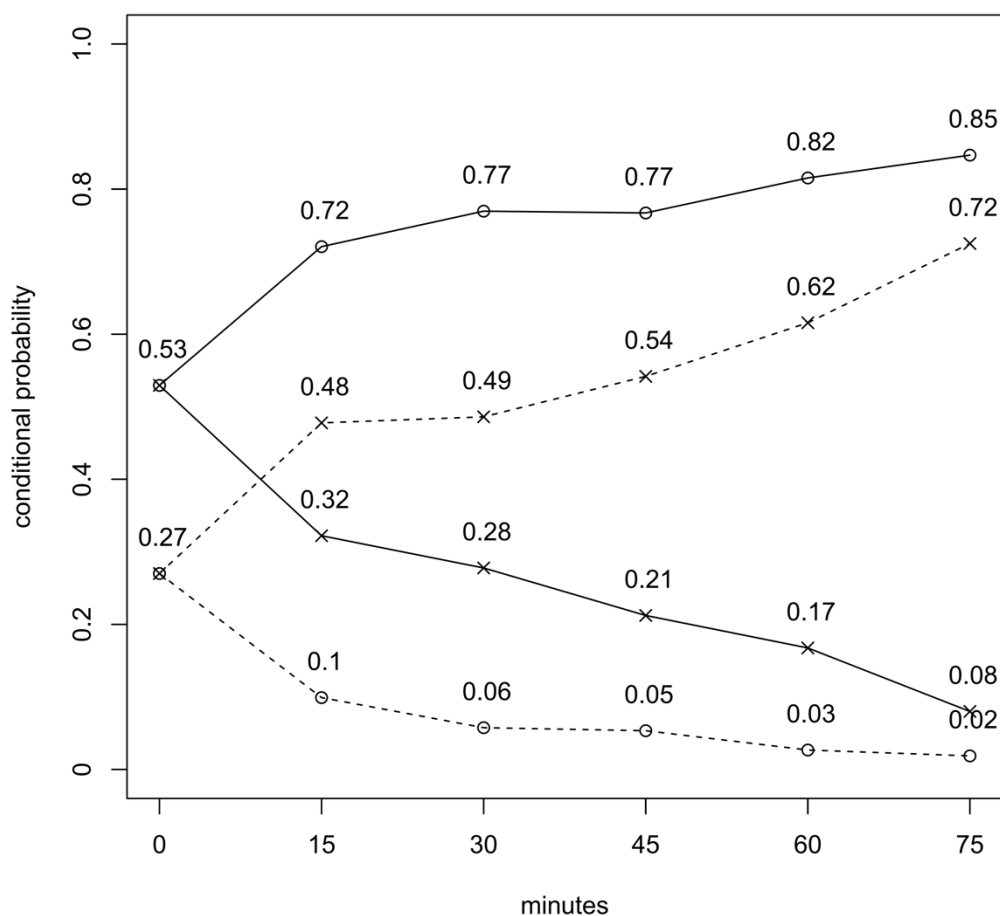
Figure 1. Initial probability (time 0) and conditional probabilities that the local team (solid line) or the visitor team (dashed line) will win the match provided that at time (15, 30, 45, 60 or 75 minutes) the locals are winning (circle) or the visitors are winning (cross).

PCA was applied in order to explore possible performance patterns in the teams. The proportion of explained variance of the first two principal components was 59% (36% for PC1 and 23% for PC2). Figure 2 and Figure 3 illustrate the results of the PCA. A typical graphical representation of a PCA is a biplot based on the first two principal components, PC1, and PC2. The biplot represents in the same plot both, the performance variables and the teams. However, to make the plot clearer, we will represent PCA with two separate graphics, one for the teams and the second for the performance variables. Dots and the variables by rows represent the individuals. The interpretation should be made as if the two plots were superimposed. In the

variables plot, rather horizontal rows are associated with the first principal component while more vertical rows are associated with the second principal component. The first principal component, PC1, mainly distinguishes those teams with large values of Counters and Penetrating Pass from those with large values of Free kicks, while the second principal component, PC2, mainly discriminates those teams with large values of Penalty kicks from those with large values of Crosses. In the teams plot we can identify some teams with a particular performance: (1) FC Dallas was characterized by large values of both PC1 and PC2, that is, large values of Counters and Penetrating Pass and low values of Crosses; (2) New York City FC stood out for large values of PC2, i.e., large values of Free kicks; (3) Los Angeles Galaxy appeared to be associated with large values of PC1, that is, large values of Counters and Penetrating Pass; (4) a group of three teams, Toronto FC, NY Red Bulls, and Columbus Crew SC, were associated with large PC1 values and small PC2 values, which means large Counters and Penetrating Pass and also large values of Crosses; finally, (5) we could identify a group of five teams with low values of PC1, i.e. large values of Free kicks and low values of Counters: Colorado Rapids, DC United, Sporting Kansas City, Real Salt Lake and Montreal Impact.
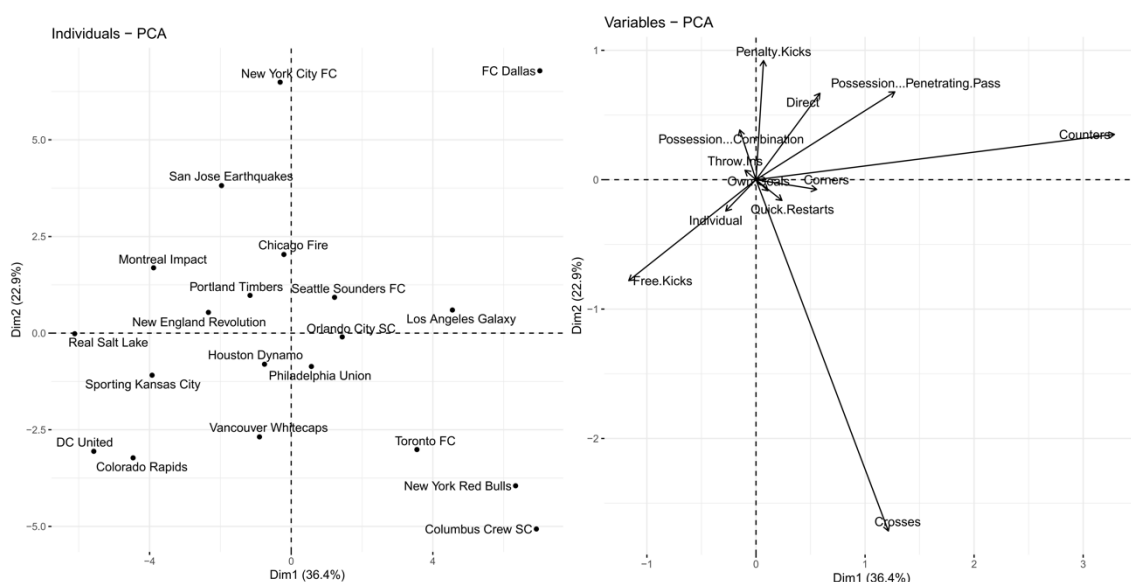

Figure 2 and Figure 3. Graphical representation of the teams (Figure 2, left) and the performance variables (Figure 3, right) projected on the two first principal components of the PCA. In the performance variable plot, rather horizontal rows are associated with the first principal component while more vertical arrows are associated with the second principal component.

In addition to the PCA, we performed a hierarchical cluster analysis (Figure 4), a mathematical method that creates groups of samples with high similarity. Each cluster can be described by the mean values of each variable, which allows teams characterization within the clusters according to similar performance. Three clusters were obtained which agreed with the groups of teams that were previously characterized in the PCA biplot: The upper cluster comprises those teams discussed in (3) and (4); the central cluster coincides with (5), and the bottom cluster is heterogeneous.
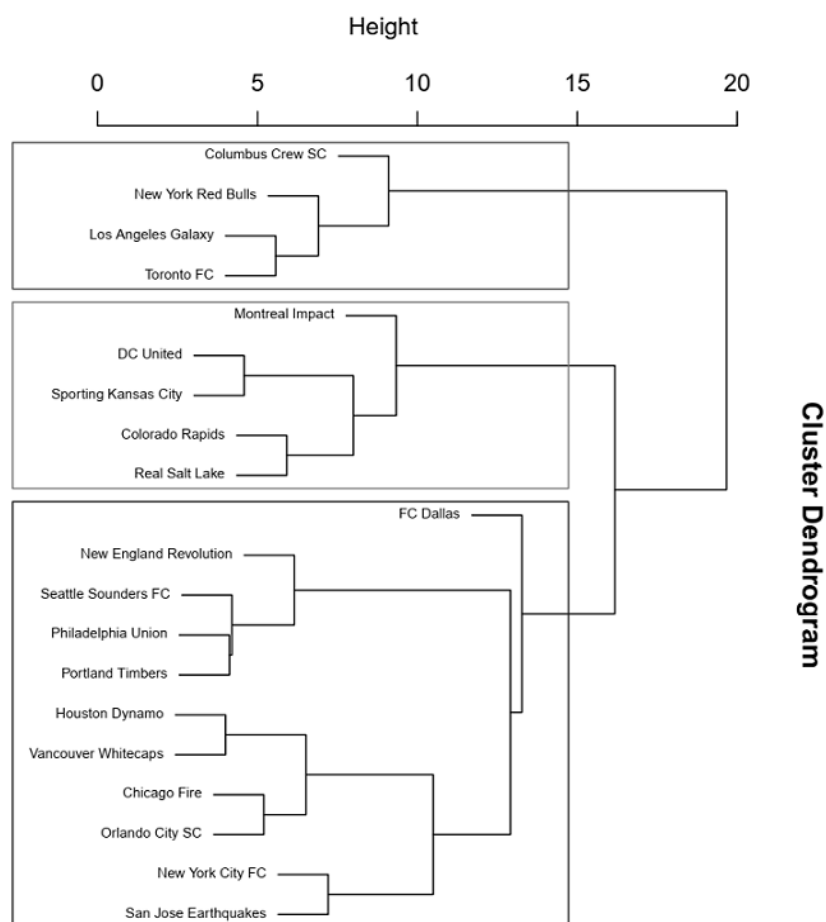
Figure 4. Hierarchical clustering or dendrogram for visualizing the similarity of the teams into three different main groups.

Complementary to this clustering method, we supply a heatmap (Figure 5) that provides information on the performance variables for each team. The darkest ones indicate either very high values, larger than the mean, or very low values of the performance variables, lower than the mean, while the lightest ones are closer to mean values. The upper cluster (comprised by Sporting Kansas City, DC United, Real Salt Lake, Colorado Rapids, and Montreal Impact) is characterized by lower values in Counters and large values of Free kicks. The third cluster (Toronto FC, NY Red Bulls, LA Galaxy, and Columbus Crew SC) is defined by large values of both Crosses and Counters. The last cluster was subdivided into two sub-clusters; low values of Crosses characterized the cluster with three teams, the bottom one (FC Dallas, NY City FC and San Jose Earthquakes) while the other sub-cluster, the second one, was rather heterogeneous. These results follow the PCA biplot.
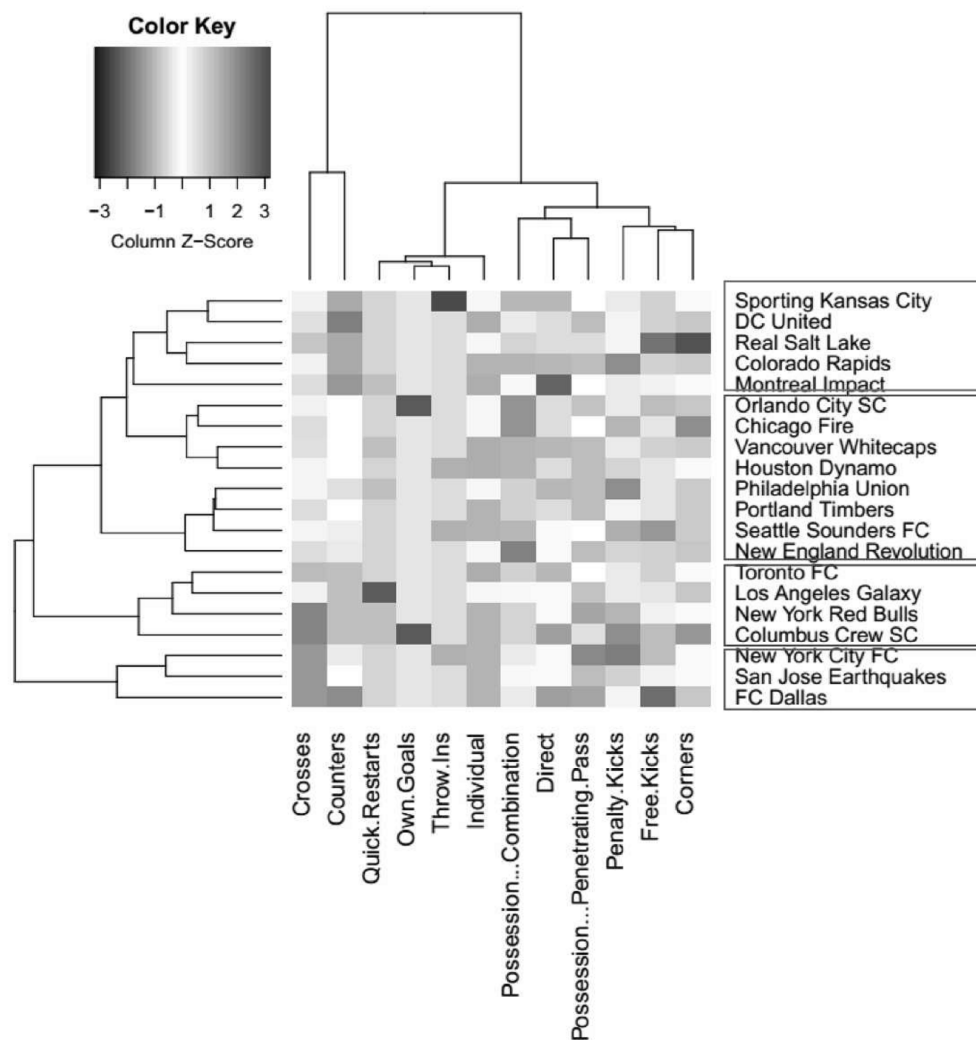
Figure 5: Heatmap with hierarchical clustering of both the teams and the performance variables.

Discussion

This study aimed to explore the importance of the scoring moment to increase or decrease the conditional probability of winning or losing in a professional association soccer match.

The main findings of the present study were the importance of playing as local or visitor and scoring or conceding first in the final score of the match. As expected, and according to Leite (2013), as the match progresses, the winning probability increases for the team who scores first, and this trend is more evident for local teams. Saavedra et al. provided similar findings in European Leagues, (2013), determining that home advantage was significant in 61% of these countries. However, after the introduction of the 3-point system, there has been a significant decrease in the importance of the playing venue (Sánchez, García-Calvo, Leo, Pollard, and Gómez, 2009). Some authors highlight that supporters and the noise coming from the crowd,

play a significant role in the psychology of the players and the final match result (Firdaus et al., 2015). In the same direction, Nevill; Balmer, and Williams (1996) and Nevill; Newell, and Williams (2002) pointed out that crowd noise significantly affected not only the players but also the referee's decisions against the visitor team, with more penalties against and more bookings than home teams.

It has been confirmed the importance of home advantage considering the number of goals scored in home matches compared with the away ones, and more goals were scored when playing at home (Firdaus et al., 2015). Other authors have studied the impact of the first goal in the final score and have found that when a team scores first, it wins the match in more than 73% of the occasions in European Championships or World Cups (Armatas and Yiannakos, 2010, Armatas, Yiannakos, Papadopoulou, and Skoufas, 2009, Leite, 2013, Mitrotasios and Armatas, 2014). In other studies, this percentage was lower but always in favour of the first scorer (Njororai, 2005). It has to be considered that during international competitions all of the teams were playing away; being specific, only one team could be considered as the home or local one, the organizer. To sum up, scoring first is a strong factor of the final outcome in soccer (Jones and Harwood, 2008) and performance after the first goal was controlled by two responses: firstly, how well was retained the advantage by the team scoring first and secondly how well was the disadvantage counteracted by the team conceding it (Njororai, 2007). Moreover, once a goal is conceded, another goal becomes more likely as the match progresses (Nevo and Ritov, 2012).

The results indicated that the number of goals scored during the second half of the matches was higher (56.7%) than in the first half, and the scoring rate was highest in the last 15 minutes of the match (23.2%). Nevertheless, it has to be stated that the last 15 minutes of each half includes the stoppage time of a few minutes and thus its comparability to the other 15 minutes intervals may be questioned. Similar findings were reported by different studies (Alberti et al., 2013, Armatas et al., 2007, Armatas and Yiannakos, 2010, Firdaus et al., 2015, Leite, 2013, Mitrotasios and Armatas, 2014, Yannakos and Armatas, 2006). These results may be explained by a decline in technical and physical performance between both halves (Rampinini, Impellizzeri, Castagna, Coutts, and Wisløff, 2009). However, and according to Williams (2003), the highest number of goals at the end of the match cannot only be explained from a physical perspective. The author suggests that lack of concentration and the deterioration in performance, due to fatigue, between defenders could also give an advantage to attackers in the final minutes. A possible explanation for this increase in goals may be the fact that technical, tactical, and psychological performance of teams can be affected by the physical performance at this point in the competition (Leite, 2013).

Using the PCA, counterattacks, and crosses were the most critical variables during open play to determine how American soccer teams scored goals. Similar results were found by different authors (Armatas and Yiannakos, 2010, Wright, Polman, Jones, and Sargeson, 2011, Yiannakos and Armatas, 2006), who concluded that most of the goals were scored after a cross (Mitrotasios and Armatas, 2014). Frequently, time pressure and the need for a better competitive result lead player to perform long passes, crosses, inside the opponent's areas (Firdaus et al., 2015).

Interestingly, there was a clear relationship between the number of counterattacks and the final position in the standings, not the same with crosses. Two of the teams who had better values in counters, New York Red Bulls and Dallas FC, ended up the regular season in the first place in their respective conferences. All the rest of the teams with higher values in counterattacks finished the regular season in playoff positions. These findings suggest that playoff teams used to counterattack more, and they felt more comfortable giving the initiative to the opposing team. Secondly, playing with this style gave many options to be in the top 6 of each conference. To counterattack effectively, a clear style of play and many hours invested in training seem to be a capital aspect, while to cross effectively, these aspects are not needed in the same proportion. This difference in the final position is not a mere product of randomness or luck, and it might have a clear link to the training processes and playing styles. Finally, it gives us an idea to understand how to prepare teams and to design proper training programmes to take advantage of counters when attacking and reduce it when defending.

The present study divided stopped balls (SB) or set pieces in four different categories: corner kicks, free kicks, throws-in, and penalty kicks. If they would have been grouped in the same category, stopped balls would have been the most common way to score goals in the MLS (30.6%), followed by counters (26.3%) and crosses (20.0%). In a similar manner, possession was divided into penetrating passes and combinations. However, in this case, even when grouping them in just one category, possession was ranked as the 4th factor in the rank to achieve a goal after stopped balls, counters, and crosses.

Another option to group variables in different attack modes used to score goals was proposed by Leite (2013). Following the authors' proposal, three options were considered: a) organized attack (OA) which included crosses, combinations, penetrating passes, direct play, individual actions, quick restarts and own goals, b) stopped balls (SB) or set pieces (corner kicks, free kicks, throws-in and penalty kicks) and c) counterattacks (CA). The highest rate of influence was found in OA with 43.4% of the variance explained by this category, followed by the SB, with 30.3%, and finally, CA with 26.3%. Different authors have obtained similar results in UEFA European Championships and FIFA World Cups (Armatas and Yiannakos, 2010, Leite, 2013, Mitrotasios and Armatas, 2014, Yiannakos and Armatas, 2006). Armatas et al., 2005 (cited in Armatas and Yiannakos, 2010) emphasized that even considering that frequency of CA in modern soccer is low, around 4.9%, this type of plays can be more effective than OA actions as 16.9% of CA lead to a goal whereas only 11.1% of OA were successful, according to their study.
The results of the present study indicate that most goals of OA were scored after a cross. This is likely influenced by different match situations and the style of play of each particular team. What is already known about possession is that losing teams tend to spend more time in possession of the ball when compared with the winning or drawing ones, probably, because they change tactics and style of play to adopt a more offensive strategies and to achieve a better result (Bradley et al., 2014, Collet, 2012).

New methods used for this analysis allowed us to understand and have a better notion of what happens in North American soccer and at the same time, opens a new way to explore and deal with similar data. All the results and information obtained from our findings can be used by coaches, technical staff members, and players to establish objectives, and with the right orientation to achieve success (Lago-Peñas et al., 2010). Our ideas can also expand the knowledge about performance indicators and help practitioners to improve their training methodologies. Our findings can also be used during specific match situations to change strategies, tactics, and the playing style to obtain a better competitive result based on the reported probabilities of winning.

Conclusions

The current study indicated that scoring first, playing as a home team always gave better chances to win the game than scoring first playing away. As the match approached the end, winning or losing was even more determinant and less reversible. Furthermore, counterattacks and crosses were the most critical variables during open play to determine how MLS teams scored goals. These findings can contribute to a better understanding of the performance indicators in professional soccer, helping coaches and practitioners to determine the right strategies and improving tactical patterns and approaches to succeed in the competition. Moreover, data obtained from this study can be used to implement the most suitable training strategies and to plan sessions in advance depending on the venue and the scoring pattern of the opposition team.

This study presents several limitations that should be considered. First, the number of matches and goals analysed it is not negligible. However, having an even larger aggregate of data could potentially add more strength to the analyses. This study includes data from the 2015 and 2016 MLS regular seasons, to include playoff data could be an interesting element to increase the value of the conclusions. The MLS is one of the most relevant soccer championships on the planet, but some of its limitations (for instance the use of salary cap) impairs the possibilities of displaying highly competitive teams when compared to those from other championships without signing restrictions. Obtaining similar data from other championships, performing similar analyses would allow comparisons and could bring interesting practical conclusions to the field.

Research complementing the study of performance factors with some other metrics such as expected goals, team ranking, positional data, or team coordination patterns should also be developed and is among our future perspectives.

Conflict of interest
The authors declare that they have no conflicts of interest relevant to the content of this manuscript.

References

Alberti, G.; Iaia, F. M.; Arcelli, E.; Cavaggioni, L., & Rampinini, E. (2013). Goal scoring patterns in major European soccer leagues. *Sport Sciences for Health*, *9*(3), 151–153.

Armatas, V.; Yiannakos, A., & Sileloglou, P. (2007). *Relationship between time and goal scoring in soccer games: Analysis of three World Cups*. *International Journal of Performance Analysis in Sport* (Vol. 7).

Armatas, V.; Yiannakos, A.; Papadopoulou, S., & Skoufas, D. (2009). *Evaluation of goals scored in top ranking soccer matches: Greek "Superleague" 2006-07*. *Serbian Journal of Sports Sciences* (Vol. 3).

Armatas, V., & Yiannakos, A. (2010). *Analysis and evaluation of goals scored in 2006 World Cup*. *Journal of Sport and Health Research* (Vol. 2).

Bradley, P. S.; Lago-Peñas, C.; Rey, E., & Sampaio, J. (2014). The influence of situational variables on ball possession in the English Premier League. *Journal of Sports Sciences*, *32*(20), 1867–1873.
https://doi.org/10.1080/02640414.2014.887850

Castellano, J.; Casamichana, D., & Lago, C. (2012). The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. *Journal of Human Kinetics*, *31*, 137–147.
https://doi.org/10.2478/v10078-012-0015-7

Collet, C. (2012). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007–2010. *Journal of Sports Sciences*, *31*(2), 123–136.
https://doi.org/10.1080/02640414.2012.727455

Federolf, P.; Reid, R.; Gilgien, M.; Haugen, P., & Smith, G. (2012) The application of principal component analysis to quantify technique in sports. *Scandinavian Journal of Medicine and Science in Sports, 24*(3), 491-499.
https://doi.org/10.1111/j.1600

Firdaus, M.; Ali, M.; Katis, A.; Patsika, G., & Kellis, E. (2015). Goal scoring charateristics in soccer: are they technique and time dependent? *Asian Pacific Journal Od Advanced Business and Social Studies*, *1*(1), 186–194.

Gómez, M. Á., & Pollard, R. (2014). Calculating the home advantage in soccer leagues. *Journal of Human Kinetics*, *40*, 5–6.
https://doi.org/10.2478/hukin-2014-0001

Harriss, D. J., & Atkinson, G. (2015). Ethical standards in sport and exercise science research: 2016 update. *International Journal of Sports Medicine*. 36, 1121–4.

Hastie, H.; Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer-Verlag.

Jackson, J. E. (1991). *A user's guide to principal components*. New York, NY: Wiley.

Jolliffe I. T. (2002). Principal component analysis, 2nd ed. New York, NY: Springer-Verlag.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions. *Series A, Mathematical, physical, and engineering sciences, 374*(2065), 20150202.
https://doi.org/10.1098/rsta.2015.0202

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254

Jones, M., & Harwood, C. (2008). Psychological Momentum within Competitive Soccer: Players' Perspectives. *Journal of Applied Sport Psychology*, *20*, 57–72.
https://doi.org/10.1080/10413200701784841

Lago-Peñas, C.; Lago-Ballesteros, J.; Dellal, A., & Gomez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. *Journal of Sports Science and Medicine*, *9*(2), 288–293.

Leite, W.S.S. (2013). Euro 2012: Analysis and Evaluation of Goals Scored. *International Journal of Sports Science*, *3*(4), 102–106.
https://doi.org/10.5923/j.sports.20130304.02

Li, Y.; Chiusano, S., & D'Elia, V. (2010) Modeling Athlete Performance Using Clustering Techniques. Proceedings of the Third International Symposium on Electronic Commerce and Security Workshops (ISECS '10) Guangzhou, P. R. China, 29-31, July,169-171.

Liu, H.; Gómez, M. A.; Gonçalves, B., & Sampaio, J. (2016). Technical performance and match-to-match variation in elite football teams. *Journal of Sports Sciences*, *34*(6), 509–518.
https://doi.org/10.1080/02640414.2015.1117121

Mitrotasios, M., & Armatas, V. (2014). Analysis of goal scoring patterns in the 2012 European Football Championship. *The Sport Journal*, (50), 1–9.

Moura, F. A.; Martins, L.E.B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences*, *32*(20), 1881–1887.
https://doi.org/10.1080/02640414.2013.853130

Nevill, A.; Balmer, N., & Mark Williams, A. (2002). The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, *3*(4), 261–272.
https://doi.org/10.1016/S1469-0292(01)00033-4

Nevill, A. M.; Newell, S. M., & Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sciences*, *14*(2), 181–186. https://doi.org/10.1080/02640419608727700

Nevo, D., & Ritov, Y. (2013). Around the goal: Examining the effect of the first goal on the second goal in soccer using survival analysis methods. *Journal of Quantitative Analysis in Sports*, *9*(2), 165–177.
https://doi.org/10.1515/jqas-2012-0004

Njororai Simiyu, W. (2005). Analysis of the goals scored at the 17th World Cup Soccer Tournament in South Korea-Japan 2002. *African Journal for Physical, Health Education, Recreation and Dance*, *10*.
https://doi.org/10.4314/ajpherd.v10i4.24678

Njororai W. (2007). Scoring goals. What the coach should know about the timing. *Soccer Journal, 11*, 34-36.

O'Donoghue, P. (2010). *Research methods for sports performance analysis*. London: Routledge.

Prabu, M.; Sudhaghar, J.; Viswajith, R.; Venkata Narsimha, I., & Srikaanth, A. K. (2019). Efficient Data Mining Methodology for Sports. *International Journal of Innovative Technology and Exploring Engineering, 8*(6S), 2278-3075.

Rampinini, E.; Impellizzeri, F. M.; Castagna, C.,;Coutts, A. J., & Wisløff, U. (2009). Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue and competitive level. *Journal of Science and Medicine in Sport, 12*(1), 227–233. https://doi.org/10.1016/j.jsams.2007.10.002

Saavedra García, M.; Aguilar, Ó. G.; Marques, P. S.; Tobío, G. T., & Romero, J. J. F. (2013). Calculating Home Advantage in the First Decade of the 21th Century UEFA Soccer Leagues. *Journal of Human Kinetics*, *38*(38), 141–150.
https://doi.org/10.2478/hukin-2013-0054

Sánchez, P. A.; García-Calvo, T.; Leo, F. M.; Pollard, R., & Gómez, M. A. (2009). An analysis of home advantage in the top two Spanish professional football leagues. *Perceptual and Motor Skills*, *108*(3), 789–797.
https://doi.org/10.2466/PMS.108.3.789-797

Shafizadeh, M.; Taylor, M., & Peñas, C. L. (2013). Performance consistency of international soccer Teams in Euro 2012: A time series analysis. *Journal of Human Kinetics*, *38*(1), 213–226.
https://doi.org/10.2478/hukin-2013-0061

Wilkinson, L., & Friendly, M. (2009) The History of the Cluster Heat Map. *The American Statistician*, *63*(2), 179-184.
https://doi.org/10.1198/tas.2009.0033

Williams, A. M. (2013). *Science and soccer: Developing elite performers* (1st edition). Abingdon, OX: Routledge.

Wright, C.; Polman, R.; Jones, B., & Sargeson, L. (2011). Factors Associated with Goals and Goal Scoring Opportunities in Professional Soccer. *International Journal of Performance Analysis in Sport*, *11*, 438–449. https://doi.org/10.1080/24748668.2011.11868563

Yiannakos, A., & Armatas, V. (2006). Evaluation of the goal scoring patterns in European Championship in Portugal 2004. *International Journal of Performance Analysis in Sport*, *6*, 178–188. https://doi.org/10.1080/24748668.2006.1186836.