The quest for the right pass: Quantifying player's decision making

Paper Track

Borja Burriel[1] and Javier M. Buldú[2,3]

[1] Barça Innovation Hub, F.C. Barcelona, Spain

[2] Laboratory of Biological Networks, Center for Biomedical Technology, UPM, Madrid, Spain

[3] Complex Systems Group & GISC, Universidad Rey Juan Carlos, Móstoles, Spain

Abstract

We propose a minimal model to quantify the extent to which players make the right decisions when making a pass just by using the *StatsBomb 360º* datasets. Our methodology evaluates the risk of losing the ball when passing and, at the same time, quantifies the gain acquired with that pass using the Expected Possession Value. Next, we define two different Decision Parameters (DPs), which account for the risk assumed and the gain obtained with a pass. The risk and gain DPs are calculated not only from the passes made by each player but also from all possible passes a player could make based on the partners' and opponents' locations and their movement abilities. In this way, we determine what are the passes with (i) the lowest risk and (ii) the highest gain at each moment (and location) of the match and quantify the players' decision making as a function of how close the actual pass was to both optimal options. Using the risk and gain DPs, we can also discriminate those players who prefer to make safer passes at the cost of having less gain and, conversely, those who take more risks to get a better outcome, as well as when and where this occurs.

Introduction

During the last decade, football analysis has been experiencing a rebirth, mainly fostered by the access to new kinds of datasets, but also thanks to the development of new (and more profound) methodological tools [1-3]. Consequently, we are gaining more and more insights about players' physical performance, tactical analysis, or the development of alternative Key Performance Indicators (KPIs), both for teams and players. On the one hand, these advances rely on the ability of capturing more and higher-quality information about what is happening on the pitch during a football match. However, what kind of information is accessible right now? Football datasets can be split into three main groups: (i) basic statistics, (ii) events and (iii) tracking data. Let us see what the main differences between them are.

Basic statistics are the most extended and accessible type of football datasets. They consist of classical football match descriptors such as the number of goals, shots, passes or possession percentage. These datasets are public and can be accessed by any football fan just consulting any media covering the selected match.
Event datasets are the next level. They consist of every single action performed by a player during a match, containing crucial information such as the corresponding temporal label, the Euclidean position, the players involved, or the outcome of the action. In this case, this information is not public but can be acquired from a diversity of providers such as StatsBomb [4], StatsPerform [5] or Wyscout [6], to name a few. Event datasets allow going one step beyond in the analysis of player and team performance. For example, it is possible to quantify the quality of a shot by using event-based parameters such as the xG (*expected Goals*), which uses large event databases to assess the probability of scoring by considering the position of the shot, the distance, the angle to the goal, and, in some advanced models, the location of the defenders [7]. At the player level, we can use event datasets to evaluate the impact of players' actions using, for example, the xT (*expected Thread*) [8] or the VAEP [9]. The xT quantifies the probability that a given action of the team with the possession of the ball will finish in a shot using Markov chains [8]. In contrast, VAEP (acronym of *Valuing Actions by Estimating Probabilities*) evaluates the short-term probability that any action (defensive or offensive) will end by scoring or conceding a goal [9]. At the team's level, we can use event datasets to construct passing networks using the number and location of passes between every pair of players, which give helpful information about (i) a team's organization during the offensive phase [10] or (ii) the way the ball is moved between regions of the pitch to identify the specific passing patterns of a team [11].

Let us move to the next level: tracking datasets. They contain the precise location of all players (even the referees) and the ball at every single moment of the match, with a spatial resolution up to a few centimetres and a typical temporal resolution of 40 milliseconds. This information takes us to another level since we can use it to analyze not only the player's position, speed, and acceleration but also any movement of players who do not have the ball, which is the strongest limitation of event datasets. It is precisely the development of off-ball metrics that have benefited the most from the use of tracking datasets. For example, pitch control models [12-15] describe the area of the pitch that each team is controlling at any moment of the match. Furthermore, we can use the alignment of the player's velocities (extracted from the tracking datasets) to evaluate the coordination of movements between teammates and opponents at different match phases [16]. Going back to the concept of player networks, which has been widely used in the framework of event datasets, it is also possible to define "tracking networks" where (i) the coordination between players or (ii) the marking carried out by the defenders lead to different kind of networks, helping to understand the tactical performance of football teams [17].

Up to now, the differences between these three kinds of datasets (statistics, event and tracking datasets) were clear. However, a new kind of dataset has been recently introduced: StatsBomb 360º. This dataset basically consists of the combination of the classical event datasets with additional tracking information. Specifically, every event recorded during the match comes with the Euclidean position of all players that appeared on camera when the action was performed, indicating the team a player belongs to and whether he/she is a goalkeeper. This "hybrid" dataset allows one step beyond the classical analysis of event datasets since we now have off-ball information. In this way, the door is open to developing new advanced metrics that require the knowledge of the player position.

In this paper, we propose to use StatsBomb 360º datasets to get some insights into the decision making of football players when making a pass. The risk assumed during a pass or, conversely, the probability that a teammate controls a pass have already been studied in the literature. In [13], the authors quantified the probability of intercepting a pass using tracking datasets. Their model analyzed the ball and player's tentative trajectories and described what passes were prone to be completed/intercepted. Furthermore, they assigned a pass value according to the proximity to the opponent's goal and suggested that this methodology could be adapted to evaluate player decision making when passing[1]. More recently, a similar approach based on the probability of interception was used to analyze the type of passes made by a team (penetrative, supportive, or backward passes), their location and the moment of the match when they are more frequent [18].

Importantly, these two previous methodologies analyzing the interception probability of a pass were based on tracking datasets. With this regard, we also developed our own methodology, in collaboration with the Sports Research Area of *LaLiga,* using a combination of event and tracking datasets, which will be published elsewhere. This methodology allows us to determine whether a player selects the most suitable pass at every moment of the match. Here, we investigate whether it would be possible to construct a (simpler) model to evaluate passes only using the StatsBomb 360º datasets. Our proposal is just a starting point since, as we will explain, it has several limitations.

Under this framework, we propose to quantify to what extent players make the right decisions when making a pass. Our model evaluates the risk of losing the ball when making a pass and, at the same time, quantifies the gain acquired with that pass using the Expected Possession Value

[1] Reference [13] captures the spirit of this paper: describing players and ball movements using dynamical models and tracking datasets. Here, the main difference is that we are restricted to use StatsBomb 360º datasets instead of full tracking.

(EPV). Next, we define two different Decision Parameters (DPs) that account for the risk assumed when making a pass and the gain obtained. The value of a pass is calculated not only for the passes made by each player but also for all possible passes a player could make based on the partners' and opponents' locations and their movement abilities. This allows determining the optimal pass at each moment of the match and quantifying the players' decisions. Using the DPs, we can also discriminate those players who prefer to make safer passes at the cost of having less gain and, conversely, those who take more risks to get a better outcome. Finally, averaging across regions of the pitch or the moment of the match allows us to determine where and when decision making is better.

Materials and Methods

1.1 Datasets

The datasets we use to illustrate the design of our model for evaluating player's decision making include 37 matches of the first division of the Spanish national league "*LaLiga".* Specifically, they are the matches[2] played by F.C. Barcelona during the season 2020/2021. Datasets have been supplied by StatsBomb and consist of contextualized events, which basically differ from the classical event datasets in the fact that, for every event, we also have the Euclidean position of all players around the ball or, more precisely, all players that appear on the camera broadcasting the match. Note that the origin of the data also comes with a substantial limitation (which will be discussed later): it may happen that the location of some players is missing. From the event datasets, we also know the name of the players making the pass and receiving the ball.

As mentioned in the Introduction, we are concerned about how players pass the ball; therefore, only the event "pass" has been introduced in the model. From all passes, we only consider those made during regular play, excluding passes from throw-ins, corners, or fouls. Furthermore, we will restrict the analysis to completed passes. The reason is that, to evaluate the performance of a pass, we need to know the origin and end of the pass, the latter not being guaranteed in the case of incomplete passes. This is another limitation to the model. Finally, only players with more than *N=100* passes are considered in the analysis. Under these assumptions, we analyzed *8822* passes made by 20 players of F.C. Barcelona.

Figure 1 shows an example of the information we use from every single pass: the starting and ending coordinates of the pass, which are indicated by a green arrow in the figure, the name of the players involved (sender and receiver) and the location of players surrounding the action. In this

[2] The datasets, supplied by StatsBomb, are all matches played by F.C. Barcelona except for F.C. Barcelona - Cádiz, whose 360º datasets were not available, i.e., 37 matches.

example, only 10 players of the team with the possession (red circles) are depicted, together with 10 opponents (blue circles).

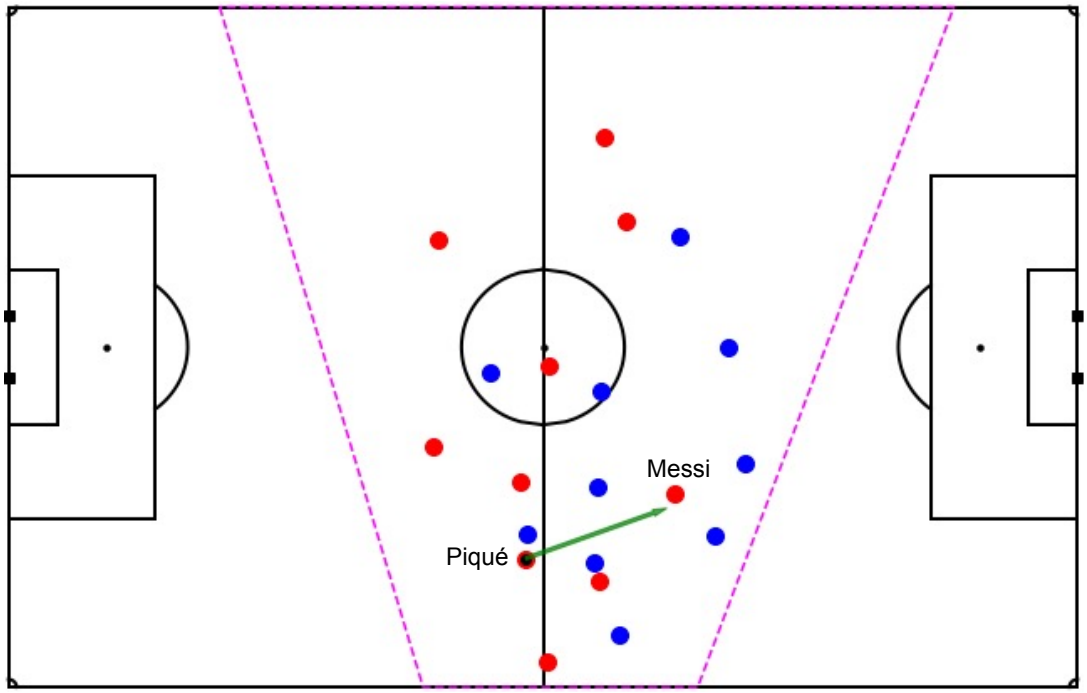

Figure 1: Datasets consist of the Euclidean position of all players appearing in the frame and the location of the pass, including the starting and ending point (green arrow) and the name of the players involved (sender and receiver). Players of the team with possession are plotted in red, while defenders are blue. Dashed lines show the limits of the area covered by the camera.

1.1 The model

"*All models are wrong, but some are useful*". This famous quotation from George Box summarizes the purpose of any mathematical model: to be useful despite its limitations. When developing a model, we must balance two opposed limits: On the one hand, we have to include as many variables as possible in order to have a "complete" model; on the other hand, we have to reduce the number of variables to facilitate the interpretation of the results, decrease the computation time and understand what variables are the most relevant.

1.1.1 Intercepting a pass

First, we are going to characterize the movement of the football's leading actor: the ball. Our primary target is to identify those regions of the pitch where the ball can be sent without being intercepted by the opponent team. Note that the ball can be captured when a rival player is placed in its trajectory and also when a rival moves to intercept it. Also, note that a pass can be sent either to a partner's feet or to a free space where he/she can control the ball before any rival.

We calculate the probability of a pass to be intercepted following a methodology similar to the one described in [13]. We divide the pith into $B = m \times n$ regions and investigate the probability that a pass sent from region b_{ini} arrives to region b_{fin}, with $b_{ini}, b_{fin} \in B$. For a completed pass, the coordinates of b_{ini} and b_{fin} are extracted from the location of the players sending and receiving the ball. However, if we want to evaluate alternative options to the actual pass, we should also consider any area of the pitch as a tentative b_{fin}, which will unavoidably increase the computation time. Once b_{ini} and b_{fin} are selected, we extract all regions $K \in B$ that the ball will traverse during its path to the receiver. Next, we identify the p opponent players that appear at the frame associated with the selected pass, $p \in P$, being P the total number of players on the pitch. Now, we evaluate the probability that any of the p rivals reaches any of the $k \in K$ regions of the ball trajectory, which will depend on the player speed and the distance to the trajectory, but also on the movement of the ball. Somehow, it is a competition between the ball and the rivals to reach any of the k regions before the other. Appendix 1 contains the details about the mathematical formulation of $T_p(k)$ and $T_{b_{ini}}(k)$ that are, respectively, the time required by every p rival and the ball to reach any of the k points of the trajectory. Once the required times are obtained, we calculate the probability that a rival p will intercept the ball as [13]:

$$\rho_{intercept}(p, b_{ini}, b_{fin}, k) = \frac{1}{1+e^{-\frac{T_p(k) - T_{b_{ini}}(k)}{\sqrt{3}\sigma(d)/\pi}}} \quad (1)$$

where $\sigma(d)$ is the variance of the expected ball velocity (see Appendix 1 for details), which depends on the distance between b_{ini} and k. To obtain Equation 1, we used a logistic distribution that allows modeling the uncertainty of the pass. Note that despite b_{fin} does not appear explicitly in Equation 1, it determines the set of k regions of the ball's trajectory.

To assign a single value to the probability of intercepting a pass, first, we assume that players move intuitively to the region k of the ball's trajectory that maximizes their probability of interception and we obtain $\rho^{max}_{intercept}(\mathrm{p}, b_{ini}, b_{fin}) = \max\{\rho_{intercept}(\mathrm{p}, b_{ini}, b_{fin}, \mathrm{k})\}$. If we consider that the probabilities of interception are independent of the players, we can obtain the probability of intercepting a pass departing from b_{ini} to b_{fin} as:

$$I(b_{ini}, b_{fin}) = 1 - \prod_{p \in P}(1 - \rho^{max}_{intercept}(\mathrm{p}, b_{ini}, b_{fin})) \quad (2)$$

In this way, the *interception parameter* I, which is bounded between the interval $[0,1]$ will be our indicator of the probability of a pass to be intercepted by a rival. Since the parameter I depends

both on the starting and ending points b_{ini} and b_{fin}, we can define, for every pass, an interception matrix $\mathbb{I} = [\mathbb{i}_{m,n}]$ of size $m \times n$, whose elements are $\mathbb{i}_{m,n} = I(b_{ini}, b_{fin})$.

Note that we can follow the same procedure considering that a teammate, instead of a rival, is the player intending to intercept the ball and obtain a *possession parameter* $P(b_{ini}, b_{fin})$ and a possession matrix $\mathbb{P} = [\mathbb{p}_{m,n}]$, whose elements are also bounded between $[0,1]$.

Finally, we define the risk matrix $\mathbb{R}$ by simply subtracting the interception matrix from the possession matrix, i.e. $\mathbb{R} = \mathbb{I} - \mathbb{P}$, and, accordingly, we obtain a risk parameter $r(b_{ini}, b_{fin}) = I(b_{ini}, b_{fin}) - P(b_{ini}, b_{fin})$. The elements of the risk matrix are now bounded between the interval $[-1,1]$ and are the indicators of which team is more prone to retain the ball during the pass. When values are close to -1, the team making the pass has a high probability of keeping the ball, while values close to 1 indicate that the pass is prone to be intercepted. It is worth noting that values close to 0 can be either because none of the teams will be able to control the ball or that both teams will have the same probability. In any case, values close to 0 reveal that none of the teams has an advantage in controlling the ball. Figure 2 shows and example of the values of the risk matrix for a pass of the red team.

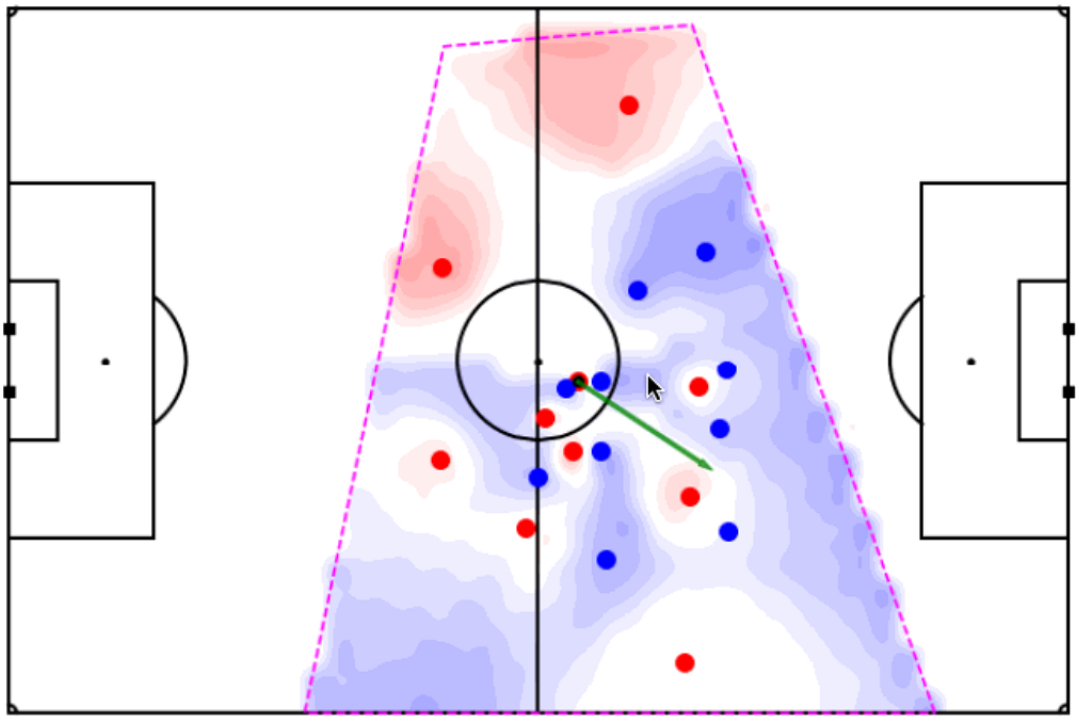

Figure 2: Example of the risk that the player making the pass assumes when passing the ball from location b_{ini} to any other region of the pitch. The green arrow shows the pass considered in the frame. Colours indicate whether the probability of intercepting the ball is higher for the red or the blue team.

1.1.1 Valuing passes

Next, we will quantify what the benefit of a given pass is. We selected the Expected Possession Value (EPV) as the variable to assess the improvement made by a pass. EPV is a parameter that quantifies, using Markov chains, the probability that an event performed at a specific location of the pitch finishes in a scoring action [19]. The pitch is divided into $m \times n$ regions and a probability associated with each region is obtained from the analysis of large event datasets obtained at different leagues and seasons. We did not extract the EPV from the StatsBomb 360º datasets, since the number of matches was not high enough. Instead, we used the EPV calculated by *Friends of Tracking* [20], which uses a pitch partition of 50×32 regions. We call $\mathbb{E}$ the matrix containing the EPV of a team. Importantly, EPV is obtained from a historical dataset containing event datasets from several leagues and seasons. Therefore, the gain matrix $\mathbb{E}$ is not specific to a given team, but on the contrary, it is the same for all teams. Despite the latter being a limitation, it will reduce the computation time for valuing passes. To quantify the gain produced by a pass, we pay attention not only to the EPV of the receiving location b_{fin}, but also to the EPV at the initial one b_{ini}. Therefore, we can define the increment of the Expected Possession Value as $\Delta EPV = EPV_{fin} - EPV_{ini}$ and use this parameter to quantify the gain (in possession value) produced by a pass.

1.1.1 Quantifying decision making

At this point, we have a methodology for evaluating the probability of a pass to be completed or intercepted. At the same time, we can measure the value of a pass. Nevertheless, how to determine that the option selected by a player is good or bad? Could we say that he/she made the best decision? This question does not have a simple answer since, in many cases, the context of the match is crucial to determine whether a pass was worth it or not. Therefore, we will define a Decision Parameter (DP), trying to be as objective as possible and always paying attention to the expected gain of a pass and the risk assumed by the player. With these premises in mind, when a pass i is made (i.e., completed), we quantify its risk $r(i)$ and its gain, defined as $\Delta EPV(i)$. Next, we calculate the value of all possible passes at the same frame, identifying the region of the pitch with the lowest risk r_{min}. We detect who is the teammate closer to the region of minimum risk and we check if the actual pass was made to this teammate. Finally, we define the *risk Decision Parameter rDP* of a player as the percentage of passes made by a player that went to the teammate with the lowest risk.

In the same way, we identify the region of the pitch with a negative value of the risk that has the highest gain ΔEPV_{max}. Note that we restrict the areas of tentative highest gain to regions whose associated trajectories are not intercepted by the rivals. As in the previous case, we identify the teammate closer to the area of highest gain and calculate the percentage of times that a player passes to this teammate. We call this parameter the *gain Decision Parameter gDP*.

1.1 Limitations

Before going to the Results and Discussion Sections, it is worth mentioning the weak points of our model to avoid misinterpretation of the results. On the one hand, we have certain limitations due to the origin of the data. We do not have the actual speed of the players, which would make the estimation of the interception times more accurate. The body orientation is also missing, a fact that cannot be solved even in the case of having access to complete tracking datasets. Furthermore, we do not have the positions of players outside the camera, which reduces the passing options and the possible interceptions made by the missing players.

Regarding the model, we are not assuming all possible ball speeds since it would exponentially increase the calculation time of the interception probabilities. As explained in Appendix 1, we have used a histogram of the expected ball speeds according to the distance of the pass, together with their corresponding variance. Furthermore, high passes are not considered. Definitely, more complete models should overcome these issues.

Finally, it is worth stressing that there is not a unique way of assessing what the optimal pass is. Further studies should investigate how adequately balance the risk assumed with the pass (i.e., decide when risk is reasonable) and the gain achieved, giving adequate weight to each of these variables.

Results

We selected F.C. Barcelona (and its players) to illustrate the outcome of our model. As the starting point, Figure 3 shows the risk (r) and the increment of the Expected Possession Value (ΔEPV) associated with all passes made by F.C. Barcelona during its match against Getafe (F.C. Barcelona 5 – Getafe 2) at season 2020/2021. Only completed passes are plotted. Dashed lines correspond to a risk r and ΔEPV equal to zero. We can observe a large cloud of points around the dashed line $\Delta EPV = 0$, indicating that most passes do not lead to a direct increment of the possession value. However, the cloud is placed below the zero-risk line, i.e., most passes have a higher probability of being completed than being intercepted. The two dashed lines divide the plot into four interesting regions. Passes with the highest ΔEPV are mainly located at the upper-right region, which corresponds to the high-value and high-risk region. This is, somehow expected since it indicates that most dangerous passes unavoidably imply a high risk. At the same time, the bottom-left region contains passes with a low risk that leaded to a decrease of possession value, something that could also be expected. The other two regions are also interesting. The upper-left one corresponds to passes with a high risk that led to a decrease of possession value. This is something not convenient for any team and, definitely, it is the region to be avoided. Finally, the bottom-right region of the plot contains low-risk passes that had a high increase of possession value. That

would be excellent for a team but, as we can observe, this kind of pass is very rare. It is the task of the defending team to avoid them.

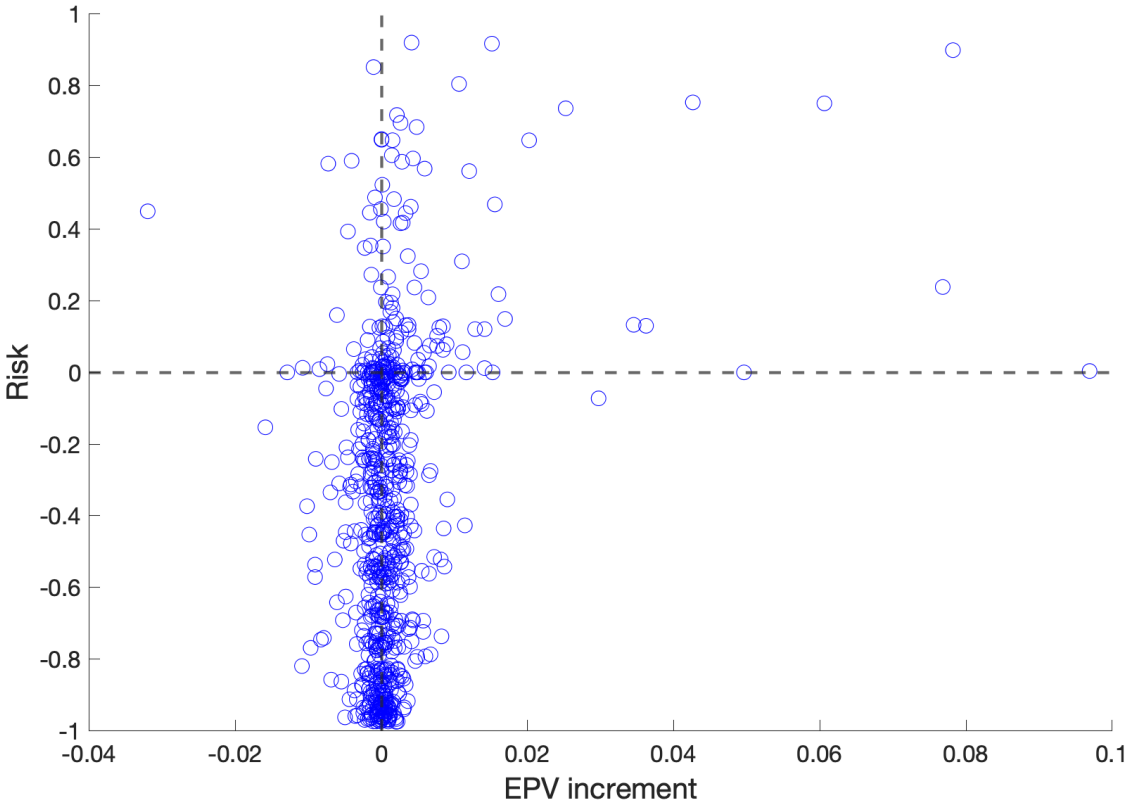

Figure 3: Risk r and increment of the Expected Possession Value ($\Delta EPV = EPV_{fin} - EPV_{ini}$) of all completed passes of F.C. Barcelona in the match against Getafe (season 2020/2021). Horizontal and vertical dashed lines indicate, respectively, zero risk and zero EPV increment. Our model assigns negative values of the risk to passes with a high probability of being completed and positive values to high-risk passes.

Now, let us analyze the risk and the EPV increment from the point of view of the player. We can do that by averaging the values of passes corresponding to each player. In Figure 4 we can observe how the closeness to the opponent's goal is strongly correlated with the risk assumed by players, indicated by a positive correlation between both variables ($R^2 = 0.754$). Only players with more than 100 passes are plotted. Interestingly, Messi is the player with the highest average risk, while Umtiti at the bottom of the distribution. However, note that we only evaluated the risk of the completed passes. Thus, Messi's highest value is not necessarily a negative indicator since it is telling us that he succeeded in completing passes with a high risk. It is also worth noting the cluster of forward players at the top of the raking while defenders and the goalkeeper are found at the bottom, which seems to indicate that our model is performing as we expected.

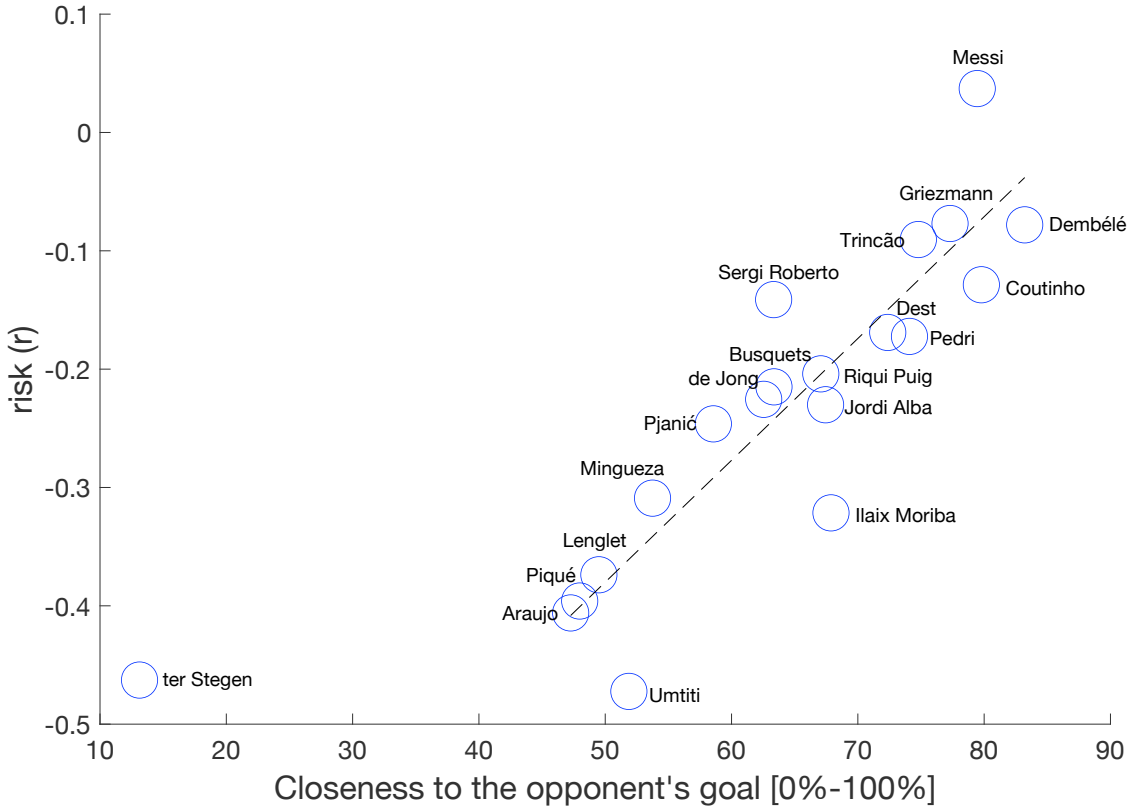

Figure 4: Player's risk r vs the closeness to the opponent's goal, which is bounded between 0 (own goal) to 100 (opponent's goal). The dashed line corresponds to a linear fit without considering the goalkeeper. We can observe a positive correlation between the risk and the closeness to the opponent's goal, with the goalkeeper as an outlier. Only players with at least 100 passes are plotted.

The model allows us to distinguish between the risk assumed when passing or receiving the ball, which could be an indicator of the player's ability to receive the ball in adverse conditions. Figure 5 shows the average risk of each player when passing or receiving the ball, respectively r_{send} and $r_{receive}$. We can observe a positive correlation between both variables ($R^2 = 0.754$). However, some players slightly deviate from the general trend. For example, Trincão is a player who receives the ball a higher risk than the risk he assumes when passing. On the contrary, ter Stegen and Umtiti are the players that receive at the lowest risk, as one may expect for the goalkeeper and the centre-back. We can also observe how Messi is the player that assumes more risks when passing, but this is not the case when receiving.

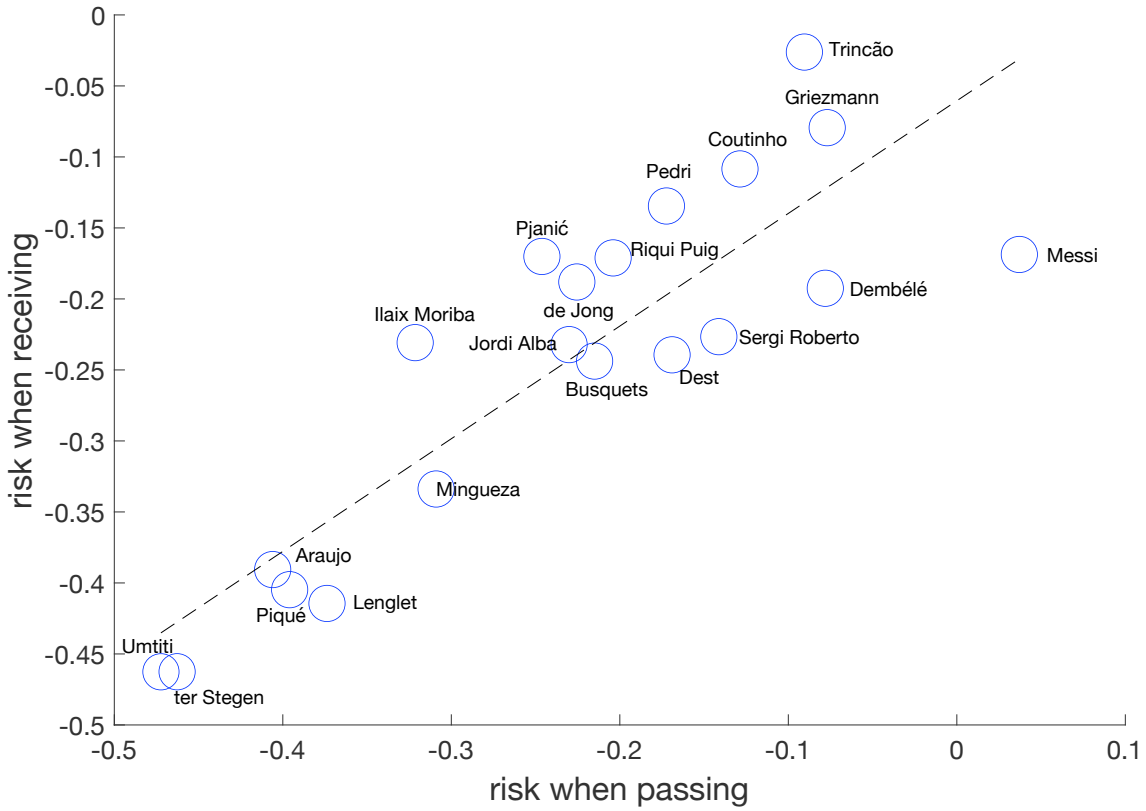

Figure 5: Risk when making and receiving a pass. The x-axis contains the average value of the risk assumed by a player when making a pass, while the y-axis refers to the average risk when the same player receives the ball. The dashed line corresponds to a linear fit. Despite there is a positive correlation, there are players that deviate from the general behavior. Only players with at least 100 passes are plotted.

Next, we are going to quantify the gain of passes made by players by computing the average ΔEPV of each player. In Figure 6 we plot the interplay between the ΔEPV of a player and the average location of his passes in terms of closeness to the opponent's goal. Despite there is a positive correlation ($R^2 = 0.239$), it is not that clear as in the previous case. Jordi Alba is the player with the highest increment of the possession value, despite not having a position so close to the opponent's goal like the forward players. In fact, forward players are the ones following Jordi Alba in the ranking, as one may expect. On the other hand, Ilaix Moriba is the player with the lowest gain, reflecting that the main purpose of his passes is not to increase the value of the possession directly.

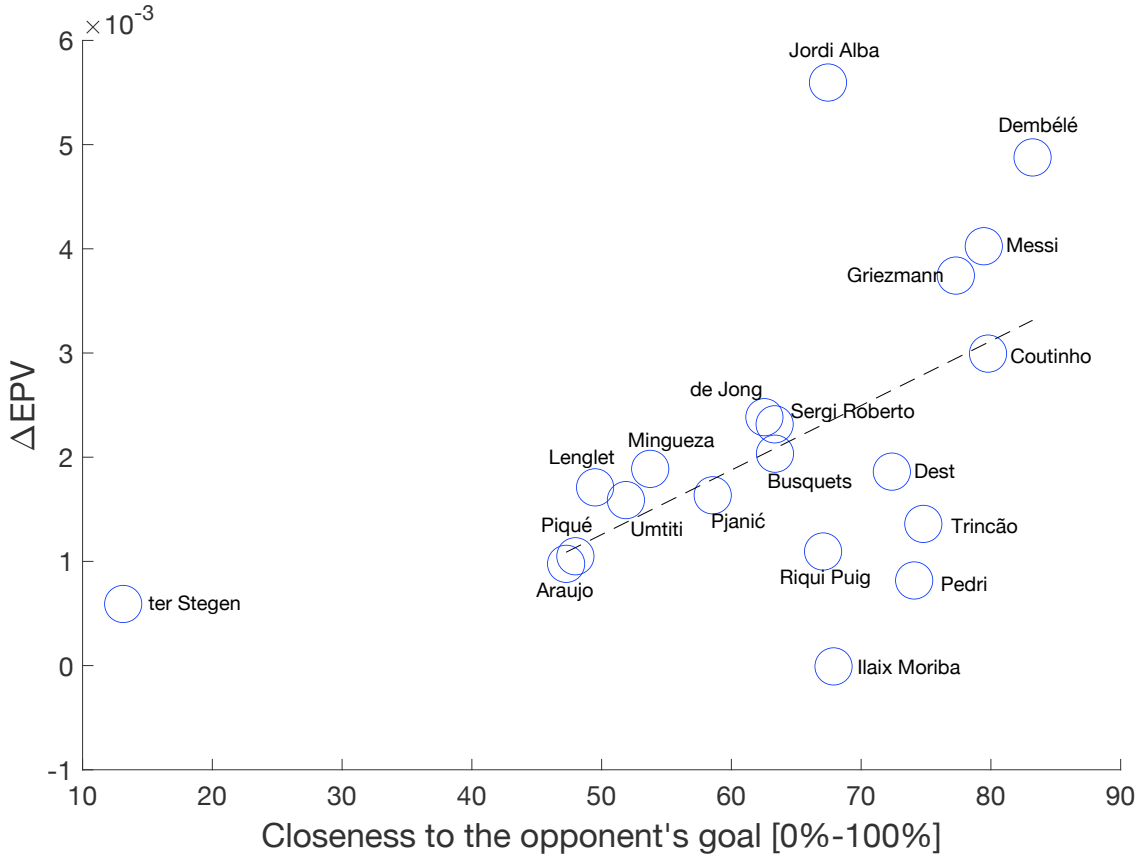

Figure 6: Increment of the expected possession value ΔEPV due to players' passes. The x-axis indicates the closeness to the opponent's goal of all passes made by each player (from 0 to 100). The y-axis refers to the average ΔEPV. Only players with at least 100 passes are plotted. The dashed line corresponds to a linear fit without considering the goalkeeper.

Now we are going to investigate how close the decision making of players was to the optimal case. As explained in the Methods Section, we will use two different decision-making parameters, one for evaluating whether a player is selecting low-risk passes and another one indicating if a player decides to pass to the teammate that increases the value of the possession the most.

Figure 7 shows the risk Decision Making (*rDP*) and its relationship with the actual risk r assumed by a player. The reason of plotting *rDP* in such a way is to show that a player can make high risk passes due to its location in the field but, at the same time, these high-risk passes could be the most conservative ones between the possibilities the player had at that specific moment. Also note that the *rDP* is normalized between 0 and 1 in a way that it directly indicates the percentage of passes where the teammate with the lowest risk was selected[3]. We can observe how Messi, Riqui

[3] For example, a *rDP=0.35* indicates that the player decides to pass the ball to the teammate with the lowest risk the 35% of the passes.

Puig and Busquets are the players with the lowest *rDP*, revealing that they are not selecting the receiver of their passes based on minimizing the risk. On the other hand, Dest, Moriba and Trincão are the players with the highest *rDP*. They are players that prefer not to risk in their passes, selecting the safer teammate in more than the 45% of their passes. We can also observe a cluster of defenders (that includes ter Stegen, the goalkeeper) with moderate values of the rDP (all of them above 0.35).

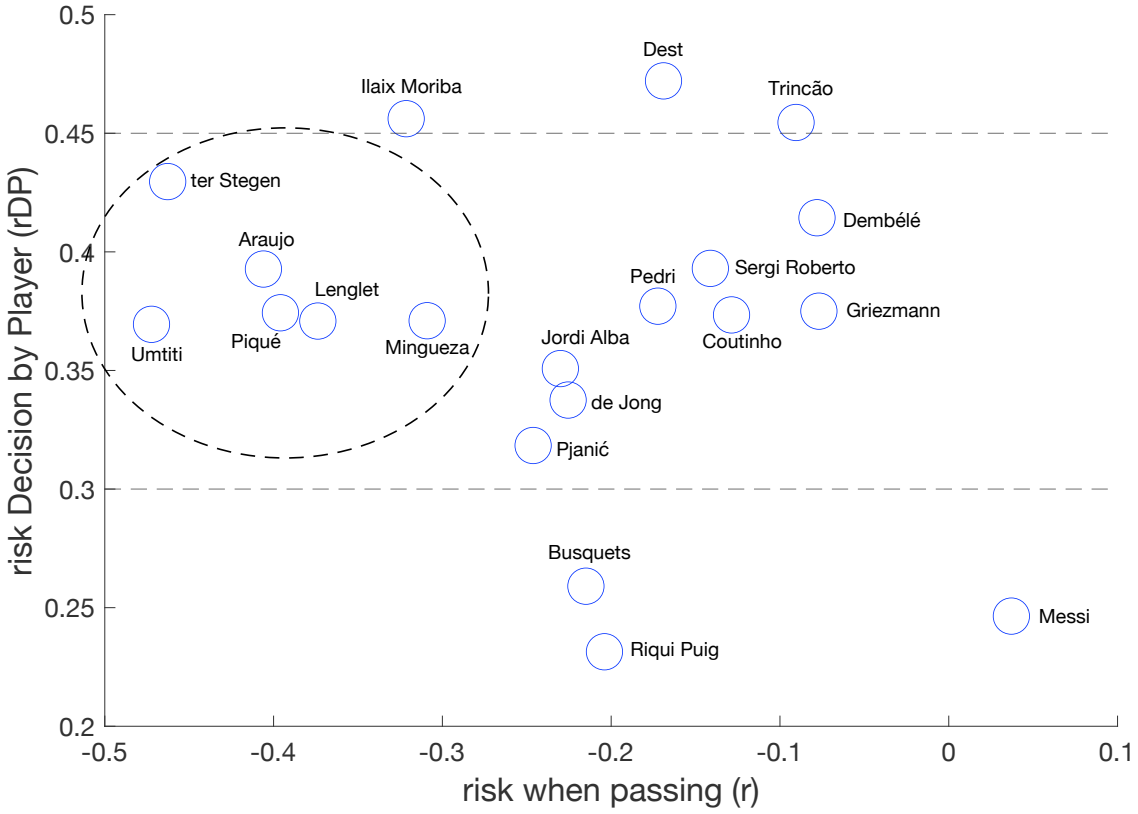

Figure 7: Risk Decision Parameter (*rDP*) accounting for the percentage of passes made to the teammate with the lowest risk. Dashed lines separate the plot into three regions: (i) below 0.3, players that decide to pass assuming high risks, (ii) between 0.3 and 0.45, players with moderate values of *rDP* and (iii) above 0.45, players that pass to the teammate with the minimum risk more than the 45% of passes. The ellipse highlights a cluster of defenders.

In Figure 8, we carry out a similar analysis paying attention to the increment of the possession value. The gain Decision Parameter (*gDP*) computes the percentage of passes that are sent to the partner who is closest to the region of highest possession value and, at the same time, with a negative risk parameter. The *gDP* is plot versus the expected possession value of the receiving location of the pass. In general terms, we can observe a positive correlation between both variables, with a $R^2 = 0.619$. Messi is the player with the highest *gDP* (~ 0.22), indicating that the 22% of his passes are sent to the teammate with the highest possession value. We can observe

that 4 forward players are placed in the top-6 ranking of the *gDP*, together with Jordi Alba and Sergi Roberto. At the lowest places of the *gDP* ranking we find Piqué and Araujo, both occupying defending positions. It is also interesting to compare the *gDP* of the defender players (indicated by a dashed ellipse). Although the average EPV of the receiving location is similar, there are difference in their *gDP*, with Lenglet having the highest one and Piqué the lowest.

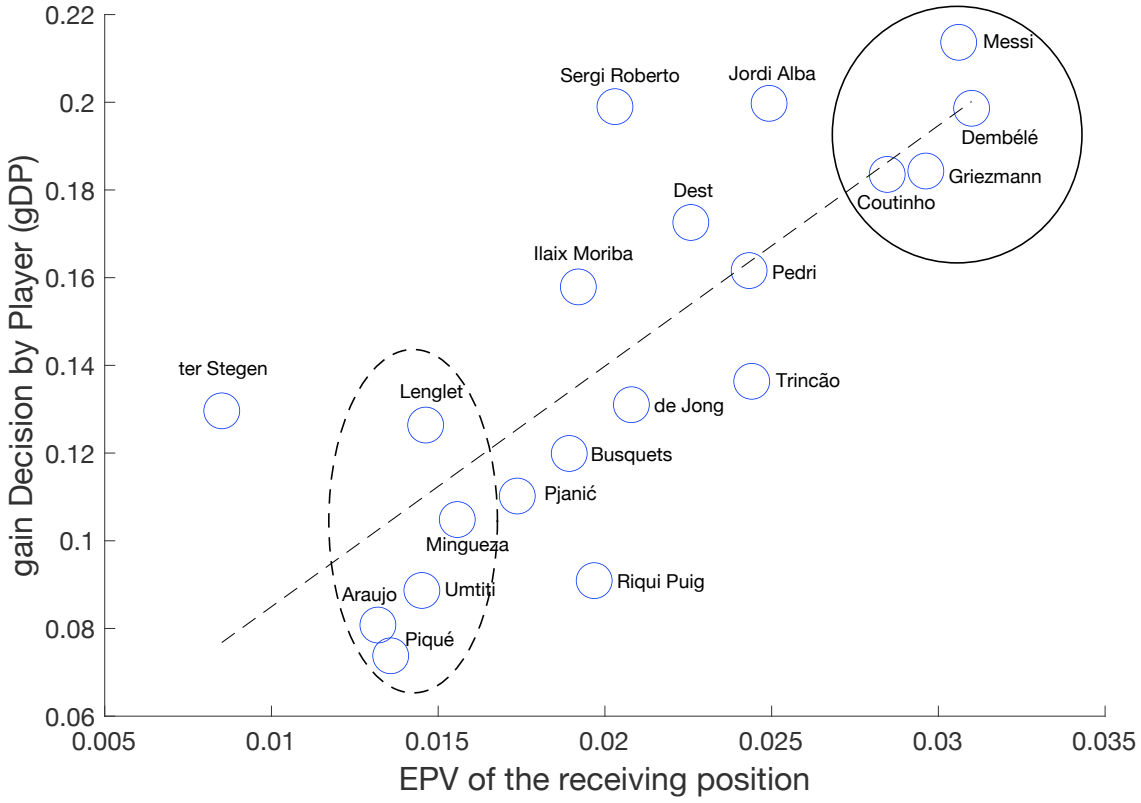

Figure 8: Figure 7: Gain Decision Parameter (*gDP*) accounting for the percentage of passes made to the teammate with the highest EPV as a function of the EPV of the receiving position (EPV_{fin}). The ellipses highlight the cluster of defenders (dashed) and forward (solid) players. The dashed line corresponds to the linear fit.

The fact that an interception probability, risk, and gain can be assigned to every pass, opens the door to different kinds of analysis beyond focusing on specific players. For example, we can analyze what regions of the pitch are related with a higher interception probability, as we show in Figure 9. We can see that regions close to the opponent's goal are the riskier; however, in this example, we can also observe that, in the own field, the sides of the pitch have a slightly higher risk than the central corridor, a piece of information that analysts and coaches could use to make tactical decisions.

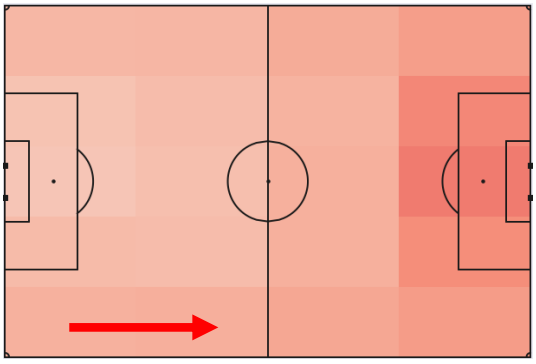

Figure 9: Spatial dependence on the Interception Parameter I. The team with the ball (F.C. Barcelona) attacks from left to right. We split the pitch into a 4×5 grid. The average interception parameter for passes made inside each region is calculated. The red colour indicates a higher probability of passes being intercepted. In this example, we considered all passes of F.C. Barcelona during 37 matches. We can see that forward positions are the ones associated with the highest interception probabilities.

Finally, what if, instead of space, we focus on time? Figure 10 shows the differences between the average values of the first and second parts of the 37 matches of F.C. Barcelona. Specifically, we plot the values of (A) the risk parameter r, (B) the increment of the possession value ΔEPV, (C) the risk decision parameter *rDP* and (D) the gain decision parameter *gDP*, averaged over each of the two periods of the match. We can observe that the first two variables point towards the same direction: on the second part of a match, players take more risks when passing the ball in order to obtain a higher EPV. Concerning the decision making, players decrease their *rDP*, i.e., decide to pass to players with higher risk more frequently and, at the same time, increase their *gDP*, selecting teammates located at regions of the pitch with higher possession value.

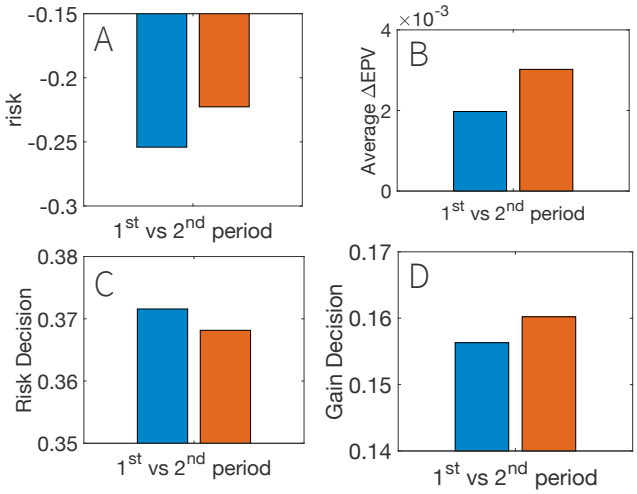

Figure 10: Differences between the first and second periods of a match. (A) risk parameter r, (B) increment of the possession value ΔEPV, (C) risk decision parameter *rDP* and (D) gain decision parameter *gDP*. Blue bars correspond to the 1st period and red bars to the 2nd one.

Discussion

In this paper, we proposed a new methodology to quantify to what extent players make the right decisions when making a pass. We first evaluated the risk of losing the ball and, at the same time, quantified the gain acquired with a pass using the Expected Possession Value. Comparing the gain and the risk, we discriminated those players who prefer to make safer passes at the cost of having less gain and, conversely, those who take more risks to get a better outcome, as well as when (1st vs 2nd parts) and where (location in the pitch) this occurs. Next, we defined two different decision-making parameters evaluating the risk assumed when making a pass and the gain obtained. With this aim, we calculated the risk and gain not only for the passes made by each player but also for all possible passes a player could make based on the partners' and opponents' locations and their movement abilities. This allowed determining the optimal pass at each moment of the match according to (i) the minimum risk and (ii) the highest gain of expected possession value. Next, players' decision making was obtained by calculating the percentage of passes that were sent to the teammate with (i) the lowest risk and (ii) the highest gain, leading to the risk and gain decision parameters, *rDP* and *gDP*, respectively.

Several aspects of the model must be discussed carefully to avoid misinterpretation of the results. On the one hand, there is not a unique way of determining what the best pass is. Our model prioritized two factors, the risk and the gain achieved, but other baselines could be chosen. For example, there are regions of the pitch, such as the first third, where the priority is reducing the risk at the expense of reducing the gain. Conversely, in the final third, the optimal pass could imply maximizing the expected gain, no matter what the risk is. The same reasoning could be applied to different moments of the match where the score is in favour or against our team. For these reasons, alternative options to adequately weight risks and gains should be explored.

On the other hand, the model has several limitations that could be overcome with more detailed datasets. It would be interesting to have access to the location of all players on the pitch, instead of those appearing in the camera. This problem could be solved by obtaining the players' location from tactical cameras covering the whole pitch. Furthermore, introducing the direction and modulus of the players' velocities would lead to better estimations of the interception probabilities. In that sense, the vector velocity could be estimated from the same video from where the player location is extracted.

Finally, there are a series of variables that could be already included in the model to make it more complete, such as (i) the possibility of making high passes, (ii) the ability of the ball to move at different speeds or (iii) the body orientation of players. The price we have to pay is an extended computation time, but we believe it would be worth including them in future versions of our

model. In this regard, we just wanted to introduce a minimal model able to get some insights about the decision making of players using StatsBomb 360º datasets.

Appendix 1

Appendix 1 contains the details about the mathematical formulation of $T_p(k)$ and $T_{b_{ini}}(k)$ that are, respectively, the time required by every p rival and the ball to reach any of the k points of the trajectory. The time taken by the ball to reach a point k depends on its speed and the distance to be covered. To avoid the simulations of all possible ball speeds, we have investigated what the typical time required by the ball to cover a certain distance is. With this aim, we used tracking datasets containing 12050 passes, which have been supplied by *LaLiga* software *Mediacoach®* [21]. Figure A1(A) shows the interplay between the distance travelled by the ball and the required time. Using this dataset, we constructed a linear model between both variables of the form $T_{b_{ini}}(k) = \beta_0 + \beta_1 d(b_{ini}, b_k)$, where $d(b_{ini}, b_k)$ is the distance between the starting point b_{ini} and the final region k, being k any point of the expected trajectory of the ball. We do not consider high passes to reduce the complexity of the model. We force $\beta_0 = 0$ (if the ball does not move, the time should be zero) and obtain $\beta_1 = 0.075$ with a linear regression. Importantly, we also calculated $\sigma(d)$, which is the variance of the travelling times for each distance. The variance contains the variability of ball speeds that can be expected and, as we see in Figure A1(B), it depends on the distance. We will use this variance as a source of uncertainty in Equation 3 of the main text.

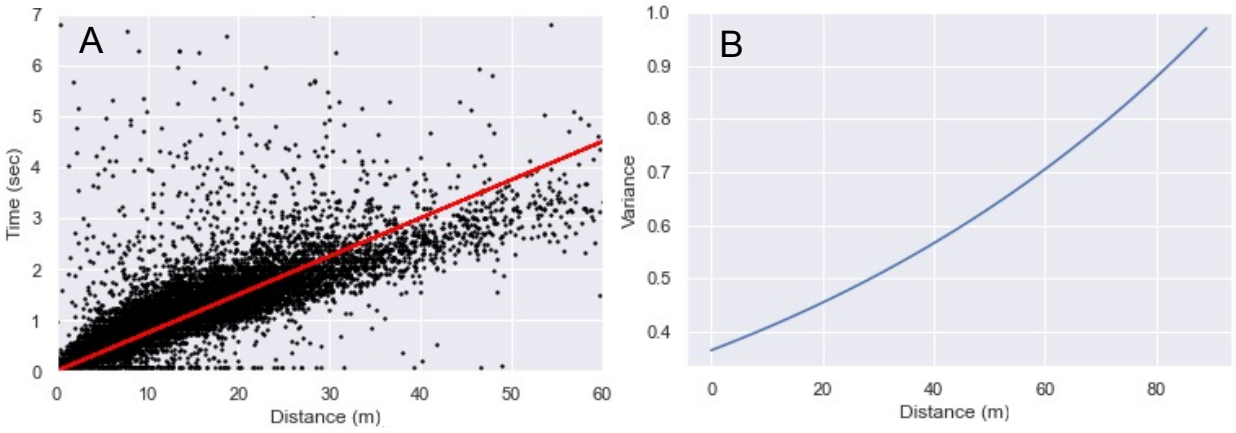

Figure A1: (A) Relation between the distance $d(b_{ini}, b_k)$ travelled by the ball and the required time $T_p(k)$. The solid red line corresponds to the linear regression $T_{b_{ini}}(k) = 0.075\, d(b_{ini}, b_k)$. In (B), the corresponding variance $\sigma(d)$, of the ball traveling times and its dependence on the distance, whose equation is $\sigma(d) = e^{-1.009+0.011\, d(b_{ini}, b_k)}$.

Concerning the time required by players to reach a region k of the pitch, we followed a model similar to the one proposed in [22]. All details about the derivation of the following equations are described in [23]. For each player p, we define a radius determining the area of the pitch that the player can reach within a particular time t. To determine this radius, we use the equations of motion proposed in [23], where a player makes a force F to increase its speed and he/she suffers a certain resistance proportional to its speed, leading to the following ordinary differential equation:

$$m\frac{d(\vec{v})}{dt} = \vec{F} - k\vec{v} \quad \text{(A.1)}$$

where k is a constant accounting for the resistance to increase the speed. To solve Equation A.1 we need to know the initial speed $\vec{v}_0$, i.e. the direction and modulus of the players' speeds. Since we do not have this information in the StatsBomb 360° datasets, we will assume that $\vec{v}_0 = 0$. With this simplification, the solution to Equation A.1 leads to a circular region where this player can arrive depending on time. This region is centered at the initial position of the player $\vec{x}_0$ when the pass is initiated and has a radius given by the expression (see [23] for details):

$$r_{int}(t) = v_p\left(t - \frac{1-e^{\alpha t}}{\alpha}\right) \quad \text{(A.2)}$$

where $\propto = k/m = 1.3$ is a constant related to the resistance to increase the speed and be $v_p = 7.8$ m/s is the highest speed a player reaches. Note that we simplify the application of Equation A.2 by assuming that all players can run at the same speed. According to Equation A.2, in case that $t = 0$, the distance that a player can reach is zero. However, the fact that a player occupies a certain space makes him/her able to intercept the ball when he/she is already in the trajectory of the ball and, at the same time, very close to the starting point. Therefore, the radius of the interception area of a player will be the maximum of:

$$\mathrm{r(t)} = \max\left\{d_o, v_p\left(t - \frac{1-e^{\alpha t}}{\alpha}\right)\right\} \quad \text{(A.3)}$$

where $d_0 = 1$ meter accounts for the interception area without moving. Equation A.3 explicitly relates the distance that a player can reach with the required time. Therefore, to obtain the time $T_p(k)$ a player needs to reach any point b_k of the trajectory of the ball we will, first, check if the region b_k is closer than a distance d_0. If it is the case, the player will intercept (block) the ball instantaneously. If it is not, we must identify at what point of the trajectory the player will intercept the ball and then get that time. To do that, we sequentially increase the value of t from $t = 0$ with an increment of $\Delta\mathrm{t} = 0.1$ seconds and obtain the value of the corresponding radius $\mathrm{r(t)}$ that

defines the area reached by the player. We repeat this process iteratively until any of the b_k regions of the ball's trajectory lies inside the area reached by the player. At this moment, we will obtain the value of $T_p(k)$.

Acknowledgments

We thank Roberto López del Campo and Ricardo Resta, from the Sports Research Area of *LaLiga*, for providing the datasets for estimating the ball speed and variance and for fruitful discussions about models evaluating player decision-making. We also thank our colleagues David Garrido and Daniel R. Antequera from the Center for Biomedical Technology for fruitful conversations. Finally, JMB would like to thank Paco Seirul.lo, from F.C. Barcelona, for inspiring conversations about the interpretation of the beautiful game.

References

[1] Gudmundsson, J., & Horton, M. (2017). Spatio-temporal analysis of team sports. ACM Computing Surveys (CSUR), 50(2), 1-34.

[2] Memmert, D., & Rein, R. (2018). Match analysis, big data and tactics: current trends in elite soccer. German Journal of Sports Medicine/Deutsche Zeitschrift fur Sportmedizin, 69(3).

[3] Cuevas, C., Quilon, D., & García, N. (2020). Techniques and applications for soccer video analysis: A survey. Multimedia Tools and Applications, 79(39), 29685-29721.

[4] StatsBomb. https://www.statsbomb.com/[accessed on 10th September 2021].

[5] Statsperform. *https://www.statsperform.com* [accessed on 10th September 2021].

[6] Wyscout. *https://www.wyscout.com/* [accessed on 10th September 2021].

[7] StatsBomb Data Case Studies: Freeze Frames And Defender Locations https://statsbomb.com/2021/03/statsbomb-data-case-studies-freeze-frames-and-defender-locations/ [accessed on 10th September 2021].

[8] Singh, K. 2019. Introducing expected threat. https://karun. in/blog/expected-threat.html. [accessed on 10th September 2021].

[9] Decroos, T.; Bransen, L.; Van Haaren, J.; and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1851–1861.

[10] Buldú, J.M., Busquets, J., Martínez, J.H., Herrera-Diestra, J.L., Echegoyen, I., Galeano, J., & Luque, J. (2018). Using Network Science to Analyse Football Passing Networks: Dynamics, Space, Time, and the Multilayer Nature of the Game. Frontiers in Psychology 9, 1900.

[11] Garrido, D., Antequera, D. R., Busquets, J., Del Campo, R. L., Serra, R. R., Vielcazat, S. J., & Buldú, J. M. (2020). Consistency and identifiability of football teams: A network science perspective. Scientific reports, 10(1), 1-10.

[12] Taki, T., & Hasegawa, J.I. (2000). Visualization of dominant region in team games and its application to teamwork analysis. In Proceedings Computer Graphics International 2000, 227-235.

[13] Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-based modeling of pass probabilities in soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*.

[14] Spearman, W. (2018, February). Beyond expected goals. In *Proceedings of the 12th MIT Sloan Sports Analytics Conference* (pp. 1-17).

[15] Fernandez, J., & Bornn, L. (2018). Wide open spaces: A statistical technique for measuring space creation in professional soccer. In Sloan Sports Analytics Conference 2018.

[16] Marcelino, R., Sampaio, J., Amichay, G., Gonçalves, B., Couzin, I. D., & Nagy, M. (2020). Collective movement analysis reveals coordination tactics of team players in football matches. *Chaos, Solitons & Fractals*, 138, 109831.

[17] Buldú, J.M., Garrido, D., Antequera, D.R., Busquets, J., Estrada, E., Resta, R. & López del Campo, R. (2020) Football tracking networks: Beyond event-based connectivity. Conference on Analytics in Sports Tomorrow 2020, ed. by F.C. Barcelona., 1-13.

[18] Gómez-Jordana, L. I., Milho, J., Ric, Á., Silva, R., & Passos, P. (2019). Landscapes of passing opportunities in Football–where they are and for how long are available. In Conference paper at Barça Sports Analytics Summit.

[19] Rudd, S. (2011). A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains. In *New England Symposium on Statistics in Sports*.

[20] Friends of Tracking. Data EPV downloaded from https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking [accessed on 10th September 2021].

[21] Mediacoach. https://www.mediacoach.es [accessed on 10th September 2021].

[22] Fujimura, A., & Sugihara, K. (2005). Geometric analysis and quantitative evaluation of sport teamwork. Systems and Computers in Japan, 36(6), 49-58.

[23] Peralta Alguacil, F. J. (2019). Modelling the Collective Movement of Football Players. PhD Thesis. Uppsala Universitet, Sweden.