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

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

Research Thesis

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Operation Research

Abir Koren

Submitted to the Senate of the Technion - Israel Institute of Technology

Shvat, 5775 Haifa February, 2015
This Research Thesis Was Done Under the Supervision of Professor Avishai Mandelbaum, from The Technion, and Professor Yaacov Ritov, from The Hebrew University of Jerusalem, in the Faculty of Industrial Engineering and Management.

The Generous Financial Help of the Technion
is Gratefully Acknowledged.

I would like to thank Professor Avishai Mandelbaum and Professor Yaacov Ritov for all their help and the effort they have put in this research. Their assistance in enabling such a unique research is sincerely appreciated. The efforts of Professor Nils Rudi from INSEAD are also gratefully acknowledged.

Many thanks are due to Mr. Nir Rikovitch for his assistance in formulating the physical model, Mr. Noam Carmel for the creation of the data-collection interface and to Mr. Or Smolnik for his invaluable assistance. Finally, I wish to thank my beloved wife Gili for her endless patience and support.
## Contents

1 Introduction .................................................. 3

2 Literature review ........................................... 4
   2.1 OR and statistical models in soccer .................. 4
   2.2 Factors associated with scoring probability .......... 6

3 Problem definition and research objectives ............... 10

4 Methodology .................................................. 12
   4.1 The models ................................................ 13
      4.1.1 Definitions .......................................... 14
      4.1.2 The solid-angle model .............................. 15
      4.1.3 The straight-path model ............................. 18
      4.1.4 Obstacles ........................................... 21
      4.1.5 Other factors ........................................ 26
   4.2 Data sources ............................................. 27

5 Physical analysis ........................................... 28
   5.1 Soccer ball motion equations ............................ 30
      5.1.1 TPBVP setup .......................................... 33
      5.1.2 Physical Constraints ................................. 34
   5.2 Trajectory analysis ...................................... 40
   5.3 Obstacle reduction ...................................... 54
      5.3.1 Z-axis analysis ...................................... 56
      5.3.2 X-Y plane analysis .................................. 61

6 Data and analysis .......................................... 69
   6.1 Data description ......................................... 71
   6.2 Data analysis ............................................ 74
      6.2.1 Non-parametric statistics ............................ 75
      6.2.2 Logistic regression ................................... 83

7 Applications and future work ............................... 95
List of Figures

1. The distance, direction and the angle of a shot. .............................................. 13
2. The solid angle of the goalmouth. ................................................................. 16
3. The output of the solid-angle model as a function of the distance of a shot. .............. 17
4. The output of the solid-angle model as a function of the angle of a shot. ................. 18
5. The contours of the solid-angle model on the soccer field. ................................ 18
6. A straight path as a representative of trajectories. ........................................... 19
7. The scoring space of a shot as a pyramid. ...................................................... 20
8. The output of the straight-path model as a function of the distance of a shot. .......... 22
9. The output of the straight-path model as a function of the angle of a shot. ............. 22
10. The contours of the straight-path model on the soccer field. ............................. 23
11. The additional horizontal area covered by the goalkeeper as a function of the motion time of the ball. ................................................................. 26
12. Motion equation solutions. ........................................................................... 35
13. Trajectories of a straight path with constant angular velocity. ......................... 38
14. Trajectories of a straight path with constant angular velocity with the initial velocity constraint. ................................................................. 40
15. Initial center of mass velocity norm and velocity angle as a function of the completion time. .................................................................................. 42
16. Velocity angle as a function of the center of mass initial velocity norm. ................ 42
17. Initial center of mass velocity norm as a function of the completion time. ............. 43
18. The effect of the $\phi$ angle on trajectory function. ......................................... 44
19. The effect of $\theta$ angle on the trajectory function. ......................................... 44
20. The effect of $\omega_z$ on the trajectory function. ............................................. 45
21. The effect of the distance on the trajectory function. ....................................... 45
22. Fitting a rational function to characterize the trajectory function. ....................... 47
23. Fitting a rational function to characterize the relation between the completion time and the velocity angle. ........................................................... 48
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>Fitting a power function to characterize the relation</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>between the line integral and the length of a straight</td>
<td></td>
</tr>
<tr>
<td></td>
<td>path.</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Fitting a linear function to characterize the relation</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>between the line integral and the length of a straight</td>
<td></td>
</tr>
<tr>
<td></td>
<td>path.</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Surface of trajectories.</td>
<td>51</td>
</tr>
<tr>
<td>27</td>
<td>Change in the trajectory function coefficient values as</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>a function of the angular velocity.</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Minimal and maximal completion times as a function of</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>the angular velocity.</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>Fitting a power function to characterize the relation</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>between the surface integral and the length of a straight</td>
<td></td>
</tr>
<tr>
<td></td>
<td>path.</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Difference between modeled and actual scoring space.</td>
<td>54</td>
</tr>
<tr>
<td>31</td>
<td>Heights of the trajectories of the representative</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>straight path.</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Heights of the trajectories of the representative</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>straight path as a function of their completion time.</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Quadratic fits for minimal trajectories.</td>
<td>59</td>
</tr>
<tr>
<td>34</td>
<td>Maximal heights of minimal trajectories.</td>
<td>60</td>
</tr>
<tr>
<td>35</td>
<td>Combining the two methods of an obstacle’s reduction.</td>
<td>61</td>
</tr>
<tr>
<td>36</td>
<td>Trajectories evading and intersecting with an obstacle</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>on the X-Y plane.</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Specifying the sets of intersecting and evading</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>trajectories, by completion time and initial velocity.</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>X-coordinate at an obstacle’s location as a function</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>of the completion time of the trajectory.</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>Intersection proportion as a function of the destination</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>of the straight path.</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Intersection proportion as a function of the destination</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>of the straight path calculated for various mid-range</td>
<td></td>
</tr>
<tr>
<td></td>
<td>obstacle’s locations.</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Intersection proportion as a function of the destination</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>of the straight path calculated for various near-edged</td>
<td></td>
</tr>
<tr>
<td></td>
<td>obstacle’s locations.</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Effective width of a defender as a function of his</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>location.</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>Obstacle widths comparison.</td>
<td>70</td>
</tr>
<tr>
<td>44</td>
<td>Obstacle width comparison for various angular velocities.</td>
<td>70</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Single-covariate logistic regressions with different model outputs</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>Comparing the model outputs to their transformations.</td>
<td>86</td>
</tr>
<tr>
<td>3</td>
<td>Comparing all the transformed model outputs in a single model</td>
<td>87</td>
</tr>
<tr>
<td>4</td>
<td>Comparing all the couples of the transformed model outputs</td>
<td>87</td>
</tr>
<tr>
<td>5</td>
<td>Incremental examinations of the two straight-path model outputs</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>Comparing the revised solid-angle transformed output to the straight path transformed outputs.</td>
<td>88</td>
</tr>
<tr>
<td>7</td>
<td>Factor effects examination.</td>
<td>89</td>
</tr>
<tr>
<td>8</td>
<td>Free-kick factor examinations.</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>Factor effects examination.</td>
<td>91</td>
</tr>
<tr>
<td>10</td>
<td>Assessment of the shot-type effect via a model with the solid-angle transformed output and the free kick factor</td>
<td>92</td>
</tr>
<tr>
<td>11</td>
<td>Separated-factors model.</td>
<td>94</td>
</tr>
<tr>
<td>12</td>
<td>Combining the transformed straight-path output with the separated factors.</td>
<td>94</td>
</tr>
<tr>
<td>13</td>
<td>Separated factors in the presence of the transformed straight-path output</td>
<td>94</td>
</tr>
<tr>
<td>14</td>
<td>Incremental examinations of the transformed straight-path model and the separated-factors model</td>
<td>95</td>
</tr>
<tr>
<td>15</td>
<td>Factor-effects examination with respect to the combined model</td>
<td>95</td>
</tr>
<tr>
<td>16</td>
<td>Comparison of common scoring-ability measures for selected Arsenal players and an average fictitious player</td>
<td>101</td>
</tr>
<tr>
<td>17</td>
<td>Comparison of additional scoring-ability measures for selected Arsenal players and an average fictitious player</td>
<td>102</td>
</tr>
</tbody>
</table>
Abstract

Soccer is considered by many to be the world’s most popular sport. The soccer industry is a billion-dollar industry that is naturally very competitive. Accordingly, in the last several years, there have been many attempts to apply scientific approaches and methods to soccer and its analysis, in order to acquire even the slightest advantage. Eventually, the objective of the game is to score more goals than the opponent and, therefore, goal scoring is the name of the game: this is the motivation for our research, in which we apply scientific methods to estimate scoring probabilities.

It is very reasonable to assume that different attempts yield different scoring probabilities—this is very intuitive when comparing extreme cases like a penalty kick versus a long-distance attempt with many players within range—and therefore we aim to investigate the effects of different characteristics on the scoring probability, and to quantify them. The variety of characteristics of scoring attempts, as well as the high variability within these characteristics, make it challenging to compare different attempts, and thus to assess the scoring probability and the factors that actually affect it.

In our research, we focus on the physical and geometric characteristics of scoring attempts, and we overcome the described challenges via a mathematical model: it gives rise to a measure that enables a quantitative comparison of scoring attempts. This is done by modeling the scoring space that is formed by an attempt, according to its physical and geometric characteristics. Our scoring space models are based on simple assumptions, backed up by previous works, that quantify the effects of the physical and geometric characteristics on the scoring probability. We also turn to detailed physical equations for soccer-ball motion, in order to calibrate the model and to adjust its parameters to increase its reliability.

Our models were validated using a data set of 982 scoring attempts, manually collected from the games of London’s Arsenal FC that were played during the 2012/13 season of the English Premier League. The model-required parameters were extracted from each attempt, and the respective scoring spaces were calculated. The statistical relation, between the scoring space and the scoring probability, was then evaluated using non-parametric estimators and a logistic
regression. The statistical analyses found the scoring space to be significant in explaining the empirical scoring probability, and furthermore allowed us to evaluate the effects of additional factors, such as type of play (set piece versus open play).

Lastly, we discuss the results and their practical significance. In addition to quantifying the effects of different factors, we apply our results in several ways that provide insights into game analysis and development of training routines. We also discuss the use of our approach and methods to investigate other related areas, like passing or dribbling. This hopefully contributes to a more complete scientific analysis of the game.
1 Introduction

Association football, commonly known as football or soccer, is a sport played between two teams of eleven players with a spherical ball; it 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 (by driving the ball into the opposing goal) than the opposing team. At the turn of the twenty-first century, the game was played by over 250 million players all around the world [2], which is why it is considered by many to be the world’s most popular sport. One of the most arguable issues in soccer regards the dominance of skill versus chance; alongside incredibly fortunate goals (as Honduras’ Valladares’ own goal against France in the 2014 FIFA’s World-Cup Valladares’ Own Goal vs France) and unbelievable misses (as Spain’s Sergio Busquets’ miss against Chile in the 2014 FIFA’s World-Cup Busquets’ Miss vs. Chile), consistency in performance is easily spotted (for example, Pep Guardiola’s Barcelona has won eight domestic and six international titles in four consecutive seasons); however, in most cases it is hard to separate chance from skill (as Colombia’s James Rodriguez’s goal against Uruguay in the 2014 FIFA’s World-Cup James’ Goal vs Uruguay). We believe that quantifying chance is the key for evaluating skill. In this research we exhibit concepts and methods to support our claim within the framework of a single, but most essential, aspect of soccer—scoring.

A scoring attempt or a shot (we shall use these terms interchangeably) 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, for example 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

3
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 statistic can provide may give insufficient insights as to 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 indicate a win. There exist many contradictory examples, in which the winning team had fewer shots on target but these shots were performed under much more favorable conditions.

The objective of our research is to evaluate and quantify the effects of the physical and geometric characteristics of a scoring attempt. Creating a proper model, taking these effects into consideration, 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 mainly handle these physical and geometric parameters, and hence the effects of other types of parameters will be eliminated using statistical methods.

In this paper we first review previous works that apply scientific methods to soccer, specifically works that suggest methods of estimation of the scoring probability. We then define the problem and describe our approaches and methodologies. Finally, we apply our models on actual scoring-attempts data to validate our hypotheses, analyze the results and discuss conclusions and practical applications.

2 Literature review

2.1 OR and statistical models in soccer

Brillinger [3] presents a review on 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 motion, 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 [3] 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 [3], 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 [3] refers to management and economics models which had evolved as a result of the understanding that soccer is an industry, and that it cannot 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 [3] 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 conditions.

In one of the early works that links statistics and soccer, Reep and Benjamin [19] investigate patterns of scored goals. In their work, they calculate shots and goal-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 [15] 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 motions 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 passes completed within the attacking motion. 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 motion, 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 motion) exhibited different results. Olsen [15] suggests that this difference may be caused by the difference in the game’s circumstances (e.g. distance from goal)
between the goal scorer and other players involved in the attacking motion. Olsen [15] derives from this analysis some practical insights and advice for soccer coaches, and concludes that extending this kind of analysis may help to gain valuable insights into many aspects of elite soccer.

Pollard [16] extends Reep and Benjamin’s [19] 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 [19] 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 [16] found that even inside the penalty area, the shots per goal ratios vary when considering a more refined segmentation of the locations of the shots. 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 [18] 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 [18] 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 of 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. [7] expand upon Pollard and Reep’s [18] 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 into two groups: factors related to the exact time of the shot and factors preceding this time. The analysis followed Pollard and Reep’s [18] definitions and methods, but was conducted on a different set of data. Consistency with Pollard and Reep’s [18] 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 [18] 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. [7] 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 their scored goals, may indicate the defensive abilities. It was shown that the sum of all shot probabilities, for some selected teams, provided a
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. [17] have also conducted a logistic regression in order to estimate the scoring probability
under various conditions. The study was based on the combined data from the two previous studies, exclud-
ing 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 [18] and Ensum et al. [7].

Wright et al. [23] 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 goalkeeper, 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, found consistency with previous works in terms of the goals
per shots ratio among several conditions mentioned above (e.g. 87% of the analyzed goals were scored inside
the penalty area, supporting a prediction in [23] of more than 70% to be scored from that area). Wright et
al. [23] 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, namely, location of the shot, goalkeeper’s position, the number of players between the location
of the shot and the goalkeeper, the type of shot and the position of feed were significant. It is then argued
that despite the significance of new factors in this model, the low value of the explained variance ($R^2 = 21\%$)
suggests that there is still plenty of work to be done in investigating the scoring probability.

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 which we divide 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 potentially 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 motion 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 outcome 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 [18]), evaluating players’ and teams’ abilities (as outlined in Wright et al. [23]), 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. [23], as well as Pollard et al. [17] and Pollard and Reep [18], 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 possible outcome, in contrast to earlier studies that excluded shots that were deflected off a defender and/or shots that were off target. Considering the different approach we present, we also have an interest in comparing our model to existing ones; such a comparison will hopefully reveal the quality of our model.

Further investigation will 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 in a given situation.
4 Methodology

The multiplicity of characteristics and their high variability make it hard to extract the relevant information that actually explains a scoring probability. Standard statistical tools enable the estimation of isolated effects (e.g. the effect of distance on scoring probability) and simple interactions. In order to overcome these limitations, we propose a model-based analysis in which we associate weights with shots in a way that reflects several factors and their relations. These weights reduce the dimension of the problem and provide a combined output, according to which we will try to explain the scoring probability. In practice, we use these outputs as the main covariates aiming to explain the outcome of scoring attempts.

Following this idea, we first explore and investigate the effects of physical and geometric characteristics and their interrelations. We then develop shot models that reflect these effects and apply them to data of actual shots. Finally, we use statistical tools to evaluate how well these models describe reality, or in other words, how well their outputs explain the outcomes of actual scoring attempts (or the empiric scoring proportion).

As a first step, we postulate some hypotheses which our model should satisfy. These hypotheses take into account the main physical and geometric factors that affect scoring. Describing these hypotheses requires some basic definitions (Figure 1). Given a scoring attempt, consider the line that connects the location of the attempt to the center of the goalmouth. We refer to the length of this line as the distance of the attempt, and to the projection of this line on the field as the direction of the shot. The angle that is formed between the direction of the shot and a line perpendicular to the goal line is referred to as the angle of the attempt. Furthermore, we refer to the area between the four lines that connect the location of the shot and the goal posts as the range of the attempt.

Using these definitions, we now formulate our basic hypotheses:

- There is an inverse relation between the distance of an attempt and its scoring probability.

- Smaller angles (in absolute value) yield higher scoring probability.

- Obstacles (i.e. defenders, other players and the goalkeeper) within the range of the attempt reduce the scoring probability.

- The longer the time it takes for the ball to arrive to the goal line the larger the area the goalkeeper can cover.
The blue line is connecting the location of the shot and the center of the goalmouth; its length, denoted by \( l \), is the distance of the shot; the dashed yellow line that connects the location of the shot to the center of the goal is the direction of the shot. The angle that is formed between the direction of the shot and the other dashed yellow line, which is perpendicular to the goal line, is the angle of the shot, denoted by \( \alpha \).

These hypotheses are both intuitive and supported by previous research, as discussed in our literature review section.

Our hypotheses explain some physical and geometric factors separately. A deeper observation reveals that when considering the location of the attempt, a slight change affects both the distance and the angle of the shot. Furthermore, equal changes affect these parameters differently at different locations; hence we come to the conclusion that we should not handle the effects of different physical and geometric factors separately. These observations lead us to develop models that evaluate the scoring space that is formed in an attempt. In our context, scoring space corresponds to the set of feasible scoring options—more on that below. This concept also allows us to take into account the other basic hypotheses very intuitively; obstacles will reduce the scoring space in a way that reflects the scoring space they eliminate (which may be determined through the traveling time of the ball).

### 4.1 The models

As mentioned, we wish to model scoring attempts in a way that will reflect the feasible scoring space formed by an attempt. Measuring this space or representatives of this space allows us to compare different attempts. The physical and geometric nature of such models captures the effects of the discussed parameters, as well as the relations between them. The question that still remains is how well does the feasible scoring space of
an attempt explain scoring and scoring probability—we shall perform this evaluation empirically, and it will be one of the major outputs of this research.

Aiming for an evaluation of the scoring space, we apply two approaches—the first quantifies the area of the goalmouth which is feasible for scoring, while the second quantifies the trajectories that are feasible for scoring. Following these approaches, we have developed two mathematical models of a scoring attempt. In order to describe these models formally (which we shall do in the next subsection), we first introduce some definitions.

4.1.1 Definitions

*Cartesian coordinate system* - we wish to quantify the properties of an attempt relative to the goal it aims at. Thus, the coordinate system is determined with respect to the appropriate goal (the goal to which the attempt was made). Given a shot, we consider a 3-dimensional Cartesian coordinate system as follows:

- The origin \((0, 0, 0)\) is the location of the shot.
- The X-Y plane is parallel to the plane of the actual field (in most cases, it is the same plane).
- The Y-axis points from the location of the shot to the goal line and it is perpendicular to it.
- The Z-axis is naturally perpendicular to the X-Y plane and its positive direction corresponds to height.
- The X-axis is parallel to the goal line and its positive direction is determined according to the right-hand rule.

We note that on grounds of simplicity and consistency of presentation, in some graphs presented in this paper the origin of the coordinate system is set to be at the center of the goal line, and the positive direction of the Y-axis is pointing towards the penalty point.

*Polar coordinate system* - we define the polar coordinate system in a standard way, with respect to the Cartesian coordinate system defined above. Given a point in space, this means:

- \(\vec{r}\) is the vector that points from \((0, 0, 0)\) to the given point.
- \(\phi\) is the angle that is formed between the projection of \(\vec{r}\) on the X-Y plane and the positive direction of the X-axis.
• $\theta$ is the angle that is formed between the projection of $\vec{r}$ on the Y-Z plane and the positive direction of the Z-axis.

**Goalmouth** - the rectangular area of the face of the goal. A goal is scored when the ball crosses this area. Its length is 7.32 meters and its height is 2.44 meters.

**Location of a shot** - relates the three coordinates from which the shot was taken.

**Destination of a shot** - the three coordinates at which the ball ends its motion. Since we discuss scoring attempts, we assume that this motion ends when the ball reaches the goal line, that is, the ball’s location share the same y coordinate as the goal line.

**Range of a shot** - the 3-dimensional space within the four lines that connect the location of a shot with the vertices of the goalmouth.

**Completion time** - the time it takes the ball to travel from the location of the shot to its destination.

**Solid angle** - The solid angle of an object from a certain location is its projection on the unit sphere centered at the given location (Figure 2).

### 4.1.2 The solid-angle model

In this model, we quantify the scoring space by the area within the goalmouth which is feasible for scoring. To do so, we use the geometric measure "Solid Angle" by calculating the solid angle of the goalmouth from the location of the shot (Figure 2). This measure takes into consideration the distance of the goalmouth, as well as its orientation relative to the location of the shot.

We consider the area of the goalmouth which is not covered by any obstacle (relative to the location of the shot) as the area that is feasible for scoring. This approach allows one to handle obstacles within the range of the shot in an intuitive way that corresponds to our hypotheses.

Mathematically, the solid angle is calculated via a double integral over the domain $S$ (which in our case is the goalmouth), and defined as

$$\Omega = \int \int_{S} \frac{\vec{r} \cdot dS}{r^3} = \int \int_{S} \sin(\theta) \ d\theta \ d\varphi,$$

where $\vec{r}$ is a vector directed from the point of view (in our case the location of the shot) to points within the surface (in our case a point within the goalmouth); $dS$ is an element of the surface $S$ directed along the local
The displayed sphere is a unit sphere centered at the location of the shot, and the green marked area on the sphere is the projection of the goalmouth on the sphere.

normal vector; and finally \( \theta, d\theta \) and \( d\phi \) corresponds to the polar coordinates, as defined earlier. Naturally, the second representation is more constructive: in order to compute the solid angle of the goalmouth area, which is feasible for scoring in a certain shot, all that we need is to determine the ranges of the respective angles.

When there are no obstacles in the range of a shot, we actually calculate the solid angle of the goalmouth. Practically, given a shot at \((0,0,0)\), which sets the center of the goal line to be at \((x,y,z)\), the solid angle calculation is done according to

\[
\Omega = \int_{\phi_1}^{\phi_2} \int_{\theta_1(\phi)}^{\theta_2(\phi)} \sin(\theta) \, d\theta \, d\phi. \tag{2}
\]

In this equation, \( \phi_1 \) is the angle to the goal post with the larger \( x \) coordinate, and it is calculated by \( \phi_1 = \arctan \left( \frac{x + 2.32}{y} \right) \); similarly \( \phi_2 \) is the angle to the other goal post, and it is calculated by \( \phi_2 = \arctan \left( \frac{x - 2.32}{y} \right) \).

Both functions \( \theta_1(\phi) \) and \( \theta_2(\phi) \) represent the range of \( \theta \) as a function of \( \phi \) and they are given by \( \theta_1(\phi) = \arccot \left( \frac{(z+2.44) \cdot \cos(\phi)}{y} \right) \), \( \theta_2(\phi) = \arccot \left( \frac{z \cdot \cos(\phi)}{y} \right) \). These calculations are discussed in detail in Appendix A.

When there are obstacles within the range of a shot, one must eliminate the solid angle of the area they cover since this area is not feasible for scoring. To do so, we use the additive property of the solid angle measure: it allows one to calculate the solid angles of the areas that the obstacles cover within the range of the shot, and to simply subtract them from the solid angle of the goalmouth. Since we restrict our computations to areas covered by obstacles within the range of the shot, the solid angle of the covered areas will be at most the solid angle of the goalmouth. In situations where there is more than one obstacle in the range of the shot,
The set of locations that were tested had a constant angle of 0 radians.

there may be intersections of the areas that they cover; in such cases, the intersected areas are reduced only once; equivalently, one subtracts the union of the obstacles area. The way we consider different obstacles (i.e. goalkeeper and other players) and the areas they cover will be discussed in detail later, in the obstacles subsection.

In order to assess properties of the model, we perform two kinds of tests that correspond to two of our basic hypotheses. First, we test the model output as a function of the distance of the attempt from goal. This test is performed by calculating the output of the model for a series of attempts with locations that vary from (0, 0, 0) to (0, 30, 0). This way, the angle of the attempts is constant for all of these shots, while the distance varies from 0 to 30. The output of this test (Figure 3) is consistent with our hypotheses, exhibiting an inverse relation between the distance of the shot and the output of the model. Similar results were found for tests with different constant angles.

The second test evaluates the effect of the angle of the shot on the output of the model. In this test, a series of shots with the same distance (16m) but with varying angles (variation between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$) are tested. Once again, consistency with the hypotheses was found, and the output of this test (Figure 4) reaches its maximum when the angle is 0 (i.e. the shot is made in front of the goal center). Furthermore, there is a natural symmetry between positive and negative angles. Similar results were found with different constant distances.

These tests assess the effects of the distance and the angle of a shot on the output of the model, separately. A natural way of describing the model and its output with respect to these effects and their correlations, is via
The set of angles that were tested had a constant distance of 16 meters (i.e. in front of the goal center).

Contours of the output of the model. The contour plot (Figure 5) exhibits the same trends for distances and angles as the individual tests, and its immediate interpretation is that shots from different locations on the same curve have the same model output, namely the same scoring probability. It is important to emphasize that the contour plot corresponds to situations where there are no obstacles within the range of the shot.

4.1.3 The straight-path model

As mentioned before, we aim to model the scoring space that is formed by an attempt. While the solid-angle model evaluates the feasible scoring area within the goalmouth, an alternative approach may handle the set of feasible scoring trajectories of an attempt. A feasible scoring trajectory is a route between the location of a shot and a point within the goalmouth that satisfies the motion laws of physics of a soccer ball. Following
As we will discuss later in detail (in the physical analysis chapter), analyzing soccer ball trajectories may be highly complicated. Even so, we wish to follow the aforementioned approach in our model of a scoring attempt, but still retain the ability to analyze it simply. To achieve these objectives we reduce the complexity of the scoring space characterization by dividing the set of trajectories to subsets, and handling representatives of these subsets. Practically, given a shot location, we classify the feasible scoring trajectories according to their destination within the goalmouth. This way every subset of trajectories consists of all the trajectories that begin at the location of the shot, and end at the same point within the goalmouth. We represent each of these subsets by a line that connects the location of the shot with the correspondent point within the goalmouth; we define a line like this as a straight path (Figure 6). In other words, a straight path represents all the feasible scoring trajectories that have the same origin and destination coordinates.

Applying this method allows us to evaluate the feasible scoring space of a shot according to the feasible straight paths that are formed by the shot. In this context, a straight path that intersects with an obstacle is not feasible. This perspective characterizes the feasible scoring space as a pyramid (when no obstacles in the shot range are considered), whose apex is at the location of the shot and its base is the goalmouth (Figure 7). Naturally, two questions arise: How to evaluate the scoring space using straight paths and how to relate the properties of the represented set of trajectories to those of the straight path.

Since the scoring space is actually formed by the straight paths, its evaluation should be based on the
Figure 7: The scoring space of a shot as a pyramid.

The pyramid contains all the straight paths that were formed given the location of the shot.

accumulation of them. Due to the fact that we are still handling an uncountable set of elements (straight paths in a given shot), this accumulation will be performed via integration. Answering the second question (relating the properties of the subset of trajectories to its representative straight path) will provide the mathematical functions on which the integration will be performed. In other words, through weighting straight paths according to the properties of the subsets of trajectories that they represent, we can evaluate the scoring space via integration over these weights.

In the context of the described setup, the relevant characteristics of a straight path are its length and orientation. To describe the orientation of a straight path mathematically, we reconstruct an idea from the solid-angle model, and characterize it according to the inner product between its direction vector and the normal of the goalmouth. As examined before, this characterization is consistent with our hypotheses. Furthermore, our hypotheses lead us to give weight to a straight path in an inverse relation to its length. Following these guidelines we get

$$S_p = \int \int_S W(r) \cdot \frac{\hat{r} \cdot dS}{r^3} = \int \int_S \sin(\theta) \cdot W(r) d\theta d\phi$$

(3)

where $S_p$ is our notation for the output of the straight-path model, $W(r)$ is the length weight function that, given a shot at $(0,0,0)$ that sets the center of the goal line to be at $(x,y,z)$, can be expressed as $W(y, \theta, \phi)$. We emphasize that we also follow the solid angle formulation in the division of the aforementioned inner product by $r^3$; this is done to achieve independence between the orientation element and the length element (as exhibited in the latter formulation).

This formulation allows us to use the function $W(r)$, not only to give weight according to the length of
the straight path, but also according to its type; this is done by giving different weights to kicks and headers. Furthermore, the presented formulation actually generalizes the solid-angle model, and characterizes a direct mathematical relation between the two models (i.e. the solid-angle model is the straight-path model with $W(r) = 1$).

Throughout the research, we have investigated several different weight functions ($W(r)$) which will be discussed in detail in the physical analysis section. As a starting point, we have decided to use a linear function that satisfies an inverse relation with the length of the straight path; namely, a function of the form $W(r) = b - a \cdot r$. Having no actual constraints over the function selection, we have set the value of “$a$” to be 1. The value of “$b$” was set to be 50 for kicks and 16.5 for headers, postulating that straight paths of the respective types which are longer than these distances should have zero weight; in other words, we treat long distance goals as outliers.

Practically, given a shot at $(0, 0, 0)$, which sets the center of the goal line to be at $(x, y, z)$, and when there are no obstacles within the range of the shot, the straight-path model calculation is done according to

$$S_p = \int_{\phi_1}^{\phi_2} \int_{\theta_1(\phi)}^{\theta_2(\phi)} \sin(\theta) \cdot W(r) d\theta d\phi.$$  \hspace{1cm} (4)

Since the model is calculated via integration, it satisfies an additive property. This allows handling obstacles the same way as in the solid-angle model—evaluating their weight according to model and subtracting it from the total weight of the attempt. Once again, the actual calculation considers paths that intersect with more than one obstacle, and takes into consideration the total weight of the attempt as an upper bound for the subtraction.

We have assessed the properties of the model using the same tests that were performed on the solid-angle model; namely by evaluating the output of the model for varying distances with constant angle, varying angles with constant distance, and finally by plotting model output contours (Figures 8-10). Once again, and due to the way we built the model, consistency with our hypotheses was exhibited.

### 4.1.4 Obstacles

Both models take into account the presence of obstacles in the range of the shot. Since both models satisfy an additive property, the calculation is done simply by subtracting the respective measure of the scoring
Figure 8: The output of the straight-path model as a function of the distance of a shot. The set of locations that were tested had a constant angle of 0 radians.

Figure 9: The output of the straight-path model as a function of the angle of a shot. The set of angles that were tested had a constant distance of 16 meters (i.e. in front of the goal center).
space that is covered by the obstacle. The practical issue that still requires discussion is the magnitude of the subtraction for obstacles.

It is easy, but essential, to realize that there is a fundamental difference between field players and goalkeepers as obstacles. While a goalkeeper may actively attack the ball during its motion, field players are usually more limited in their actions against a shot. Furthermore, goalkeepers are allowed to use their hands to stop the ball, and thus can expand the area they cover by jumping and diving towards the ball. In other words, the field players are more static obstacles while goalkeepers are more dynamic. Naturally, these observations lead us to model the two types of obstacles differently—a static model for field players and a dynamic one for goalkeepers.

Considering the above, as well as some computational aspects, we have decided to model field players, within the range of the shot, as static rectangles. Though the average shoulder width of athletes is around 38 cm [14] and the average height is around 182 cm (2014 World-cup players’ heights), we set the dimensions of the rectangle to be $45 \times 200$ cm; this is done in order to consider some additional areas that may be covered due to slight motion and jumps towards the ball. The rectangles are considered to be on the ground (i.e. the lower base of the rectangle intersects with the field plane), and perpendicular to it; that is, given a shot at $(0,0,0)$ that sets the center of the goal line to be at $(x,y,z)$, the Z-coordinates of the bases will be $z$ and $z + 2$ respectively. Assuming that obstacles aim to cover as much area as they can, we consider the orientation
of the rectangles to be perpendicular to the direction of the shot. We note that in practice, field players in
the range of the shot can be categorized according to their team; that is, while an opposing team player’s
objective is to prevent the player from scoring, a teammate will try to help the attempting player. Our
perception of the field player as a static obstacle neglects this classification and handles all the field players
within the range of the shot in the same manner (the data collection was performed accordingly).

While for the field player model we take into account an additional area to compensate for minor motions,
in the case of the goalkeeper we wish to model the motion explicitly. Modeling the motion of a goalkeeper
is a vast topic that can be studied extensively and be the subject for many researches; since our focus is on
the estimation of the scoring probability, we have turned to previous works for assistance in constructing a
simple model that will satisfy our needs. The main idea of such a model is to relate the size of the goalkeeper
(i.e. the area he covers) to the time he has to respond. To do so, we consider the goalkeeper as a rectangle
with some initial dimensions that expands according to the shot properties.

The width of the initial rectangle is set to be 38cm (similar to the width of a field player) and its height
to 188cm—on average, goalkeepers are taller than field players [2014 World-cup players’ heights]. To
determine the size of the additional area, we consider the time it takes the ball to arrive to the goalkeeper’s
area (i.e. the ball motion time), as well as the goalkeeper’s area-covering pace. We define the area-covering
pace to be the speed at which the goalkeeper covers an area in space (usually, such area that needs to be
protected). The evaluation of the additional area is done simply by multiplying the motion time of the ball
by the area-covering pace. The motion time of the ball is calculated according to its distance from the
goalkeeper, at the moment of the shot, divided by the average velocity of the ball (25m/sec for kicks and
9m/sec for headers [20]). To evaluate the area-covering pace of the goalkeeper during the motion time of the
ball, we first need to characterize his motion.

Generally, the motion of a goalkeeper, after a shot was made toward the goal he protects, can be divided
into two parts: anticipation and diving. The anticipation motion consists of the preparations for a diving
towards the ball; during this part the goalkeeper tries to set himself at the best location for diving. The diving
motion is the active action of jumping and stretching the body towards the ball. This division is natural in
terms of the area-covering pace. The diving motion is more effective, in terms of the area-covering pace, but
limited in terms of time—according to Kerwin and Bray [11] the diving motion time of a goalkeeper ranges
from 500ms to 700ms. Thus, we assume that if the time a goalkeeper has to react is longer than the required time for diving motion, he will first perform an anticipation motion and then dive; otherwise, all the time he has will be used for the diving motion. In our model we take 600ms as the maximal diving motion time.

Following the above, we realize that there are several area-covering paces to consider. Due to the difference in the motion properties, we handle vertical and horizontal components separately. The vertical component is achieved only during the diving motion. Since the modeled goalkeeper has the initial height of 188cm, there are only 56cm left in order to cover the whole height of the goalmouth. Considering the 600ms the goalkeeper has during the diving motion, we get a vertical pace of 0.933 m/sec. This pace is within the range of the vertical takeoff velocity of a goalkeeper’s diving motion, as reported by Suzuki et al. [21]. Note that we consider the height of 2.44m to be the maximal height that can be covered by the model; so this is the area that is covered by our model even for shots that yield reaction times which are longer than 600ms.

For the horizontal component of the area-covering pace, we rely on some guidelines from previous works. We assume that the anticipation motion is usually at jogging speed since during this motion the goalkeeper needs to allocate most of his attention to track the motion of the ball. According to Di Salvo et al. [6], the jogging speed of a goalkeeper ranges from 7.3 km/h to 14.4 km/h; we choose the average (10.85 km/h which equals 3.01 m/sec) as the horizontal pace for the anticipation motion.

Kerwin and Bray [11] reported a maximal reach of 2.84m during a diving motion that lasted 500 – 700ms. In their work, the goalkeeper was located at the center of the goal line prior to the shot, so we assume that his initial horizontal dimension was half of his shoulder width (i.e. 19cm); this means that there were actually 2.65m covered during a period of 600ms, which yields a horizontal pace of 4.42 m/sec for the diving motion.

This analysis allows us to calculate the additional horizontal range that can be covered by the goalkeeper for any amount of time, as exhibited in Figure [11]. The two different slopes correspond to the two horizontal paces, where the 600ms point is the transition point; for times which are longer than 600ms, the additional horizontal range is composed of the 2.65m covered within the 600ms of the diving motion and an additional range that was covered during the anticipation motion. To ease our practical calculation, we prefer using a single simple function for the goalkeeper model. Since the function should correspond to the structure of the 2 linear lines, we fit a power function which satisfies a decreasing rate of the area-covering pace. Denoting
Figure 11: The additional horizontal area covered by the goalkeeper as a function of the motion time of the ball.

The blue squares represent the area according to the two-part linear model, and the red line is the fitted power function.

\[ f(t) = 3.845 \cdot t^{0.8296} \]

which sets \( g(t) \)—the horizontal dimension of the goalkeeper as a function of the response time—to be

\[ g(t) = 0.38 + 3.845 \cdot t^{0.8296}. \]

### 4.1.5 Other factors

As discussed, our models evaluate the scoring space that is formed by a scoring attempt, and by that enable the quantification of the main factors (according to our perception) that affect the scoring probability. Having said that, there are other factors which are not taken into account by the models, or that are not modeled explicitly, that may affect the scoring probability. Since our objective is to estimate the scoring probability we wish to consider these factors; this will be done simply by adding them as additional covariates for the statistical model.

Namely, the factors that may be considered are:

- Time - the official game time at which the shot was made.
• Half - on which half of the game was the shot made.

• Team - the team to which the attempting player belongs.

• Home - indicates whether the attempting player belongs to the home or away team.

• Player - who is the player that attempted scoring.

• Type of shot - indicates whether the attempt was a kick or a header.

• Height - the height of which the shot was made.

• Contact - indicates ball control of the attempting player prior to the shot.

• Space - the area the attempting player had to perform the shot.

• In box - indicates whether the shot was made within the 16 meter box.

• Free kick - indicates whether the attempt was made directly out of a free kick or during open play.

• Penalty - indicates whether the shot was a penalty kick.

While the effects of some of these factors may be interpreted intuitively (e.g. the shot type), the rest may just point out some statistical effects (e.g. difference in scoring probabilities between halves). The characterization of these factors depends on the data set we use, and therefore some of them may be excluded due to data constraints.

4.2 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 types of data are not sufficient for our research needs. In all of the previous works mentioned ([7], [15], [16], [17], [18], [19], [23]), the data collection was performed using observers (sometimes the authors themselves), that used different types of notation-systems to record the events of interest, 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 (Figure 45). 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 previously inputted pitch measures. Validity and reliability issues would still be a matter of concern, but using this interface will help to estimate locations in a consistent way.

There has been recent major progress in the area of soccer data collection [3]. 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. [22]. These high levels of correlation, as well as other reliability scores that were measured by Di salvo et al. [22], indicate that this kind of data is reliable and reflects very accurately the motion of the game’s participants. Such a data base will suffice for the needs of our research, and we did aim to achieve it but, unfortunately, we did not manage to do so. We emphasize that this failure in obtaining such data does not fail the research. Instead, it resulted in recording the data ourselves, using our described interface.

5 Physical analysis

Useful models must simplify reality, and naturally this holds true for our models as well. In our models we tried to capture the relevant factors that, according to our hypotheses, affect the scoring probability, and to do so as simply as possible in order to avoid the actual complexity of actual soccer ball motion. In both of our models, the solid-angle model and the straight-path model, we implicitly postulate direct aiming; namely, we do not consider the curves that are formed in an actual soccer ball motion. Addressing our objectives
with a more accurate model can become very complicated and much less insightful. Therefore, our purpose in this section is to investigate and analyze the actual motion of a soccer ball, in order to adjust our models in ways that correspond to this analysis.

Performing an analysis of a motion of an object requires physical tools—namely, motion equations. There are various equations that characterize the same motion; the complexity of these equations is derived from the selection of physical elements and forces that are taken into account. We found no need to develop a novel physical model and accordingly, we resort to previous research. The model we chose to use provides high precision, but remains practical in terms of calculation; we describe it in detail in the next subsection.

Using physical motion equations to describe the motion of a soccer ball, after a shot was made, allows us to analyze the behavior of trajectories. There are two aspects of our models that can be naturally addressed by such an analysis:

1. Weighting of a straight path - as mentioned, a straight path represents all the trajectories that correspond to its origin and destination. We wish to equip a straight path with a weight that will reflect the properties of the trajectories it represents. Analyzing the properties of the set of these trajectories can assist in identifying and applying proper weight.

2. Accommodating obstacles - as discussed, in both models we consider obstacles simply by subtraction of the scoring space they eliminate. The evaluation of the subtracted scoring space in both cases is consistent with the evaluation of the total scoring space of the model; that is, under a postulation of direct aiming. Since trajectories are actually curved, they do not correspond completely to our direct aiming models. Practically, this means that the area we subtract may represent trajectories that do not intersect with it and not represent trajectories that do intersect with it. Therefore, an accurate analysis of trajectories in the presence of obstacles can lead to adaptations in the way we handle obstacles.

Following all the above, we first define the motion equations of a soccer ball and present a method to numerically solve them. The following subsections will discuss the analyses of the solutions of these equations in the context of the two topics presented above: weights and obstacles.
5.1 Soccer ball motion equations

Applying a physical model to describe the motion of an object, requires mapping and understanding of the physical forces involved. Obviously, the forces that the attempting player applies to the ball are significant in this context. These forces affect the ball by changing the velocity of its center of mass (which describes the rate of the change in the position of the ball), as well as the angular velocity (which describes the rotation of the ball). Moreover, gravity and aerodynamic friction are two additional forces affecting the ball’s trajectory.

We now turn to previous research in order to understand how these forces are accounted for in a systematic model. According to H الرسميز et al. [10], the discussed factors can be characterized by three aerodynamic forces: the air drag ($\vec{F}_d$), Magnus force ($\vec{F}_m$) and the weight ($m\vec{g}$). The air drag ($\vec{F}_d$) acts against the flight direction, while the Magnus force ($\vec{F}_m$) is perpendicular to the plane containing the ball’s velocity vector and the rotation axis of the spinning ball. In some papers, such as Asai et al. [1], Myers & Mitchell [13] and Griffiths et al. [9], the Magnus force is split into lift and lateral components. Both drag and Magnus forces are aerodynamic forces that reflect the effects of the airflow on the ball; accordingly, both of them are affected by the size of the ball (reflected by the ball cross-sectional area), the density of the air and the ball velocities. The actual values of the drag and Magnus forces are proportional to these effects, multiplied by an appropriate coefficient (drag coefficient or Magnus coefficient) which reflects the needed adaptations for the calculation. The values of these coefficients may vary as a function of some environmental conditions (such as altitude and humidity), but for our needs we will use constant standard values for these parameters. Namely we set the following values:

- Ball cross-sectional area - according to FIFA, the official circumference of the ball should be between 68 and 70 cm [5]; we set it to be 69 cm and extract the proper radius, which yields a cross-sectional area of $0.0379m^2$.

- Air density - we set the air density to be the density at sea level, which is $1.225 \text{ kg/m}^3$ [4].

- Drag and Magnus coefficients - as a matter of fact, the values of these coefficients are not constant and they vary as a function of several parameters. According to Asai et al. [1], under standard conditions (that is, not under critical conditions), the drag coefficient values vary around 0.4 (this is consistent with the average of the range of coefficients presented by رسميز et al. [10]) and so we set it to be
the constant value 0.4. Asai et al. [11] also exhibit a range of calculated Magnus coefficients affected by various conditions. The calculated value of these coefficients ranges from 0.1 to 0.3 and, therefore, we set our Magnus coefficient to be 0.2.

One last parameter to be considered in modeling the flight of a soccer ball is the angular velocity decay rate. In fact, according to Herzer et al. [10], there are only few studies that examined the decay rate of a spinning soccer ball, but it is common to use an exponential decay rate to simplify the actual complexity of this factor. Goff and Carr$^i$e [8] assumed a 10\% drop rate for the angular velocity and thus we set this parameter to be 0.1.

These forces and parameters can be formalized as follows:

\[
\vec{x}(t) = \begin{bmatrix}
\vec{p}(t) \\
\vec{v}(t) \\
\vec{\omega}(t)
\end{bmatrix},
\]

\[
\dot{\vec{x}} = \begin{bmatrix}
\vec{v}(t) \\
\vec{g} - \frac{C_M}{m} (\vec{\omega}(t) \times \vec{v}(t)) - \frac{1}{2} \rho C_D A \vec{v}(t) \|\vec{v}(t)\| \\
-\alpha \vec{\omega}(t)
\end{bmatrix}.
\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vec{p}(t)$</td>
<td>Ball’s center of mass position</td>
<td>meters $^T$</td>
</tr>
<tr>
<td>$\vec{v}(t)$</td>
<td>Ball’s center of mass linear velocity</td>
<td>meters/seconds $^T$</td>
</tr>
<tr>
<td>$\vec{\omega}(t)$</td>
<td>Ball’s angular velocity about its center of mass</td>
<td>revolutions/seconds $^T$</td>
</tr>
<tr>
<td>$\vec{g}$</td>
<td>Gravity vector</td>
<td>meters/seconds$^2$ $^T$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Angular velocity decay rate</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>Air density</td>
<td>$\text{kg/m}^3$</td>
</tr>
<tr>
<td>$C_D, C_M$</td>
<td>Air drag coefficient and Magnus effect drag coefficient</td>
<td></td>
</tr>
<tr>
<td>$A$</td>
<td>Ball cross-sectional area</td>
<td>meters$^2$</td>
</tr>
</tbody>
</table>
The initial and final conditions are:

\[
\vec{x}_0 = [p_{x0}, p_{y0}, p_{z0}, v_{x0}, v_{y0}, v_{z0}, \omega_{x0}, \omega_{y0}, \omega_{z0}]^T \\
\vec{x}_f = [p_{x_f}, p_{y_f}, p_{z_f}, v_{x_f}, v_{y_f}, v_{z_f}, \omega_{x_f}, \omega_{y_f}, \omega_{z_f}]^T.
\]  

(7)

Since the angular velocity \(\vec{\omega}\) is decoupled from the rest of the state space, we can explicitly solve the differential equation associated with it:

\[
\vec{\omega} = [\omega_{x0}, \omega_{y0}, \omega_{z0}] e^{-\alpha t}.
\]  

(8)

Substituting (8) into (6) and introducing \(t\) as another state space variable to make the model autonomous, we define a new state space model:

\[
\vec{z}(t) = \begin{bmatrix} \vec{p}(t) \\ \vec{v}(t) \\ t \end{bmatrix}, \\
\dot{\vec{z}} = \begin{bmatrix} \vec{g} - \frac{C_m}{m} (\vec{\omega}_0 \times \vec{v}(t)) e^{-\alpha t} - \frac{1}{2} \rho C_D A \vec{v}(t) \| \vec{v}(t) \| \\ \vec{v}(t) \| \vec{v}(t) \| \\ 1 \end{bmatrix},
\]  

(9)

where the initial and final conditions are modified as follows:

\[
\vec{z}_0 = [p_{x0}, p_{y0}, p_{z0}, v_{x0}, v_{y0}, v_{z0}, 0]^T \\
\vec{z}_f = [p_{x_f}, p_{y_f}, p_{z_f}, v_{x_f}, v_{y_f}, v_{z_f}, T]^T \\
\vec{\omega}_0 = [\omega_{x0}, \omega_{y0}, \omega_{z0}].
\]  

(11)

In this formulation, the motion time \(t\) is the independent variable. Given a proper set of values of the independent variables, these equations yield a unique solution. Different selections yield different solutions; however the physical nature of the problem imposes certain physical bounds. Specifically, assume that the maximal linear velocity the ball is kicked at is \(\|v\|_{\text{max}}\). Suppose the distance (closest) from the kicking point to the goal frame is \(d_0\). Thus, the minimal amount of time to score is \(t_{\text{min}} = \frac{d_0}{\|v\|_{\text{max}}}\) and the maximal is \(t_{\text{max}} = \frac{\bar{d}_{\text{max}}}{g}\). Both of these are theoretical bounds since the minimum assumes a straight line trajectory (no accelerations) and the maximum assumes an extreme parabolic trajectory utilizing the kinetic energy of the ball to remain airborne. These considerations bound the time interval in which later search is conducted to
This will be discussed further in the constraints subsection together with additional physical constraints that must be imposed in our problem.

### 5.1.1 TPBVP setup

The presented formulation provides a framework of differential equations to describe a soccer ball trajectory. Setting various combinations of initial and final conditions yields different trajectories. In practice, this can be used to generate trajectories by setting the initial conditions vector and (numerically) simulating the trajectory according to the equations; the outcome of such trajectories, in terms of scoring, is unknown a priori. This method may provide some information regarding the behavior of trajectories, but since we wish to focus only on feasible scoring trajectories, applying it may be exhaustive and not efficient.

One way to restrict our analysis to only relevant trajectories, is to solve the equation in a Two Point Boundary Value Problem setup. According to Youdong et al. [12], the problem of solving a system of ordinary differential equations is a TPBVP, when there are two values of the independent variable at which conditions are specified. Considering $2m$ unknown parameters ($m$ parameters that correspond with each of the independent variable values), a selection of $m$ of them is enough to provide a unique solution (if one exists), and to find the values of the rest of the unknown parameters.

Applying this setup to our equations, we get $t$ as the independent variable, and $m = 9$ unknown parameters $(p_x, p_y, p_z, v_x, v_y, v_z, \omega_x, \omega_y, \omega_z)$ at each point of $t$. Following our concepts and models, it is natural to consider the “two points” in terms of location, that is the shot origin and a point within the goalmouth. In fact, according to our definitions, the shot origin is always at $(0, 0, 0)$, and thus one is left with the selection of the three coordinates of a point within the goalmouth. Considering the discussed time interval limitations, we set the independent variable at the shot origin to be zero (i.e. $t_0 = 0$) and at the shot destination to be $t_f \in [t_{\text{min}}, t_{\text{max}}]$. For any point within the goalmouth, the selection of $t_f$ will yield a different solution (in our case a trajectory). Naturally, this approach leads one to choose the coordinates of the origin and destination of the shot as six of the unknown parameters to be set. Thus, there are three more conditions out of the remaining parameters (namely $v_{x0}, v_{y0}, v_{z0}, v_{xf}, v_{yf}, v_{zf}, \omega_{x0}, \omega_{y0}, \omega_{z0}$) to set, in order to get a unique solution.

We choose to set the 3 components of the angular velocity $(\omega_{x0}, \omega_{y0}, \omega_{z0})$ to obtain a solution. This is done since the angular velocity can be constrained easily—particularly in terms of the point of origin—under
the following observations. If we consider each of the components of the angular velocity separately, we can
discuss three different types of motion. Actually, discussing the angular velocity is more natural under a
transformation of the $X$ and $Y$ components. $\tilde{X}$, is a transformation of $X$ such that the direction of the shot
is the positive direction of $\tilde{X}$. $\tilde{Y}$ is a transformation of $Y$ which is perpendicular to the direction of the shot
on the $X - Y$ plane, where the positive direction of $\tilde{Y}$ is set by the right-hand rule. We emphasize that this
transformation does not affect the angular velocity vector, which is solved in isolation.

An angular velocity directed along the $\tilde{X}$ axis characterizes motion which is typical for a spinning bullet
or a thrown football (American football); this kind of ball motion is not common in soccer. Myers and
Mitchell \cite{13} state that when the spin axis is the $Z$-axis, the spin is a pure side-spin, and when the spin
axis is the $\tilde{Y}$-axis it is a pure top-spin. According to Myers and Mitchell \cite{13} and Hӓrzer et al. \cite{10}, the
average shot actually spins around an axis which is a combination of $Z$ and $\tilde{Y}$ components. A top-spin is
typical for lob shots—in a lob shot, the player attempts to arch the ball over the goalkeeper (and/or other
players)—and these kinds of attempts have different characteristics than standard attempts (e.g. lower center
of mass velocity); furthermore, they are usually exercised in some unique situations. Therefore, and in order
to reduce the dimensionality of our problem, we choose to consider shots that spin around the $Z$-axis solely
(i.e. shots with pure side-spin). In practice, this means that we set the angular velocity to be $(0, 0, \omega_z)$.

Eventually, summarizing all the above, to generate a single trajectory given the location of the shot at
the time $t_0 = 0$, we set the additional conditions as follows: the coordinates of a selected point within the
goalmouth (i.e. the destination of the shot), the final time $t_f$ (i.e. the completion time) and the $Z$ component
of the angular velocity $\omega_z$. The actual solution for this setup is performed numerically using Mathworks’
Matlab (from here on will be referred as Matlab). Examples of different sets of conditions that yield different
trajectories are exhibited in Figure \cite{12}

### 5.1.2 Physical Constraints

The described method enables one to solve the motion equations in order to generate trajectories. A different
selection of parameters will yield different trajectories. However, since this is a mathematical model, it may
generate solutions that do not necessarily correspond to physically feasible trajectories. Since our problem
emerges from a real-world situation, there are some additional natural limitations that can further restrict our
The graph exhibits 3 different trajectories: the blue and the red trajectories share the same origin (0,20,0) and destination (2.5,0,1) and differ by the selection of the completion time (1.5sec for the blue and 0.8sec for the red) and the initial angular velocity ($5\text{revs/sec}$ for the blue and $-5\text{revs/sec}$ for the red). The conditions of the green trajectory are (-5,15,1) for the origin, (0,0,0) for the destination, 2.5sec for the completion time and $0.1\text{revs/sec}$ for the initial angular velocity.

analysis. To apply these limitations to our mathematical model, we simply constrain its parameters.

The natural first step, in limiting model parameters, is applying constraints on the parameters that are set to get a solution. Following our previous discussion this means to constrain the range of the coordinates of the two points (origin and end), the completion time and the angular velocity.

- **The coordinates of the origin of the shot**: Considering the $X - Y$ plane, the natural limitations for these parameters are derived from the dimensions of the field. The limitations regarding the $Z$-axis are derived from the players’ physical limitations (a player cannot kick the ball from a height he cannot reach) and type of shot selection (players usually decide whether to kick or to head the ball according to its height). Actually, following our basic definitions, the coordinates of the origin of the shot are always $(0,0,0)$; however, the selection of the shot origin is reflected in $(x,y,z)$—the coordinates of the center of the goal line (as discussed earlier)—and therefore we constrain these coordinates instead.

The official length of a soccer field ranges between 90 and 120 meters, and the range of its official width is 45 – 90 meters [5]; for our analysis, we will use a length limit of 50 meters for kicks and 16.5 meters for headers (following a preceding assumption) and width limit of 45 meters. Neglecting
some possible spectacular kick attempts, we limit the height (that is, the range of the Z-axis) of which kicks are made to the range of 0 − 1m above the ground, and the height of headers to the range of 1 − 2.44m from the ground (the upper bound reflects the height of the goalmouth). Taking all of these into account, we limit the coordinates of the center of the goal line to be

\[
\{(x, y, z) \mid -22.5 \leq x \leq 22.5, 0 \leq y \leq 50, -1 \leq z \leq 0\}
\]

for kicks, and

\[
\{(x, y, z) \mid -22.5 \leq x \leq 22.5, 0 \leq y \leq 16.5, -2 \leq z \leq -1\}
\]

for headers. Note that the range of the x coordinate is symmetric relative to 0 due to the way we define the coordinate system.

- **The coordinates of the destination of the shot**: As mentioned, we are only interested in trajectories that could yield a goal, and therefore the destination of the shot is limited to be within the goalmouth. Given a shot that sets the center of the goal line at \((x, y, z)\), and considering the dimensions of the goalmouth, we limit the coordinates of the destination of the shot to be

\[
\{(u, v, w) \mid x - 3.66 \leq u \leq x + 3.66, v = y, z \leq w \leq z + 2.44\}.
\]

- **Completion time**: As discussed earlier, we constrain the completion time according to theoretical lower and upper bounds. In fact, since we set the time at the moment of the shot to be \(t_0 = 0\), this constraint is reflected in the value of \(t_f\). Mathematically, we set

\[
t_f \in [t_{min}, t_{max}],
\]

where the values of \(t_{min}\) and \(t_{max}\) are determined according to the selection of the origin and destination coordinates, and the maximal velocity—as described at the beginning of this chapter.

- **Angular velocity**: Griffiths et al. [9] reported a range of 4 − 11 \(\text{rev/sec}\) (revolutions per second) for a free kick. Since in a free kick, the kicking player will probably try to bend the ball (i.e. generate an extreme curved trajectory) in some direction, we assume that there will be a higher angular velocity than in a shot made in an open play. Therefore, and since we also consider free kicks in our analysis, we set the upper bound of the angular velocity to be 11 \(\text{rev/sec}\) for kicks; the lower bound is set to be 0 \(\text{rev/sec}\),
postulating that there are shots with negligible angular velocity. The different kinematic characteristics of headers (e.g. shorter distances, less control of the ball), as well as computational considerations, lead one to model headers to have no angular velocity. Following our previous assumptions, we set the range of the angular velocity to be

$$\{(0, 0, \omega_z) \mid 0 \leq \omega_z \leq 11\}$$

for kicks and $$(0, 0, 0)$$ for headers.

Another natural limitation that we wish to consider in our solutions is the center of the mass velocity limit. Shewchenko et al. [20] noted that, for top football players, the center of mass velocity peaks at a range of $17 - 33 \text{ m/sec}$ for kicks, and a range of $6.3 - 12 \text{ m/sec}$ for headers. We therefore set the upper bounds of the norm of the center of mass velocities to be $$||\vec{v}_{\text{max}}|| = 33 \text{ m/sec}$$ for kicks and $$||\vec{v}_{\text{max}}|| = 12 \text{ m/sec}$$ for headers. Naturally, considering the norm of the center of mass velocity vector we get $$||\vec{v}|| \geq 0$$. Unlike the previous parameters, the velocity of a shot is given post-hoc, and therefore the natural way to handle its constraints is by elimination of trajectories that yielded velocities that are higher than the upper bound (Figure 13).

At this point, we wish to assess the validity of the discussed constraints, that is to evaluate their effect in term of the actual problem. To do so, we reduce our discussion to two points (origin and destination), within the permitted ranges, which can be selected differently each time. Once the two points are selected, we further reduce the magnitude of the problem by setting a single angular velocity. This leaves us with the completion time as the last parameter for selection. Accordingly, the trajectories are generated by solving the equations for any completion time, within the permitted range (in practice we discretize the range), and they are checked with respect to the initial velocity constraint. An example of the outcome of such a procedure can be seen in Figure 13.

Performing this routine for several different selections of origins, destinations and angular velocities reveals trajectories that in some points rise to extreme heights (30 meters and more). Such trajectories, though satisfying the constraints, are not common in soccer—in particular, when lobbed shots are excluded from the discussion—and therefore we seek for an additional constraint to eliminate them.

Naturally, given an origin and destination of a shot, there is a direct relation between the height achieved in a trajectory and the $Z$ component of the initial center of mass velocity (namely $v_z$). However, due to the variety of allowed distances, the absolute value of $v_z$ does not necessarily correspond to extreme trajectories.
Figure 13: Trajectories of a straight path with constant angular velocity.

All of the trajectories are represented by the same straight path (origin at (0, 16, 0) and destination at (0, 0, 0)) and share the same angular velocity ($\omega = 2.5 \text{ revs/sec}$) but are completed within different times within the permitted time range (discretization step of 0.1 seconds). The norm of the initial center of mass velocity of the red trajectories exceeds the upper bound of $33 \text{ m/sec}$ and therefore they are eliminated and not considered in further calculations.
Since we wish to constrain the trajectories in a general way that will not depend on the other constraints, we focus on the ratio between the vertical and horizontal components of the center of mass velocity, that is \( \frac{v_z}{\sqrt{v_x^2 + v_y^2}} \). This ratio can be expressed by the angle between these two components; we refer to this angle as the \textit{velocity angle} and denote it by \( \theta_v \). The mathematical definition of this angle is

\[
\theta_v = \arctan \left( \frac{v_z}{\sqrt{v_x^2 + v_y^2}} \right).
\]

(12)

As a matter of fact, it might be reasonable not to impose any constraint on \( \theta_v \) since one can lift the ball from the ground in any angle. However, by setting an upper bound for \( \theta_v \), we can control the heights achieved by trajectories. Therefore, we choose to classify trajectories according to their aiming type, that is \textit{direct aiming} or \textit{indirect aiming}. Hence, an upper bound for \( \theta_v \) should separate direct aiming trajectories from indirect aiming ones.

Considering all the above, the natural selection for the upper bound is an angle of 45˚, which in basic kinematic models is the angle of maximum range, and provides the desired separation between the two types of aiming. This classification is based on theoretic value, while in practice, using complicated models as ours, the actual classification threshold is lower. This means that some trajectories, that should be classified as indirect aiming, might be taken into account in our calculations. This is acceptable since, as mentioned, the way we classify the trajectories is based on a rough estimation.

The calculation of \( \theta_v \) of a trajectory is based on its initial velocities. As mentioned, these parameters are revealed post-hoc, and therefore we must once again test the adherence to the constraint after the trajectory was generated, in order to decide whether to eliminate it or not. Applying this constraint reduces the set of eligible trajectories exhibited in Figure 13 to be the set of trajectories in Figure 14. By repeating this procedure for a selection of various origins, destinations and angular velocities, we can now analyze the behavior of trajectories and relate this behavior to other parameters of shots.

In the following subsections, we describe the analysis of trajectories with respect to the physical equation parameters and their constraints; then we exhibit the applications of this analysis to our models by characterizing straight paths and assessing the effects of obstacles on the scoring space.
Figure 14: Trajectories of a straight path with constant angular velocity with the initial velocity constraint.

The set of presented trajectories is the set of eligible trajectories presented in Figure 13. The green trajectories yield an initial \( \theta \), that exceeds the upper bound of 45\(^\circ\) and therefore are excluded from further calculations.

5.2 Trajectory analysis

Solving the equations using the TPBVP setup is very natural in the context of the straight-path model. In fact, given a shot location and a point within the goalmouth (that defines a straight path), one can generate all the trajectories between these two points. Following the concepts of the straight-path model, in this part we wish to analyze the properties of the set of trajectories represented by a straight path and to relate them to its length. The following analysis focuses on kicks and lays the foundations for a similar analysis for headers. In practice, we apply concepts and results from this analysis to construct a weight function for headers instead of performing the same analysis with different parameters.

An alternative way to perceive the system of equations, in our context, is as a "black-box" function—from the domain of the parameters that are predetermined, to the range of the parameters that are given by the solution. Given two points of interest (origin and destination), this function accepts the completion time and the angular velocity as an input and yields the initial velocity as an output, or if we follow our framework, the norm of the initial velocity and its angle. Denoting \( \chi \) as this function, we can express this relation via

\[
\chi : T \times \Omega_e \rightarrow V \times \Theta_e,
\]

(13)
where $V$ stands for the norm of the initial velocity. Due to the discussed constraints we can reduce the domain to be

$$\tilde{T} \times \tilde{\Omega}_z = \{(t, \omega) \mid (t, \omega) \in T \times \Omega_z, t_{\min} \leq t \leq t_{\max}, 0 \leq \omega \leq 11\};$$

(14)

the constraints on the range of the function can be then expressed via

$$\tilde{V} \times \tilde{\Theta}_v = \{(v, \theta) \mid (v, \theta) \in V \times \Theta_v, 0 \leq v \leq 33\}$$

(15)

and we can use (15) to further reduce the domain by

$$\tilde{T} \times \tilde{\Omega}_z \cap \chi^{-1}\left(\tilde{V} \times \tilde{\Theta}_v\right),$$

(16)

where $\chi^{-1}\left(\tilde{V} \times \tilde{\Theta}_v\right)$ represents the set of origins of $\tilde{V} \times \tilde{\Theta}_v$. The last expression (16) is actually a representation of the set of trajectories of a straight path. This representation is not constructive and we wish to identify an alternative one, that will enable simple calculations and analyses.

The first step in seeking such an alternative is to try to identify trends and relations between parameters by plotting the function (where the output values are calculated numerically using the equations). We start as simply as we can, by examining the function only according to the variation in completion time, while the value of $\omega_z$ is constant (and satisfies the constraints). This allows us to present a plot (Figure 15) in three dimensions. It is noticeable that Figure 15 actually presents a two-dimensional curve (as a function of the completion time); furthermore, the curve exhibits a one-to-one relation between $v$ and $\theta_v$ (this relation is also exhibited in Figure 16). Thus, we realize that there is a redundancy in describing the output of the function in two dimensions. We therefore reduce the dimensions of the output to a single dimension by choosing to characterize it only by the norm of the initial velocity.

In Figure 17, we exhibit the final alternative representation, that is the norm of the initial center of mass velocity as a function of the completion time; we refer to this representation as the trajectory function. This representation allows us to examine the effect of various parameters on the properties of the function by comparing the trajectory functions achieved for different combinations of parameters. Figures 18-21 exhibit the effects of changing the parameters: origin and destination of a shot, and the initial angular velocity.

The effects of the variation of the origin and destination of a shot, can be translated to the effects of the distance (the length of the straight path) and the angles ($\theta$ and $\phi$) of the shot. Each of these shot characteristics was examined separately, that is several values of the examined parameter were tested while
Figure 15: Initial center of mass velocity norm and velocity angle as a function of the completion time.

This curve connects the mentioned parameters as extracted from a straight path with the following parameters: origin (0,16,0), destination (0,0,0) and angular velocity of $2.5\text{revs/sec}$.

Figure 16: Velocity angle as a function of the center of mass initial velocity norm.

The graph describes the same values as Figure 15 without the dimension of completion time. It is noticeable that there is a one-to-one relation between the parameters.
Figure 17: Initial center of mass velocity norm as a function of the completion time.

The graph describes the same values as Figure 15 without the dimension of velocity angle. This is actually our alternative way to describe all the trajectories represented by a straight path for a constant angular velocity.

others remained constant through each test (this includes the constant angular velocity). In Figure 18, we present an example of the test for varying values of $\phi$; all the different tested values of $\phi$ yielded the exact same curve (so actually the graph appears as a single curve). This result is consistent in all of the tested combinations of the other parameters, and it corresponds to the fact that there are no forces in the $X-Y$ plane that depend on a certain direction within the plane (the properties within the coordinate system are not affected by rotation around the $Z$ axis).

Similar tests were performed in order to test the effects of varying values of $\theta$ (Figure 19) and of $\omega_z$ (Figure 20) on the trajectory function curves. Unlike the results achieved in the tests for $\phi$, here there are slight differences between curves with different values of $\theta$ and $\omega_z$, where the significance of these changes varies with distance but remains minor (especially with respect to the differences caused by the distance). The most significant effect on the function curves was that of the distance of the shot. It is easy to see that the curves in Figure 21 become shorter and consist of higher values of velocity as the distance increases. Similar results were found in all the other combinations of parameters that were examined.

Following these results, and despite the fact that the angles of the shot do have some effect on the
Figure 18: The effect of the $\phi$ angle on trajectory function.

A range of values of $\phi$ was tested, while the other parameters were held constant (straight path length of 20 meters, $\theta = 0^\circ$ and $\omega = 5$ revs/sec), and yielded the exact same curves. The graph indicates that the trajectory function is not affected by the change of $\phi$.

Figure 19: The effect of $\theta$ angle on the trajectory function.

A range of values of $\theta$ was tested, while the other parameters were held constant (straight path length of 20 meters, $\phi = 0^\circ$ and $\omega = 5$ revs/sec). The curves exhibit very similar trends and the differences between them are minor.
Figure 20: The effect of $\omega_z$ on the trajectory function.

A range of values of $\omega_z$ was tested, while the other parameters were held constant (straight path length of 20 meters, $\phi = 0^\circ$ and $\theta = 0^\circ$). The curves exhibit very similar trends and the differences between them are minor.

Figure 21: The effect of the distance on the trajectory function.

A range of values of distances was tested, while the other parameters were held constant ($\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5 \text{ revs/sec}$). The differences between the curves are significant, and it is clear that longer distances yield shorter curves.
function, we have decided to consider the distance of the shot as the main factor affecting the function, and thus we characterize a straight path according to this distance alone. This decision is consistent with our modeling approach, where the orientation of the straight path is already modeled via the elements of the solid angle. Therefore, all that is left is to quantify the effect of the distance of a shot on the function, and by that its effect on the set of trajectories.

Going back to the definitions of the trajectory function, we realize that each point on the curve actually represents a different trajectory, characterized by a unique combination of completion time and initial velocity norm. Since our objective is to evaluate the cardinality of a trajectory set, we wish to measure the set of points within the curve. Due to the continuity of the curve, this can be reflected simply by its length. Therefore, we turn to the natural way of measuring a length of a curve—the line integral.

Practically, performing such a calculation, requires a mathematical characterization of the curves (by fitting a known function to describe it). Several functions were experimented with, and eventually we chose a rational function of the form

$$V(t) = \frac{p_1 \cdot t^3 + p_2 \cdot t^2 + p_3 \cdot t + p_4}{t + q_1},$$  \hspace{1cm} (17)$$
due to its excellent fit as presented in Figure 22 ($R^2 > 0.99$). The use of this function yields the line integral

$$\int_C f(V(t)) \, dl = \int_{t_{\text{min}}}^{t_{\text{max}}} \sqrt{1 + \left( \frac{p_1 \cdot t^2 \cdot (3 \cdot q_1 + 2 \cdot t) + p_2 \cdot t \cdot (2 \cdot q_1 + t) + p_3 \cdot q_1 - p_4}{(t + q_1)^2} \right)^2}; \hspace{1cm} (18)$$

here $f(V(t)) = 1$, since we assume a uniform density of trajectories along the curve. This integral is solved numerically using Matlab, but in order to perform this calculation one must set the limit values of the integral—in our case in terms of completion time (namely $t_{\text{min}}$ and $t_{\text{max}}$). The lower limit, or the lowest completion time, is the completion time that is achieved when the initial velocity norm is equal to its upper bound (33 m/sec). The upper limit is achieved when the maximal value of $\theta_v$ is obtained. Finding these values is a matter of finding the roots of an equation, that is, finding the values of the completion time that yield these velocity values. This requires mathematical representations to relate the center of mass velocity norm and $\theta_v$ to the completion time. The first relation is established via (17) so basically we wish to solve the equation

$$V(t) = 33.$$

(19)
This fit was made for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5\text{revs/sec}$. Similar results with different coefficients were obtained when performing the same tests but with a different selection of parameters.

Following a similar procedure to characterize $\theta_v(t)$ we get

$$\theta_v(t) = \frac{p_1 \cdot t^3 + p_2 \cdot t^2 + p_3 \cdot t + p_4}{t^2 + q_1 \cdot t + q_2}$$

(20)

as seen in Figure 23 and thus the upper limit is achieved through

$$\theta_v(t) = 45$$

(21)

where both of these equations are solved numerically using Matlab.

At this point, we have the means to relate the length of a straight path to the cardinality of the set of trajectories it represents (for a given angular velocity). This is done by performing the described calculations for a set of lengths ranging from 0.5 to 50 meters; namely, for each length, the trajectory function is generated numerically (by solving the motion equations for the proper set of parameters), then a rational function is fitted and the line integral is calculated. The values of the calculated integrals are then presented as a function of the length for which they were calculated (Figure 24). As expected, since the curves are getting shorter when the length of the straight path is getting longer, the values of the calculated integrals exhibit an inverse relation to the length of the straight path. This result reinforces our hypothesis of an inverse relation.
Figure 23: Fitting a rational function to characterize the relation between the completion time and the velocity angle.

This fit was made for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5\text{rev/sec}$. Similar results with different coefficients were obtained when performing the same tests but with a different selection of parameters.

between the distance of a shot and its scoring probability (less trajectories for greater distances means less opportunities to score).

Actually, the described function (Figure 24) establishes a mathematical relation between the cardinality of the set of trajectories of a straight path and its length. Once again, we fit a function to characterize this relation. Among several functions that were tested, we have selected a power function of the form

$$f(r) = a \cdot r^b + c,$$

and when we use the average value of the angular velocity (for the numerical solution of the motion equations within the described procedure) we get $a = -1.6$, $b = 0.69$, $c = 30.63$ (Figure 24). Naturally, due to the discussed properties of this function, we consider this function to be a candidate for the straight-path weight function $W(r)$ (see (3) and (4)).

Another interesting observation, regarding the calculated curve, is that excluding the values of short distances, gives rise to a linear trend. Fitting a linear function (Figure 25) to describe this relation (using all the values) we get

$$\tilde{f}(r) = 27.32 - 0.43 \cdot r,$$
Figure 24: Fitting a power function to characterize the relation between the line integral and the length of a straight path.

This fit was made for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5.5 \text{revs/sec}$. Similar results with different coefficients were obtained when performing the same tests but with a different selection of parameters.

which is close to our first candidate for the straight-path weight function ($W(r) = 50 - r$) up to a multiplicative constant (approximately 2). In fact, it means that a linear fit to this output is equivalent to the use of the first candidate since the integration over the weight function is a linear operation and the constant will only affect the scale of the covariates. This observation supports our primary intuition regarding the way straight paths should be weighted. Furthermore, it suggests that a linear model is sufficient and therefore we will not investigate further the behavior of header trajectories; instead, we will use a linear weight function and estimate its coefficients—as a starting point (prior to this estimation) we shall use the intuitive coefficients divided by two (the multiplicative constant).

As seen in Figure 20, performing the same analysis with various values of the angular velocity, yields similar trends with slight differences. The similarity of the results led us to add the dimension of the angular velocity to the analysis. In fact, considering a two-dimensional domain yields a trajectory surface instead of a trajectory function (Figure 26). Respectively, the cardinality of the set of trajectories is reflected by the area of the surface and therefore its evaluation is done using a surface integral. This calculation requires a mathematical formulation of the surface which we now pursue.
Figure 25: Fitting a linear function to characterize the relation between the line integral and the length of a straight path.

This fit was made for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5.5\text{revs/sec}$. Similar results with different coefficients were obtained when performing the same tests but with a different selection of parameters.

Since different values of the angular velocity exhibit similar trends—that can be formulated similarly, as discussed—we have examined the change in the coefficients of these formulations as a function of the angular velocity (Figure 27). This allows us to once again fit mathematical models to these functions—in fact, a second-degree polynomial characterizes all of these functions in a satisfactory manner ($R^2 > 0.975$), and provides symmetry with respect to 0 angular velocity (corresponding to the expected symmetry in the results when only the direction of spin changes). These polynomials can now describe the change in coefficients as the angular velocity changes, and thus enables one to impose a mathematical representation of the surface.

Eventually the fitted formulation of the surface was of the form

$$V(t, \omega) = \frac{(p_{11} \cdot \omega^2 + p_{12} \cdot \omega + p_{13}) \cdot t^3 + (p_{21} \cdot \omega^2 + p_{22} \cdot \omega + p_{23}) \cdot t^2}{t + (p_{51} \cdot \omega^2 + p_{52} \cdot \omega + p_{53})} + \frac{(p_{31} \cdot \omega^2 + p_{32} \cdot \omega + p_{33}) \cdot t + (p_{41} \cdot \omega^2 + p_{42} \cdot \omega + p_{43})}{t + (p_{51} \cdot \omega^2 + p_{52} \cdot \omega + p_{53})}$$

(24)

which is actually based on (17) where the coefficients are functions of the angular velocity. The calculation
Each point on the surface is representing a single trajectory characterized by its completion time, angular velocity and the norm of its initial center of mass velocity. This surface was extracted for a straight path with length of 20 meters and $\theta = \phi = 0^\circ$.

of the surface integral is performed according to

$$\int \int_S f dS = \int \int_S f (V(t, \omega)) \cdot \left\| \frac{dV}{dt} \times \frac{dV}{d\omega} \right\| dtd\omega = \int_0^{t_{\max}(\omega)} \int_{t_{\min}(\omega)}^{11} \sqrt{\left( \frac{dV}{dt} \right)^2 + \left( \frac{dV}{d\omega} \right)^2 + 1} dtd\omega; \quad (25)$$

here again we assume uniform density of trajectories within the surface and hence $f (V(t, \omega)) = 1$; $t_{\min}(\omega)$ and $t_{\max}(\omega)$ are second-degree polynomials that characterize the range of eligible time as a function of the angular velocity (Figure 28). Both of these functions are characterized accurately by a second-degree polynomial function ($R^2 > 0.975$).

Relating the surface integral to the length of the straight path is done by repeating the calculation in (25) for a set of different lengths. The outcomes are presented as a function of the length of the straight path (Figure 29). It is easy to see that this curve exhibits a trend which is similar to the one in Figure 24 and the difference between them is in terms of scale. This observation is validated when we fit a power function (Figure 29), which yields

$$f(r) = a \cdot r^b + c, \quad (26)$$

where $a = -17.04$, $b = 0.69$, $c = 334.7$. This result provides a power coefficient (formulated as “b”) which is similar to the one in the line integral power fit (22), but with different scale coefficients (formulated as
Figure 27: Change in the trajectory function coefficient values as a function of the angular velocity.

Each of the subplots exhibit the change in the respective coefficient of the trajectory function \(^{17}\) as the values of the angular velocity changes. All of the plots were extracted for a straight path with a length of 20 meters and \(\theta = \phi = 0^\circ\).

Figure 28: Minimal and maximal completion times as a function of the angular velocity.

Both of these graphs were generated for a straight path with a length of 20 meters and \(\theta = \phi = 0^\circ\). The fit of the maximal completion time is less accurate due to the fact that the maximal completion time values are extracted numerically using the relation between the completion time and the \(\phi\) angle, while the minimal completion time values are extracted using the relation between the completion time and the initial center of mass velocity norm. Nonetheless, both fits yield an \(R^2\) value which is greater than 0.99.
Figure 29: Fitting a power function to characterize the relation between the surface integral and the length of a straight path.

This fit was made for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, and $\theta = 0^\circ$. Similar results with different coefficients were obtained when performing the same tests but with a different selection of parameters.

“$a$” and ”$c$”), which are approximately the coefficients fitted in (22) multiplied by 11 (the size of the angular velocity range) . Naturally this function is also a candidate for the straight-path weight function.

Both of the discussed candidates were tested similarly to the first, intuitive, weight function (see Figures 3-5 and 8-10). They both yielded results which were similar to the intuitive weight function. Considering this similarity, we can further reduce the number of weight functions according to other properties. In fact, we have seen that the weight function on the trajectory surface, is approximately the weight function that is based on the trajectory curve with different scaling. We will therefore analyze the data using only one of them. Since we assume no angular velocity for headers, choosing the weight function based on the trajectory curve (single angular velocity) makes more sense. Summing it all up, we are left with two possible weight functions:

- **Linear:** $W_1(r) = (50 - r) \cdot I_{\text{Kick}} + (16.5 - r) \cdot I_{\text{Header}}$

- **Power:** $W_2(r) = (-1.6 \cdot r^{0.69} + 30.63) \cdot I_{\text{Kick}} + (8.25 - 0.5 \cdot r) \cdot I_{\text{Header}}$.

The final selection of the weight function is described in the next chapter, where we present our statistical analyses.
5.3 Obstacle reduction

One major simplification in our models is neglecting the curvature of actual trajectories. Though this simplification is practice-based, it may still have a significant effect on the validity of our models. An example of such a situation is exhibited in Figure 30: in this example, our models yield a weight of 0, while it is easy to see that there exist feasible scoring trajectories. A similar problem, with a less significant effect, may occur in free kicks, which are frequent in soccer. In this subsection, we investigate and analyze the effect of obstacles, within the range of a shot, on the feasible trajectories, and apply it in our models. The analysis is done in the context of the straight-path model since we have established that the solid-angle model can be perceived as a special case of the straight-path model with a constant weight function equal to 1. Furthermore, we apply this analysis to kicks exclusively, since we assume that the different characteristics of headers (namely, lower angular velocities and higher origins) makes the discussed phenomenon less significant.

It is very natural to postulate that the location of an obstacle, relative to the location of the shot, is a significant factor in determining the effect of the obstacle on the scoring probability. This can be easily justified by thinking of an obstacle located very close to the location of the shot, which eliminates essentially
all the feasible trajectories. In fact, this factor is reflected in our models, as obstacles eliminate less straight paths and solid angles, as their distance from the location of the shot increases. However, we believe that the location of the obstacle should also be considered with respect to its distance from the goalmouth—a factor that is not taken into account in our models. This is reasonable since, when considering a straight path, all of its trajectories start at the same point, then expand in space and ultimately converge to a single point. An intuitive clarification of this concept arises from considering the fact that an obstacle located on the goal line will eliminate all the trajectories of the straight paths whose destinations are within the area it covers.

Considering all the above, we wish to find some kind of a correction to adjust the way that our models accommodate obstacles. The first step in obtaining such a correction, is a systematic analysis of the actual effect of an obstacle on the set of feasible trajectories under various conditions and locations. Comparing the outcomes of the two approaches (models versus accurate analysis) will enable the evaluation of the needed correction. Since, as we have emphasized, we wish to remain in the framework of our models, we choose to compensate for the miscalculation of the models by reducing the dimensions of the obstacles—that is, consider smaller obstacles. Obstacles can appear in the range of the shot in various ways. Specifically, the areas they cover may vary according to the setup of the attempt. In order to simplify the analyses (of how the presence of an obstacle effects trajectories) we will perform them under arbitrary conditions and apply their results for any given situation.

Considering the physical factors that are modeled via the motion equations, we realize that the curvature of a trajectory can arise in any of the three Cartesian dimensions. However, a quick examination reveals some essential differences between the characteristics of the problem in the Z-axis, and those of the problem in the X-Y plane. In addition to the difference in the physical forces that appear in these two spaces, another major aspect of the difference is in the motion ability of the obstacles. While in the X-Y plane the obstacle motion towards the ball is only time limited, the motion along the Z-axis is much more constrained. Moreover, considering field players, the X-Y motion is not modeled while in the Z-axis case it is modeled through the dimensions of the player. Finally, and with high significance, when considering the X-Y plane, there might be some straight paths that do not intersect with the obstacle but do represent trajectories that intersect with it. This situation is not possible in terms of the Z-axis; that is, if a straight path passes above an obstacle, it does not contain any trajectory that intersects with the obstacle. In accordance with all the above, we divide
our analysis into these two spaces: $Z$ and $X-Y$.

5.3.1 Z-axis analysis

The purpose of the Z-axis analysis is to evaluate the trajectories that circumvent an obstacle in terms of the Z-axis (i.e. pass it from above), which are represented by straight paths that do intersect with the obstacle. This will assist in correcting the models (or practically the obstacle size), so they will correspond to the actual set of intersected trajectories or to its weight.

In order to keep this analysis in one dimension, we consider a set of straight paths that differ among each other only in their $Z$ coordinate of the intersection with the obstacle. The natural choice is obviously the mid-line of the obstacle, that is the line that connects between the centers of the two bases of the obstacle that are perpendicular to the Z-axis. This selection provides the desired differentiation between the straight paths, a differentiation that can practically be expressed in terms of the $\theta$ angles of the straight paths. As discussed previously, the differences between sets of trajectories caused by differences in the $\theta$ angle are negligible, and thus we can reduce our analysis to a single straight path—we choose a representative straight path to be perpendicular to the Z-axis and we refer to it as *the representative straight path*.

The analysis focuses on obstacles that are within the range of the shot. Moreover, we reduce our analysis to shots that set the center of the goal line to be $(0, y, 0)$ (i.e. attempts that are made in front of the center of the goal line), and to obstacles such that the center of their lower base is at $(0, u, 0)$ (where $0 \leq u \leq y$), and that are parallel to the goalmouth. These simplifications are legitimate in light of our previous findings which revealed that the main factor that characterizes the difference between straight paths is their length, while other factors were found negligible. Furthermore, handling obstacles in such a constellation ensures a representative straight path, the destination of which is within the goalmouth (actually its destination is the center of the goal line). Eventually, the degrees of freedom in this analysis are the length of the representative straight path and the value of $u$; these parameters actually reflect both the distance from the location of the shot and from the goalmouth—as desired.

In this particular analysis, evading an obstacle is a matter of the height of the trajectory when its $Y$ coordinate reaches the value of $u$. Therefore, given the locations of an attempt and an obstacle, we wish to examine the heights of the trajectories of the representative straight path at the location of the obstacle.
The curves display the motion of the ball from the straight path origin to the location of an obstacle for trajectories that differ only in their completion time. They were generated for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$, $\omega = 5.5 \text{revs/sec}$ and $u = 5$ (that is, the obstacle is located 5 meters from the location of the attempt).

(Figure 31). Plotting the heights against the completion time of the trajectories (Figure 32) we notice a positive correlation between completion time and height. Similar results with different numerical values were obtained when performing the same tests with a different selection of locations. The relation between the height and the completion time determines that the trajectory that achieves the minimal height is the one with the minimal completion time—we refer to this trajectory as the minimal trajectory. Therefore, by generating the minimal trajectory, we can calculate the minimal height achieved at any obstacle’s location. Actually, the minimal height achieved at a location may be interpreted alternatively—the range from the top of the obstacle down, in which intersected straight paths represent only trajectories that evade the obstacle. In other words, if a straight path intersects with the obstacle at a point in that range, then even the minimal trajectory will evade the obstacle, and hence all the trajectories of that straight path will not intersect with the obstacle. Therefore, at any location, the height of an obstacle can be reduced by the height of the minimal trajectory at that location.

In practice, we seek a systematic way to calculate this reduction without having to solve the motion equations. To do so, we first wish to impose the minimal trajectory into a mathematical framework. Con-
Figure 32: Heights of the trajectories of the representative straight path as a function of their completion time.

It is clear from this graph that the higher the completion time of a trajectory is, the higher is its $Z$-coordinate at the same $Y$-coordinate (this result is consistent for any selected location of the obstacle). These heights were generated for the following parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$, $\omega = 5.5 \text{ rev/s}$, and $u = 5$ (that is, the obstacle is located 5 meters from the location of the attempt).

Considering the described setup, the location of the obstacle can be reflected via the progress along the Y-axis. This allows us to plot the height of the minimal trajectory as a function of the length covered (Figure 33), and furthermore to fit a polynomial function to the plot. A set of minimal trajectories with distances within the range of $0.5 - 50$ meters were generated. A second-degree polynomial $(Z(y) = a \cdot y^2 + b \cdot y + c)$ was fitted and provided a very good fit ($R^2 > 0.99$). However, following our choice of the representative straight path, we fitted a second-degree polynomial with $c = 0$ (Figure 33). The selection of this model allows us to calculate the coefficients in a simpler manner. If we note the total distance covered (in terms of the $Y$-axis) by $l$ and the maximal height achieved by $m$, we get that

$$a = -\frac{4 \cdot m}{l^2}, \quad b = \frac{4 \cdot m}{l}$$

which is derived by setting $Z(l) = 0$ and $Z \left( \frac{l}{2} \right) = m$. This result enables us to fit a polynomial to characterize the minimal trajectory according to the values of $l$ and $m$ alone. Since the physical models relate these two parameters, we can, in fact, reduce this problem by expressing one of them in terms of the other.

Actually, the generation of the minimal trajectories allows us to track their height at each point and
Minimal trajectories of straight paths of several distances were generated and functions of the form \( Z = a \cdot y^2 + b \cdot y \) were applied. It is noticeable that this formulation provides a reasonable fit for all of these distances. The trajectories were generated for the following parameters: \( \phi = 0^\circ \), \( \theta = 0^\circ \) and \( \omega = 5.5 \text{rev/sec} \).

Therefore to extract their maximal height (Figure 34). In fact, extracting the heights for the set of different lengths allows us to fit a third-degree polynomial to describe the relation between the maximal height of a minimal trajectory and its distance. Eventually, the fitted polynomial is given by

\[
m(l) = 3.265e^{-5} \cdot l^3 + 1.374e^{-4} \cdot l^2 + 0.014 \cdot l - 0.039
\]

and combined with (27) we can characterize the minimal trajectory of any distance at any point. For instance, if we wish to calculate the height achieved by the minimal trajectory that covers a distance of 20 meters after 5 meters, we set \( l = 20 \), and applying (28) we get the maximal height to be \( m = 0.5544 \). Applying this result to (27) we get \( a = -\frac{4m}{l^2} = -0.0055 \), \( b = \frac{4m}{l} = 0.111 \), and therefore the desired height is achieved by \( h = -0.0055 \cdot 5^2 + 0.111 \cdot 5 = 0.416 \).

Even after we perform the reduction in the obstacle’s dimension, there are still some trajectories that evade the obstacle from above that are represented by straight paths that do intersect with the obstacle. Though the discussed analysis can be extended in some ways to quantify the weight of these trajectories, and to further reduce the dimensions of the obstacle, we choose not to do so. This choice is made since much uncertainty is involved in our analyses. It is mainly expressed in terms of the parameters and constraints.
Figure 34: Maximal heights of minimal trajectories.

Each of the points represent the maximal height of the minimal trajectory of the respective length. The trajectories were generated with the following parameters: $\phi = 0^\circ$, $\theta = 0^\circ$ and $\omega = 5.5 \text{revs/sec}$.

chosen in the solutions of the motion equations. In light of this uncertainty, we believe that this general application is sufficient. However, the discussed methods and analysis, may lay the foundation for a more accurate evaluation when the overall uncertainty is reduced.

Finally, in order to apply the results of this analysis to our models, we must compare them to the way the models handle obstacles, and to add the needed corrections in a proper way. Actually, there are many possible situations in which our models will not consider the whole height of an obstacle (Figure 35); in these situations, the straight paths that connect the locations of the shot to the upper frame of the goalmouth (points of the form $(x, 0, 2.44)$), intersect with the obstacle in heights that are lower than its maximal height. Therefore the effective height of the obstacle is determined by the highest point of intersection. This reduction is obviously consistent with the models. However, when considering the curvature of actual trajectories, this is not necessarily the case. The reason for that is that even if an area was eliminated due to a lower intersection point, this area may still intersect with trajectories that are represented by straight paths within the effective height of the obstacles. For instance, if the minimal height for a certain location is 30 cm, and the effective height of the obstacle at the same location is 150 cm (that is, a reduction of 50 cm by the model), then the trajectory with minimal height represented by a straight path that intersects with the obstacle at 150
This plot represents a kick made at a distance of 20 meters from the center of the goal line with an obstacle (a field player) located 9 meters from the location of the attempt. The red line is the straight path that connects the location of the attempt to the center of the upper base of the goalmouth; it intersects with the obstacle at a height of 1.098 meters—this is the obstacle’s effective height according to our models. The green curve is the minimal trajectory of the red straight path; since it does not evade the obstacle the latter is not reduced further than the model reduction.

cm height will not evade the obstacle. Therefore, the total reduction of the height of an obstacle is determined according to the maximal reduction between the two discussed reductions.

5.3.2 X-Y plane analysis

As explained, the major difference between the analysis of the X-Y plane and the Z-axis is expressed through the fact that, in the X-Y plane there may be straight paths that do not intersect with the obstacle but do represent trajectories that intersect with it. This leads us to aim for an analysis of the effect of an obstacle on all the straight paths formed by the attempt. Therefore, the analysis of the effect of obstacles in the X-Y plane should be done based on an alternative approach, and by using different methods and tools.

Since we implicitly postulate that the effect of an obstacle within the Z-axis is independent of its effect on the X-Y plane, and due to complexity concerns, we limit our analysis to straight paths of a single Z-coordinate. Moreover, we perform the analysis in the same setup that was used for the Z-axis analysis, that
is considering a kick taken in front of the center of the goal line and an obstacle, which is parallel to the
goalmouth, located in various points between the location of the shot and the center of the goal line. This
framework actually sets our analysis to be one-dimensional, with variation only along the X-axis—reflected
by the selection of the destination of the straight path within the goalmouth (in the range $-3.66 \leq x \leq 3.66$
)—while the Y-coordinate and the Z-coordinate remain constant.

Given an obstacle, for each straight path in our analysis, we wish to evaluate the set of represented tra-
jectories that evade the obstacle, and the set of those that intersect with the obstacle. Solving the motion
equations, using the TPBVP setup, allows us to define stopping terms, that is a set of conditions that termi-
nates the calculation (or the progress of the trajectories) once satisfied. Following our setup and definitions,
we naturally set the location of the obstacle (actually the Y-coordinate of its location) as a stopping term.
Solving the equations using Matlab also allows us to track the values of the trajectory at any point along the
range between its origin and its end (where the stopping term was satisfied). In particular, this enables us to
extract the value of the X-coordinate of the trajectory at the time when the condition is satisfied, and by this
to determine whether the trajectory evaded the obstacle or intersected it (Figure 36). Considering the analysis
setup, for a certain location of an obstacle (in terms of the Y-coordinate), the obstacle covers a range within
the X-axis according to its width and symmetry respectively to the plane in which $X = 0$; a field player for
instance will cover the range of $-0.225 \leq x \leq 0.225$. Thus, a trajectory intersects with the obstacle if the
value of its X-coordinate at the time when the stopping condition is satisfied is within the range covered by
the obstacle.

The discussed routine enables a classification of the trajectories as desired. Furthermore, if we classify
the trajectories of a straight path and examine their completion time and initial velocity parameters (according
to our alternative representation of trajectories), we can notice that the set of the intersected trajectories is
continuous in these terms (Figure 37). Actually, if we plot the trajectory curve of a straight path, we notice
a continuous curve formed by the intersected trajectories. This observation naturally leads us to evaluate the
cardinality of the set of intersected trajectories by calculating the line integral along the partial curve, formed
by these trajectories.

The calculation of this line integral is done similarly to the prior calculation of the same type. In fact,
since we also wish to compare the cardinality of the set of intersected trajectories to the cardinality of the set
Figure 36: Trajectories evading and intersecting with an obstacle on the X-Y plane.

The red curve is a trajectory that intersects with the obstacle (located 3 meters from the location of the shot) while the green one evades it in terms of its X-coordinate. Both of the trajectories are represented by the same straight path and were generated using the parameters: straight path length of 20 meters, $\phi = 0^\circ$, $\theta = 0^\circ$, and $\omega = 5.5 \text{revs/sec}$.

Figure 37: Specifying the sets of intersecting and evading trajectories, by completion time and initial velocity.

The presented curve corresponds to our alternative trajectory representation, and the types of the trajectories are separated via different colors: blue for evading trajectories and red for intersecting trajectories. All the trajectories were generated for the following parameters: shot origin at $(0,0,0)$, shot destination at $(-2,20,0)$, center of obstacle’s lower base located at $(0,8,0)$ and $\omega = 5.5 \text{revs/sec}$. 
Figure 38: X-coordinate at an obstacle’s location as a function of the completion time of the trajectory.

Each of the blue points represent the X-coordinate achieved at the location of the obstacle of the trajectory with the corresponding completion time. The red curve is the quadratic fit applied to characterize the points. All the trajectories were generated for the following parameters: shot origin at (0,0,0), shot destination at (−2,20,0), center of obstacle’s lower base located at (0,8,0) and \( \omega = 5.5 \text{revs/sec} \).

of all the trajectories of the straight path, we use the same function to characterize the curve mathematically (17), which yields the same formulation (18) for the line integral. Hence, the only missing parameters needed to perform the calculation are the boundaries of the integration, which are actually the edges of the range of the completion time of the intersected trajectories. Turning to our previously-used methods, we fit a function to describe the relation between the completion time of the trajectory and the value of its X-coordinate at the moment of the stopping term satisfaction (Figure 38). A quadratic function provided a very good fit \( (R^2 > 0.99) \) for all the examined combinations of distances and obstacle locations (the required values of the stopping terms), and the respective roots were found numerically (using Matlab), as in (19) and (20). We note that in some cases there is a partial range of the obstacle with which trajectories do not intersect; in these cases, the theoretical completion time boundaries exceed the range of the eligible completion time (according to the previously-discussed constraints). This means that the proper completion time constraint becomes active prior to intersecting with a part of the obstacle, and therefore the integration uses the constraint boundary instead.

At this point, we have established a way to measure and evaluate the cardinality of the set of trajectories
For each possible destination of the straight paths within the goal line, the proportion of trajectories that intersect with the obstacle was calculated. All the trajectories were generated for the following parameters: shot origin at \((0, 0, 0)\), center of goal line \((0, 20, 0)\), center of obstacle’s lower base located at \((0, 18, 0)\) and \(\omega = 5.5\text{revs/sec}\).

that intersect with a given obstacle. Actually, we evaluate the effect of an obstacle on a straight path by the proportional difference between the cardinality of the set of intersected trajectories, and the cardinality of the set of all the trajectories that are represented by the straight path. This measure allows us to compare the effect of an obstacle on the range of straight paths of interest. Practically, this means that, given an obstacle’s location, we can now get a full characterization of the magnitude of its effect on the straight paths that end in the range of \(-3.66 \leq x \leq 3.66\) (Figure 39). This procedure allows us to investigate the effect of an obstacle as a function of its location (Figures 40, 41). We note that if we had set an angular velocity of the same value but with opposite signs, we would have obtained mirror images of the plots.

The described procedure reveals patterns in the effect of an obstacle on straight paths. Analyzing the patterns, we can divide the range of the location of the obstacle into two groups: the obstacle is close to one of the edges (the location of the shot or to the goalmouth) or the obstacle is in midway between them. When the obstacle is somewhere midway between the location of the shot and the goalmouth (Figure 40), it is noticeable that as the distance of the obstacle from the location of the shot is bigger, the range of straight paths that are affected becomes smaller. Furthermore, these plots exhibit a pattern in which the magnitude
Figure 40: Intersection proportion as a function of the destination of the straight path calculated for various mid-range obstacle’s locations.

Each curve describes the intersection proportion as a function of the destination of the straight path for a given obstacle’s location (within the mid-range of possible locations). All the trajectories were generated for the following parameters: shot origin at \((0,0,0)\), center of goal line \((0,20,0)\) and \(\omega = 5.5 \text{rev/sec}\).

of the effect (measured by the proportion of the intersected trajectories) increases to a maximum, and then drops back to zero. Finally, we notice that the effected range gets closer to the range of x-coordinates of the obstacle \((-0.225 \leq x \leq 0.225)\) as the obstacle’s location is closer to the goalmouth. Actually, these trends also appear within the other group (Figure 41), in which the obstacle is closer to the edges, but due to the high density of trajectories in these ranges, many straight paths exhibit the maximal intersection proportion, and the decreasing part of the pattern does not occur immediately after the increasing part. Moreover, when the obstacle is close to the location of the shot, the increasing part is censored—that is, its theoretical beginning is out of the range of the goalmouth.

In order to quantify these phenomena, we wish to evaluate the total effect of an obstacle in a certain location, that is to define a single measure that reflects the effect of the obstacle on the whole range of the straight paths. Since the value, measured for each straight path, reflects a proportion, we can calculate the total effect by evaluating the area under the graph. Such a calculation requires integration over the range. Due to the complex structure of these graphs, we choose to calculate it numerically. The numerical calculation is done by fitting an interpolant model (a non-parametric model) and calculating the integral numerically—this
Figure 41: Intersection proportion as a function of the destination of the straight path calculated for various near-edged obstacle’s locations.

Each curve describes the intersection proportion as a function of the destination of the straight path for a given obstacle’s location (within the edges of possible locations). All the trajectories were generated for the following parameters: shot origin at (0, 0, 0), center of goal line (0, 20, 0) and $\omega = 5.5\text{revs/sec}$.

is done using special features within Matlab.

The achieved measure actually reflects the effect of the obstacle as if it was located on the goalmouth. We therefore refer to this measure as the obstacle’s effective width. In fact, using this measure makes the effect in different locations of the obstacle comparable, and thus allows us to examine the effect of an obstacle as a function of its location (Figure 42). As expected, the effective width as a function of the location exhibits a decreasing trend starting at 7.32m—when the obstacle is very close to the location of the shot it eliminates all the trajectories and therefore, its effective width is equal to the total width of the goalmouth—and ending at 0.45m—when the obstacle is located at the goalmouth, each straight path intersected with it is completely eliminated and thus its effective width is equal to its total width.

As stated before, we wish to compare the effect of the obstacle, as measured via its effective width, to the one reflected via our models. Following the above, we need to transform the latter to reflect the effect in terms of the width of the goalmouth. The nature of our models allows us to perform this transformation easily using proportional considerations. Given a location of a shot and an obstacle located between the location of the shot and the goalmouth (in a manner that is consistent with our previously-described setup),
The effective width was calculated for a discretization of the possible locations of a defender for the following parameters: shot origin at \((0, 0, 0)\), center of goal line \((0, 20, 0)\) and \(\omega = 5.5 \text{revs/sec}\).

one can form a triangle by connecting the edges of the obstacle’s base to the location of the shot. The lines, that connect the obstacle’s base to the location of the shot, can be extended until they reach the goal line. This forms another (larger) triangle whose base is actually the desired transformation. Due to the fact that the base of the obstacle is parallel to the goal line, the triangles are proportionate and the following equation holds:

\[
\frac{x}{w} = \frac{y}{u},
\]

(29)

where \(x\) is the unknown size of the base of the larger triangle, \(w\) is the obstacle’s width (or the size of the base of the smaller triangle) and \(y\) and \(u\) reflect the distances of the center of the goal line and the center of the obstacle’s lower base respectively (as defined previously). Therefore, the proportionate size can be calculated via \(x = \frac{yw}{u}\). While the maximal cover in terms of the goalmouth is bounded by 7.32 meters, the base of the larger triangle is theoretically unbounded. In cases where the obstacle’s cover is bigger than the goalmouth, the models consider only the part of the obstacle that covers the goalmouth, and therefore the
actual calculation is achieved via

\[ x = \min \left( \frac{y \cdot w}{u}, 7.32 \right). \]  

(30)

Finally, we compare the obstacle’s effective width to its transformed width according to our models (Figure 43). Surprisingly, the widths are almost identical for the majority of the ranges of the obstacle’s locations. In fact, there is a significant difference between the widths within a minor range, in which the location of the obstacle is close to the location of the shot. This difference is in fact caused by our definitions, particularly the way we define the effective width of an obstacle. The effective width considers trajectories that satisfy two properties: representation by a straight path and intersection with the obstacle. The discrepancy occurs when the obstacle is close to the location of the shot; then there are trajectories that intersect with the obstacle, but end at a point outside of the goalmouth (and therefore are not represented by a straight path, and excluded from our analyses); these trajectories are not considered in the calculation of the effective width. This does not happen when the obstacle is far enough so that all of the trajectories that intersect with it are actually feasible for scoring. This phenomenon can also be observed in the intersection proportion versus destination graphs (Figure 41): it seems like the increasing section is partial, and can be continued within the range which is not feasible for scoring. The magnitude of this phenomenon is associated with the curvature of the trajectories—since curved means not straight—that can actually be expressed by the constant value of the angular velocity. Repeating the described scheme for various angular velocities justifies this claim (Figure 44).

Following the above, we realize that a different setup (in which the obstacle is not on the line that connects the location of the shot and the center of the goal line) may yield different outcomes, with different ranges of effect and therefore, the results may not be generalizable. However, since we did notice a strong similarity to the way the models accommodate obstacles, we have decided not to apply any correction to the width of the obstacle.

6 Data and analysis

As stated, the approach of this research in estimating the scoring probability is to model the physical and geometric factors, that characterizes scoring attempts, to examine their correspondence to the real world, and to quantify their relation to the scoring probability. Up to now, we mainly focused on describing the
Figure 43: Obstacle widths comparison.

The blue curve exhibits the effective width of the obstacle as defined previously, being a function of the location of the obstacle; the red curve is the obstacle’s width according to the models. The curves were calculated for the following parameters: shot origin at \((0, 0, 0)\), center of goal line \((0, 20, 0)\) and \(\omega = 5.5 \text{revs/sec}\) (the angular velocity does not affect the way the models calculate the width).

Figure 44: Obstacle width comparison for various angular velocities.

The red curve is the obstacle’s width according to the models: the rest of the curves exhibit the effective width of the obstacle, as defined previously for several values of angular velocity. It is clear that the difference between the model curves and the effective width grows bigger with the magnitude of the angular velocity. The curves were calculated for the following parameters: shot origin at \((0, 0, 0)\), center of goal line \((0, 20, 0)\) and varying angular velocities (namely: \(\omega = 0.1 \text{revs/sec}\), \(\omega = 5.5 \text{revs/sec}\) and \(\omega = 11 \text{revs/sec}\)).
essence of the models, the way they were developed, and how they quantify the physical and geometric factors assumed to effect the scoring probability. These discussions, and the respective analyses were mainly theoretical. In order to examine our models empirically, we apply the models on data of actual scoring attempts; for each attempt the relevant parameters are extracted and the attempt is then weighted according to the calculations of the models. Eventually, each of the attempts consist of new attributes containing the outputs of the models for the parameters of the attempt. This allows us to statistically evaluate the relation between the outputs of the models, scoring, and empirical scoring probabilities.

In this chapter we describe the data set we have used, as well as the statistical methods we applied in order to estimate the scoring probability and to quantify its relation to the outputs of our models.

6.1 Data description

Though there are many companies that possess and handle data that suits our needs, we were not able to attain a decent data set for the research from any company. Therefore, as in previous researches, we decided to collect data ourselves. In order to maintain the research within the scope of the factors of interest (physical and geometric), we need to control the rest of the factors. Thus, we aimed to collect a data sample which is as homogeneous as possible, in terms of factors which are external to the model. Specifically, we decided to collect data from matches of a single soccer team, played within a single tournament, during one season. For this matter, we collected the scoring attempts made in the games of London’s Arsenal FC, played in the English Premier League, during the 2012/2013 season.

Three types of attributes were collected for each attempt: physical and geometric inputs, as required for the models, outcome attributes for the probability estimation, and finally, for each attempt, several objective measurable attributes were collected, enabling an examination of effects which are external to the models. Namely, each record contains the following attributes: attempting player’s location (in 3 dimensions), locations of field players within the range of the shot, goalkeeper’s location, type of shot (kick\header), goal outcome, frame outcome, official game time of the attempt, the half of the game in which the attempt was made, team identity (Arsenal\other), mode of attempt (free kick\open play), attempting player’s identity (player’s number for arsenal players and 0 for players of other teams), the round of games (the English premier league consists of two rounds of games), attempting player domestication (home or away), and ball
Figure 45: Data-collection interface.

The left part of the interface demonstrates half a soccer field on which pins can be located; the right part is basically a form in which the proper parameters can be set.

control (indicating whether the attempting player touched the ball prior to the attempt).

The data collection was done using an interface specially developed for us by Noam Carmel (Figure 45). The interface visualizes half a soccer field in which pins, that represent different types of players (one for the attempting player, one for the goalkeeper and ten for other field players), can be positioned according to the scoring attempt to extract their locations. The rest of the interface contains a form that allows setting the other parameters. The process of data collection was done by viewing the game at $\times 4$ speed and marking the times of the scoring attempts; then, for these times, a second view was performed, in which the attempt was observed several times and from different angles (if possible) in order to reconstruct it on the interface.

A total of 38 games were viewed, from which a data set of 982 scoring attempts was extracted. The distribution of several of the dichotomous attributes of the data is exhibited in Figure 46. The most significant result derived from these distributions is a scoring ratio (goals/attempts) of 0.1059, which is consistent with the 0.1 scoring probability that was found in all the previous works, as described in the literature review chapter [7,15,16,17,18,19,23]. Rather than that, these distributions reveal several trends: an approximate
ratio of 1 to 5.5 headers per kicks, a relative rareness of free kicks, a significant advantage for Arsenal in the attempts made versus the attempts conceded (which is consistent with the fact that Arsenal finished in fourth place that season) and minor advantages in the number of attempts for the home teams and for the second halves.

It is reasonable to expand the discussion, considering the nature of our data, with the distributions of the location attributes. However, since these attributes are less informative separately, in the context of the scoring probability, and furthermore, our models actually connect these attributes, it is more natural to discuss them together and to present them as they reflect by the outputs of our models. This is done by creating new attributes containing the outputs of the models—by applying the models on each of the records—and examining their distributions (Figure 47). A quick look at the histograms reveals a majority of observations with low output value, and a relatively long tale of the distribution with few observations. This may imply on a nonuniform scale, so to overcome this we transform the output using log transformations for all of the outputs (Figure 48). Once again, the majority of observations is at low values, but the transformation sharpens the distinction between different values. From here on we use the transformed values of the model outputs as our main covariates and we refer to them as the model outputs or the transformed model outputs.
6.2 Data analysis

In our research, we implicitly assume that a scoring attempt is a Bernoulli experiment, and thus can be characterized compactly by a single parameter—success (or scoring) probability denoted by $p$. The immediate difficulty, that rises from this assumption, is the paltry information received by the output of the experiment; this becomes even more significant considering the fact that even an experiment with very low value of $p$ can yield a success and vice versa. Another essential assumption in our research is that the value of $p$ changes with different values of the physical and geometric factors; this assumption increases the complexity of the problem. In fact, our models allow us to estimate the values of $p$ as a function of the physical and geometric factors through the outputs of our models. Though the use of the models reduces the complexity of the problem, we are still facing a complicated estimation task, in which any value of model’s output may be associated with a different value of $p$. Ideally, it would have been advantageous to have many observations for each of the values of the outputs of the models, so that the values of $p$ could be estimated separately. This is obviously impossible due to the continuity of the output values.

All of these difficulties are clarified when observing a plot (Figure 49) of the dependent variable (goal/no-goal) against the independent value (the transformed output of the models), aiming to visualize the relation
Figure 48: Histograms of the transformed models outputs.

between these two variables. Actually, the plot exhibits spreads of observations (in terms of the outputs of the models) for both of the values of the dependent variable. It is easy to see that there is no dichotomous classification of the dependent variable according to the independent variable (which is reasonable according to our assumptions), and furthermore it is hard to identify trends that may explain the variability of the data. All in all, this plot does not provide clear information and more complicated methods are needed to extract and evaluate the desired relations.

Following the above we now turn to some appropriate statistical methods and tools. As a first step, we use non-parametric statistics in order to get some idea about the relation between the variables, their relation to the scoring probability and the ways they may be characterized. Then we will examine these relations using a standard model for such situations (explaining a binary variable via various kinds of variables), and evaluate the goodness of these methods.

6.2.1 Non-parametric statistics

As discussed previously, we wish to estimate the scoring probability for any given value of the model outputs, but due to the continuity of the range of values this is not possible. A naive way to overcome this situation is by a discretization of the problem; that is, by dividing the ranges of the model outputs into distinct intervals, and estimating the probability within each of these intervals separately. Though, in some way, applying
Figure 49: Model output versus scoring.

Each point in the graph is an observation, characterized by its transformed model output value and its scoring output. The graph reveals differences in the distribution of the model output for the different scoring categories.

such a method satisfies the need of estimating a parameter under varying environments, it suffers from many problems.

Generally, transforming a continuous variable into a discrete one involves a loss of information. The nature of this loss corresponds to a tradeoff between the proximity of the new, discrete, variable and the continuous variable, to the effectiveness of the new variable (in terms of distinction between different values). This tradeoff can be reflected via the interval division: while a division into many small intervals may maintain most of the properties of the original variable, it will probably contain many intervals with insufficient number of observations within them; alternatively, a division to a small number of intervals may provide a good estimation within each interval but significant trends may be lost. Moreover, applying a discrete segmentation upon a continuous range may result in situations in which some observations are closer to observations within a different interval than to observations within the same interval. In other words, applying such a method may be more appropriate where the data exhibit a natural divisive pattern—which is not the case for our data. Nevertheless, we have applied this method on our data trying several types of division. Figure 50 exhibits such an attempt; despite the discussed problems, one can conclude from this graph that there is a positive correlation between the range of values of the model outputs and the scoring probability,
The range of the model output (calculated using $W_2(r)$) was segmented according to the distribution segmentation, and for each segment the scoring proportion was calculated (red dots). The histogram reflects frequencies relative to the maximal frequency bin, in order to achieve a scale which is similar to the proportion scale. An increase in the scoring proportion is noticeable.

as the scoring proportion rises with the output range.

In order to overcome the discussed difficulties, and to extract all the information from our covariates (the outputs of the models) in a continuous fashion, we use smoothing estimators instead of the discrete alternative. The main idea of smoothing is to avoid a division of the range of the covariate by weighting observations, according to their distance from the value of interest. So actually, instead of estimating the probability according to a point-wise or piece-wise average, we can consider all the observations, with respect to their proximity to the desired investigated value. In fact, using such a method allows us to estimate the scoring probability for any given value of the covariate—including values with no observations—and by that to estimate the probability as a continuous function of the outputs of the models.

We perform the smoothing using a Nadaraya-Watson estimator, which is common and standard. This estimator postulates a normal distribution of the observations centered at the value of interest, and weighs the observations according to the density obtained at a value equivalent to their distance from the value of
interest. Mathematically this estimator is formulated via

\[ \hat{p}(x) = \frac{\sum_{i \in I} \phi \left( \frac{x_i - x}{h} \right) \cdot y_i}{\sum_{i \in I} \phi \left( \frac{x_i - x}{h} \right)}, \tag{31} \]

where \( x \) is the value of interest, \( x_i \) is the value of the output of the model for the \( i^{th} \) observation, \( I \) is the index set corresponding to the observations, \( y_i \) is the outcome (goal/no-goal in our case) of the \( i^{th} \) observation, \( \phi (\cdot) \) is the normal density function and \( h \) is a smoothing parameter known as the bandwidth. Actually, the bandwidth role can be interpreted as the continuous analog to the role of the division selection (though, unlike the histogram, there is no arbitrary location of the division) in the discrete method; the size of the bandwidth controls the range of values in which observations will affect the scoring probability of the value of interest. Understanding that, we once again point out the tradeoff that occurs when selecting the size of the bandwidth: a selection of a high value of bandwidth will consider many observations of high weight, at any point of interest, and will result in a smooth estimator with minor trends (Figure 51); a selection of a low value of bandwidth will practically consider few observations at each point of interest and thus will provide an estimator with a rapid, not necessarily consistent, change of values with many curves and trends (Figure 51).

Naturally, we wish to find the balance in the discussed tradeoff; to do so we need to measure the properties of the estimator for any selected bandwidth value in a comparable way. A standard measure for such estimators is calculated by their error, given by

\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{p}(x_i))^2, \tag{32} \]

where \( n \) is the total number of observations (\( n = |I| \)). A standard way to apply this measure is by using the Leave One Out Cross Validation method. The LOOCV method assesses the forecasting abilities of the estimator, by systematically estimating the predicted value for each observation, based on the rest of the observations. This way we can compare different bandwidth selections, and select the bandwidth value which provides the lowest forecasting error (Figure 52). Selecting the bandwidth of minimal error yields an estimator that balances the aforementioned tradeoff (Figure 53). Using the transformed data is actually equivalent to using a varying value of bandwidth, since for a given bandwidth value, the range of the relatively significant observations is determined according to the value of interest with respect to the logarithm function.
Both of the curves were estimated for the transformed output of the straight-path model using $W_2(r)$ weight function. The difference between these two curves of extreme bandwidth values is reflected by their sensitivity to local trends.

Each point along the curve reflects the LOOCV error achieved for the given bandwidth; the curve exhibits a minimum point. The forecasting was done using the transformed solid angle output with $W_2(r)$ weight function.
Figure 53: Smoothed scoring probability with LOOCV bandwidth.

All three graphs exhibit a generally smoothed curve with few bumps. They all suggest a positive correlation between the models’ output and the scoring probability.

Applying the LOOCV method (Figure 53), we select bandwidths that provide smooth estimators of the scoring probability. The curves support our hypothesis that there is a positive correlation between the outputs of the models and the scoring probability, as it is noticeable that as the outputs become higher so does the scoring probability. However, none of these estimators exhibit a strictly increasing trend, and they all contain ranges which present local negative correlation with the model outputs; this can be significantly observed in the solid angle-based estimator, that exhibits a decreasing trend for model output for high values. This inconsistency can be associated with the minority of observations in that range, but requires further investigation, which will be discussed later.

Rather than estimating the scoring probability and relating it to the outputs of our models, the smoothing estimators allow us to examine the effects of the external factors on the scoring probability. Testing these effects is done by dividing the observations into distinct groups according to the values of the examined factor; for each value of the external factor, the estimator is applied on the proper data, and compared to the estimators generated for the other values of the factor.

The purpose of using such a method is to achieve a general insight on the influence of some factors, rather than determining the significance of their effects. In fact, examining the data according to this procedure revealed some differences between the estimators of different values (Figure 54); these differences appear
The estimators were generated for the straight-path model using $W_2(r)$ weight function. The bandwidth value for all estimators is the LOOCV bandwidth based on all the observations.

mainly at high values of the model output, and can be associated with the minority of observations at those ranges. The most significant difference occurred when the effect of the shot type was tested (Figure 55): for all the examined model outputs there were differences between the estimators for kicks and headers. Particularly, the most significant difference occurred for the solid angle-based estimators, which is reasonable in light of the use of the same weight function for the different types of shots in this model. In all cases, the magnitude of the difference is maximal in a range that corresponds to the ranges in which the total estimators exceeded their increasing trend. Actually, these differences between kicks and headers may provide a good explanation for these trends.

As discussed earlier, the weight functions for headers were determined according to some general guidelines, without accurate physical analysis. This may be a reason for the discrepancies that occur between the kicks and headers estimators. One way to overcome this situation is by examining other weight functions for headers. We will, however, maintain the guidelines achieved in the physical analysis by reducing our search to linear functions. In fact, this will be done by estimating the coefficients for the headers weight functions. Since the model outputs are calculated numerically, and are not represented mathematically as a function of the weight functions, it is complicated to approach this estimation problem as a standard optimization problem. Within the boundaries of our research we choose to apply a simple solution to overcome
As mentioned, the linear structure of the weight functions allows us to ignore multiplication of the functions by constants—it only affects the values of the probability coefficients. In fact, this means we can reduce the problem to a single parameter search. Aiming for consistency with the weight functions of kicks, we choose to set the multiplicative parameter correspondingly. In practice this means that the weight functions are

\[
W_1(r) = (50 - r) \cdot I_{\text{Kick}} + (b_1 - r) \cdot I_{\text{Header}}
\]

\[
W_2(r) = (-1.6 \cdot r^{0.69} + 30.63) \cdot I_{\text{Kick}} + (b_2 - 0.5 \cdot r) \cdot I_{\text{Header}},
\]

where \(b_1\) and \(b_2\) are the linear parameters we wish to find. Since these parameters affect the error only for headers, we evaluate the error separately for the header observations. The search was done for values that are in the proximate range of the initial values (\(b_1 = 16.5, b_2 = 8.25\)) and the achieved local minimums (Figure 56) suggest that these ranges ([5, 28] and [2.5, 14] respectively) were sufficient. Eventually, the estimated coefficients were \(b_1 = 12.67\) and \(b_2 = 6.91\) and we use these values from here on.

Despite these discussed corrections, there have not been dramatic changes in the smoothed estimator graphs (Figure 57), and there are still discrepancies between the kicks and headers estimators. Considering the different characteristics of kicks and headers, it is reasonable to consider two different estimators for these
Figure 56: Estimating the “b-coefficients” of the headers weight functions.

Each point on these curves represent the prediction error obtained using the estimator that is based on a header weight function with the respective b coefficient.

two types of attempts. In any case, the effect of the type of attempt will be tested within the framework of the logistic regression. We note that there may be some other methods that might have reduced the difference between the kicks and headers values (such as defining the error relative to the kicks estimator) but in the scope of this research we did not explore them.

At this point, after revealing the correlation between the model outputs and scoring probability, and understanding some of the factors that affect it, we wish to actually evaluate the scoring probability for any given scoring attempt. This is done using the Logistic regression as described in the following subsection.

6.2.2 Logistic regression

The use of a non-parametric regression, to estimate the scoring probability, provides an overview on the relation between the scoring probability and the model outputs. The estimators are fitted based on local information, and are very sensitive to the actual data set. Adding the bandwidth as another degree of freedom increases the exposure of this method to overfitting (that is the dependency of the fitted estimator to the specific data set). Seeking a more robust method to estimate the scoring probability, we turn to a stan-
Figure 57: Comparison of smoothing estimators.

The graphs exhibit the differences between the smoothing estimators, obtained using the original header coefficients as weight functions, and those obtained when these coefficients were estimated. In all the graphs the blue curves are the estimators obtained using the original values and the red curves are obtained using the estimated coefficients values. While the estimators of the estimated coefficient yield a slightly smoother estimator, they still present a discrepancy compared to the kicks estimators.

The standard model for describing the effects of covariates on a binary variable in a probabilistic context—namely, Logistic regression.

The Logistic regression model, similar to our perception, is based on the assumption that the dependent variable is a dichotomous result of a probabilistic experiment. Specifically, when handling a binary dependent variable, the model handles the values of the dependent variable as if they were outputs of Bernoulli experiments. Accordingly, the regression aims to describe the probability of success of the dependent variable as a function of the independent variables, or mathematically

\[ P(Y = 1|X), \]

where \( Y \) is the dependent variable and \( X \) represents a vector of independent variables. The conditioned probability can be described as a function of \( X \), \(( p(x) = P(Y = 1|X = x) )\) and thus the regression actually aims to characterize this function. An implicit assumption, on the basis of this model, is a general correlation (positive or negative) between the values of the independent variables and the success probability of the dependent variable. This assumption supports our hypothesis of a positive correlation between the model...
### Table 1: Single-covariate logistic regressions with different model outputs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Chi statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straight path with $W_1$</td>
<td>78.51</td>
<td>$4.0316e-19$</td>
</tr>
<tr>
<td>Straight path with $W_2$</td>
<td>77.61</td>
<td>$6.3386e-19$</td>
</tr>
<tr>
<td>Solid angle</td>
<td>53.85</td>
<td>$1.1013e-13$</td>
</tr>
<tr>
<td>Transformed straight path with $W_1$</td>
<td>157.31</td>
<td>$2.1992e-36$</td>
</tr>
<tr>
<td>Transformed straight path with $W_2$</td>
<td>157.46</td>
<td>$2.0445e-36$</td>
</tr>
<tr>
<td>Transformed solid angle</td>
<td>105.58</td>
<td>$4.5965e-25$</td>
</tr>
</tbody>
</table>

outputs and the scoring probability. Model-wise, this assumption is reflected by the selection of the logit function for this model. Mathematically, the $X - Y$ relation is modeled according to

$$ p(x) = P(Y = 1 | X = x) = \frac{e^{B^T X}}{1 + e^{B^T X}}, $$

where $B$ is the vector of coefficients. The use of this function ensures the assumed monotony as well as a probabilistic output range (i.e. the range of the logit function is $[0, 1]$).

Eventually, the output of an estimation based on this model is the vector of estimated coefficients. The values of the estimated coefficients reflect the magnitude of the effects of the different covariates, and enables a quantitative analysis. In fact, using the logistic regression framework allows us to compare models (different sets of covariates) and examine the significance of effects, rather than just provide estimated probability values. Respectively, we will first use the logistic regression to quantify the relation between the outputs of the model and the scoring probability and then compare them. Next, we will examine the effects of other factors on the scoring probability aiming to obtain the best model. Naturally, we are also interested in comparing this model to previous models that were discussed in the literature review section.

As a first step, we tested the significance of each of the model outputs and their transformations as single covariates to explain scoring; that is, six different models—each containing a single covariate—were examined. The results of these tests (Table 1) determine that each of the outputs and the transformed outputs are significant in explaining the scoring probability when tested separately.

Even though the major differences between the statistics of some of the outputs may suggest an advantage of one output over the other, this requires a deeper examination and comparison. A primary comparison of several different outputs can be done by evaluating a model that consists of all of them and testing the significance of each of them. We first performed this comparison for couples of outputs and their log-
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>-1.9245</td>
<td>2.9389e-53</td>
</tr>
<tr>
<td>Model output</td>
<td>-0.0031</td>
<td>0.8684</td>
</tr>
<tr>
<td>Transformed model output</td>
<td>1.0629</td>
<td>2.3571e-18</td>
</tr>
</tbody>
</table>

**Straight path with $W_1$**

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>-1.3532</td>
<td>1.1743e-21</td>
</tr>
<tr>
<td>Model output</td>
<td>-0.0083</td>
<td>0.79</td>
</tr>
<tr>
<td>Transformed model output</td>
<td>1.1046</td>
<td>5.6472e-19</td>
</tr>
</tbody>
</table>

**Straight path with $W_2$**

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>0.7124</td>
<td>0.0344</td>
</tr>
<tr>
<td>Model output</td>
<td>-0.1066</td>
<td>0.8025</td>
</tr>
<tr>
<td>Transformed model output</td>
<td>0.9056</td>
<td>2.2556e-14</td>
</tr>
</tbody>
</table>

**Solid angle**

| Table 2: Comparing the model outputs to their transformations. |

The next reasonable step involves a comparison between the transformed outputs of the different models. We begin with a model that contains all the transformed outputs. Though this model was found significant (p-value<0.0001) all of its coefficients were found insignificant (Table 3), that is, none of the covariates is significant in the presence of the other covariates. These results lead us to perform a similar comparison for couples (Table 4). Though all the models were found significant (p-value<0.0001), the results reveal that $X_3$ is insignificant in the presence of each of the other outputs while in these models $X_1$ and $X_2$ are significant; this implies that the transformed output of the straight-path model may explain the scoring probability better than the transformed output of the solid-angle model. The test that examined $X_1$ and $X_2$ together found that both of the covariates are insignificant in the presence of the other. This means that this test cannot determine whether one of these outputs is better than the other.
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>-1.4953</td>
<td>0.1420</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.3028</td>
<td>0.8867</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.7632</td>
<td>0.7371</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.0141</td>
<td>0.9454</td>
</tr>
</tbody>
</table>

Table 3: Comparing all the transformed model outputs in a single model

<table>
<thead>
<tr>
<th>Significance of/In the presence of</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>-</td>
<td>0.8995</td>
<td>1.6850e-09</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.6672</td>
<td>-</td>
<td>8.5145e-10</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.7984</td>
<td>0.9975</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4: Comparing all the couples of the transformed model outputs

Considering the last result, another kind of test is required in order to figure out if there is a significant difference in the way they explain the scoring probability. Such a test can be done by an incremental examination. An incremental examination is done by comparing the statistical deviances of two models that satisfy a containment property—the set of covariates of one model is contained in the set of covariates of the second model. The deviance difference follows a Chi distribution, and its number of degrees of freedom is determined according to the difference between the size of the set of covariates of the two models. In our case, we compare the model that contains the two outputs ($X_1$ and $X_2$) to each of the models that contain only one of them. The results of these tests (Table 5) did not find an incremental significance of one of the outputs over the other; this means that one cannot conclude whether any of these outputs is better than the other.

According to the previous discussed comparisons, it may be implied that the solid-angle model is incompetent relative to the straight-path model. This is reasonable considering the fact that the solid-angle model does not discriminate between kicks and headers (in terms of their weight functions). Understanding that, before we exclude the solid angle output from our further calculations, we wish to examine the transformed

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>Chi statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>506.2628</td>
<td>0.1617</td>
<td>0.6876</td>
</tr>
<tr>
<td>$X_2$</td>
<td>506.1178</td>
<td>0.0167</td>
<td>0.8972</td>
</tr>
<tr>
<td>$X_1, X_2$</td>
<td>506.1011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Incremental examinations of the two straight-path model outputs
output of the solid angle together with a binary variable indicating whether an attempt is a kick or a header. The model that contains $X_3$ and the shot-type variable was found significant (p-value<0.0001) with significant coefficients (all p-values<0.0001); furthermore an incremental test reinforced these tests determining that the model that additionally contains the shot-type variable adds to the explanation of the scoring probability (p-value<0.0001). In order to be able to compare this whole model to the other outputs we define a new variable which is the linear combination of the variables according to the coefficients—namely, its the inner product of the vector of coefficients and the variables. We denote this new variable as $\tilde{X}_3$.

The new variable was compared to $X_1$ and $X_2$ using the same comparison methods. The results (Table 6) suggest that there are no significant differences between the new variable $X_1$ and $X_2$ in explaining the scoring probability, and in fact, we cannot conclude whether one of these variables is better than the others. In order to be sure that the lack of discrimination between shot types has indeed created the primary differences between the model outputs, we tested models with the shot-type variable additional to $X_1$ and $X_2$. In both cases, the shot type variable was insignificant (p-value=0.9507 and p-value=0.9157 respectively).

The next step after comparing and analyzing the differences between our main covariates is testing the effects of the other factors. Since we have not found significant differences between the main covariates, we postulate that the other factors affect them similarly; thus we chose only one of the main covariates to

| Table 6: Comparing the revised solid-angle transformed output to the straight path transformed outputs. |

<table>
<thead>
<tr>
<th>Significance of/In the presence of</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td></td>
<td>0.8995</td>
<td>0.2634</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.6672</td>
<td></td>
<td>0.2919</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.3898</td>
<td>0.5737</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incremental test for $X_1$ and $X_3$ models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$X_1$, $X_3$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incremental test for $X_2$ and $X_3$ models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$X_2$, $X_3$</td>
</tr>
</tbody>
</table>
Each value represents the significance of the respective factor within the model that contains the factor and $X_2$.

participate in the factors examination. We selected the transformed output of the straight-path model with $W_2$ weight function ($X_2$), since this output in some way contains most of the knowledge we possess (considering the physical analysis). The primary examination of the other factors was done using models that contain $X_2$ and a single additional factor—one model for each factor. The time factor (indicates the official game time in which the attempt was made) was also tested as a categorical variable, dividing the time line of a match to six distinct periods of fifteen minutes. These tests have found that only the free kick factor has a significant effect (Table 7). An incremental examination for this factor has also determined its significance (p-value=0.0006). We note that all the tested models were found significant (p-value<0.0001) and that $X_2$ was found significant within all of them (p-value<0.0001).

Despite these results, we were afraid that the significance of the free kick factor is mainly affected by the inclusion of penalties; this matter required a deeper investigation. Penalty kicks differ greatly from other scoring attempts exhibiting a scoring proportion that varies between 75% and 80% [18, 23], relative to the 10% average considering all attempts. Our data consist of sixty-three free-kick attempts of which twelve are penalty kicks; among the penalty kicks nine ended in a goal, which yields a consistent 75% scoring proportion. The other fifty-one free-kick scoring attempts resulted in five goals which yields an approximately 10% scoring proportion which is consistent with the general scoring proportion. Considering that, we wanted to distinguish between penalty kicks and free kicks from outside of the box in order to analyze whether a finer definition of free kicks adds to our model, and improves its ability to explain the scoring probability.

A primary examination for this purpose was done by excluding all the penalty kicks from the data, and rechecking the effect of free kicks. The examined model (containing $X_2$ and the free kick factor) was found significant (p-value<0.0001) and so were the free kick coefficient and the incremental examination (p-value=0.0313, p-value=0.0431 respectively); this means that the effect of free kicks is in fact significant even

<table>
<thead>
<tr>
<th>Factor</th>
<th>Time</th>
<th>Discrete time</th>
<th>Half</th>
<th>Side</th>
<th>Free kick</th>
<th>Contact</th>
<th>Home</th>
<th>Round</th>
<th>Player</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.2873</td>
<td>0.3569</td>
<td>0.5212</td>
<td>0.6009</td>
<td>0.0029</td>
<td>0.9225</td>
<td>0.6694</td>
<td>0.6634</td>
<td>0.8601</td>
</tr>
</tbody>
</table>

Table 7: Factor effects examination.
when the penalty kicks are excluded. The next step was to separate the free-kick factor into two variables (with respect to all the attempts), one indicating whether the attempt was a penalty and the other indicating whether the attempt was a non-penalty free kick. In fact, this allows us to examine the effects of three factors “Free kick” (denoted as $Y_1$), “Free kick outside the box” (denoted as $Y_2$) and “Penalty kick” (denoted as $Y_3$). We first examined each of them in addition to $X_2$ and discovered that rather than the significance of all the models (p-value<0.0001), all of them were significant as a part of the combined model as well as in the incremental examination.(Table 8). Another examination found the combination of both $Y_2$ and $Y_3$ significant in the presence of $X_2$ (p-value=0.0312 and p-value=0.0065 respectively), where the model itself was significant (p-value<0.0001) as well as the $Y_2, Y_3$ increment (p-value=0.0019). Since the new variables were set up in a way that their summation is actually the original free kick variable (i.e. $Y_1 = Y_2 + Y_3$), examining all the variables together will reflect this relation and will not add any information. Instead, we wish to examine whether the separated variables add information over the original free-kick variable. This kind of examination is done by testing the models that contain each of the separated variables in addition to the original free kick variable. Both $Y_2$ and $Y_3$ were found insignificant in the presence of the original free kick variable (p-value=0.3702 in both cases). Following all the discussed results we choose to stick with the original free-kick variable, and we shall use it from now on.

Commonly, after evaluating the effects of each factor separately one may consider performing a stepwise regression in order to find a significant combination of covariates to explain the dependent variable. Having only the free-kick factor as a significant covariate in the presence of $X_2$ we decide to first check whether any of the other covariates is significant in the presence of both $X_2$ and the free-kick covariate. Since none of the other factors was found significant in this constellation (Table 9), we stop the search for a better model aiming to explain scoring probability with additional factors. We noted that all of these tested models were found significant (p-value<0.0001) and so were $X_2$ (p-value<0.0001) and $Y_1$ (p-value<0.001) within them.

Following all of these analyses, we chose the model with the transformed output of the straight-path

<table>
<thead>
<tr>
<th></th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value in combined model</td>
<td>0.0029</td>
<td>0.0323</td>
<td>0.0066</td>
</tr>
<tr>
<td>p-value of Incremental exam</td>
<td>0.0006</td>
<td>0.0442</td>
<td>0.0031</td>
</tr>
</tbody>
</table>

Table 8: Free-kick factor examinations.
model with the $W_2$ weight function, and the free-kick factor as covariates, to be our final model. Now we are able to calculate the specific probability of each attempt, and to interpret the effect of each of the covariates according to their coefficients. Mathematically, the selection of this specific model translates to

$$P(Y = 1 | X = x) = \frac{e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}}{1 + e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}}, \quad (36)$$

where, according to the estimations, $\beta_0 = -1.4871, \beta_1 = 1.1053, \beta_2 = 1.3828$. These coefficients can be interpreted in terms of scoring odds. Odds is defined as the ratio between the probability of an event to occur and its probability not to occur, namely $P(Y = 1) / P(Y = 1)$; since we discuss a conditional probability, we can consider the odds as a function of the conditioned variables, that is

$$\text{odds}(x) = \frac{P(Y = 1 | X = x)}{1 - P(Y = 1 | X = x)}, \quad (37)$$

where $X = x$ represents a certain combination of our covariates. In fact, if we apply (36) into (37) we get

$$\text{odds}(X_2, Y_1) = \frac{e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}}{1 + e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}} = \frac{e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}}{1 + e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}}, \quad (38)$$

which allows us to evaluate the marginal effect of each of the covariates. The marginal effect of a covariate is measured by the change of odds achieved as a result of a one-unit increase of one unit of a single variable. For instance, if we increase the value of $X_2$ by one unit, the respective change of odds is achieved through

$$\frac{\text{odds}(X_2 + 1, Y_2)}{\text{odds}(X_2, Y_2)} = \frac{e^{\beta_0 + \beta_1 (X_2 + 1) + \beta_2 Y_1}}{e^{\beta_0 + \beta_1 X_2 + \beta_2 Y_1}} = e^{\beta_1}. \quad (39)$$

So practically the marginal effect of the covariates is reflected via their coefficients. Going back to our specific coefficients we get that $\beta_2 = 1.3828$ means marginal effect of $e^{1.3828} = 3.99$, that is the odds for scoring are approximately four times higher if an attempt, weighted with the same model output, was made within a free kick. Similarly, the marginal effect of the transformed output of the straight-path model, using $W_2$ weight function, is $e^{1.3828} = 3.02$. Since the transformation of the model output is non-uniform—that is,
### Table 10: Assessment of the shot-type effect via a model with the solid-angle transformed output and the free kick factor.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>$\beta_0 = 2.6666$</td>
<td>$3.3387e-09$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$\beta_1 = 1.4319$</td>
<td>$1.9713e-26$</td>
</tr>
<tr>
<td>Free kick</td>
<td>$\beta_2 = 1.3151$</td>
<td>$5.6715e-04$</td>
</tr>
<tr>
<td>Shot type</td>
<td>$\beta_3 = -2.3440$</td>
<td>$2.6559e-09$</td>
</tr>
</tbody>
</table>

The magnitude of the transformation changes with respect to the transformed value—we cannot evaluate the marginal effect of the original model output generally, but only ad hoc, with respect to a specific examined value.

Another marginal effect we are interested in investigating, is the type of shot effect. The differences between the shot types in the straight-path model are expressed via the use of different weight functions. The complexity of the calculations of the straight-path model makes it very difficult to assess the difference in outputs, as a result of the shot-type effect exclusively. In order to overcome this limitation, we turn to the solid-angle model. As discussed, the transformed output of the solid-angle model performed similarly to the other transformed outputs, when tested with the shot-type factor. This type of model allows us to evaluate the marginal effect of the shot type directly from the values of the coefficients (since it consists of a separate coefficient for the shot-type factor). Thus, we fit a model with the transformed output of the solid-angle model, the free-kick factor and the shot-type factor; the model was found significant and so were its coefficients (Table 10). Analyzing the coefficients of this model, in comparison to the coefficients of the previously-selected model, reveals a strong resemblance between the free-kick factor coefficients (and thus between their marginal effects), which according to the discussed similarity between the model outputs is expected. The value of the shot-type coefficient ($\beta_3 = -2.344$) determines a marginal effect of $e^{-2.344} = 0.0959$; since the shot-type covariate equals one when the attempt is a header this means an approximately ten times higher odds for kicks over headers with the same model outputs. Separating this marginal effect from the transformed output of the model we get a higher coefficient for the latter, resulting in a $e^{1.4319} = 4.2$ marginal effect.

One final examination we wish to perform, is comparing our model to models of previous works. In the literature review chapter, we discussed some works in which a logistic regression was applied. Each of
the models described in those works is slightly different from the others, but they all share the covariates—distance and angle. Some of the previous models also consider the defenders in the range of the shot and the presence of the goalkeeper. Due to different notational systems and differences in collected data (in terms of additional factors) we choose to compare our model to a model which is a bit different from all of these models but reflect their main ideas and concepts. Namely, we use our data to extract the following factors:

- Distance - the length of the line that connects the location of the shot to the center of the goalmouth (Figure 1).
- Angle - the absolute value of the angle that is formed between the line that connects the location of the shot and the center of the goal line, and a line perpendicular to the goal line (Figure 1).
- No. defenders - the number of players within the scoring space pyramid.
- Goalkeeper presence - a binary variable indicating whether or not the goalkeeper is within the scoring space pyramid.
- Shot type - a binary variable indicating whether the attempt was a kick or a header.

These factors, though not interpreted exactly as in previous works, reflect the separated factors approach which is common among all the previous works discussed. A model, containing all of these factors, was tested and was found significant (p-value<0.0001) and so were each of the factors within the model (Table 11). The comparison to our model was first done by testing all of these factors together with \( X_2 \). This combined model was found significant (p-value<0.0001) and moreover \( X_2 \) was found significant together with the distance, the angle and the goalkeeper presence factors (Table 12). Considering these results we decided to first examine each of the factors in the presence of \( X_2 \); though all of these tested models were found significant (p-value<0.0001), none of the factors was found significant in the presence of \( X_2 \) (Table 13). This, once again, implies that a stepwise regression would have ended in a single covariate model containing only \( X_2 \).

Since we actually want to compare models, we wish to handle these factors as a whole, therefore, we create a new variable which is a linear combination of the factors, according to the coefficients that were found when the factors were tested together (Table 11). Comparing this new variable (denoted as \( Z \)) to
Table 11: Separated-factors model.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Intercept</th>
<th>Distance</th>
<th>Angle</th>
<th>No. defenders</th>
<th>Gk presence</th>
<th>Shot type</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>3.6776</td>
<td>-0.2037</td>
<td>-1.6409</td>
<td>-0.2572</td>
<td>-1.5969</td>
<td>-1.4888</td>
</tr>
</tbody>
</table>

Table 12: Combining the transformed straight-path output with the separated factors.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Intercept</th>
<th>X2</th>
<th>Distance</th>
<th>Angle</th>
<th>No. defenders</th>
<th>Gk presence</th>
<th>Shot type</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>1.3308</td>
<td>0.6363</td>
<td>-0.0861</td>
<td>-0.8599</td>
<td>-0.1544</td>
<td>-1.1082</td>
<td>-0.6204</td>
</tr>
</tbody>
</table>

Our model is done using a model that contains them both; this comparison yields a significant model (p-value<0.0001) in which both X2 and Z are significant (p-value=0.0006 and p-value=0.0029 respectively). We further examined this result using two incremental tests comparing the model with the two covariates to each of the models that contains only one of them; both of the tests found the incremental addition significant (Table 14). The meaning of these results, is that these two covariates are different and explains the scoring probability in an independent way, as opposed to models that contained two different model-based covariates. Here the covariates do not compete; instead their combination adds to the explanation of the dependent variable.

Following that, we chose this combined model over the other models; we thus needed to recheck the other factors in the presence of the new model. In all these tests the models were found significant (p-values<0.0001) and so were X2 (p-values<0.0008) and Z (p-values<0.003), but once again only the free-kick factor was found significant (Table 15). A further examination of each of the other factors in the presence of X2, Z and the free-kick factor did not reveal any significant factor (p-values>0.3). An incremental examination was also done to test the significance of the free-kick factor with respect to the X2, Z model, and it was indeed found significant (p-value=0.0002). Hence, our final model consists of the transformed output of the straight-path model with the W2 weight function, the linear combination of the separated factors and

Table 13: Separated factors in the presence of the transformed straight-path output.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Distance</th>
<th>Angle</th>
<th>No. defenders</th>
<th>Gk presence</th>
<th>Shot type</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.3326</td>
<td>0.5679</td>
<td>0.3386</td>
<td>0.1392</td>
<td>0.9157</td>
</tr>
</tbody>
</table>
Table 14: Incremental examinations of the transformed straight-path model and the separated-factors model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>Chi statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₂</td>
<td>506.1178</td>
<td>8.7496</td>
<td>0.0031</td>
</tr>
<tr>
<td>Z</td>
<td>512.3762</td>
<td>15.0079</td>
<td>1.0706e-04</td>
</tr>
<tr>
<td>X₂, Z</td>
<td>497.3682</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15: Factor-effects examination with respect to the combined model.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Time</th>
<th>Discrete time</th>
<th>Half</th>
<th>Side</th>
<th>Free kick</th>
<th>Contact</th>
<th>Home</th>
<th>Round</th>
<th>Player</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-value</td>
<td>0.3476</td>
<td>0.4336</td>
<td>0.6384</td>
<td>0.7116</td>
<td>0.0001</td>
<td>0.9143</td>
<td>0.5390</td>
<td>0.5544</td>
<td>0.9555</td>
</tr>
</tbody>
</table>

the free-kick factor, which can be mathematically presented as

\[ P(Y = 1|X = x) = \frac{e^{B^TX}}{1 + e^{B^TX}} = \frac{e^{\beta_0 + \beta_1 X_2 + \beta_2 Z + \beta_3 Y_1}}{1 + e^{\beta_0 + \beta_1 X_2 + \beta_2 Z + \beta_3 Y_1}}, \tag{40} \]

where \( \beta_0 = 1.3457, \beta_1 = 0.5636, \beta_2 = 0.5476, \beta_3 = 1.7138 \) and all of them are significant (p-value=0.0115, p-value=0.0014, p-value=0.0009 and p-value=0.0001 respectively). Following our previous explanations this means marginal effects of \( e^{0.5636} = 1.76 \) for \( X_2 \), \( e^{0.5476} = 1.73 \) for \( Z \) and \( e^{1.7138} = 5.55 \) for the free kick factor.

Using this selected model, we can now calculate the scoring probability for each of our collected scoring attempts. This enables us to assess the distribution of the scoring probability of scoring attempts (Figure 58). The distribution of the scoring probability suggests that most attempts are made in situations where the probability of scoring is low (around 0.1) while only a few attempts (reflected by the small tail of the distribution) are made at a high scoring probability. This observation raises the question as to why players make these low probability attempts—a question that requires further investigation. However, we do discuss implementable aspects of the scoring-probability estimation in the following chapter.

7 Applications and future work

The presented research concerns the estimation of scoring probability in soccer. It involves a detailed analysis of some of the major factors that affect the scoring probability, and rather than just suggesting an estimator, it also lays the foundation for a more precise estimation of the scoring probability, as well as other soccer-related measures. Throughout the research we have covered various topics; some of them may be further investigated in future research. Such potential topics may include:
Figure 58: Scoring probability histogram and smoothed distribution.

Both charts reflect the distribution of the scoring probability among the sample of attempts.

- **Physical analysis** - this subject can be further investigated simply by adding a detailed analysis for headers. Another natural way to deepen this analysis is by applying more accurate models and supporting them with actual motion data of a soccer ball. This may assist in setting more accurate parameters and thus, in reducing the involved uncertainty.

- **Data bases** - as mentioned, this research was validated using manually-collected data, while there are data sources that may enable deeper, more precise conclusions (due to their accuracy and volume). Using digitally-collected data may also allow evaluation and analysis of additional factors, such as the space that the attempting player has, or the actions made prior to the shot.

- **Statistical analysis** - though the selected methods and models provided the desired results, other statistical tools may be applied and reveal additional insights and effects.

- **Second-order factors** - naturally, the described models do not consider all the elements that may affect the scoring probability; however, adopting our suggested approach can derive secondary factors as the goalkeeper’s range of sight (which was crucial in Neymar’s goal versus Croatia in the 2014 FIFA
However, even after it is further investigated, the estimation of the scoring probability is only the tip of the iceberg as far as a scientific analysis of soccer is concerned. Nevertheless, this research provides opportunities to affect the game, both by implementing its core concepts for investigating other aspects of the game (such as the probability of a successful pass), and by direct applications derived from the ability to measure the scoring probability. In this chapter, we present two examples of such applications, highlighting their benefits to soccer coaches and players.

### 7.1 Improving scoring abilities

One of the major advantages in estimating the scoring probability via our models, is the ability to perform sensitivity analysis; that is, analyzing the outcomes of minor changes of inputs and their effect on the scoring space and the scoring probability. In fact, perceiving the scoring probability as an essential soccer measure suggests a continuous monitoring of it, regardless of the actual execution of a scoring attempt; that is, we can answer the question “what is the scoring probability at a certain moment, had the player decided to execute an attempt?”. This continuous evaluation over time enables an assessment of the decision making process of soccer players, both in terms of analyzing the best decision for the examined situation, and for evaluating the actual decision made by the player. Performing such analyses empirically requires a much richer database than ours. However, such questions can be discussed in a more theoretical fashion.

Obviously, a soccer move analysis requires additional measures rather than the scoring probability alone. Since at this point we can only evaluate the latter, we explore a situation in which the scoring probability plays the main role; a one-on-one situation qualifies as such. The title “one-on-one” refers to a situation in which a player possesses the ball past the last defence line of the other team, so that he is facing only the goalkeeper as an obstacle in the way to scoring. One-on-one situations are not uncommon in soccer games, and though many of the elements that characterize them may vary greatly, their main course of events is usually the same. Though actions like passing the ball to another teammate, or trying to dribble with the ball past the goalkeeper, may be possible, and sometimes even a better choice, an analysis of a one-on-one situation can be viewed, from the player’s side, as looking for the best moment to shoot. Indeed, this is a continuous game between the player and the goalkeeper, in which the player aims to increase his probability
Figure 59: Scoring probability in a hypothetical one-on-one situation.

Each point on the curve represents the scoring probability, as a function of the distance between the player and the goalkeeper.

to score and vice versa. In practice, the player and the goalkeeper try to achieve these objectives by closing the distance between them, and choosing the time of action (shooting in the case of the player vs. attacking the ball or jumping in the case of the goalkeeper). Measuring the scoring probability continuously during such a situation can provide an insightful assessment of the decisions to be made.

The value of analyzing a one-on-one situation (in terms of the scoring probability) can be well demonstrated in a simplified problem setup. Consider a one-on-one situation in which the player achieves ball possession twenty-five meters from the goal, in front of the center of the goal line ($\phi = 0$), and dribbles on a straight line towards the center of the goal line, while the goalkeeper is stationed five meters from the goal line. We estimate the scoring probability over a discretization of the twenty meters range, covered by the player until he reaches the goalkeeper’s location. Plotting the scoring probability against the distance between the player and the goalkeeper (Figure 59) yields a continuous curve that exhibits a maximum point.

The maximum point suggests that there is an optimal shooting point (or moment), to which the attempting player should aim. In fact, following the main concepts of our models, we realize that there are two opposing factors that affect the scoring probability over the covered range. On the one hand, the player is approaching the goal and thus increasing his scoring space; on the other hand, the distance from the goalkeeper reduces
constantly, which increases the scoring space that is eliminated by the goalkeeper. The scoring probability curve actually reflects the tradeoff between these two factors. Considering the trends of the curve with respect to the distance between the players, we conclude that, at the beginning of the move, the increased scoring space achieved by approaching the goal is more significant than the space eliminated by the goalkeeper; this lasts up to the maximum point, after which the significance of these two factors switches.

The tradeoff, that we have revealed in a theoretical experiment, does occur in the reality of one-on-one situations—as an example we have analyzed an actual situation played by Maccabi Haifa’s Shimon Abuhazira which can be found at One-on-One with Analytics. Accordingly, it is reasonable to train players to seek the maximum point in a general way, which will allow them to make a better decision regardless of the specific setup of the one-on-one situation. Actually, awareness of the tradeoff and its associated factors depends on a player’s cognitive abilities. Such abilities are trainable, and can in fact be trained via a computer software (https://www.intelligym.com/). The understanding of this process, and the factors that affect it, provide a sound basis for designing such a software-based trainer for this particular purpose.

7.2 Player evaluation

The scoring probability, while being a mathematical estimator, reflects the difficulty of scoring; that is, a high scoring probability characterizes easy-to-score attempts and vice versa. Assigning attempts with probability-based weights provides additional information and improves the analysis of the scoring situations. Analyzing sets of scoring attempts related to a certain player or a team, will therefore yield insights regarding their abilities.

The common way of evaluating a player’s scoring abilities (rather than a general impression), is based on two main parameters: the number of goals scored by the player, and the number of goals per attempt made by the player. These measures do reflect some aspects of the scoring ability, but considering the rarity of attempts and goals, they provide merely a partial view. Adding the probability-based weights to scoring attempts can provide information regarding the quality of attempts made by a certain player or a team, and improves the assessment of their scoring capability.

The first aspect of the additional information provided by the scoring probability, helps in characterizing the quality of the attempts made by a player (or a team). A natural way to assess this quality is by the
distribution of the scoring probability of a player’s attempts. This distribution reflects the quality of a player’s
decision making, as manifested through the scoring probability of his attempts. A high density of attempts
with high scoring probability could imply effective decision making, but may also reveal good positioning
ability; on the other hand, it may also be interpreted as hesitation. Once again, more information is needed in
order to understand the complete picture (for example, situations in which the player decided not to shoot),
but a soccer coach definitely gains information about his players from such an analysis, specifically when
comparing the distributions across players.

The other aspect of a player assessment, as revealed by the scoring probability, reflects how well a
player converts a scoring opportunity into goals. Having a method of estimation (that does not depend on
a specific player) of the scoring probability, allows a comparison of the performance of a player against
expected performance. A simple way to measure a player’s conversion rate is by comparing the number
of goals he scored to the expected number of goals through his attempts. Taking the view that scoring
attempts are independent Bernoulli experiments, the expected number of goals is simply the sum of the
scoring probabilities over all of these attempts. Dividing the number of scored goals by the expected number
places the scoring ability of a player on a scale relative to the average; such a measure can highlight players
who are very accurate and score goals from difficult situations, as well as players who tend to miss even
from promising situations. This approach naturally does not tell the whole story, but it does provide some
significant insights.

The above two aspects provide a framework for analyzing players’ scoring abilities, and can easily be
refined through additional variables (such as differentiating kicks from headers to evaluate these two types of
abilities separately), or by seeking patterns within subsets of attempts (like estimating the scoring probability
distribution of attempts which were scored). In order to demonstrate the types of outputs and conclusions
that can be drawn from such analyses, we now use our data of Arsenal players to perform a case study.
Each attempt in our data contains the identity of the attempting player; Arsenal players are identified by
their official number and players of their opponents are all identified with the number zero. Since we have
data from only a single season, we have restricted the scope of the case study to players who have attempted
more than fifteen times and scored five or more goals. This yielded a set of six Arsenal players: Arteta
(identified by the number 8), Podolski (identified by the number 9), Giroud (identified by the number 12),
Table 16: Comparison of common scoring-ability measures for selected Arsenal players and an average fictitious player.

The number in brackets is the player’s official number or, in the case of the fictitious player, the identification number.

Walcott (identified by the number 14), Cazorla (identified by the number 19) and Gervinho (identified by the number 27). When considering all the attempts and goals of the opponent players together they also satisfy the case study prerequisite. Accordingly, we have analyzed an additional, fictive, player which represents the opponent players, and performs as a control group to which the Arsenal players can be compared; in other words, the additional player reflects the performance of all the opponent players who attempted scoring against Arsenal, and considered as the average player.

The initial comparison was done with respect to the common measures of scoring abilities: number of attempts, number of goals and goals per attempt (Table16). A quick analysis of the results suggests that all the tested Arsenal players perform better than the average other player, in terms of the conversion of attempts to goals. Rather than that, it seems like Arteta has the best scoring abilities in terms of conversion, and though attempting the lowest among the list of players, he scored a respectable number of goals. Furthermore, when considering all of the columns of Table 16, different types of players are revealed.

We now execute the analysis by calculating the expected number of goals for each player and then the ratio between the goals scored and the expected number of goals. These two measures give a different perspective on the examined players (Table17). The expected number of goals gives an overall idea regarding the quality of the attempts made by the players, whereas the ratio between the scored and expected goals gives a normalized comparison to the conversion-of-attempts-to-goals ability. The most significant insights, revealed by applying these measures to our set of players, are in terms of the conversion ability. While prior to this analysis, it seemed that all the tested players perform better than the average player; the current
Table 17: Comparison of additional scoring-ability measures for selected Arsenal players and an average fictitious player.

The number in brackets is the player’s official number or, in the case of the fictitious player, the identification number.

analysis suggests that Arteta and Giroud’s conversion abilities are lower than those of an average player. This difference is significant in the case of Arteta, who according to the goals-per-attempts measure, had the best performance, while after the extended analysis he is second to the worst.

While it may seem reasonable to assess the quality of attempts, made by a player, via calculating the average scoring probability of his attempts (which is equivalent to dividing his expected number of goals by his number of attempts), we now address this issue in a refined manner—namely, we estimate the distribution of the scoring probabilities of the players based on his actual attempts. In fact, this estimated distribution, rather than just the calculation of its average, provides a more refined view of all the attempts, allowing the identification of trends and phenomena. Examining the scoring-probability distribution of the players within our case study (Figure 60), we easily identify a general distribution pattern (similar to the one revealed in Figure 58), in which most of the attempts are of low probability and a small tail characterizes the attempts of high probability.

Though adhering to the common pattern, the distribution of the scoring probability of Arteta exhibits a major peak around the probability of 0.8. This peak actually completes the picture regarding this player, who is, in fact, the penalty taker of Arsenal: it explains the discrepancy between his performance according to the naive goals-per-attempt measure vs. the goals-per-expected-goals measure; the high value of his goals-per-attempt is caused by the high scoring probability of penalties, an aspect that is considered in the goals-per-expected-goals measure. Further analysis regarding Arteta could distinguish between scoring-ability measures for regular attempts and penalties separately.
The dashed blue curve represents the distribution of the average fictitious player while the other curves correspond to the 6 chosen Arsenal players.

Leaving out Arteta’s distribution, there are some differences among the other players, though not significant and without any immediate conclusions. Assessing the distributions of the other players (excluding Arteta) through their survival functions (Figure 61) adds a different perspective, and allows a better comparison of players. In order to understand the meaning of the survival function, we let $X_i$ denote a random variable indicating the scoring probability of an attempt made by the $i^{th}$ player; thus, $X_i$ is distributed according to the respective player’s scoring-probability distribution, as characterized in Figure 60. The survival function is defined as usual:

$$S_{X_i}(x) = P(X_i \geq x),$$

(41)

where $x \in \mathbb{R}$, but in our case can be limited to $x \in [0, 1]$. Translating this mathematical expression back to the context of our discussion, it practically means: what is the probability that the $i^{th}$ player will make an attempt with a scoring probability higher than $x$. Therefore, if a survival function of a certain player dominates a survival function of a different player—that is, at each point the first function has higher values than the second (stochastic order)—it can be concluded that the first player’s attempts enjoy higher probabilities than those of the second. Most of the survival functions in our case study do not achieve a complete domination among them, but we do notice that Giroud, Gervinho and Podolski’s functions dominate the average player’s
Figure 61: Scoring-probability survival functions of different players.

The dashed blue curve represents the survival function of the average player while the other curves display the survival function of the 6 chosen Arsenal players.

function at almost all points, while Cazorla and Walcott’s functions are dominated by the average player’s function for medium-high values of scoring probability.

This type of analysis does not aim at determining whether one player is better than the other; alternatively, it suggests a decomposition of scoring abilities according to different measurable aspects. For instance, comparing Podolski, Giroud and Walcott using all the aspects and information, one can deduce that:

- **Giroud** - a player that enjoys many attempts of relatively high scoring probability. However, his conversion rate of attempts to goals is lower than the average. In other words, it seems that he is good in creating scoring situations, but his accuracy needs improvement.

- **Podolski** - a conservative and efficient player in terms of selecting his attempts; this is demonstrated by a relatively low number of attempts with high scoring probability and a high rate of goals scored per expected goals.

- **Walcott** - a player that yields many attempts of low scoring probability but has a great conversion rate. In other words, in order to score he needs only a slight opportunity, and it seems that correspondingly, a slight opportunity is enough for him to actually attempt.
These three players exhibit different profiles with different advantages and drawbacks. As mentioned, these profiles do not determine who is better, but they allow a coach to understand his players’ qualities so that he is able to train them properly and to fit them better to positions and game strategies.

7.3 Discussion

The scoring probability in soccer, rather than just an answer to the question “what is the likelihood to score a goal?”, is an essential measure that assists in analyzing game plays and players. Obviously, the scoring probability alone does not provide all or most of the information that is needed to understand, measure and analyze the game, but it is most certainly one of the foundations that would support these objectives. Accordingly, our research, in addition to providing an estimator of the scoring probability, also sets guidelines and offers methods and tools for further research of additional measures, such as the passing probability (the probability that an attempted pass will succeed) or dribbling probability (the probability that a dribble between two points will be completed).

Estimation of such measures can naturally rely on the principles and methods of our research. The general approach of a mathematical model that quantifies the relevant factors and their correlations will provide a useful framework for such estimations, which could be then implemented following our statistical approach. Such a process is not easy to implement and much further research is needed. Nevertheless, we believe that our research establishes the feasibility of such an implementation based on the value in the measures described above.

Identifying and validating the basic measures that characterize the main actions in soccer will provide the basis for a more complex, more event-oriented analysis. On top of analyzing each of these measures separately—which definitely adds information, as demonstrated via the analyses based on the scoring probability—combining them together will allow a broader perspective on events, game plays and planned moves. For instance, let us focus on a player who possesses the ball at a certain moment. One can describe his options compactly as: shooting, passing or dribbling. Having the proper tools for analyzing the relevant possibilities would enable the analysis of this possession, especially in terms of decision making. Considering such a possession as an event may be the basis for systematic description and analysis of planned moves and game plays, which can lead in turn to assessments of formations and tactics.
Ultimately, the objective of analyzing soccer scientifically is to serve the needs of managers, coaches and other decision makers, by providing them with quantitative answers to their questions. Such an objective requires the final conclusions to be easily communicated and understood. This requires an understanding of the perspective of those who make the calls, and naturally an applicable research must accommodate inputs from these people. Our conceptual structure, in which relevant measures serve as building blocks for more complex measures that, in turn, will become higher level building blocks, is definitely a natural constructive process towards the goal.

A Geometric and Trigonometric calculations

In this section, we describe in detail the geometric calculations that served us throughout the research. These are simple and standard calculations based on straightforward geometric and trigonometric concepts. The main issue discussed here is the transformation of the geometric input from a Cartesian coordinate system to the Polar coordinate system and its actual application within our models.

As described in this paper, both of our models (the straight-path model and the solid-angle model) are calculated with the origin of a shot and the goalmouth taken into consideration. In fact, since the solid-angle model can be considered as a straight-path model with a constant weight function \( W(r) = 1 \), both of the models are actually calculated as an integration of the straight paths (with the respective weight function) over the scoring space pyramid. Considering the modeled factors and their “real world” context (specifically distance and angle) and the way we define the weight functions, it is natural to perform the integrations using the polar coordinate system. Since transforming from a Cartesian coordinate system to the Polar coordinate system is common and standard, we focus on the specific calculations of the integration limits.

Consider a shot taken from a certain point on the field with no obstacles in the range of the shot (Figure 62). Following our definitions, the origin of the shot is set at \((0,0,0)\) (also denoted as point \(A\)) while the center of the goal line is set at \((x,y,z)\) (also denoted as point \(B\)). Let \((a,y,b)\) be a point within the goalmouth, or alternatively be the end point of a straight path formed by this attempt (we also denote this point as point \(C\)). The projection of this straight path on the \(X-Y\) plane is the line that connects \((0,0,0)\) and \((a,y,0)\) (also denoted as point \(D\)). One last item of interest is the point \((0,y,0)\), which we denote as point \(E\).

According to these notations and our basic definitions (as appear in Section 4.1.1) we note that:
Figure 62: Geometric example of a shot.

The blue line \((AC)\) is the straight path. The dashed lines \((AD\) and \(AE)\) are its projections. Point \(A\) is the origin of the shot and point \(B\) is the center of the goal line.

- The angle \(\angle DAE\) is actually the angle \(\varphi\) (Figure 63).
- The angle \(\angle CAD\) is actually the angle \(\frac{\pi}{2} - \theta\) (Figure 64).
- The line \(AC\) is actually \(\vec{r}\) and respectively \(|AC| = |r|\).

Since \(\angle AED = \frac{\pi}{2}\), the following two trigonometric relations

\[
\tan(\varphi) = \frac{DE}{AE}
\]

and

\[
\cos(\varphi) = \frac{AE}{AD}
\]

apply. Accordingly, by simple manipulations of (42) and (43), we get

\[
|AD| = \frac{y}{\cos(\varphi)}
\]

and

\[
\tan(\varphi) = \frac{a}{y} \Rightarrow \varphi = \arctan\left(\frac{a}{y}\right).
\]
Figure 63: Geometric example of a shot - taking the X-Y plane view

Figure 64: Geometric example of a shot: Y-Z plane view
Similarly, since $\angle ADC = \frac{\pi}{2}$, the trigonometric relations

$$\tan\left(\frac{\pi}{2} - \theta\right) = \frac{DC}{AD}$$

and

$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin(\theta) = \frac{AD}{AC}$$

hold true. Accordingly, by the same manipulations over (46) and (47) and application of (45) we get that

$$|r| = |AC| = \frac{y}{\cos(\psi)} = \frac{y}{\cos(\psi) \cdot \sin(\theta)}$$

and

$$\tan\left(\frac{\pi}{2} - \theta\right) = \cot(\theta) = \frac{b}{y \cdot \cos(\psi)} = \frac{b \cdot \cos(\psi)}{y},$$

which implies that

$$\theta = \arccot \left( \frac{b \cdot \cos(\psi)}{y} \right).$$

These equations relate the Cartesian coordinates and the Polar coordinates of a single straight path given an attempt. In order to evaluate the total scoring space of an attempt, integration over all of the straight paths is required, which can be achieved by bounding the limits of the integration parameters. Following all the above calculations, it is clear that the relevant parameters to be bound are the angles $\psi$ and $\theta$ (specifically, $r$ as expressed in terms of these angles, and $y$ which is a given value when considering an attempt). Setting the limits of these parameters is done simply by applying the limits of the $a$ and $b$ parameters to (45) and (50) respectively.

Considering our parametrization, the feasible range for the parameter $a$ is

$$\left[x - \frac{7.32}{2}, x + \frac{7.32}{2}\right],$$

since 7.32 meters is the width of the goalmouth and $x$ is the midpoint parameter. Thus, the range of the angle $\psi$ is given by

$$\left[\arctan \left( x - \frac{7.32}{2} \right), \arctan \left( x + \frac{7.32}{2} \right)\right].$$

Similarly, and since the total height of the goalmouth is 2.44 meters, we get that the range of the parameter $b$ is

$$[z, z + 2.44],$$
where in this case \( y \) is used as one of the boundaries. Once again we transform this range to terms of the angle \( \theta \), which yields

\[
\left[ \arccot \left( \frac{z + 2.44 \cdot \cos(\varphi)}{y} \right), \arccot \left( \frac{z \cdot \cos(\varphi)}{y} \right) \right].
\] (54)

These ranges appear in (2) and (4) and are applied throughout the thesis. Since, according to (48) the size of \( r \) also depends on the angles \( \theta \) and \( \varphi \), setting these ranges also bounds the value of \( r \). Similarly, this type of calculation is also applied when considering obstacles in the range of the shot—since they are handled as rectangular surfaces—and by that we actually cover the whole geometric-calculation aspect of our research.

References


