

A New Dataset for Exploring Actions of Italian Football Matches

Mattia Cefis¹, Rodolfo Metulini², Maurizio Carpita¹

¹*Department of Economics and Management, University of Brescia*

²*Department of Economics, University of Bergamo*

Abstract

We introduce an original dataset for football analytics, adopting detailed event and performance data from the Italian Serie A season 2022/2023. The dataset integrates information from multiple sources, including *whoscored.com*, *understat.com*, and *sofifa.com*, through an advanced ETL process. It includes over 8,400 shot actions with variables such as pass details, player roles, pitch coordinates, and performance indices derived via a PLS-SEM approach. Preliminary analyses highlight the potential of the dataset to uncover insights into player performance and team strategies. This resource aims to support advanced research in football analytics, including extensions to the expected goals framework and evaluations of player contributions during actions.

Keywords: Football Analytics, Player Performance, Shot Action Data, Expected Goals

1 Introduction

Football, known as soccer in North America, New Zealand, and Australia, is widely recognized as the most popular sport in the world. According to estimates from the International Federation of Football Association (FIFA), more than 250 million people of all ages and levels, both men and women, play football in more than 200 nations. Given its important appeal, football clubs are increasingly managed like real firms, as a consequence, the decision-making phase is becoming strategic for some crucial areas. For that reason, data-driven approaches are now fundamental to help coaches, staff, scouts, and team managers in their decision-making processes [1]. Football analytics research is focused on several key areas; these include, for example, the study of player performance, employing exploratory methods to create rankings [2; 3], the refinement of advanced metrics such as expected goals [4; 5], and prediction of injury risk using machine learning algorithms [6].

However, a notable challenge is obtaining homogeneous data from a single source, requiring merging data from multiple sources [5; 7]. This work aims to introduce an innovative dataset covering the season 2022/2023 from the Italian Serie A. This dataset integrates event data from all the matches, specifically all the actions leading to shots, and players' performance data. We will provide a detailed explanation of the entire dataset and present

preliminary descriptive statistics; it has the potential to serve as a foundation for significant future analyses and developments.

In the following sections, we describe the ETL process and the merging of data from different sources, along with the variables used and constructed in this work.

2 The dataset creation

We merged data from multiple sources to create our action dataset from the Serie A Season 2022/2023. Specifically, for event data related to actions and passing networks, we employed a web scraping procedure from *whoscored.com*, a renowned football statistics website. Additional event data concerning the context of actions and expected goals was obtained via web scraping from *understat.com*, utilizing the R package `worldfootballR`. Meanwhile, performance data was collected from *sofifa.com*: these data, provided by Electronic Arts (EA, *easports.com*) experts, combine the subjective evaluations of over 9,000 scouts, coaches, and season-ticket holders into ratings for over 18,000 players. In this case, they have been used to create some composite indicators to sum up and measure players' performance starting from the 29 Key Performance Indicators (KPIs) for movement players and 31 KPIs for goalkeepers via a PLS-SEM approach [3]. Since these performance data are constantly updated [8], we computed the last composite indicators available before the match date for each player. The pitch coordinates based on Understat and Whoscored rules were adopted (width as y coordinate and length as x coordinate, each one as a percentage of the pitch, considering the offensive phase from left to right, see Figure 1).

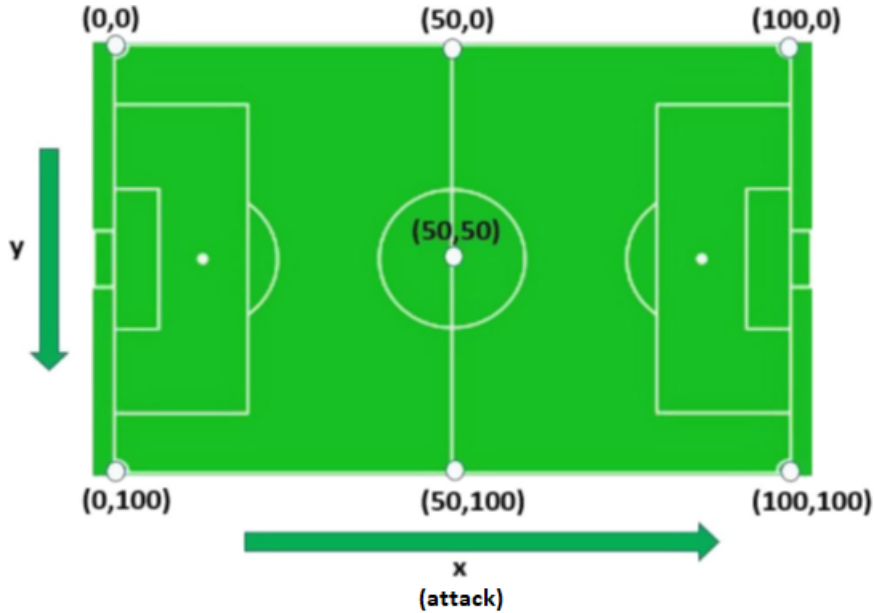


Figure 1: Normalized x and y coordinates on the pitch.

The final dataset consists of 8,421 shot-actions with the following variables:

- Event features from Whoscored: the number of passes (**passNb**) and the number of players involved in each action (**playersNb**); then, the average pass distance, the location (with x and y coordinates) of the first and last pass on the pitch, the players id and the role which are involved in the actions and the timing during the game of each action.
- Event features from Understat: some circumstance features, such as the match name and date, the shooter, the assist-man, and the team names involved in the actions; the shot location on the pitch and its expected goal, the action situation (e.g. open play, free kick, penalty, etc.), the shot type (e.g. right, left, head or other), the final result of each shot (goal or not).
- Performance feature from Sofifa: for each action, **PI_NoShooter** is the variable representing the mean of the offensive performance index of the players involved in the action, excluding the shooter. Note that for the actions with only one player involved, the **PI_NoShooter** corresponds to that player's offensive index.

3 Descriptive statistics related to players involved in actions

Table 1 provides a summary about the features of the dataset. Specifically, Table 1a shows details on the continuous variables, while Table 1b focuses on the categorical variables.

Table 1: Feature descriptions and statistics for the sample of 8,421 shot-actions from the Italian Serie A season 2022/2023 (in bold the "derived" features).

(a) Continuous Features Descriptions and Descriptive Statistics.

Feature	Description	Source Dataset	Mean	St. Dev.	Skewness	Kurtosis
X	X coordinate of the shooter	Understat	85.31	7.39	-0.41	-0.82
Y	Y coordinate of the shooter	Understat	50.84	12.42	0.00	-0.50
xG	Expected Goal	Understat	0.11	0.16	2.72	7.37
first_pass_x	X coordinate of the first action pass	WhoScored	51.15	27.12	0.10	-0.94
first_pass_y	Y coordinate of the first action pass	WhoScored	50.79	29.34	-0.04	-0.93
last_pass_X	X coordinate of the last action pass	WhoScored	85.41	7.29	-0.40	-0.82
last_pass_y	Y coordinate of the last action pass	WhoScored	50.87	12.46	0.00	-0.54
passNb	Number of passes	WhoScored	6.47	5.85	2.00	6.44
playersNb	Number of players	WhoScored	4.87	2.52	0.39	-0.79
avg_pass_distance	Mean passes distance in the action	WhoScored	27.28	10.60	1.39	2.99
PI_NoShooter	Mean players-action offensive performance index	Sofifa	84.81	9.39	-1.36	4.17

(b) Categorical Features Descriptions and Frequencies.

Feature	Categories	Source Dataset	freq(1)	freq(2)	freq(3)	freq(4)
h_a	Home(1) Away(2)	Understat	0.54	0.46	-	-
result	Goal(1) No Goal(2)	Understat	0.10	0.90	-	-
situation	Open Play(1) Free Kick(2) Penalty(3) Others(4)	Understat	0.73	0.09	0.01	0.17
shot type	Right(1) Left(2) Head(3) Others(4)	Understat	0.47	0.33	0.19	0.01

Now, we present a brief exploration of the valuable descriptive insights provided by this new dataset. In Figure 2, a stacked bar chart illustrates the distribution of player roles across the entire dataset based on the number of players involved in each action. For the most frequent actions (involving from 2 to 6 players), midfielders account for the highest percentage. In other cases, defenders dominate, except for actions involving a single player, where, as expected, forwards represent the largest share.

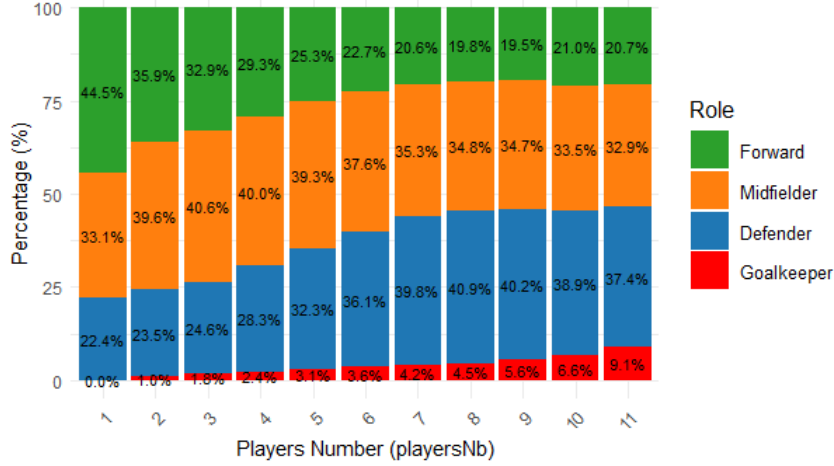


Figure 2: The role percentage distribution across the number of players involved in the shot-actions in the Italian Serie A 2022/2023.

Additionally, we provide a focused analysis (Figure 3) comparing two teams: Juventus and AC Milan. The percentage of players involved by role in shot-actions was examined (513 actions for Juventus and 536 for AC Milan). Juventus (Figure 3a) predominantly involved midfielders in their actions, with nearly 44% of their actions including at least one midfielder. In contrast, AC Milan (Fig. 3b) exhibited a more balanced distribution among the three macro-roles (35%, 32%, and 29%), suggesting a style of play that emphasizes cooperation among roles.

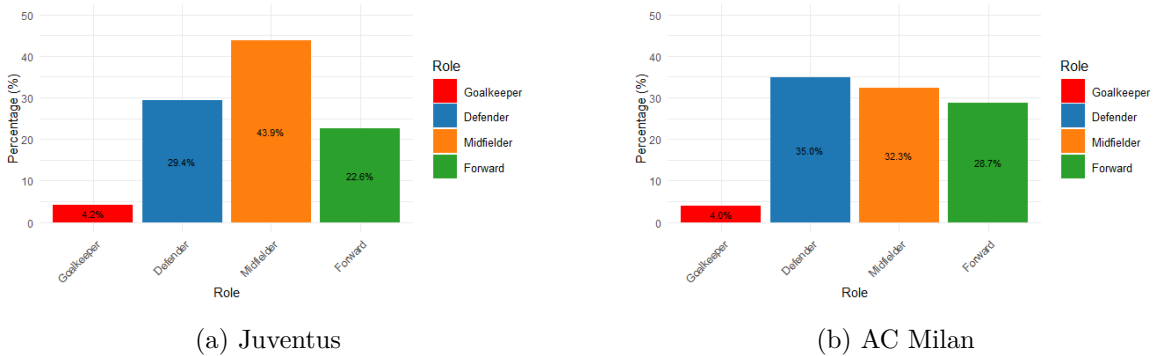


Figure 3: The percentage of players involved by role in shot-actions for two teams of Italian Serie A 2022/2023.

4 Potential dataset applications and final insights

By this work we created an original dataset for football analytics, providing a detailed explanation and presenting preliminary descriptive statistics, and the first idea for future works is to update it in order to obtain data also for other seasons and leagues, or expanded it with additional features. The examples presented above are just a few insights that can be derived using this data set, but many other interesting things can be derived from it.

For example, given the classical concept of expected goal used by media and team staff to measure shot quality, assigning to each shot a real number between 0 and 1 (which represents the probability of resulting in a goal) based on this new dataset, the idea could be to extend the concept to an "expected goal action" framework that measures the contribution of each player involved in an action. The works by Metulini and Gnecco [9] and Metulini and Cefis [10] represent, for basketball and soccer, respectively, a first attempt. However, in this latter study, the network of players involved in the action that leads to a shot is completely ignored. This dataset allows us to fill this gap. For example, Figure 4 illustrates a goal-scoring action from the AC Milan vs Udinese match. In this network graph, yellow arrows represent passes, the dashed arrow indicates a player's run, and the dots are the players on the pitch (including names and players' offensive index -POI-), color-coded by their roles (in this case two forwards and one defender). Using expected goal action as a target function, it would be intriguing to quantify the individual contributions of all players involved in the action.

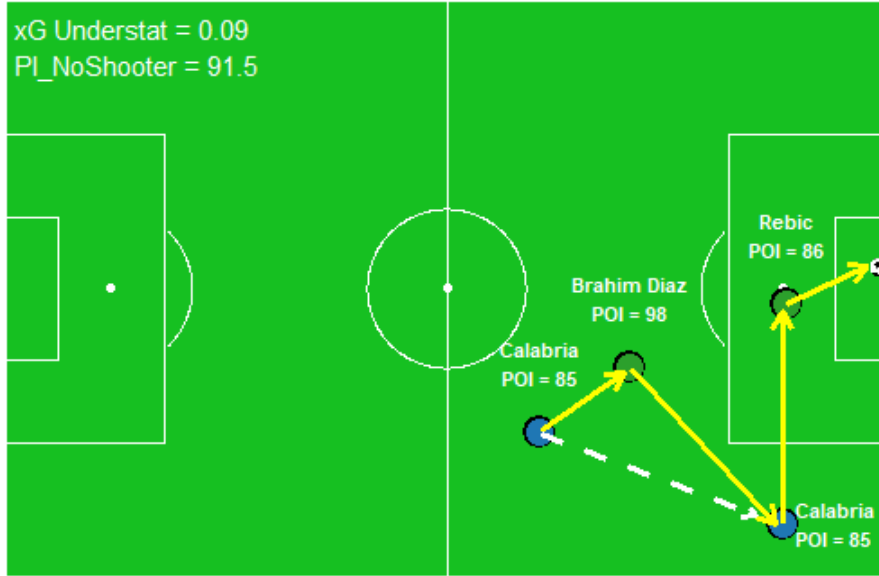


Figure 4: AC Milan vs. Udinese 4-2, Italian Serie A 2022/2023: network plot showing the action leading to a goal for AC Milan made by A. Rebic.

References

- [1] Mattia Cefis. Football analytics: a bibliometric study about the last decade contributions. *Electronic Journal of Applied Statistical Analysis*, 15(1):232–248, 2022.
- [2] Maurizio Carpita, Enrico Ciavolino, and Paola Pasca. Playersâ role-based performance composite indicators of soccer teams: A statistical perspective. *Social Indicators Research*, 156:815–830, 2021.
- [3] Mattia Cefis and Maurizio Carpita. The higher-order pls-sem confirmatory approach for composite indicators of football performance quality. *Computational Statistics*, 39(1):93–116, 2024.
- [4] Mattia Cefis and Maurizio Carpita. A new xg model for football analytics. *Journal of the Operational Research Society*, pages 1–13, 2024.
- [5] Mattia Cefis and Maurizio Carpita. Accuracy and explainability of statistical and machine learning xg models in football. *Statistics*, pages 1–20, 2024.
- [6] Haniyeh Mohammadi, Raheleh Ghaffari, Abdolreza Kazemi, David G Behm, and Mahdi Hosseinzadeh. Evaluation of the lower extremity functional test to predict lower limb injuries in professional male footballers. *Scientific Reports*, 14(1).
- [7] Alexander Scholtes and Oktay Karakuş. Bayes-xg: player and position correction on expected goals (xg) using bayesian hierarchical approach. *Frontiers in Sports and Active Living*, 6:1348983, 2024.
- [8] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory model analysis: explore, explain, and examine predictive models*. Chapman and Hall/CRC, 2021.
- [9] Rodolfo Metulini and Giorgio Gnecco. Measuring players’ importance in basketball using the generalized shapley value. *Annals of Operations Research*, 325(1):441–465, 2023.
- [10] Rodolfo Metulini and Mattia Cefis. Some novelty on the xg model for football analytics. In *EURO 2024 Conference Handbook Abstracts: 33rd European Conference on Operational Research (EURO XXXIII)*, pages 248–249, 2024.