Estimating Goal-Scoring Probabilities in Soccer, Based on Physical and Geometric Factors

אמידת הסתברויות הבקעת שערים בכדורגל, בהתבסס על מאפיינים פיזיקליים וגיאומטריים

M.Sc. Research Proposal

Abir Koren

Advisors: Prof. Michal Penn, Prof. Yaacov Ritov

The Faculty of Industrial Engineering and Management

Technion- Israel Institution of Technology

February 2013

Contents

1 Introduction

Association football, commonly known as football or soccer, is a sport played between two teams of eleven players with a spherical ball. At the turn of the twenty-first century, the game was played by over 250 million players all around the world[1], which is why it is considered by many to be the world's most popular sport. The game is played on a rectangular field of grass or green artificial turf, with a goal in the middle of each of the short ends.

The objective of the game is to score more goals than the opposing team. Scoring is achieved by driving the ball into the opposing goal. A scoring attempt or a shot is a move in the game aiming to score, usually by kicking or heading towards the goalmouth. Scoring attempts differ from each other by various parameters such as: the position from which the attempt was made, the positions of players (and goalkeeper) in the area between the attempt position and the goal, the type of execution (a header versus a kick), the attempting player's abilities, the pressure turned against the attempting player (by an opposition player), the weather, the stadium in which the game takes place, the current score and many more. These parameters have different properties and thus one can make a general classification of them to: (a) Match characteristics—reflecting the differences between matches such as home or away match, the type of pitch, the playing teams' league ranks and so on; (b) Player characteristics—reflecting the scoring attempt performer's abilities; (c) Physical and geometric characteristics—reflecting the differences in the state of the game at the moment of the shot such as the location from which the attempt was made, the ball speed and so on. The last group's parameters can sometimes be easily interpreted. Using such interpretations enables a comparison between attempts made under highly different circumstances such as shooting a penalty kick against shooting from thirty meters with many players in the range between the origin of the shot and the goal. Nevertheless, this comparison becomes much harder when the differences in the characteristics of the shot are not so significant.

In mainstream soccer statistics, one merely counts scoring attempts while the only differentiation among them is whether an attempt went on or off target (i.e. the differentiation is outcome based). The kind of information that this sort of statistics can provide may give insufficient insights on a team's performance in a game. When the differences in these measures,

between two playing teams, is extreme, one can assume that the team with the higher measure played better. But this need not be the case and, furthermore, having the higher measure in these parameters does not necessarily indicates a win. There exist many contradictory examples, in which the winning team had fewer shots on target but these shots were performed under much better conditions.

The objective of our proposed research is to evaluate and quantify the effects of the physical and geometric characteristics of a scoring attempt. Creating a proper model regarding these effects can make different attempts comparable. Evaluations of players, evaluations of teams and even analyses of planned moves can be achieved by relying on scoring attempts analysis. For this purpose, the model will relate mainly to these physical and geometric parameters, and hence the effects of other types of parameters will be eliminated using statistical methods.

This research proposal is arranged as follows: in Section 2 we present a brief literature review. In Section 3 we introduce the problem and our research objectives. Our proposed methodology for solving the problem is presented in Section 4. In Section 5 we describe the data sources we plan on using.

2 Literature review

2.1 OR and statistical models in soccer

Brillinger [2] presents a review of a variety of disciplines that have been applied in order to analyze soccer or certain aspects of the game. In the review, different types of available data and descriptive statistical applications are described, and used as an example of the clear insights into the game achievable using simple methods and proper data (e.g. the discovery of the home advantage). It is suggested there that the existence of such data holds great potential in terms of analyzing games. Stochastic models are presented as a method that helps to describe "Between games" relations as well as "Within game" relations. Among the "Between games" models, the number of goals scored by a team in a game, the number of passes in a successful passing movement, the final score of a match (considering relevant covariates) and the number of points collected in a tournament, are discussed while the "Within games" models handle the progress of a single game in terms of evaluating the probabilities of game events (such as

scoring, substitutions and a red card), and estimating their effect on the game.

Brillinger[2] also presents various models, in which different methods were applied, for analysis of ranking and ranking methods of soccer teams, according to several measures of their performances. Ranking is then presented as one of the factors that may affect the scheduling of tournaments, as well as other external factors that may have an influence on the scheduling, such as: traveling distances (of competing teams), budgets (of competing teams and the organizers of the tournament), schedules of teams from the same city, referees' time and even TV schedules.

Game theory models, according to Brillinger[2], provide a general framework to study soccer, though it is sometimes difficult to apply them due to the game's dynamics. The presented models analyze technical game theory problems, such as penalty shooting and point gaining, and general aspects of game theory, analyzing different tactics and tactical decisions.

Furthermore, Brillinger[2] refers to management and economics models which had evolved as a result of the understanding that soccer is an industry, and that it can not rely on traditional methods, and needs some proper managerial tools in order to be organized. The models define the management of a soccer team as the providers of the super-structure, and relate to their specific roles. The implementation of methods following this kind of definition, and other basic managerial methods, such as objective functions, loss functions and profit/revenue analysis, aims to improve the team's performance in several aspects.

The review[2] is recent, hence we refer the reader to it for further references on OR and statistical models in soccer.

2.2 Factors associated with scoring probability

The subject of estimating the probability of scoring, as well as evaluating the effect of different factors on scoring, has been addressed from different aspects. The levels of analysis vary greatly, ranging from a simple search for patterns among scored goals and among the events prior to them, to the developments of formulas, that can estimate scoring probability under various situations.

In one of the early works that links statistics and soccer, Reep and Benjamin[8] investigate patterns of scored goals. In their work, they calculate shots and goals related ratios. These ratios reflect various kinds of relations between shots, goals, and some of their characteristics, such

as the field zone they were made in or the type of event that generated the attack that ended in the shot. The results exhibited consistency among several different tournaments. Furthermore, it was found that on average it takes ten shots to score a goal.

Olsen[4] analyzes the goals scored in the 1986 Mexico World-Championship. It is claimed there that the tournament was influenced by certain climatic conditions, which led to a use of attacking and defending styles, that were expected to yield a small number of scoring opportunities. The fact that there was a significantly large number of goals scored in the tournament, indicated a high level of attacking skill, and was an incentive to analyze the goals scored and the attacking movements leading to them, in terms of space and time. The time aspect was represented in terms of the number of touches on the ball, by the goal scorer and the assisting player, as well as the number of the passes completed within the attacking movement. The space aspect was represented by the amount of space (i.e. the distance from the nearest defender) the goal scorer and the assisting player had, at the moment of the shot and pass respectively. Other parameters, reflecting the origin of the attacking movement, were also taken into consideration. The analysis was based on descriptive statistics, and exhibited variations in both the number and the percentage of goals scored under different levels of each parameter. It was found that most goals were scored after one touch of the ball by the goal scorer, suggesting that a faster, or more immediate, attempt lowers the defense's ability to respond and, thus, increases the probability of scoring. Furthermore, it was found that most goals were scored within a distance of sixteen meters from the goal, and about twenty percent of the goals were scored under strong defensive pressure. The analysis that was based on the parameters of events prior to the shot (reflecting the assisting player time and space and the origin of the attacking movement) exhibited different results. Olsen[4] suggests that this difference may be caused by the difference in the game circumstances (e.g. distance from goal) between the goal scorer and other players involved in the attacking movement. Olsen[4] derives from this analysis some practical insights and advices for soccer coaches, and concludes that extending this kind of analysis may help to gain valuable insights into many aspects of elite soccer.

Pollard[5] extends Reep and Benjamin's[8] early work by investigating the relation between shots and goals with respect to the locations of the shots. Using a data-base covering 3,931 shots (of which 394 were goals), goals to shots ratios were calculated for shots from outside

the penalty area, shots from inside the penalty area, penalty kicks and for the whole set without separation. The ratio that was calculated for all the shots exhibited consistency with Reep and Benjamin's[8]work, yielding a 1 goal per 10 shots ratio, while the other ratios, that were based on the separate data, established a 1 goal per 6.5 shots ratio for inside the penalty area, 1 goal per 46.1 shots ratio for outside the penalty area and 1 goal per 1.2 shots ratio for penalty kicks. Furthermore, Pollard[5] found that even inside the penalty area, the shots per goal ratios vary when considering a more refined segmentation of thelocations of the shotn. Based on these findings, it is concluded that there is an inverse ratio between the distance from goal and the scoring probability.

Pollard and Reep[7] propose a method that enables a comparison between different soccer tactics and strategies by estimating their effectiveness. The main idea is to quantify and estimate the outcome of a team's possession (i.e. a continuous period of time in which the ball remains with one team). Since goals and even shots are relatively rare events in a soccer game, a use of them as outcome parameters is insufficient, and thus a more sophisticated measure was developed. The developed estimator is based on the estimated probability of scoring, and for that reason a proper estimate of the scoring probability was needed. Following the assumption that the scoring probability varies greatly with location as well as other quantifiable factors, a logistic regression was conducted over a set of 489 team possessions that resulted in a shot of which 47 were goals. The logistic regression was applied using the goal/no-goal as the dependent variable, while representatives of the location of the shot, the number of ball touches—made by the player taking the shot—prior to the shot, the distance from the nearest defender and the type of play leading to the shot (i.e.open or set play), were selected as explanatory variables. Furthermore, the data was divided into groups according to the type of shot (i.e.a kick, a header or a penalty kick) in order to test whether the explanatory variables have different effects under different types of shot.

Pollard and Reep[7] found that the significant factors affecting the scoring probability in all sets of data were the location (in terms of distance and angle) and the type of play leading to the shot; the other variables, as well as the interactions between them, were found insignificant. Furthermore, there was a significant difference in the values of the coefficients between the kicks and headers, in particular, the distance from the nearest defender was found significant

for kicks but not for headers. In addition to determining the significant factors, the logistic regression also provides a formula that enables one to estimate the scoring probability for any shot, as a function of its significant factors. Having this analysis carried out enabled the authors to compute the mentioned team-possession outcome estimator. The fact that the scoring probability can be estimated under various circumstances, enables the outcome estimator to relate to the same factors, and hence an improvement in the scoring probability estimate will refine the outcome estimator as well.

Ensum et al.[3] expand upon Pollard and Reep's[7] approach by adding factors relating to the goalkeeper's position, the number of players between the shot taker and the goal, the pitch position from which the possession was regained (i.e.where the attack originated) and whether the shot followed a cross. The factors were divided to two groups: factors related to the exact time of the shot and factors preceding this time. The analysis followed Pollard and Reep's[7] definitions and methods but was conducted on a different set of data. Consistency with Pollard and Reep's[7] findings was exhibited for both kicked and headed shots. The factors reflecting the location and the distance from the nearest defender for kicked shots were found significant as well as the factors reflecting the location for headed shots. Except for these factors, the findings of this study were not consistent with Pollard and Reep's[7] work. It was suggested that the discrepancies may be a result of differences in the data recording methods and differences in the analyzed factor combinations, as well as other circumstantial differences. Moreover, it was found that most factors preceding the shot were not significant, and hence the importance of measuring factors as close as possible to the moment of the shot was concluded.

In addition, Ensum et al.[3] suggest an application of the discussed results for estimating a team's performance: The summation of the scoring probabilities achieved by a team in a certain game (or a number of games) and the comparison of this sum to the actual number of goals scored, may indicate the team's as well as specific players' scoring abilities; similarly, the summation of the scoring probabilities achieved by the opposing team and the comparison of this sum to the actual number of goals scored by the opposing team, may indicate the defensive abilities. It was shown that the sum of all shot probabilities, for some selected teams, provided good estimate of the total number of goals that were expected to be scored. It is finally suggested that a more detailed study, based on broader data sets, may provide more accurate information,

and hence contribute to performance analysis as well as implementations of effective training methods.

Pollard et al.[6] have also conducted a logistic regression in order to estimate the scoring probability under various situations. The study was based on the combined data from the two previous studies, excluding headed shots, penalty kicks and direct shots from a free kick. It was decided to analyze the data according to the factors that were common to both studies which covered the location of the shot, the distance from the nearest defender, the type of play leading to the shot (i.e. open or set play), the number of touches of the ball by the player prior to the shot, whether the shot followed a cross and the zone of origin of the move leading up to the shot. The logistic regression was again conducted with goal/no-goal as the dependent variable. The final model consisted of the factors reflecting the location and the distance from the nearest defender, which exhibited a significant effect on the probability of a shot becoming a goal. It is concluded that even though other factors still require further investigation, these findings can still be useful for several analytical purposes, as outlined in Pollard and Reep[7] and Ensum et al.[3].

Wright et al.[10] analyzes data that was collected from one full season of the English FA Premier League; it contained 1,788 attempts of which 169 were goals. In order to investigate the factors associated with an attempt on goal, recorded games were coded, extracting key behaviors and events including: the position of the shot, the shot type, the initiation of attack type of event, the number of players between the location of the shot and the goal keeper, the goalkeeper's position, the type of feed leading up to the shot, the position of feed, the reception of the feed type of event and the number of passes leading up to the shot. Analyzing the data, using descriptive statistics, exhibited consistency with previous works in terms of the goals per shots ratio among different levels of some of the mentioned factors (e.g. 87% of the analyzed goals were scored inside the penalty area, supporting a prediction in[10] of more than 70% to be scored from that area). Wright et al.[10] also investigated the data by conducting a logistic regression, using goal/no-goal as the dependent variable and seven covariates consisting of twenty-seven predictor variables. It was found that different levels of the factors: location of the shot, goalkeeper's position, the number of players between the location of the shot and the goal keeper, the type of the shot and the position of feed were significant. It is then argued that

despite of the significance of new factors in this model, the low value of the explained variance ($R^2 = 21\%$) suggests that there is plenty of variability to be investigated, which has not been accounted for.

3 Problem definition and research objectives

We define a scoring attempt or a shot as a player's attempt to drive the ball into the opposing team's goal. The shot can be made using any part of the body (apart from the hand due to the rules of the game) located at any position on the pitch. A shot has many characteristics; we divide them into three groups: (a) match characteristics - reflecting external differences such as the playing teams, the current score, whether the game is at home or away and so on; (b) personal characteristics - reflecting the differences in the abilities of the player who performs the shot such as kicking accuracy, heading accuracy and kicking power; (c) physical and geometric characteristics - reflecting the differences in the state of the game at the moment of the shot, such as the location at which the shot was made, the part of body used for the shot, the location of other players in the range between the ball and the goalmouth at the moment of the shot, the position of the goalkeeper and the official time of the shot. An important property of the parameters of the last group is that they can be easily and accurately measured, which potentialy makes them objective parameters.

Despite this variation in characteristics, any shot has only 5 possible outcomes: (a) it may be deflected off another player; (b) it may be deflected or caught by the goalkeeper; (c) it may hit one of the posts; (d) it may complete its movement without hitting the goalmouth (usually outside of the pitch); (e) it may be a goal scored. In many cases, the outcome of a shot may combine some of these outcomes (e.g. the ball deflects off a player into the goal). This raises the need for a dichotomous definition of the outcome, which will enable one to classify any shot to one group (there will not be intersections between the groups). Moreover, in order to achieve these classification abilities, outcomes must be well defined so that the classification will be as objective as possible.

Considering these attributes, one may regard a shot as a "black box" with its characteristics as inputs and the outcome as an output. Modeling and analyzing this "black box" may reveal the

connections between the inputs and outputs, and thus may support estimation of the different outcomes probabilities regarding a given shot. In particular, having a proper estimator of the scoring probability may have a major impact on soccer analysis; evaluating playing strategies (as exhibited in Pollard and Reep[7]), evaluating players' and teams' abilities (as outlined in Wright et al.[10]), analyzing matches off-line and improving decision making of both players and coaches, are only a few examples of areas that may benefit from such an estimate.

Wright et al.[10], as well as Pollard et al.[6] and Pollard and Reep[7], have found that the factors position (i.e. distance and angle from goal) and space (i.e. the distance from the nearest defender) are significant in explaining the dependent variable goal/no-goal when conducting a logistic regression procedure. Furthermore, they claim that the probability of scoring decreases when the distance increases, when the angle (relative to a vertical line from one of the posts) increases and when the distance from the nearest defender decreases. The factors—position of goalkeeper, type of play (set play or open play) and type of shot—were handled differently among the studies, so the different results regarding them imply that these factors should be taken into consideration. Despite these findings, the fraction of the explained variance was low, which indicates that these models are amenable to improvements.

In our research, we plan to develop a model that will assist in estimating the scoring probability, under any objective and measurable circumstances. While existing models evaluate the above-mentioned factors and simple correlations among them (i.e. in terms of using multiplications of them as covariates), we seek to model these factors and their correlations in a way that will reflect their physical meaning. Using such interpretations will allow us to apply the model to the dependent variable on-target/off-target as well. Furthermore, we will take into consideration shots that resulted in any of the possible outcomes, in contrast to earlier studies that excluded shots that deflected of a defender and/or shots that were off target. Considering the different approach we present, we also have interest in comparing our model to existing ones; such a comparison will hopefully reveal the quality of our model.

Further investigation would be considered regarding the forecasting of a match score, using an estimate of aggregated probabilities (i.e. the sum of scoring probabilities gained in a match) that were calculated according to our model. We will also consider the implementation of the main ideas to other events of the game, particularly passing. Moreover, sensitivity analysis of

the probability values as a function of the input factors may be useful in answering some basic questions such as finding the optimal location of a shot under a given situation.

4 Methodology

We aim to model a scoring attempt in a way that will reflect its physical and geometric characteristics and their correlations as well as their effect on the scoring probability of the attempt. We wish to create a model as general and as objective as possible. In order to reduce subjectivity, we will regard only well defined and measurable aspects; in practice, this means that the model will be based mainly on the spatial temporal coordinates of the players involved in the shot (the attempting player, the defenders and the goalkeeper). The possible outcomes of a shot will also be defined using a dichotomous definition, eliminating any aspects of subjectivity. The output of the model will then be tested as a covariate of two dependent binary variables: goal/no-goal and on/off target. In this way we will ensure an objective model in terms both of inputs (explanatory variables) and outputs (dependent variables).

We believe that the physical and geometric properties of a shot affect the shot outcome mainly through the effective goal size or effective paths toward the goal. Here we define *effective* (goal size or paths) as feasible for scoring; for instance, if a player stands at the goalmouth, a part of the goal is blocked and also some paths toward the goal are eliminated, and hence this part of the goal and the eliminated paths are not feasible for scoring, that is the effective goal size and the effective paths toward the goal are reduced. Thus, the essence of a model, aiming to capture the physical and geometric effects on the outcome of a shot, would be based on a calculation of the effective goal size or the effective paths of the ball toward the goal.

We will consider two model representations, the first quantifies the effective paths toward the goal—we will refer to it as *the path representation*—and the second quantifies the effective goal size—we will refer to it as *the goal size representation*.

For the path representation, we define *a possible trajectory* as a feasible ball trajectory that starts at a given ball location (defined by the three dimensions of the ball-center) and ends at a point within the goalmouth. Since a goal is achieved when the ball passes through the goalmouth, we assume that if a goal was scored, then the ball traveled via a possible trajectory.

Any possible trajectory can be represented by the straight line connecting its origin to its end as shown in Figure 1; we define this line as *a straight path to the goal* (or *a straight path*). Therefore, a straight path represents all the possible trajectories that start and end at the same coordinates as the straight path itself. A smaller set of straight paths in a certain situation represent a smaller set of possible trajectories from the position of the shot to the goalmouth and hence a lower probability of scoring.

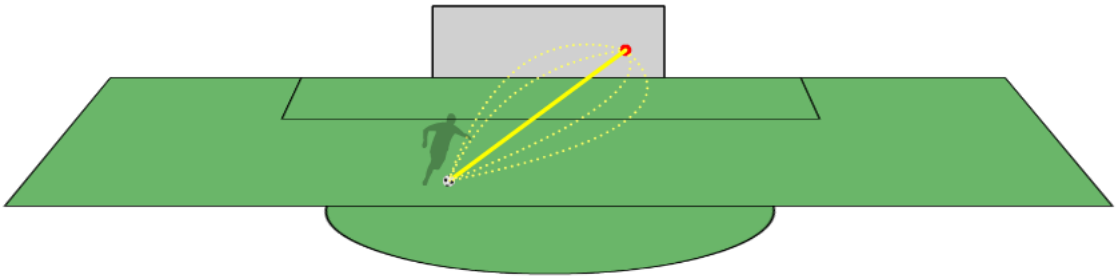

Figure 1: Possible trajectories and a straight path. The thin dashed curves are some of the possible trajectories from this given shot origin and the thick line is the straight path that represents them.

Following these definitions, when considering a certain origin of a shot and no players in the range between the ball and the goalmouth, we can relate to the pyramid, formed by the location of the center of the ball (in three dimensions) as the apex and the goalmouth as the base, that contains all the straight paths from this location as a representative of all of the possible trajectories as presented in Figure 2; we will refer to this pyramid as the *path pyramid*. Since the path pyramid represents all the possible trajectories of a shot, we shall relate to it as the shot representative and the model calculations will relate to the path pyramid.

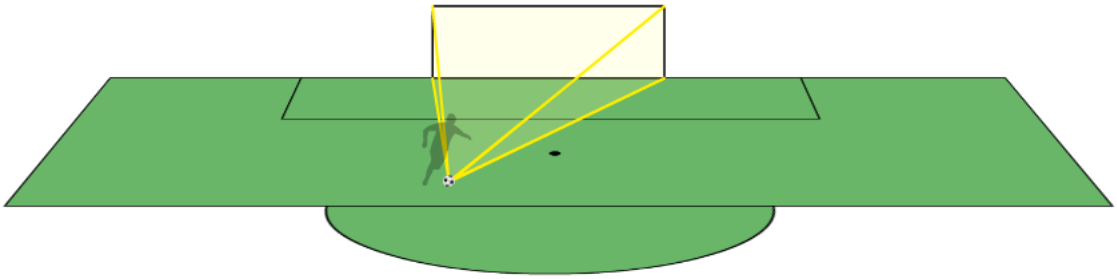

Figure 2: A path pyramid. The path pyramid is formed by the location of the center of the ball as the apex and the goalmouth as the base.

The goal size representation will be based on the solid angle (i.e. the projection of a specified object on a unit sphere centered at the point of observation) of the goalmouth, as measured from the location of the center of the ball as described in Figure 3. The solid angle considers the distance of the measured object—the solid angle decreases when the distance of the measured object increases—as well as its orientation—the measured object projection on a unit sphere changes according to its relative orientation to the point of measure.

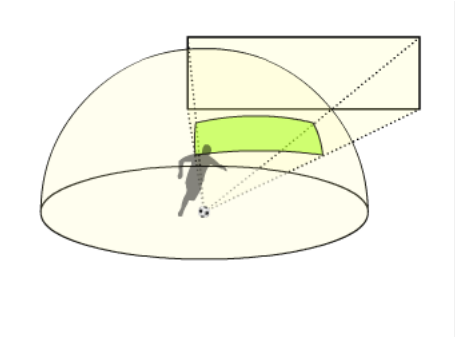

Figure 3: The solid angle of the goalmouth. The presented sphere is a unit sphere centered at the location of the shot and the marked area on the sphere is the goalmout projection on the sphere.

Given a shot we will refer to the length of the straight path that connects the origin of the shot with the center of the goalmouth as the *distance of the shot*, and the angle of this straight path (relative to a line that is vertical for the goal line) will be referred to as *the angle of the shot* as shown in Figure 4. When considering the solid angle of the goalmouth as measured from a shot location, the solid angle decreases when the distance of the shot increases; and it decreases when the angle of the shot increases.

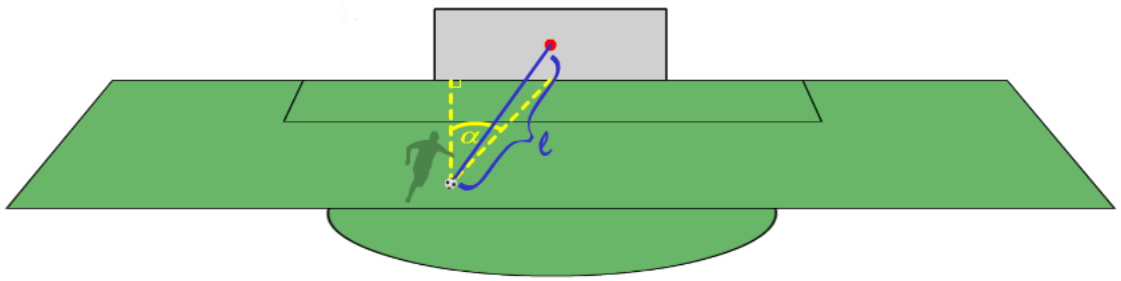

Figure 4: The distance and the angle of a shot. The blue line is the straight path connecting the location of the shot and the center of the goalmouth; its length, denoted by l, is the distance of the shot, and the angle of its projection on the ground relative to a line that is vertical to the goal line, denoted by α, is the angle of the shot.

Considering these two representations, we shall investigate the effects of the physical and geometric factors on a shot and its probability to be scored as a goal, in terms of each of the described representations. A detailed analysis of the effects of these factors would be based on the following observations:

- There is an inverse ratio between the length of a straight path and the probability of shooting the ball in the trajectories the path represents. In other words there is an inverse ratio between distance and accuracy.
- Since the path pyramid represents all the straight paths of a given shot, accumulating the paths of a certain shot will involve integration over the base of the pyramid. This kind of integration will depend on the vertical and horizontal angles of the apex of the path pyramid.
- Obstacles (i.e. defending players, other players in the range and the goalkeeper) inside the path pyramid eliminate straight paths and also reduce the solid angle of the goalmouth.
- The length of a straight path and the velocity of the ball determine how long it will take the ball to travel from the origin of the shot to the goalmouth; this is also the time the goalkeeper has to get to the ball or to cover more of the goalmouth area (by reducing the effective goal size or eliminating more straight paths).

These observations actually describe physical and geometric interpretations of the characteristics of a shot. In our research, we shall explore different implementations of these observations on the path representation as well as the goal size representation. The main issues to be examined will adress: (a) The quantification of a single straight path—which function will describe the mentioned inverse ratio in the most accurate way? (b) The quantification of obstacles inside the path pyramid—how to quantify the effective goal size and paths towards goal reduction caused by each player? (c) Approximations—the classical modeling question, trying to find a balance between the simplicity of the model and its relation to reality. We shall investigate these issues for each of the two representations properly. Answering these questions will provide models relating to the physical and geometric factors, as desired.

We assume that modeling these parameters and observations properly can capture and explain the main differences between different shots; in particular a model based on these definitions can also explain the difference between kicked and headed shots as well as the difference between shots from a set play to shots from an open play. Moreover, we believe that every attempt (regardless of its outcome) gives us information, and may assist in evaluating the goodness of the model and hence, unlike previous studies, we will consider every possible shot. Combining the ability to explain the difference between various types of shots with the consideration of every shot, will allow our model to be general and thus compatible with a diverse types of soccer analyses.

In each of the representations, the model will provide a numerical output based on the discussed parameters. Using a logistic regression, we will test the significance of these outputs in explaining the goal/no goal dependent variable and the on/off target dependent variable separately. The logistic regression will further enable us to estimate the on-target and scoring probabilities under any given situation, and to estimate the magnitude of probability change as an outcome of a change in the inputs. Further analysis will investigate the correlation between these two dependent variables. We shall then be able to produce scoring and target hitting probability-contours as well as performing related analysis (e.g. the optimal location for shooting under a given situation) using these probabilities.

5 Data Sources

Common soccer statistics and data are available in many websites including www.uefa.com, www.fifa.com, www.soccerbase.com, www.soccerway.com, and www.soccerpunter.com. As mentioned before, these type of data are not sufficient for our research needs. In all of the previous works mentioned[3, 4, 7, 5, 8, 10], the data collection was performed using observers (sometimes the authors themselves), that used different types of notation-systems to record the events of interests as well as their characteristics. This method has some disadvantages relating the data reliability and validity; nevertheless, using this method will provide us with the desired data.

In order to conduct a more precise data collection, we have developed a computer interface

flash-based tool. The interface presents a half soccer-field screen and allows an observer to locate representatives of the players (i.e. the kicker, other players in the range between the location of the shot and the goal, and the goalkeeper) at any location on the half field. The interface then converts these locations to actual coordinates according to the pre-inputed pitch measures. Validity and reliability issues would still be a matter of concern, but using this interface will help to estimating locations in a consistent way.

There has been recent major progress in the area of soccer data collection[2]. Companies such as Prozone-Amisco, Match Analysis and Sport-Universal SA collect near-continuous high frequency digital spatial-temporal data. The data collection is based on an array of video cameras set up at stadiums. By signal processing, the spatial-temporal coordinates of the changing locations of the players on the field, the ball, and the referees, are extracted. This type of data allows a direct tracking of quantities and events such as tackles, crosses, distances covered, key moments and actions, possessions, passes, interceptions, runs with the ball, fouls, penalty kicks, challenges, entries into the opponent's area, ball touches, blocks, forward passes, long balls, high intensity running, ball velocity and accuracy, and balls received. Moreover, tracking data, collected by Prozone, showed a very high correlation with actual measures of the tracked activities as described in Di Salvo et al.[9]. These high levels of correlation, as well as other reliability scores that were measured by Di salvo et al.[9], indicate that this kind of data is reliable and reflects very accurately the movement of the game participants. Possession of such a data base will suffice for the needs of our research and we do aim to achieve it. We emphasize that a failure in obtaining such data will not result in a failure of our proposed research. Indeed, this will result in the recording of the data ourselves, using our described interface.

Glossary

***cross*:** delivery of a ball from either side of the field across to the front of the goal.

***Set-play*:** situation when the ball is returned to open play following a stoppage.

***Free-Kick*:** method of restarting play in a game of association football following a foul. The free-kick can be direct (from which one may score directly) or indirect (from which one may not score directly, that is another player other than the kicker must touch the ball before a score) according

to the type and position of the foul. Opponents must remain 9.15 meters (10 yards) from the ball until the ball is in play.

***Penalty-kick*:** type of direct free kick, taken from 12 yards (approximately 11 meters) out from the goal. All players other than the defending goalkeeper and the penalty taker must be outside the penalty area until the ball is kicked.

References

[1] Federation International Football Association et al. Fifa big count 2006: 270 million people active in football, 2008.

[2] D.R. Brillinger. Soccer/world football. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

[3] J. Ensum, R. Pollard, and S. Taylor. Applications of logistic regression to shots at goal at association football. In *Science and football V: the proceedings of the Fifth World Congress on Science and Football*, page 214. Routledge, 2005.

[4] E. Olsen. An analysis of goal scoring strategies in the world championship in mexico, 1986. *Science and football*, pages 372–376, 1988.

[5] R. Pollard. Do long shots pay off? *SOCCER JOURNAL-BINGHAMTON-NATIONAL SOCCER COACHES ASSOCIATION OF AMERICA-*, 40:41–43, 1995.

[6] R. Pollard, J. Ensum, and S. Taylor. Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. *International Journal of Soccer and Science*, 2(1):50–55.

[7] Richard Pollard and Charles Reep. Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(4):541–550, 1997.

[8] C. Reep and B. Benjamin. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General)*, 131(4):pp. 581–585, 1968.

[9] D.S. Valter, C. Adam, M.N. Barry, and C. Marco. Validation of prozone: A new video-based performance analysis system. *International Journal of Performance Analysis in Sport*, 6(1):108–119, 2006.

[10] C. Wright, S. Atkins, R. Polman, B. Jones, and L. Sargeson. Factors associated with goals and goal scoring opportunities in professional soccer. *International Journal of Performance Analysis in Sport*, 11(3):438–449, 2011.