uc3m | Universidad Carlos III de Madrid

Master Degree in...
Academic Year (e.g., 2018-2019)

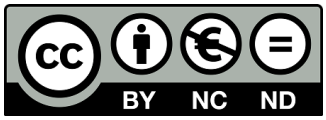
Master Thesis

Data Analytics in Football: Pitch Control Models and Beyond

Alvaro Novillo Correas

1st Tutor complete name
2nd Tutor complete name
Place and date

AVOID PLAGIARISM
The University uses the **Turnitin Feedback Studio** program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.



CONTENTS

DEDICATION

LIST OF FIGURES

1. INTRODUCTION

The digital revolution is currently one of the most significant challenges of our time, altering numerous aspects of society. Football, in particular, has also been influenced by this transformation. Technological advancements and digitalization have resulted in a swift upsurge in the number of measuring devices, data collection and volumes of data. The leading data companies worldwide, including IBM, Intel, SAP and Microsoft, are vying for superior data analytics tools and leveraging sports as an example domain to showcase their products and brand power [Footballytics, 2021].

The practice of data analytics in football has a long history, dating back to the post-World War II era, when data collection and analysis was undertaken manually using pencil and paper [Footballytics, 2021]. It wasn't until Moneyball was published in 2003 that significant progress began to emerge: The book, "The Art of Winning an Unfair Game" introduced sports analytics to a broader audience. It illustrated the use of data analytics in identifying undervalued players and constructing a successful team. Since then, data analytics has become an integral component of sport, football inclusive [Footballytics, 2021].

One of the best examples of data analytics being applied to sports is basketball. Teams use data to analyze player performance, identify strengths and weaknesses, and develop strategies to win games [Sarlis and Tjortjis, 2020]. They use in-memory analytics, visualization, the cloud, mobility, camera footage, and sensors to transform their game. This performance analyses are of vital importance to a team, aiming to reduce expenditure, enhance team worth and refine processes across all levels and segments of operations. The German Football Association (DFB) and the National Basketball Association (NBA) are two unique cases of digital transformation from the sports world. Successful teams turn player performance data into action and gain a competitive advantage.

Over the last years, football analytics has gained significant popularity, aiming to delve deeper into the game by utilizing advanced data analysis techniques to optimize team and player performance.

The main objective of this master's thesis is to enhance understanding and performance in football through the use of Data Analytics. The master's thesis includes a literature review of the field, alongside the commonly found data types within this industry and the main metrics used to analyse player and team performance, focusing on tracking data and Pitch Control models [Spearman, 2018]. Following the initial review of the field's state of the art, we propose a new methodology for quantifying the effectiveness of the offside strategy of teams and players using Pitch Control models. Our study defines a new performance parameter, called *Offside Control*, which quantifies the amount of threat posed by the attacking team or player beyond the offside line.

In the study, we will compute both effective and ineffective Offside Control at a rate of 2 frames per second for 100 matches from LaLiga 2019-2020, resulting in a total of 1,251,934 frames analysed. This will allow us to characterize successfully the Offside Control of 442 players in total.

Our proposed methodology aims to contribute to this growing body of knowledge. Analyzing vast amounts of tracking data from LaLiga matches, we seek to uncover patterns in player and team behavior, shedding light on the tactical nuances that underlie successful offside strategies.

All data processing and modeling in this project has been made using python. You can find the source code of the project inside the GitHub repository[1]

[1] All the source code is inside the code folder at https://github.com/AlvaroNovillo/master_thesis.git

2. DATA ANALYTICS IN FOOTBALL

When discussing sports analytics in football, the first metric that often springs to mind is the Expected Goals (xG) ratio. This statistical indicator is a predictive Machine Learning (ML) model used to assess the likelihood of scoring for every shot made in the game. In the context of each shot, the xG model computes the scoring probability, leveraging a set of event parameters.

Wyscout xG model, for example, encompass the shot's spatial coordinates, the assisting player's position, the striking player's use of foot or head, the type of assist involved, the occurrence of a dribble by either a field player or the goalkeeper immediately preceding the shot, whether the shot arises from a set piece, whether it transpires during a counterattack or in a transitional phase of play, and the subjective assessment of shot danger as determined by a designated tagger. The amalgamation of these parameters serves as the foundation for training the xG model using historical Wyscout data, culminating in the prediction of the likelihood of a given shot resulting in a goal [Wyscout, 2023].

The probabilities range from 0 to 1. Thus, a shot with an xG value of 0.1 has a 10% chance of being scored. Penalties have a fixed xG value of 0.76.

Fig. 2.1 provides a visual representation of the cumulative development of expected goals (xG) during the Eibar - Malaga match, which took place on January 15th, 2023 in Spain's second division. Each data point on the graph corresponds to a shot made by both teams over the course of the game, offering a comprehensive overview of the evolving scoring opportunities and outcomes throughout the duration of the game.

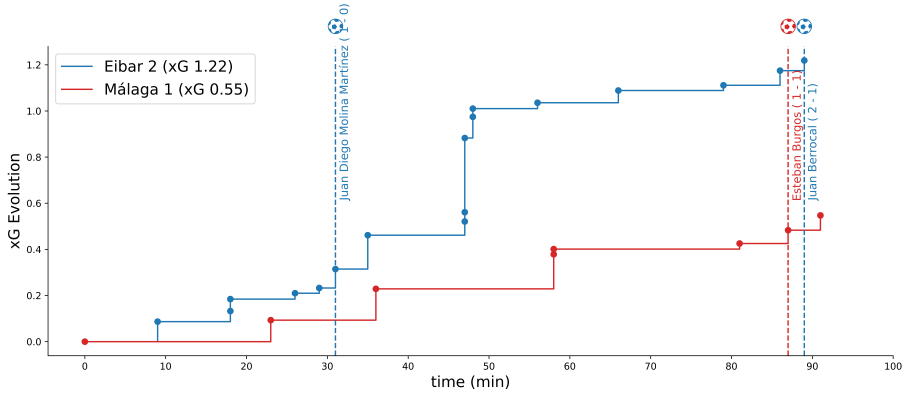


Fig. 2.1. Cumulative development of expected goals (xG) during the Eibar-Malaga match, held on January 15th in Spain's second division. Each point denotes a shot made by both teams throughout the game. Vertical dashed lines indicate the goal scored, displaying the player and the corresponding score at that specific moment of the match.

Expected Goals (xG) have revolutionized the analysis of football by quantifying the quality of scoring opportunities. However, it is important to consider other variables such as player positioning, velocity, passing accuracy, defensive pressure, and tactical formations to gain a broader understanding of the sport. Using the game presented earlier as an example, the next chapter will introduce the broad field of football analytics.

Looking at an xG evolution figure, such as Fig. 2.1, and solely focusing on shot probabilities while disregarding the spatial distribution of shots and occasions feels like merely scratching the surface of what sport analytics can offer to football.

To illustrate the spatial distribution of shot locations taken by both teams during the game, we can create a shot map for each shot. In Fig. 2.2, the size of each data point corresponds to the expected goals (xG)

generated for the respective shots, providing insights into the perceived scoring potential. Goals scored are visually highlighted with straight lines, indicating the trajectory the ball followed as it found its way into the opponent's net. Bellow each shot map, a plot of the net can be also found, where goals are represented by football balls, and blocked shots by shadowed points. This detailed analysis not only enhances our understanding of scoring opportunities but also sheds light on the tactical strategies employed by both teams, player positioning, and defensive vulnerabilities. Analyses such as the one above are carried out using the most common source of data in football: **Events** datasets.

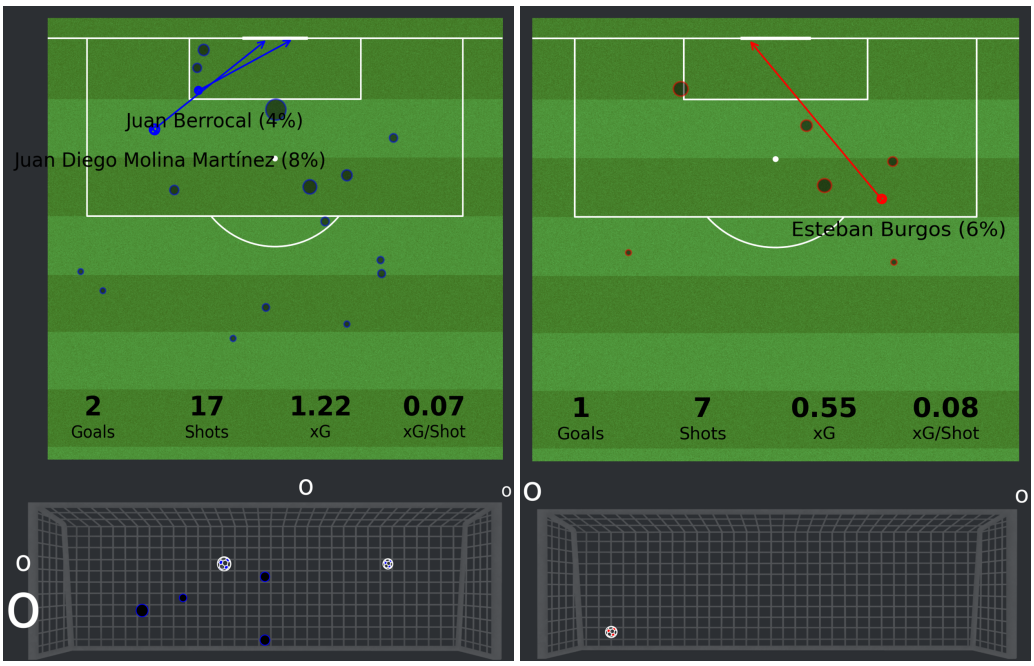


Fig. 2.2. Shot map of the Eibar (blue, left) - Malaga (red, right) football match. The locations of the points indicate where shots were taken. The size of each point is proportional to the expected goals (xG) generated. Shots that resulted in goals are depicted with a straight line, representing the path the ball took to enter the opponent's net.

2.1. Events data

`Event data` describes specific, human-defined events during a match, including passes, shots, and fouls. It is captured by human annotators from various providers. However, this manual process is time-consuming and typically requires three individuals:

The data collection process is carried out by professional video analysts (known as operators), who are specialists in football data collection, using proprietary software (the tagger). The tagger has undergone several years of development and improvement and is regularly updated to ensure the highest level of performance is achieved. To ensure accurate data collection when tagging events in soccer matches, three operators are assigned: one per team and one supervising the output of the entire match. This process is based on analysis of the tagger and soccer match videos. When near-live data delivery is necessary, a team of four operators may be utilized, with one operator dedicated to hastening the collection of complex events that require additional, specific attributes or a quick review [Pappalardo et al., 2019].

This type of data structure can be used in a number of ways: it can be used to measure team performance through general statistics extracted from event datasets, such as goals, fouls, xG, etc. It can also be used to create advanced analysis of the team using ensembles of mathematical tools.

The analysis of the match is furthered through the use of graph theory, [Buldú et al., 2018], [Álvaro Novillo et al., 2024]. Combining different elements of the events dataset, we can create a graph corresponding to the passing network of each team, allowing us to understand the passing structure of both teams.

Figs. 2.3 illustrate the passing networks observed in the Eibar versus Málaga football match, providing insight into the passing interactions and tactical strategies used by both teams. The nodes in the graphs represent individual players who participated in the match for each team. The nodes are sized according to their degree, which represents the amount of ingoing and outgoing passes. The node position corresponds to the average passing position of each player. Substitutes are represented by yellow nodes, and links are created if there have been at least 5 passes made in that direction between two players. The edge's width corresponds to the amount of passes made in that direction between the two players.

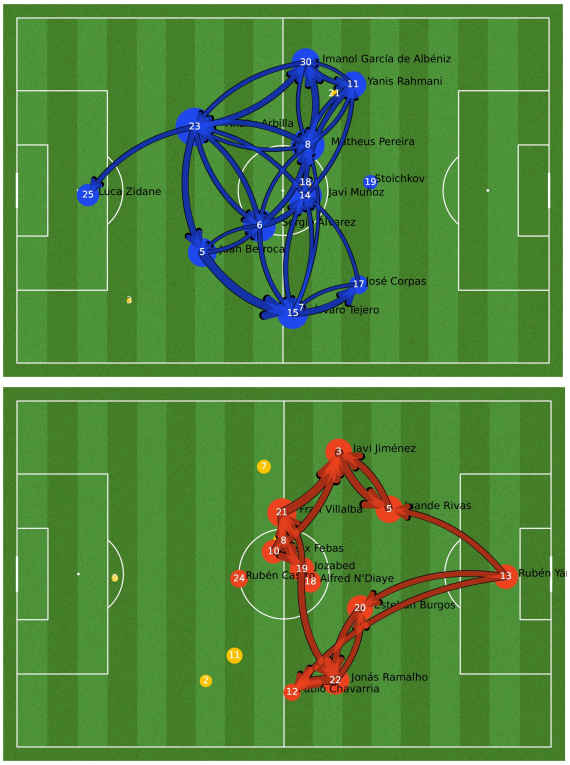


Fig. 2.3. Representation of the Eibar (blue, left), Málaga (right, red) passing networks of the match Eibar - Málaga. Nodes represent players, edges represent passes between players. The position of the players in the field is their average passing position. The size of the nodes reflects the number of ingoing and outgoing passes (i.e. node's degree), while the size of the edges is proportional to the number of passes between the players. Substitutes are represented in yellow. A connection is set if those players share at least 5 passes. The edge's width is proportional to the amount of passes made in that direction between the two players.

Analysis as the former can be conducted *in real-time*[2] during the match using appropriate data sources. Additionally, we could examine Eibar's macro situation during the 2022-2023 season to better comprehend how this micro-statistics contribute to the overall perception of the team.

Fig. 2.4 presents the expected goals (xG) produced by Eibar in all matches played against their opponents. It is noticeable that Eibar has generated a higher xG when playing at their Home stadium, on average. In Fig. 2.5 an overview of Eibar's performance against other teams in the Second Division is presented. It can be observed that Eibar ranks third in generating xGs against their opponents.

[2]Opta uses a combination of human annotation, computer vision, and AI modelling to offer real-time data at various levels of detail based on customer requirements. In our situation, the data feed updates itself when an event such as a goal, foul or pass occurs. Otherwise, it updates every 90 seconds. [StatsPerform, 2023]

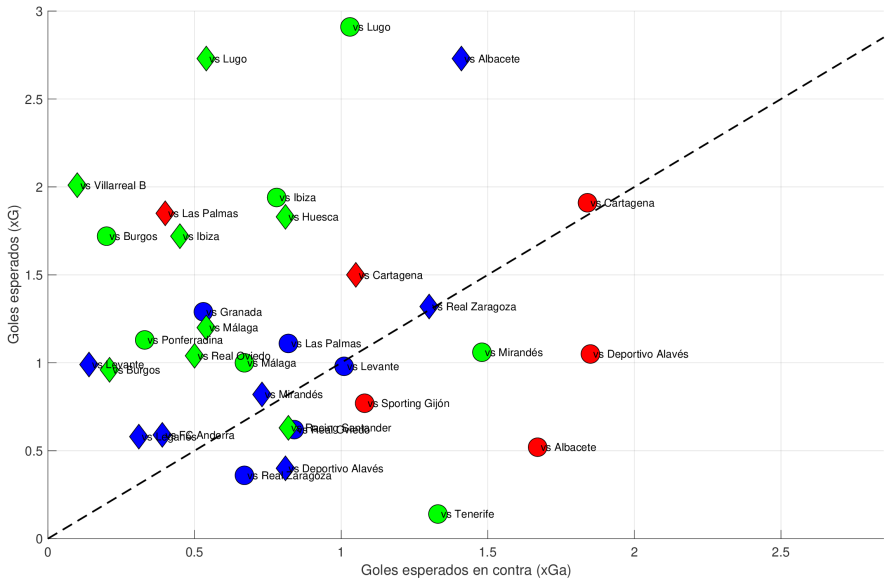


Fig. 2.4. Expected Goals (xG) and Expected Goals Against (xGA) per match. Codes: Home Matches (Diamonds), Away Matches (Circles), Wins (Green), Draws (Blue), Losses (Red). Matches above the dashed lines represent those matches where Eibar has generated more xG than the opponent.

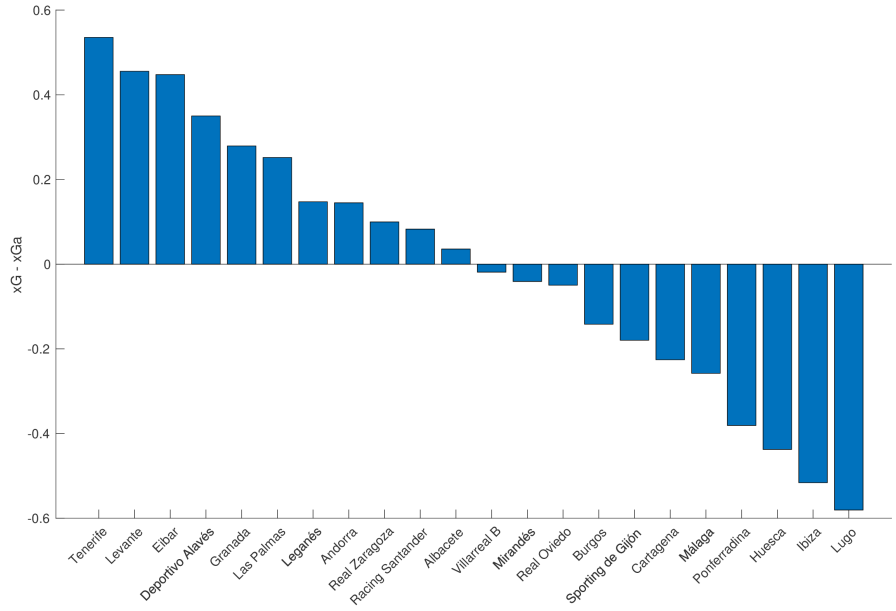


Fig. 2.5. Ranking of the average differences in Expected Goals Scored (xG) Minus Expected Goals Conceded by Opponents (xGa) per team.

We've just discussed some of the many statistics that can be inferred from this data sources to characterize the team performance, such as possession, pressure, duels, fouls, etc. This type of datasets are both easily manageable, and provide teams with useful information. Due to limitations in space and scope, however, we are unable to provide a more in-depth analysis of these measures

Although event datasets supply beneficial information regarding the team's overall performance, deeper scrutiny can be conducted via tracking data, which consists of the players' and ball's position and movement

during the match. Tracking data can offer additional insight into both the physical and tactical aspects of the game.

2.2. Tracking data

`Tracking data` offers a more comprehensive perspective than event data by providing access to information on all players, their trajectories, and velocities. This allows for the analysis of off-ball players and team dynamics, resulting in a more nuanced understanding of the game.

There are two main techniques for obtaining tracking data, which decide its classification: **Image detection algorithms** extract players' positions from the match broadcast and infer locations of concealed players, whereas **optical tracking** employs a specialized camera system installed on the field to record players' data. Our research will concentrate on the latter method, as it offers more precise and statistically informative data.

Our tracking data has been provided by Mediacoach®. They utilises the Tracab Optical Tracking system to obtain on-the-pitch player positions. This multi-camera system captures each player's position at 25 frames per second. The system consists of three units, each with a resolution of 1920x1080 pixels, producing a panoramic picture that generates a stereoscopic view for triangulating the players and ball. In case of a temporary loss of any location, a skilled operator adjusts the players' positions. The datasets obtained by the Mediacoach® system have been validated in advance using GPS [Felipe et al., 2019].

Fig. 2.6 contains a frame from the tracking dataset of an Atlético de Madrid (Blue) - Getafe (Red) game from the Spanish 2019 League. The ball is shown as a black dot. Referees are shown as yellow squares. Purple arrows represent the speed vectors of the players.

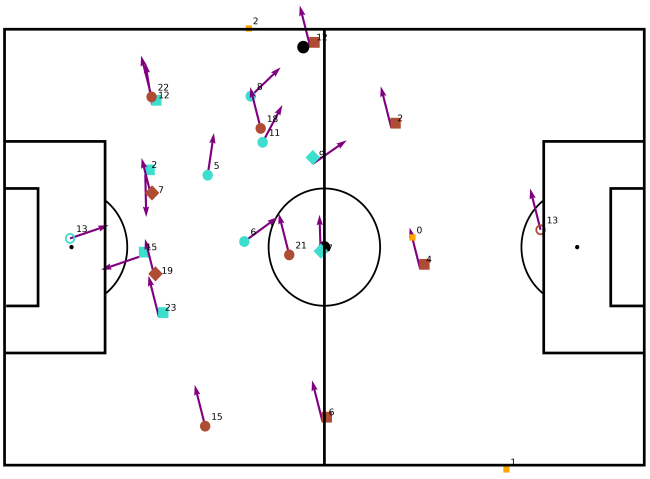


Fig. 2.6. A frame of tracking data from a football match. The home team is shown in blue, the away team in red. The ball is shown as a black dot. Referees are shown as yellow squares. Purple arrows represent the speed vectors of the players.

Traditionally, all football statistics have been produced using event datasets. In this respect, tracking datasets have been crucial in developing new ways of measuring team and player performance. With this type of information, we are not just limited to ball-related events such as passes, goals, etc. tracking datasets contain the precise location of players and the ball during a match. [Garrido et al., 2022] showed that the correlation between heatmaps, Fig. 2.7, made with event datasets and those made with tracking datasets is low. Event

heatmaps show where a player has performed more actions. On the other hand, tracking heatmaps provide information about a player's position without adding information about how relevant his about how relevant their performance was.

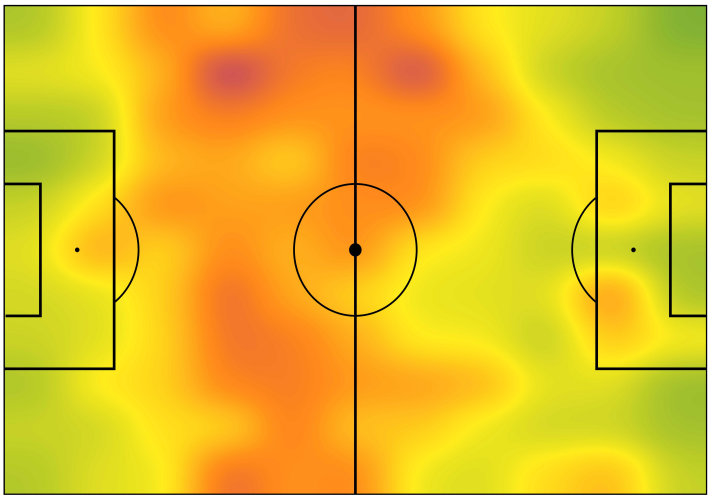


Fig. 2.7. Heatmap of the ball position during the Atlético de Madrid - Getafe game under study. Note that we always keep the direction of play from left to right, so the home team will always be placed on the left side of the field and the away team on the right.

Thus, tracking datasets are crucial for developing new ways of measuring team and player performance, as they provide us with global information about the game.

This data-driven approach enhances the understanding of the sport and its strategic nuances, fostering a deeper appreciation of the game's intricacies. As a result, tracking data has become a valuable tool in football analysis, providing insights into player performance and team strategy. These advanced metrics can be applied to specific games, as in the analysis presented earlier (Fig. 2.7), or to an ensemble of them to provide a comprehensive view of general player behavior under different parameters. In the following section, we will introduce two main frameworks from which our present a selection of these metrics, both physical and tactical, to provide context for our proposed offside control metric across multiple games.

After the initial review of the state of art of football analytics, we will introduce the methodology behind our proposed metric to characterize offside strategy's effectiveness. In this chapter, we will combine both physical metrics and tactical models derived from Tracking datasets to build the proposed metric: *Offside Control*

3. METHODOLOGY

As discussed earlier, tracking datasets provide the position of all players and the ball throughout the match with a temporal resolution of 25fps. This enables us to estimate the players' covered distance, speed, and acceleration. The potential for extracting information from tracking datasets that is useful for football analytics extends beyond variables related to players' physical performance. Many tactical metrics has been implemented to decode how the intricate movements of players translate to the soccer field.

These models provide a scientific perspective for analysing player positioning, decision- making, and team dynamics, illuminating the complex interactions that occur during a match.

Pitch Control is one of the most relevant tactical metrics used to analyse player positioning, decision-making, and team dynamics during a soccer match. It combines player position and speed with mathematical models that simulate ball and player movement [Spearman, 2018].

This master's thesis proposes using Pitch Control Models to evaluate how football teams interact with the offside line when attacking and defending. Before that, we need first to define and understand Pitch Control Models, including their construction and implementation.

3.1. Pitch Control Models

The *Pitch Control (PC)* at a given location represents the probability of a player or team gaining control of the ball if it moves directly to that location. *PC* models simulate the dynamics of the ball and the players to evaluate which player would control the ball if it moves to any location on the pitch at any moment. The model captures not only the players' current position, but also their movement. When players are running at high speeds, they are more likely to control the space they are moving into rather than the space they currently occupy.

To construct this model, we must calculate the following for a given location on the pitch:

- How long it would take for the ball to reach to the position of interest (from its starting position).
- How long would it take for each player to get to that position.
- What is the total *probability* that each team will control the ball *after* both the players and the ball have arrived at the desired position?

In the model, the ball is set to move at a constant speed of $v_b = 54\ km/h$, which is approximately the average speed of the ball in the game (See Fig. 3.1)

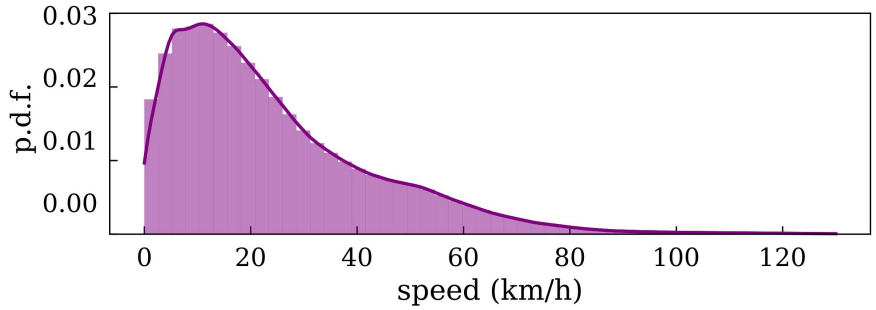


Fig. 3.1. p.d.f of the ball speed over a 100 matches from LaLiga 2019-2020 season.

Therefore, the time taken to arrive at the location of interest can be easily calculated as $t_{b,arr} = \Delta x_b / v_b$, where Δx_b is the distance between the initial and final positions of the ball.

When considering how long it will take the players to reach the target position, given their initial position and speed, players are assumed to only have a maximum speed of $v_{max,p} = 18\ km/h$, which corresponds to the 95 percentile of the average speed of the players in the game (See Fig. 3.2). This upper limit should not be misunderstood as the maximum speed at which players can move, but rather as an estimate of the maximum speed at which they are likely to move when trying to control the ball

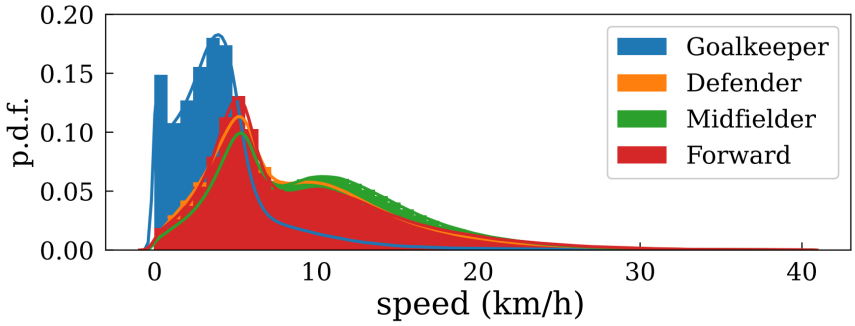


Fig. 3.2. p.d.f of the players speed over a 100 matches from LaLiga 2019-2020 season.

To compute the player's expected arrival time, $\tau_{exp}(\vec{r}; t_r)$, we use a simple approximation consisting of a two-step process:

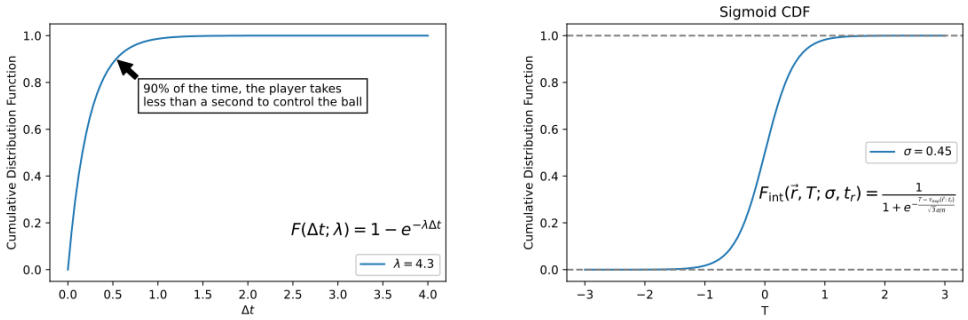
1. There is an initial *reaction time*, assumed to be of $t_r = 0.7$ seconds for every player. This is approximately the time it takes a player moving at maximum speed to come to a complete stop. During this reaction time, we assume that players continue to move along their current trajectory without changing speed or direction (reaching a position $\vec{r}_{react}$).
2. After this time, we assume that the player runs directly towards the ball at his maximum speed of $v_{max,p}$.

$$\tau_{exp}(\vec{r}; t_r) = t_r + \frac{|\vec{r} - \vec{r}_{react}|}{v_{max,p}} \tag{3.1}$$

Once we computed the time it takes for the ball and the players to get to the target location, we need to look at how long it will take each player to control the ball. To do so, we will assume that controlling the ball is a stochastic process that follows an exponential distribution with a fixed rate λ, with units of $1/s$. Thus, for any differential time Δt that a player is near the ball, he has a probability of $\lambda \cdot \Delta t$ of controlling the ball.

So far, the model assumes that we know exactly when each player will arrive at the target location. However, we introduce some uncertainty, labelled σ, in the arrival time of the players. The reason for including such temporal variability in our model is to account for some effects that have not been explicitly modeled, such as player effort. Thus, the probability of a player intercepting the ball at time T is given by the cumulative distribution function of the sigmoid distribution (See Fig. 3.1).

\begin{figure}[H]



{

}

\caption{{Cumulative distribution function of the time to control the ball.} \end{figure}

Both λ and σ has been selected according to [Spearman et al., 2017], where they model passes as a Bernoulli trial, with probability mass function

$$P(k \mid \sigma, \lambda, x) = \begin{cases} 1 - p \text{ for k} = 0 \\ p \qquad\qquad\qquad \text{for k} = 1 \end{cases} \tag{3.2}$$

where $k \in [0, 1]$ is the outcome of the pass. Then, the likelihood of a set of parameters, σ and λ, given outcome k and the start of the pass x is:

$$\mathcal{L}(\sigma, \lambda \mid k, x) = P(k \mid \sigma, \lambda, x) \tag{3.3}$$

Then, maximizing the product of the likelihood for each pass for a training sample P, the best fit is found at $\sigma = 0.45 \pm 0.05$ s and $\lambda = 4.30 \pm 1.14$ s^{-1}[3]

$$\min_{\sigma, \lambda \in \{\mathbb{R}, \mathbb{R}\}} \left\{ - \sum_{i \in P} \log \left[\mathcal{L}\left(\sigma, \lambda \mid k_i, x_i\right) \right] \right\} \tag{3.4}$$

[3] See [Spearman et al., 2017] for further details

BIBLIOGRAPHY

J. M. Buldú, J. Busquets, J. H. Martínez, J. L. Herrera-Diestra, I. Echegoyen, J. Galeano, and J. Luque. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.01900. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01900.

J. L. Felipe, J. García-Unanue, D. Viejo-Romero, A. Navandar, and J. Sánchez-Sánchez. Validation of a video-based performance analysis system (mediacoach®) to analyze the physical demands during matches in laliga. *Sensors (Basel, Switzerland)*, 19, 2019. URL https://api.semanticscholar.org/CorpusID:202746198.

Footballytics. Data analytics in football, 2021. URL https://www.footballytics.ch/post/data-analytics-in-football.

D. Garrido, B. Burriel, R. Resta, R. L. del Campo, and J. M. Buldú. Heatmaps in soccer: Event vs tracking datasets. *Chaos, Solitons & Fractals*, 165:112827, 2022. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2022.112827. URL https://www.sciencedirect.com/science/article/pii/S0960077922010062.

L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6, 10 2019. doi: 10.1038/s41597-019-0247-7.

V. Sarlis and C. Tjortjis. Sports analytics — evaluation of basketball players and team performance. *Information Systems*, 93:101562, 2020. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2020.101562. URL https://www.sciencedirect.com/science/article/pii/S0306437920300557.

W. Spearman. Beyond expected goals. 03 2018.

W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop. Physics-based modeling of pass probabilities in soccer. 03 2017.

StatsPerform. Opta data from stats perform, 2023. URL https://www.statsperform.com/opta/.

Wyscout. Wyscout data glossary, 2023. URL https://dataglossary.wyscout.com.

Álvaro Novillo, B. Gong, J. H. Martínez, R. Resta, R. L. del Campo, and J. M. Buldú. A multilayer network framework for soccer analysis. *Chaos, Solitons & Fractals*, 178:114355, 2024. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2023.114355. URL https://www.sciencedirect.com/science/article/pii/S0960077923012572.