The harsh rule of the goals: data-driven performance indicators for football teams

Paolo Cintia
Department of Computer Science
University of Pisa, Italy
Email: paolo.cintia@isti.cnr.it

Luca Pappalardo
Department of Computer Science
University of Pisa, Italy
Email: lpappalardo@di.unipi.it

Dino Pedreschi
Department of Computer Science
University of Pisa, Italy
Email: pedre@di.unipi.it

Fosca Giannotti
Institute of Information Science and Tecnologies
National Research Council (CNR), Italy
Email: fosca.giannotti@isti.cnr.it

Marco Malvaldi
Institute of Information Science and Tecnologies
National Research Council (CNR), Italy
Email: marcoampelio@hotmail.com

***Abstract*—Sports analytics in general, and football (soccer in USA) analytics in particular, have evolved in recent years in an amazing way, thanks to automated or semi-automated sensing technologies that provide high-fidelity data streams extracted from every game. In this paper we propose a data-driven approach and show that there is a large potential to boost the understanding of football team performance. From observational data of football games we extract a set of pass-based performance indicators and summarize them in the H indicator. We observe a strong correlation among the proposed indicator and the success of a team, and therefore perform a simulation on the four major European championships (78 teams, almost 1500 games). The outcome of each game in the championship was replaced by a synthetic outcome (win, loss or draw) based on the performance indicators computed for each team. We found that the final rankings in the simulated championships are very close to the actual rankings in the real championships, and show that teams with high ranking error show extreme values of a defense/attack efficiency measure, the Pezzali score. Our results are surprising given the simplicity of the proposed indicators, suggesting that a complex systems' view on football data has the potential of revealing hidden patterns and behavior of superior quality.**

I. Introduction

Sports analytics in general, and football (soccer in USA) analytics in particular, are attracting wide interest from a long time ago. Already in the early 1950s Charles Reep collected football statistics by hand to suggest that "the key to scoring goals and winning games was to transfer the ball as quickly as possible from back to front", thereby indirectly starting the long-ball movement in English football [1][2].

In the recent years football statistics have evolved in an amazing way, thanks to automated or semi-automated sensing technologies that provide high-fidelity data streams extracted from every game, based on video recordings by different cameras or observations by various kinds of fixed and mobile sensors. There are now professional statistical analysis firms like ProZone [3] and Opta [4] which provide data to football clubs, coaches and leagues, who are interested in such services to ensure they can remain in control of their performances and results as much as possible, by monitoring their players and opponents. Fan engagement is another driver of football analytics: more and more statistics and visualizations are being made available for enjoyment, either to back up a viewpoint in a friendly bar conversation, or to challenge a friend's opinion. A large number of websites also make use of football statistics to produce critical analyses, insights and scoring patterns of their own, such as EPLIndex.com and WhoScored.com.

However, despite the increasing wealth of data, a data scientist's view on the state-of-the-art of football analytics cannot avoid to notice that this wealth has been exploited to a limited extent so far. There is not yet a consolidated repertoire of statistics that are accepted as reference indicators for the various facets of team performance. Even more importantly, there is very limited work on adopting the powerful tools of data mining and network analytics, despite the evidence that two football teams and a ball in a game represent a highly complex system, whose global behavior depends in subtle ways on the dynamics of the interactions among each of the 23 components (not to mention the referees!).

Our aim here is precisely to show how by adopting a data-driven approach there is a large potential to boost the understanding of team performance, since even simple indicators that we propose reveal as surprisingly accurate predictors of team success across an entire season. Our idea is based on capturing crucial aspects of the *passing behavior* of a team from observational data of a football game. From a list of events occurred in the game (passes and goal attempts) we first define for every team a set of pass-based performance indicators, each capturing a different aspect of the passing behavior of the team. We then summarize all these indicators into a single value – the H indicator – representing the passing behavior of a team. We observe a strong correlation among the indicators and the success of a team and therefore perform two analyses on high-fidelity event data for every game played in one season of four European football leagues, almost 1,500 games involving globally 78 teams. First, we investigate the difference in the value of H indicator of the teams according to the outcome of a game, discovering that wins, losses and draws of the home team are characterized by typical ranges of values of the H indicator. We then construct a repertoire of classifiers to predict the outcome of a football game from the

history of performance of the two teams, obtaining an accuracy higher than models which do not use performance information in the learning phase.

In the second analysis, we conduct a computationally intensive experiment consisting in a complete simulation of each of the four national championships – England's Premier League, Spain's Liga, Germany's Bundesliga, and Italy's Serie A. The outcome of each game in the championship is replaced by a synthetic outcome (win, loss or draw) based on the performance indicators computed for each team. We found that the final rankings in the simulated championships are very close to the actual rankings in the real championships, and that the final standings emerge quite early during the season, especially for the top positions. In the case of the German Bundesliga we find a correlation of ≈ 0.9 between simulated and real rankings, a value that is really surprising given the simplicity of the proposed indicators. The strongest European teams present the highest values of our performance indicators and the simulation predicts their position in the final standings with high precision. We also characterize each team's playing style during a game by a defense/attack efficiency rate – the Pezzali score – discovering that teams for which the simulation overestimates or underestimates the position in the final standings show extreme values of such efficiency rate.

The lesson learnt is that football analytics has only begun to scratch the surface in the quest to understand, measure and predict performance. Despite many studies find that randomness has a strong role in football games [5], our indicators have proven to be a good proxy of the performance of a team. If simple indicators such as those introduced here exhibit surprising connections to the success of teams, then probably a complex systems' view on football data has the potential of revealing hidden patterns and behavior of superior quality.

II. Football Data

We have data about the games of four major European leagues – Germany, England, Spain, Italy – in the season 2013/2014. The Italian, Spanish and English leagues have 20 teams each playing 38 games, the German league has 18 teams each playing 34 games. In total our dataset stores information about 1,446 football games. A football game is described by a sequence of events on the pitch (passes and goal attempts), with a mean of 450 events per game and a total of ≈600,000 events in our dataset (see Table I). Each event consists in the following information: the timestamp of the event, the player who generated the event, the position of the ball on the pitch when the event is generated, the position of the ball on the pitch when the event ends, the outcome of the event (successful or unsuccessful). Note that a successful goal attempt has to be intended as a goal. Table II gives some examples of events occurred during a game in the Spanish league: the event "pass" identifies a successful pass made by Lionel Messi at position (65.4, 20.2) of the pitch; the event "goal attempt" at minute 55:00 indicates a successful goal attempt (a goal) made by Cristiano Ronaldo.

Since each event specifies the destination point on the pitch, the data allow us to reconstruct the ball trajectory during the game. However we do not have direct information about the destination player, i.e. the player to which the pass is directed.

TABLE I. Size of our dataset of football games.

Season 2013/2014		
leagues	4	Germany, England, Spain, Italy
teams	78	20 England, Spain, Italy - 18 Germany
games	1,446	360 games per league in average
events	600,000	450 events per game in average

TABLE II. Example of events during the game Real Madrid-Barcelona (Spanish league).

event	time	player	origin	destination	outcome
pass	17:24	Messi	(65.4, 20.2)	(67.8, 44.1)	successful
attempt	18:12	Messi	(98.4, 15.0)	(118.7, 15.0)	unsuccessful
pass	45:00	Bale	(78.56, 12.2)	(78.5, 36.0)	successful
attempt	55:00	Ronaldo	(89, 45)	(100, 45)	successful
⋮	⋮	⋮	⋮	⋮	⋮

We infer this information by sorting all the events by time and making a spatial agglomeration where we split the pitch into zones of size 11m×6.5m (100 zones in total). Then, given a player A who generates a pass event p toward zone (x, y) at time t, if player B generates an event from zone (x, y) at time $t' > t$ we assume that player B is the destination player of the event, otherwise we discard the event p. This step allows us to reconstruct the movements of the ball between players during the game, which can be represented as a player passing network, i.e. a weighted network where nodes are players and weighted edges represent movements of the ball between players (see Figure 1) [6][7].

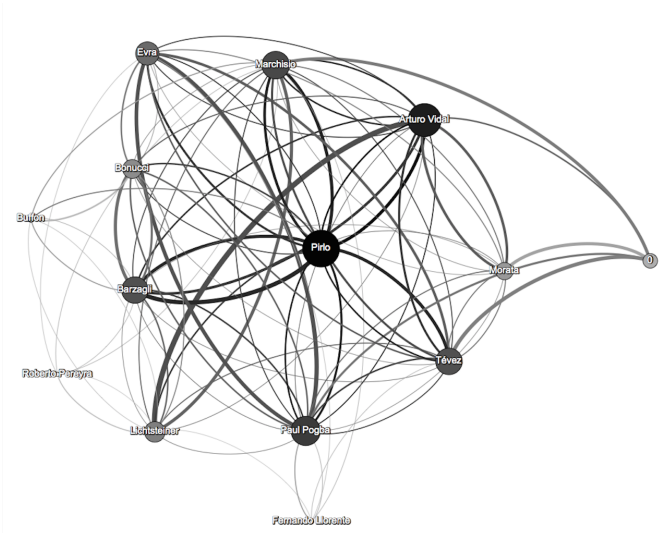

Fig. 1. A representation of the player passing network of Juventus FC extracted from a game in season 2014/2015. Nodes represent players, edges represent passes between players. The size of a node is proportional to the number of ingoing and outgoing passes the player managed during the game; the size of an edge is proportional to the number of passes between the players during the game. Node 0 indicates the opponent's goal, edges ending in node 0 represent goal attempts.

From a first exploratory analysis of our dataset we find a clear correlation between the average amount of passes made by a team during the season and (i) total goals scored, (ii) total goal attempts, and (iii) points obtained in the final rankings (Figure 2a–c). This suggests that the passing activity of a

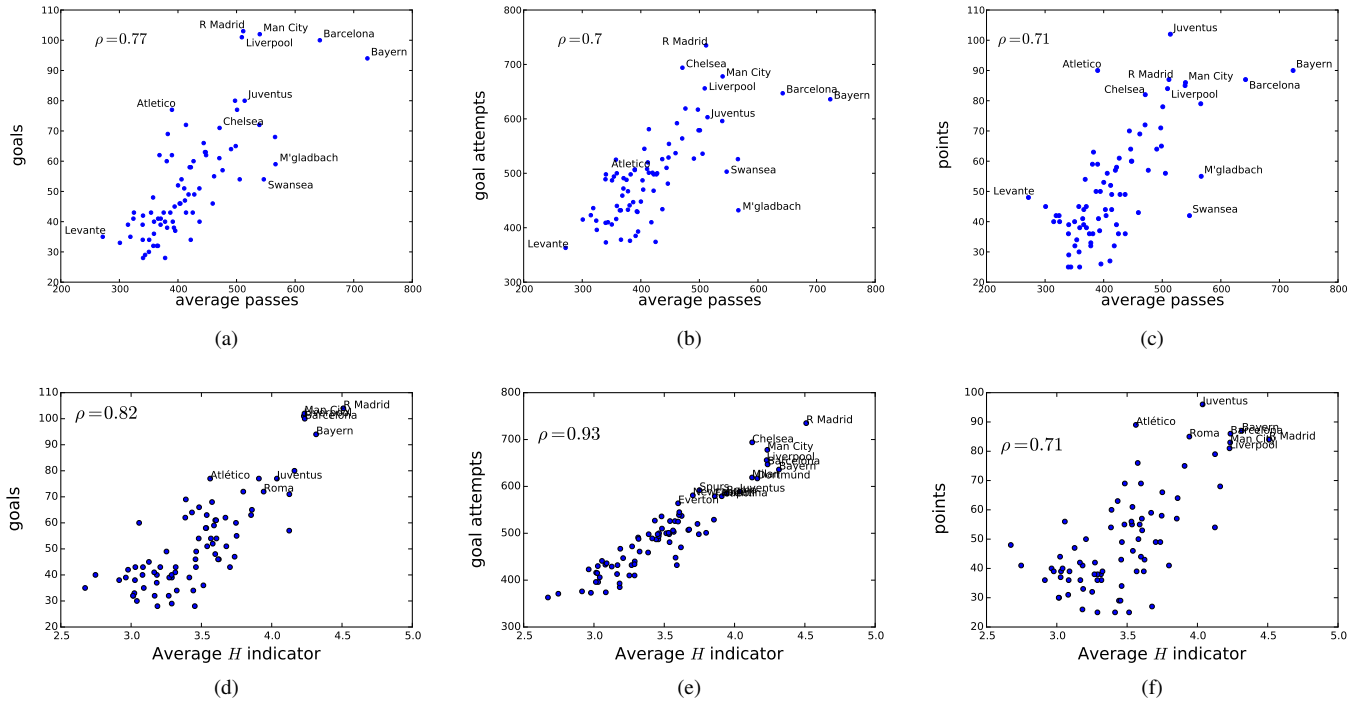

Fig. 2. **First row:** The correlation between teams' average amount of passes and their success in national leagues. Each point represents a team in the four major leagues and indicates the correlation between the average number of passes and (a) the total number of goals scored during the season; (b) the total number of goal attempts during the season; (c) the total points gained at the end of the season. We observe in all the cases a strong correlation (ρ indicates the Pearson correlation coefficient) suggesting that the passing activity of a team is a key feature for its success. The strongest European teams (winner of national leagues or with good results in European cups) show a high average number of passes together with many goals, attempts and points at the end of the season. Some outliers also emerge which do not follow the clear trends: they produce low passing activity achieving a considerable amount of goals/attempts/points (Levante, Atlético Madrid), or they produce high passing activity but a few goals/attempts/points (Swansea, Borussia Mönchengladbach). **Second row:** The correlation between teams' average H indicator and their success in national leagues. Each point represents a team in the four major leagues and indicates the correlation between the average H indicator and (d) the total number of goals scored during the season; (e) the total number of goal attempts during the season; (f) the total points gained at the end of the season. We observe strong correlations between H indicators and goals scored, goal attempts and points gained by the teams in the four leagues.

team is related to its success during the competition, as teams with high passing activity tend to score more goals, to have more goal opportunities, to gain more points. In particular the strongest European teams, i.e. the winners of national leagues or with good performance during the European cups (Barcelona, Real Madrid, Manchester City, Bayern München, etc.) show a high average number of passes together with many goals, attempts or points gained (Figure 2a–c). However, some teams do not follow the clear trends: they either produce low passing activity achieving a considerable amount of goals/attempts/points (Levante, Atlético Madrid) or they produce high passing activity but a few goals/attempts/points (Swansea, Borussia Mönchengladbach). In general Figure 2a–c tells us that the amount of passes produced by a team, a proxy for its ball possession during the games, is linked to its success during the competition. It makes sense therefore to describe the performance of a team during a game in terms of its passing behavior and to define performance indicators based on the passing activity produced during the games. Starting from these observations we investigate other aspects related to the passing behavior of a team and extract several performance indicators from the football data.

III. Indicators of Team Performance

Many aspects characterize the passing behavior of a team. The first one is certainly the amount of passes w introduced before, a measure of the total passing volume generated by a team during a game. Figure 2a–c suggests a clear trend: the higher the value w of a team the more it scores goals and gains points during the competition. Nevertheless this simple indicator, though useful, gives only a partial picture of a team's passing behavior.

The distribution of passes over its players gives a different and fundamental point of view on the passing behavior of a team. While in some teams a few key players manage the majority of passes during a game, other teams prefer to distribute the possession more equally on all the players (think about Barcelona FC and "tiki-taka"). We capture the distribution of a team's passes over its players by defining two indicators: (i) the average amount μ_p of passes managed by players in the team during the game; (ii) the variance σ_p of the amount of passes managed by players in the team during the game. These indicators can be easily computed from the player passing network introduced in Section II: the weighted degree (in-degree + out-degree) of a node indicates the volume of passes the player manages during the game. The mean weighted degree of the network μ_p hence measures the mean

players' passing volume of the team in the game. Indicator σ_p is instead the variance of players' passing volume: the higher its value the higher is the heterogeneity in the volume of passes managed by the players. A high value of σ_p means a coexistence of players which manage many passes and players with low pass activity during the game.

The distribution of passes over the zones of the pitch is another key aspect of a team's passing behavior. To capture this aspect we build a *zone passing network*, where nodes are zones of the pitch and an edge (z_1, z_2) represents all the passes performed by any player from zone z_1 to zone z_2. The zones are obtained by a spatial agglomeration splitting the pitch into zones of size 11m×6.5m, 100 zones in total. Figure 3 clarifies the concept showing a zone passing network extracted from a game of FC Barcelona. On the zone passing network we define two indicators: (i) the average amount μ_z of passes managed by zones of the pitch during the game; (ii) the variance σ_z of the amount of passes managed by zones of the pitch during the game. A high σ_z means a coexistence of "hot" zones with high passing activity and "cold" zones with low pass activity during the game. Low values of σ_z indicates a more uniform distribution of the passing activity across the zones of the pitch.

Finally we combine the five indicators by their harmonic mean $H = 5/(1/w+1/\mu_p+1/\sigma_p+1/\mu_z+1/\sigma_z)$ to summarize the passing behavior of a team into a single value. For each game in the four leagues we compute the six indicators for both the home team and the away team. Figure 4 shows the ten teams with the highest average H indicator, computed across all the games in the season. We observe that the Champions League winner Real Madrid is the strongest European team according to our performance indicator. Figure 2d–f and Table IIIb clearly show that the H indicator of a team is better correlated with its success (goals, attempts, points) than the mere amount of passes (indicator w), highlighting the usefulness of the defined indicators in capturing important aspects of the performance of football teams.

TABLE III. **(A)** THE PASS-BASED INDICATORS USED IN OUR STUDY. **(B)** CORRELATIONS BETWEEN INDICATORS AND EVENTS.

measure	description
w	total passing volume
μ_p	mean players' passing volume
σ_p	variance of players' passing volume
μ_z	mean zones' passing volume
σ_z	variance of zones' passing volume
H	combination of above measures

(a)

indicator	goals	attempts	points
w	0.76	0.69	0.71
μ	0.63	0.82	0.71
σ	0.68	0.81	0.57
μ_z	0.71	0.57	0.40
σ_z	0.45	0.57	0.40
H	**0.82**	**0.93**	**0.71**

(b)

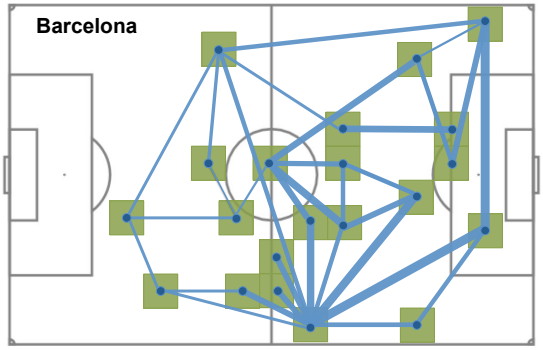

Fig. 3. A zone passing network extracted from a portion of a game of Barcelona (Spanish league) in season 2013/2014. Nodes are zones on the pitch, edges represent passes performed by any players between two zones, the size of an edge is proportional to the number of passes between the zones. Here we observe the presence of a dominant zone (in the bottom part of the figure) where most of the passes take place.

	team	mean *H*	league
1	Real Madrid	4.51	SPA
2	Bayern München	4.31	GER
3	Barcelona	4.23	SPA
4	Manchester City	4.23	ENG
5	Liverpool	4.22	ENG
6	Borussia Dortmund	4.16	GER
7	Chelsea	4.12	ENG
8	Milan	4.12	ITA
9	Juventus	4.03	ITA
10	Roma	3.94	ITA

Fig. 4. The ten teams with the highest average H indicator computed across all the games in the season. We observe that the Champions League winner Real Madrid is the strongest team according to the H indicator. In the H ranking the national league winners are second (Bayern München), fourth (Manchester City), ninth (Juventus) and 31th (Atlético Madrid).

IV. TEAM PERFORMANCE ANALYSIS

We deeply investigate at what extent our performance indicators are descriptive of the success of a team by performing two types of analyses on our football data. For each game in our dataset we compute the six indicators for both the home team and the away team. To include some additional and not explicit information, such as the attack strategy of a team and the defense efficiency of the opponent, we compute the pass-based indicators also on a subset of passes, i.e. the passes of the team composing a chain that actually led to a goal attempt[1].

In the first analysis we investigate the difference in the values of H indicator of the two teams according to the outcome of a game (home team wins, away team wins, or draw). In other words we split all the games of a league into three groups: all the games where the home team wins, all the games where the away team wins, and the games resulted in a draw. For each group we plot the mean value of H indicator of the home team against the value of H indicator of the away

[1]We performed all the analyses showed in this paper using indicators computed considering both all the passes of a team and the subset of passes that compose a chain leading to a goal attempt. The results are similar even though the indicators defined on the subset of passes that lead to goal attempts show better correlations. For space reasons we present in this paper results relative to indicators computed on goal attempts chain passes only.

team (Figure 5). From the plot a clear result emerges: the home team is more likely to win when its H indicator is higher than the opponent, it is more likely to lose when its H indicator is lower then the opponent, a draw is more probable if the difference in the H indicator is the range $[0, 0.5]$ (we find similar results by using the other five indicators).[2] We also observe that in general home teams have higher pass activity than away teams (most of the games are under the bisector of the plot) and that away teams need in general a pass activity slightly higher than home teams to win a game. We report that in the 73% of home wins the home team has a H indicator higher than the away team, while in the 51% of home losses the home team has a H indicator lower than the opponent.

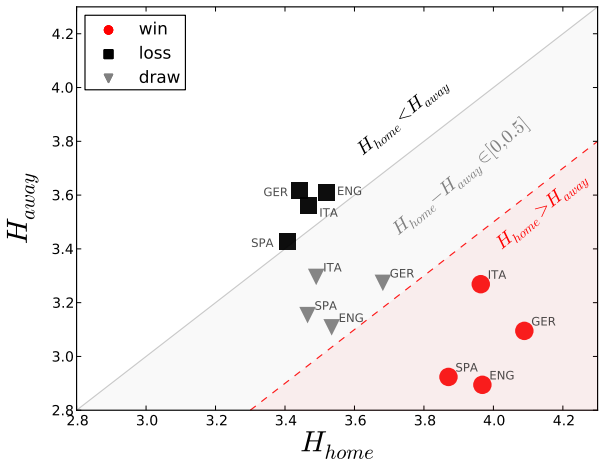

Fig. 5. Average values of H indicator of home teams and away teams for games where the home team wins (red circles), the away team wins (black squares), or there is a draw (grey triangles). Each point represents a league. We observe that home teams tend to win when their H indicator is higher than away teams' one, they tend to lose when the H indicator is lower than the away teams' one. Draws are more probable when the difference between home teams' H indicator and away teams' indicator are in the range $[0, 0.5]$.

In the second analysis, we study how the performance of teams changes as the season goes by (Figure 6). We observe that the teams qualified to the Champions League (the first three or four in the ranking) show the highest mean values of H indicator during the course of the season. The dominance of the strongest teams is immediately evident in the plots: they show the highest mean values of H indicator since the first games of the season (Figure 6). This result suggests that the pass-based indicators can be also used to predict the outcome of a football game based on the history of performances produced by the teams during the past games. We investigate this aspect by constructing a repertoire of classifiers to predict the outcome of a football game based on the performance indicators of the two teams in the *past games*. To do that we build, for each league, a dataset where every football game is an observation each consisting in six features. These features are the simple exponential smoothed means of performance indicator values produced by the teams in their past games. The target value for each observation is the game's outcome, with three possible values: 1 indicates a win of home team, 2 indicates a win of away team, 0 indicates a draw. For each classifier we use k-fold cross validation ($k = 10$) to validate the accuracy of the classifiers[3]. Table IV shows the accuracies achieved by the classifiers on the four leagues. We observe for the German league a maximum accuracy of 0.60 obtained with the K-Nearest Neighbor classifier ($k = 10$), while for the other leagues the Random Forest classifier outperforms the other models reaching accuracy values of 0.58, 0.53 and 0.55 for English, Spanish and Italian league respectively. Hence, the pass-based indicators allow us to accurately predict more than half of the games in a league, a significant improvement with respect to baseline classifiers which do not use performance information in the learning phase reaching a maximum accuracy of 0.45.[4] In particular, for the German league the K-Nearest Neighbor classifier ($k = 10$) accurately predicts around the 80% of the victories by the home team, the 60% of victories by the away team, and the 20% of draws.

TABLE IV. ACCURACY IN THE PREDICTION OF FOOTBALL GAMES

classifier	Germany	England	Spain	Italy
KNearestNeighbor	**0.60**	0.55	0.51	0.52
Logistic Regression	0.53	0.57	0.52	0.53
Decision Tree	0.54	0.56	0.50	0.53
SVM	0.53	0.57	0.52	0.53
Naive Bayes	0.50	0.56	0.49	0.50
Random Forest	0.57	**0.58**	**0.53**	**0.55**
baseline	0.45	0.45	0.45	0.45

V. SIMULATION OF MAJOR FOOTBALL LEAGUES

Figure 5 shows that the difference in the H indicator of two teams in a game is descriptive of the relative performance of the teams: the higher the H indicator of a team, the higher is its probability to win the game. Starting from this result we try to address the following issue: Can we detect the winner of a game just by observing the passing activity of the teams during the game? In other words, we forget about the goals scored and we want to detect the winner on the only basis of teams' passing activity observed during the game.

To answer this question, we perform the following experiment. For each game we compute the six measures defined in Section III both for the home team and the away team. We then simulate the outcome of the games in the season, round by round, in the following way: given home team t_1, away team t_2 and indicator x (w, μ_p, σ_p, μ_z, σ_z, or H), if the difference $0 \leq x(t_1) - x(t_2) \leq \epsilon$ we set the outcome as a draw, otherwise we assign the victory to the team with the highest x. We set the bounds for a draw according to Figure 5 which suggests $\epsilon = 0.5$. As done in official football leagues, the winning team gains three points, the losing team gains no points, both teams gain one point in case of draw. Finally according to the simulated outcomes we compute a ranking of the teams round by round, till obtaining a final simulated ranking of the entire season. Table V compares the final simulated ranking with the final actual ranking of the German league. We observe a good

[2]The draw range is the same for μ_p, μ_z, σ_p and σ_z, while for indicator w the draw range is $[0, 13]$.

[3]We use the Python library scikit-learn [8] to instantiate and validate the classification models.

[4]This is the best accuracy across three dummy classifiers that: (i) always predict the most frequent label in the training set; (ii) generate predictions by respecting the training set's class distribution; or (iii) generate predictions uniformly at random.

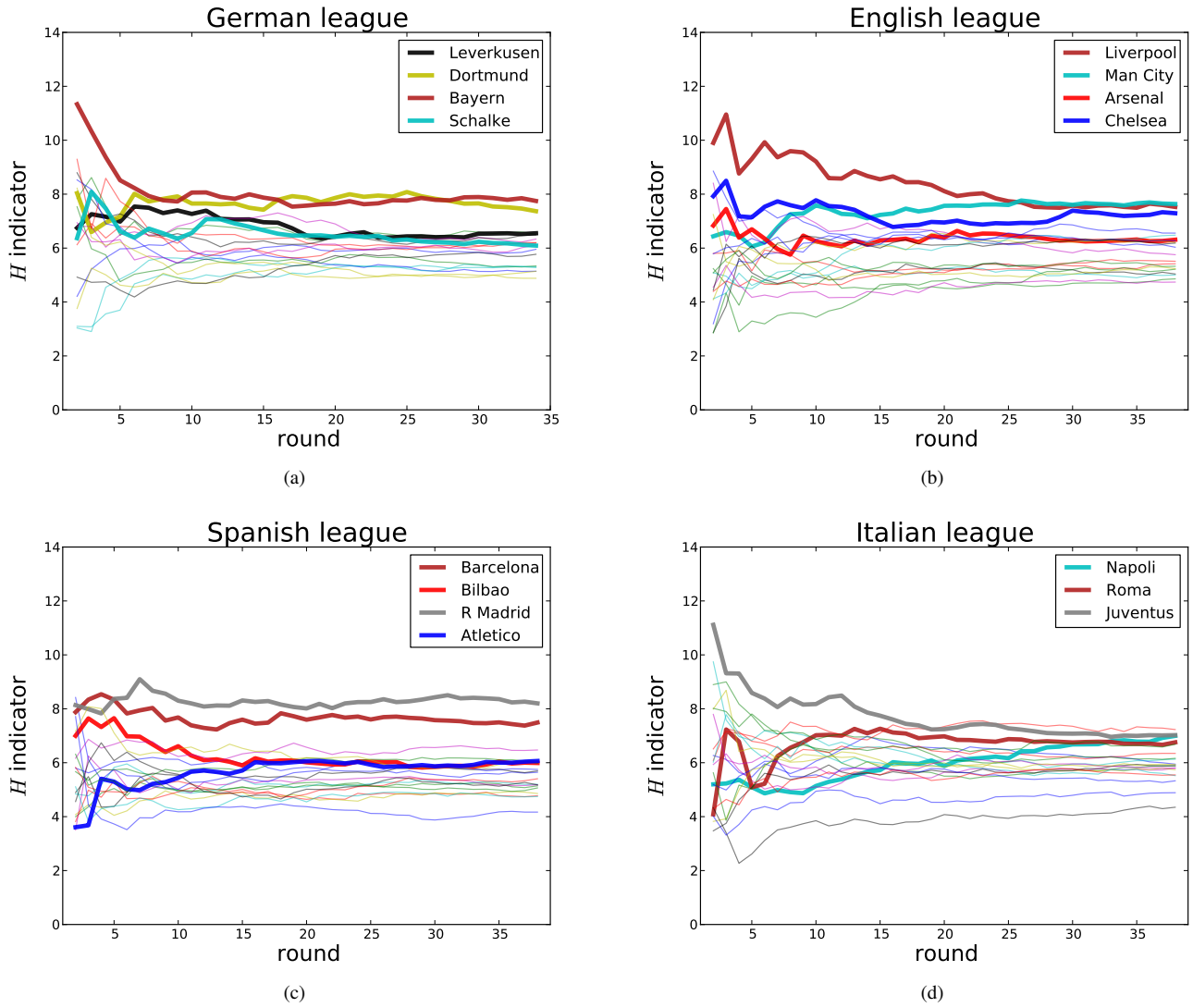

Fig. 6. Evolution of H indicator of teams in German (**a**), English (**b**), Spanish (**c**) and Italian (**d**) league. We highlight the teams which achieved the qualification to Champions League at the end of the season. The teams qualified for the Champions League show the highest mean values of H indicator during the course of the season, with a dominance immediately evident since the first games of the championships.

agreement between the two rankings especially for the teams at the top of the ranking: three on four of the teams qualified to the Champions League are predicted in the exact position (Bayern München, Borussia Dortmund and Bayer Leverkusen).

We compute the correlation between the simulated ranking and the actual ranking round by round and study how it evolves over time. Figure 7 shows the evolution of the correlation as the season goes by in the four leagues. The correlations between the simulated ranking and the real one stabilize as the season goes by, reaching values at the end of the season of 0.89 (German league), 0.84 (English league), 0.84 (Italian league) and 0.66 (Spanish league). We note that indicator w, with the exception of Italian league, produces the worst simulation highlighting the usefulness of the other network-based indicators in describing teams' performance.

We observe that the ranking error, i.e. the difference between points gained in real championship and in simulated championship, is close to zero for the majority of teams in the four leagues (a peaked distribution with average 0). However, for some teams the ranking error is either high or low, meaning that the simulation overestimates or underestimates the success they achieve in the championship. For example in Spanish league Real Betis got in the simulated ranking 33 points more than it actually achieved, while Levante got 27 points less. To better understand the source of the error we investigate the characteristics of teams that resulted in a high ranking error.

VI. Pezzali score

We observe that the teams with high ranking error show unique patterns with respect to their attack and defense efficacy. We recall what is popularly known as *the harsh rule of the goals*, introduced by Pezzali [9]. The statement is: *It's the harsh rule of the goals: you play a great game but if you don't have a good defense, the opponent scores and then wins.* Starting from this insight, for each team and each game we compute a defense/attack efficiency rate:

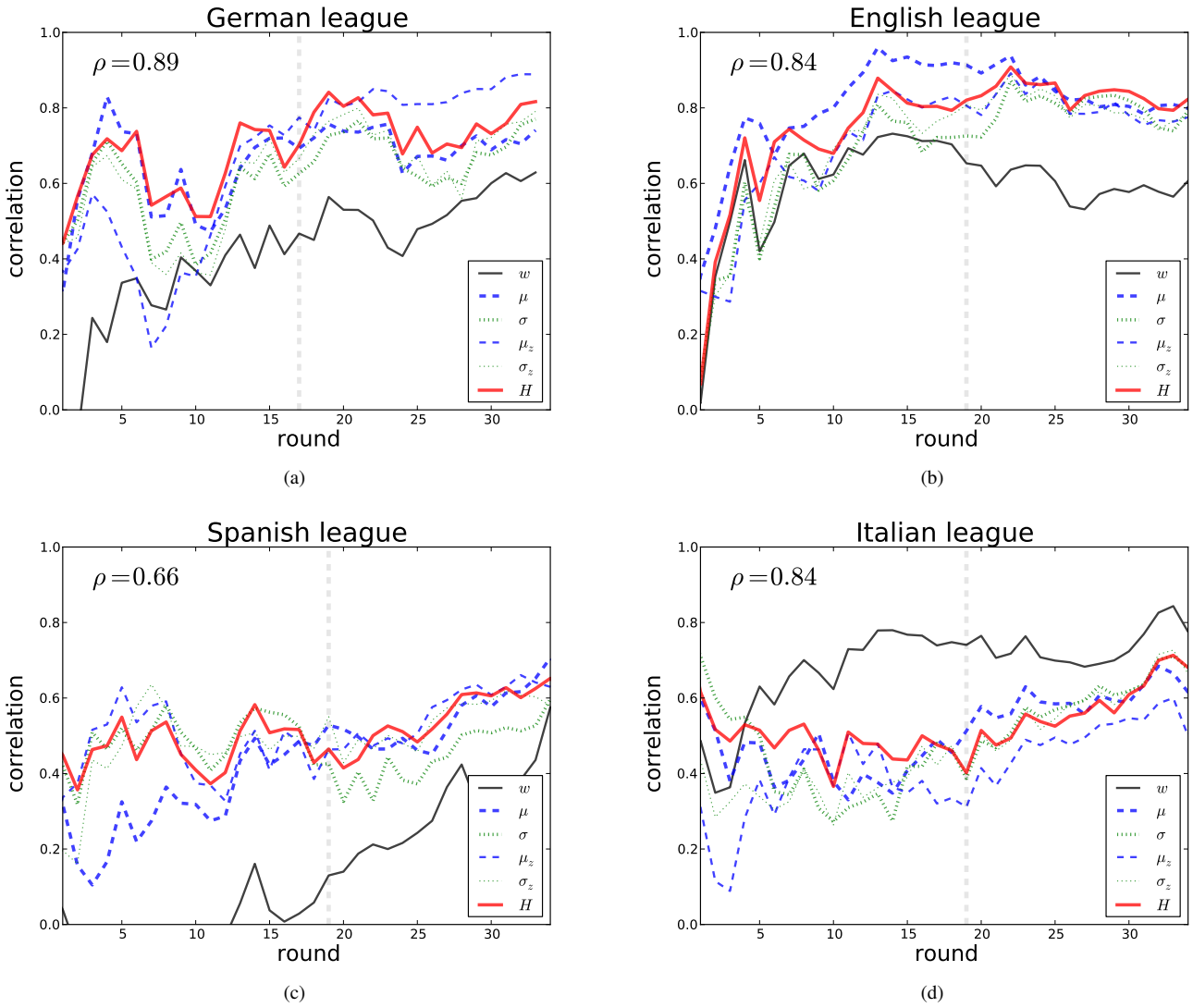

Fig. 7. The correlation (Spearman coefficient) between the simulated ranking and the actual ranking according to the six performance indicators, for (**a**) German league, (**b**) English league, (**c**) Spanish league, (**d**) Italian league. The blue dashed lines indicate the correlation using indicators μ_p and μ_z, the green dashed lines indicate the correlation using indicators σ_p and σ_z, the red solid line indicates the correlation using H indicator. The vertical grey dashed line indicates the half of the season. The value ρ indicates the correlation reached at the end of the season.

$$Pezzali\ Score(team) = \frac{|goals(team)|}{|attempts(team)|} * \frac{|attempts(opponent)|}{|goals(opponent)|}$$

Given a football game, the Pezzali score of a team is high when the team is highly effective both in attack and defense, i.e. it needs few attempts to score a goal (low ratio goals/attempts) and the opponent needs many attempts to score a goal (high ratio goals/attempts). Conversely, the Pezzali score of a team is low when the team is ineffective both in attack and defense: it needs many attempts to score a goal while the opponent needs few attempts to score a goal. From Figure 8 it is evident how the simulation overestimates the points gained by teams with a low Pezzali score, while it underestimates the number of points for teams with high Pezzali score. Real Betis, for example, has the lowest Pezzali score w.r.t. all the teams in Spanish League and presents the highest ranking error according to our simulations (Figure 8, on the left). In other words Real Betis *suffers* the harsh rule of the goals: though it produces considerable passing activity, it is not effective in scoring and easily concedes goals to opponents. Bologna in the Italian league is another example of this behavior (Figure 8, on the left): its passing behavior led the simulation to wrongly overestimate its success of 20 points, that have been enough to save Bologna from the relegation to the second division it actually reported. The actual success of other teams, conversely, is underestimated by our simulation. An example is Hellas Verona (Figure 8, on the right): although its passing activity suggests poor attack performances, Hellas Verona shows a high attack efficiency allowing its forwarder Luca Toni to score 20 goals (second best scorer of the tournament). Hellas Verona, Genoa and Borussia Möchengladbach (see Figure 8, on the right) are typical examples of teams which *impose* the harsh rule of the goals: they achieve high score efficiency while conceding very few goals to the opponents. Interestingly, the strongest European teams lie in the middle of these two extreme behaviors (see Figure 8, the insets). They have an average Pezzali score and the simulation predicts the

TABLE V. SIMULATED RANKING AND ACTUAL RANKING OF GERMAN LEAGUE (SIMULATION ON H INDICATOR).

simulated ranking		real ranking	
Bayern	95	**Bayern**	90
Dortmund	75	Dortmund	71
Wolfsburg	62	Schalke	64
Leverkusen	59	Leverkusen	61
Augsburg	54	Wolfsburg	60
Hoffenheim	54	Mönchengladbach	55
Hannover	49	Mainz	53
Schalke	47	Augsburg	52
Hertha	43	Hoffenheim	44
Mönchengladbach	42	Hannover	42
Mainz	40	Hertha	41
Hamburg	40	Werder	39
Stuttgart	38	Freiburg	36
Frankfurt	34	Frankfurt	36
Nürnberg	29	Stuttgart	32
Braunschweig	26	Hamburg	27
Freiburg	24	Nürnberg	26
Werder	22	Braunschweig	25

points they achieved in the final ranking with high precision (ranking error close to zero). This suggests that, according to our performance indicators, those teams behave in a peculiar way: they produces high passing activity during the games (high H indicator), they exploit many of the numerous goal attempts they create, their defense strategy is effective hence not allowing opponents to score easily.

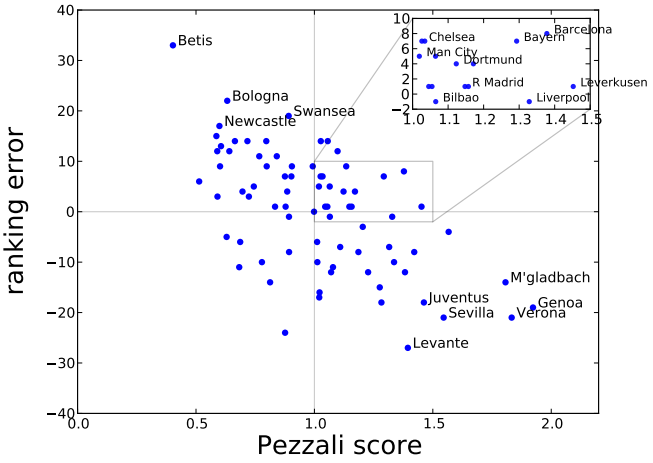

Fig. 8. Ranking error as a function of Pezzali score. Each point represents a football team, on the x axis the mean Pezzali score of the team during the season, on the y axis the difference between the points in the final ranking according to the simulation and the real points they achieved in the real championship. Teams with extreme values of Pezzali score present high ranking error (Betis, Bologna on the left, Hellas Verona, Genoa and Borussia Möchengladbach on the right). The insets highlights the presence of a peculiar zone where the top teams lie.

VII. RELATED WORKS

Data Science have entered the world of sports during the past decade and increased its pervasiveness as the technological limits were pushed up [10]. Individual and team sports are highly dependent from data: from professional cyclists to amateurs, all sportsmen are collecting data from easy-to-get monitoring devices. Cintia et al. [11][12][13] developed a first large scale data-driven study on cyclists' performance by analyzing data about workout habits of 30,000 amateur cyclists, downloaded from a popular fitness social network application. The analysis revelead that cyclists' wourkouts and performances follow a precise pattern, thus discovering an efficient training program completely learned from data. NBA basketball league is monitored in every possible dimension: Performance Efficiency Rating [14] is nowadays a stable and well known measure, able to assess players' performance by combining the manifold type of data gathered during every game (i.e. pass completed, shots achieved, etc.). In the context of tennis, Terroba et al. presented a pattern discovery exploration to find common winning tactics in tennis matches [15]. Smith et al. propose a Bayesian classifier for predicting baseball awards, prizes assigned to the best pitchers in the US Major Baseball League. The model is correct in the 80% of the cases, highlighting the usefulness of underlying data on describing sports results and performances [16].

Advances in computer vision and image processing opened up a wide research area focused on positional data from team sports, such as football, hockey and basketball. The possibility to observe a football game by means of players' spatio-temporal positions introduced a new scenario in the Data Science world as data mining theories and methods can be applied on these new data sources. The behaviors, strategies and decisions of two football teams during a match have attracted the attention of scientists since the past decade [17]. Borrie et al. [18] used T-Pattern detection to find similar sequences of passes from games. Gudmunsson and Wolle [19] analyzed and clustered players' sub-trajectories using Frechét distance as similarity measure. The same authors encoded and mined typical sequences of passes by using suffix trees [20]. Still looking at the problem from a data mining perspective, Bialkowski et al. [21] extracted players' roles over time by clustering spatio-temporal data on players' positions during a game. Gyarmati et al. [22] mined frequent motifs from teams passing sequences in order to classify team playing style. They discovered that FC Barcelona, the most awarded team in the last decade, has unique passing strategy and playing style. Tamura and Masuda [23] used Japanese and German football data to investigate correlates between temporal patterns of formation changes across games and game results. They found that teams and managers tend to stick to the current formation after a win and switch to a different formation after a loss, showing a win-stay lose-shift behavior.

Taki and Hasegawa [24] introduced a geometric model named *dominant region*, based on Voronoi spatial classifications [25]: in such model the football pitch is divided into cells owned by the players that reach every point of the cell before any other player. The concept of dominant region is further developed by Fujimura and Sugihara [26] who defined an efficient approximation for region computations. On the top of this model, Gudmunsson and Wolle [20] built a passes analysis

based on passing options computations; such investigation revelead the ability of a player to enforce and maximize the dominance of his team. Horton et al. examined another branch of passes classification research area: in [27] they performed a supervised learning of passes efficiency by involving domain expert to rate the features of a pass between two players. Lucey et al. [28] exploited shots to make a similar analysis: the result of their work is a shot outcome prediction method which considers strategic features (i.e. defender positions) extracted from spatio-temporal data.

Another approach to the problem of football data analysis is based on network theory. Players can be easily identified as nodes of a network where a pass between two players represents a link between the respective nodes. The first attempt in this field is the one by Peña and Touchette [6] who analyzed the matches of FIFA 2010 World Cup adopting typical network analysis tools. As a result, they highlighted how the two teams that reached the final (Spain and Netherlands) show the two highest values of average clustering; in other words, network representation of Spain and Netherlands playing strategies revelead a high tendency of their nodes to cluster together, forming communities of high connected – thus extremely ball exchanging – nodes. Similar conclusions are reported by Clemente et al. [29]: after a density evaluation of teams' playing networks, they show that network metrics can be a powerful tool to assess players connections, strength of such links and therefore help support decision and training processes. Cintia et al. [7][30] exploit passing networks to predict the outcomes of football games, showing that a network-based approach is more effective on long-running competitions like national leagues.

VIII. Conclusion and Future Works

In this work we do a further step towards the understanding of football through Data Science by performing a data analysis of 1,446 football games in the four major European leagues. We first show that the passing activity of a team is related to its success during the competition, and then extract six indicators measuring different aspects of a team's passing activity. We use these pass-based indicators to describe the performance of a team during a game. To investigate how much our indicators are descriptive of a team's performance we perform two types of analyses. First, we investigate the typical differences in the value of indicators of the teams according to the outcome of a game, thus constructing a set of classifiers to predict the outcome of a game based on the value of the indicators of the two teams in *previous* games. Second, we simulate all the games in the four leagues by computing, for each round in the season, a ranking of the teams according to the outcomes of the simulated game. We observe that the correlation round by round between the actual ranking and the simulated ranking increases as the season goes by, reaching maximum values of 0.89. The simulated ranking is particularly reliable for the top teams, i.e. teams qualified to Champions League. We show that teams with high ranking error (difference between official and simulated ranking) present two extreme, opposite behaviors: they have either high Pezzali score or low Pezzali score, a measure of defense/attack efficiency. Top European teams lie in the middle of the two extreme behaviors, presenting average Pezzali score values and high values of our performance indicators. Our results suggest that the proposed indicators are powerful tools to describe to performance of a team.

Passing activity, however, is only one of the many aspects characterizing the complexity of a football team, and our model can be further improved in several ways. First of all, we include no information about the difficulty of a pass in our performance indicators. Figure 9 shows the distribution of the distance covered by successful passes (blue solid line) and failed passes (red dashed line). While we observe two peaks relative to 20 meters for both successful and unsuccessful passes, many failed passes surprisingly take place at short distances (a peak exists at two meters or so) referring to passes close to the opponent's goal. This suggests that passes, either short or long, have different difficulties depending on additional features, i.e. the zone where they take place and the direction of the pass (forward/backward). A philosophical observation: football is a game in which the probability of scoring grows as one approaches the goal, and the probability of losing the ball raises as one approaches the goal. It is mandatory to risk losses, and to lose the control of the game many times, in order to succeed. On this aspect, football resembles many other real-world situations (namely war, to state one). Our pass-based indicators can be improved by including information about the difficulty of passes according to position on the pitch, direction, and proximity to the opponent's goal.

Second, we have not included information about defensive events (tackles, goalkeeping actions, recoveries of ball and so on) because they are not available in our dataset. The story narrated in this paper shows that for a subset of teams, characterized by extreme values of Pezzali score, the only passing activity is not able to accurately represent their performance during the games. Defensive actions are indeed crucial in the strategy of a team and they can improve significantly the description of the performance of a team.

Third, while in this paper we focus on the performance of *teams*, it would be interesting to study the problem focusing on the performance of *players* in order to detect the features which identify successful players. Provided that we now have more detailed data, we plan to include the above aspects in order to refine our performance indicators.

Acknowledgment

The authors wish to thank TIM company, Mariano Tredicini and Maven 7 for supporting part of our research. We also thank Filippo Simini, Fabrizio Lillo, Daniele Tantari, Maurizio Mangione, Adriano Bacconi and Albert-László Barabási for the insightful discussions, Salvatore Rinzivillo for his support on data visualization. Thanks also to Luca De Biase, Pierangelo Soldavini, and Carlo Morosi for their interest in our work. Finally, we wish to thank Max Pezzali for the inspiration about the "harsh rule of the goals".

This work was partially funded by the European Community's H2020 Program under the funding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement #641191 "CIMPLEX: Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories" (https://www.cimplex-project.eu/).

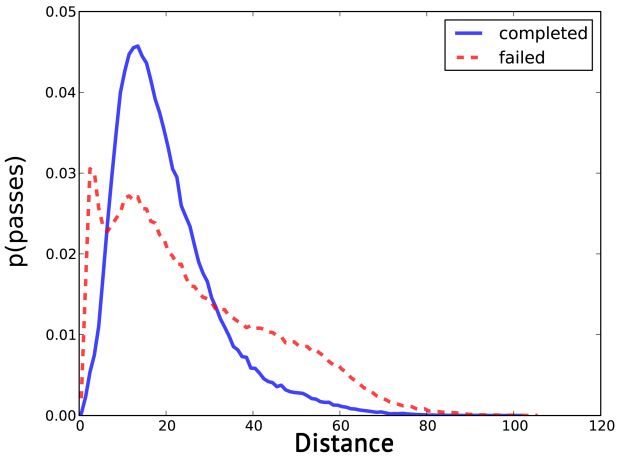

Fig. 9. Distribution of completed passes (blue solid line) and failed passes (red dashed line). Both distributions have a peak at 20 meters. The distribution of failed passes shows also a peak at 2 meters, corresponding to short distance passes close to opponent's goal.

REFERENCES

[1] C. Reep and B. Benjamin, "Skill and chance in association football," *Journal of the Royal Statistical Society*, vol. 131, pp. 581–585, 1968.

[2] C. Reep, R. Pollard, and B. Benjamin, "Skill and chance in ball games," *Journal of the Royal Statistical Society*, vol. 134, pp. 623–629, 1971.

[3] Prozone sports. [Online]. Available: www.prozonesports.com

[4] Opta sports. [Online]. Available: www.optasports.com

[5] M. Lames and T. McGarry, "On the search for reliable performance indicators in game sports," *International Journal of Performance Analysis in Sport*, vol. 7, no. 1, pp. 62–79, 2007.

[6] J. L. Peña and H. Touchette, "A network theory analysis of football strategies," *arXiv preprint arXiv:1206.6904*, 2012.

[7] P. Cintia, S. Rinzivillo, and L. Pappalardo, "A network-based approach to evaluate the performance of football teams," in *Proceedings of the Machine Learning and Data Mining for Sports Analytics workshop, ECML/PKDD 2015*, 2015.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] M. Pezzali, "La dura legge del gol," 1998.

[10] R. Schumaker, O. Solieman, and H. Chen, *Sports Data Mining*. Springer, 2010.

[11] P. Cintia, L. Pappalardo, and D. Pedreschi, "Engine matters: A first large scale data driven study on cyclists' performance," in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 147–153.

[12] ——, "Mining efficient training patterns of non-professional cyclists," in *22nd Italian Symposium on Advanced Database Systems (SEBD 2014), Sorrento Coast, Italy, June 16–18, 2014*, 2014, pp. 1–8.

[13] L. Pappalardo and P. Cintia, "We are the champions: the patterns of success in sports," http://bigdatatales.com/blog/2014/03/04/we-are-the-champions-the-patterns-of-success-in-sports-2/, March 4, 2014.

[14] J. Hollinger, "The player efficiency rating," 2009.

[15] A. Terroba, W. Kosters, and J. Vis, "Tactical analysis modeling through data mining," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2010.

[16] L. Smith, B. Lipcomb, and A. Smikins, "Data mining in sports: Predicting cy young award winners," *Journal of Computing Sciences in Colleges archive*, vol. 22, 2007.

[17] T. Reilly and A. M. Williams, *Science and soccer*. Routledge, 2003.

[18] A. Borrie, G. K. Jonsson, and M. S. Magnusson, "Temporal pattern analysis and its applicability in sport: an explanation and exemplar data," *Journal of sports sciences*, vol. 20, no. 10, pp. 845–852, 2002.

[19] J. Gudmundsson and T. Wolle, "Towards automated football analysis: Algorithms and data structures," in *Proc. 10th Australasian Conf. on Mathematics and Computers in Sport*. Citeseer, 2010.

[20] J. Gudmunsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, vol. 47, pp. 16–27, 2014.

[21] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Large-scale analysis of soccer matches using spatiotemporal tracking data," 2014.

[22] L. Gyarmati, H. Kwak, and P. Rodriguez, "Searching for a unique style in soccer," *arXiv preprint arXiv:1409.0308*, 2014.

[23] K. Tamura and N. Masuda, "Win-stay lose-shift strategy in formation changes in football," *EPJ Data Science*, vol. 4, no. 9, July 2015.

[24] T. Taki and J.-i. Hasegawa, "Visualization of dominant region in team games and its application to teamwork analysis," in *Computer Graphics International, 2000. Proceedings*. IEEE, 2000, pp. 227–235.

[25] M. De Berg, M. Van Kreveld, M. Overmars, and O. C. Schwarzkopf, *Computational geometry*. Springer, 2000.

[26] A. Fujimura and K. Sugihara, "Geometric analysis and quantitative evaluation of sport teamwork," *Systems and Computers in Japan*, vol. 36, no. 6, pp. 49–58, 2005.

[27] M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Classification of passes in football matches using spatiotemporal data," *arXiv preprint arXiv:1407.5093*, 2014.

[28] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, "quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data." MIT Sloan Sports Analytics Conference, 2014.

[29] F. M. Clemente, M. S. Couceiro, F. M. L. Martins, and R. S. Mendes, "Using network metrics in soccer: A macro-analysis," *Journal of human kinetics*, vol. 45, no. 1, pp. 123–134, 2015.

[30] L. Pappalardo and P. Cintia, "Taca la bala says the wizard: a trip into the world cup 2014," http://bigdatatales.com/blog/2014/07/12/taca-la-bala-says-the-wizard-a-trip-into-the-world-cup-2014/, 2014.