14TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 6 - 7, 2020 | BOSTON, MA
presented by ESPN

Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer

Francisco Peralta Alguacil[1,3], Javier Fernandez[2], Pablo Piñones Arce[3], David Sumpter[3,4]

Paper Track: Other sports
Paper ID: **1548690**

Abstract

Soccer has some of the most complex team movement patterns of any team sport. Recently, several measurements have been proposed for evaluating state of play and for identifying the expected value of dribbles, passes or shots [1-6]. The next step is to automatically identify the alternative actions available to players both on and off the ball. We address this challenge by defining three optimization criteria that drives the movement of players during attack. (1) *Pass probability*: A player moves to maximize the probability of pass success to either himself or to another player, e. g. by opening up a passing lane. (2) *Pitch Impact*: Occupy point on the field which is maximally dangerous. For example, a striker moves to a point directly in front of goal. (3) *Pitch Control*: Maximize the amount of space controlled by the team. Soccer players often rate their teammates in terms of their ability to anticipate the movement of the other players on the pitch a few seconds in to the future. To account for this, and building on studies of pedestrian movement, we assume players maximize their future value position on a weighted combination of these three criteria.

We then built a 'self-propelled player' model, simulating attacking roles by maximizing a weighted combination of pass probability, impact and control. We compared the simulations to player decisions during matches by top-flight men's teams of Hammarby IF and FC Barcelona. In simulations, we found that the two or three players nearest to the ball tended to optimize the product of pass probability and pitch impact. We found that simulations in which players optimized pitch control did not reliably capture the movement of players.

In a first-team coaching intervention at Hammarby, players re-watched attacking situations in which they had been involved in the form of pass probabilities, pitch control visualisations and comparisons to the simulation model. The players often agreed that the model captured complex game patterns, including attacking runs to displace defenders and pressing that narrows down the opponent's passing opportunities. The model also recommended runs that the players hadn't taken, which the players also found realistic and aided discussions. Despite the fact that discussion of models with professional players is rare, the players showed a high willingness to engage with them. We further

[1] Mathematics Department, Uppsala University
[2] FC Barcelona
[3] Hammarby IF Fotboll
[4] Mathematics Department, Uppsala University. Email: david.sumpter@math.uu.se

explored how these techniques can be used to provide automated feedback to players within the match cycle.

1. Introduction

The fundamental challenge in soccer analytics lies in understanding the collective motion of the players and the ball. Soccer is a more fluid sport in comparison to games like American football, baseball and cricket, which involve discrete 'plays'. And, with twenty-two players involved at all times, it has more moving parts than basketball or ice hockey. From a mathematical modelling point of view, this means that soccer has more degrees of freedom than these other sports, making it difficult to assess the game using one or a small number of metrics.

This challenge has, until recently, been made even more difficult by the lack of in-game data. Event data, recording the co-ordinates of the player with the ball and their actions (pass, interception etc.) has been available for almost a decade from companies like Opta and has recently been supplemented with a description of how much defensive pressure the player with the ball is under from other players, from Statsbomb. Open data sets from these companies are now available [8,9]. The best-known statistic derived from event data is expected goals, which measures the quality of chances players create [1,7]. Other more advanced metrics include expected assists, passing models that assign a value to every pass based on how much it progresses the ball [2], and possession chains which measure involvement in attacking sequences [3]. However, due to the limited nature of event data, these metrics really just measure a small portion of what occurs during a soccer game. To make this concrete, consider the fact that, during a typical match, Barcelona striker Luis Suarez has the ball for less than 90 seconds of the 90 plus minutes of match time. What Suarez, or any other player, contributes to the play---pressing, runs to open up space and tactical positioning---can't simply be measured by event data alone.

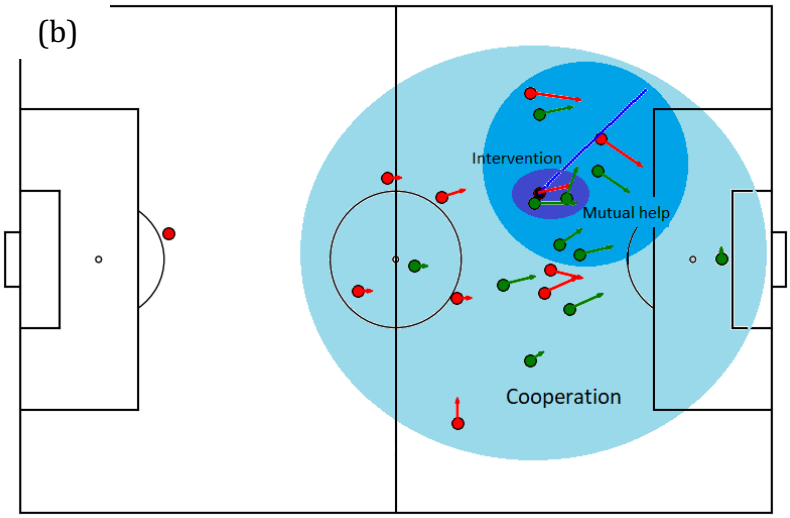
More recent work has focused on spatial-temporal tracking data of the co-ordinates on the pitch of all the players, as well as the position of the ball [4-6,10]. Even more advanced techniques are attempting to reconstruct body orientation from match video [11]. Most professional leagues now collect this tracking data for all matches and while this data still contains errors (player ids switched, players obscured and not properly tracked at certain time points), it is sufficiently accurate to be used to start to develop models.

Tracking data allows for the incorporation of much more information than event data, such as, the interceptability of any potential pass, the relative location of players according to the defending block, the degree of space control at any location, and players' velocity. One way to approach tracking data is to extend the metrics based on estimating pass values in event data. For example, Spearman developed a comprehensive game state representation using tracking data. His model combines probability of scoring from a point on a pitch, the probability the team controls that point and the probability the ball transitions to that point, providing an objective way of estimating the expected long-term value of soccer possessions [10]. Parameters for the model were then fitted from match data. Tracking data can also be used to learn directly from raw-level tracking data using deep neural networks. This data-driven approach has been adopted in both basketball [12,13] and soccer [14]. These models can produce realistic looking player trajectories and be used to automatically identify different types of formations and forms of attack (e.g. counterattacks) [15].

A data-driven approach using machine learning certainly has its advantages. However, there are other successful approaches to the study of collective motion which might prove more appropriate in soccer, in particular. In the study of collective motion of animal groups, one very successful modelling approach is commonly known as self-propelled particles. These model individuals (fish, birds or humans) in terms of a small number of principles that determine how they interact with neighboring individuals. The earliest models of this sort described fish as responding to their neighbors in three zones [16,17]. A small repulsion zone where they avoid collisions, a larger alignment zone where they move in the same direction of neighbors in that zone and an outer attraction zone in which they move toward their neighbors. Simulating these rules demonstrated how simple interaction rules showed how even very simple can produce complex patterns [18,19].

(a)

(b)



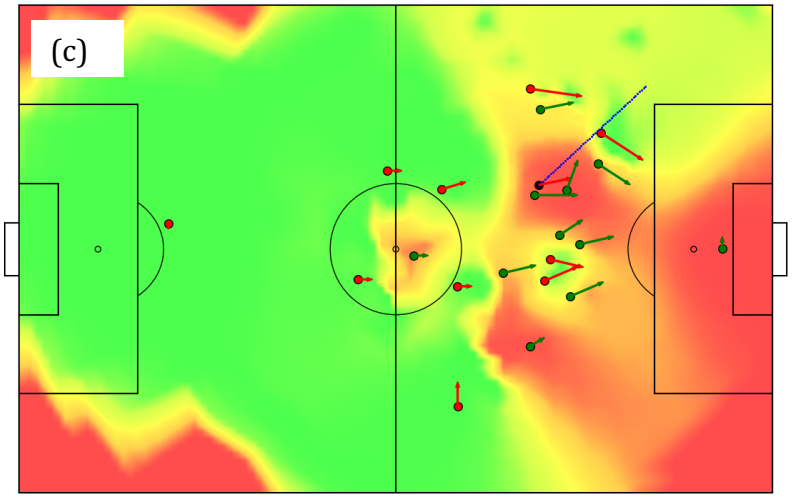


Figure 1: (a) A specific match situation in which a winger runs towards the box in order to open up space for the left back. (b) Illustration of the three zones in football for both teams. Barcelona (red) are in possession of the ball and attacking to the right. The Real Betis (green) are defending. The player with the ball and the two defenders nearest to the ball are in the Intervention zone. The players in the direction of the pass (blue line) as well as the nearest defenders are in the mutual help zone. Other players are considered to be in the co-operation zone. (c) Passing probabilities calculated for this situation. Green is high pass success, red is low success probability. See Appendix A.1 for details of how pass probability is calculated.

The power of these models in biology has been that they provide a framework for combining modelling and experiments to uncover the rules of interaction of individual animals in more and more detail [19,20]. In many cases, this has led the initial, rather naive models to be replaced by more realistic models that capture details of animal behavior [19,22]. Some of this process has involved, what would today be called, machine learning [23,24]. However, the most successful studies in collective motion have started from the viewpoint of creating a simulation model based on simple rules of motion for individuals, comparing the model outcome with experimental data and revising the model to improve understanding [18,19].

It is this modelling approach we adopt to tracking data in soccer.

2. Self-propelled player model

There is good reason to believe that a modelling approach based on self-propelled particles interacting based on zones can be fruitful in soccer. Coaches often talk about three playing zones relative to the position of the ball. Francisco Seirul-lo, at the department of methodology FC Barcelona, outlines three zones (1) make an intervention (2) provide mutual help and (3) provide co-operation [25,26]. We now make an interpretation of these zones that will allow us to link them to a self-propelled player (SPP) model (Figure 1b).

1. **Intervention zone**. This zone covers the immediate points around the ball. It includes the player with the ball and those defending players who could touch or intercept it immediately.
2. **Mutual help zone**. Players in this zone are relatively close position to the ball, but further away than the players in the intervention zone. Teammates of the player with the ball are considered inside this zone if they could receive pass within a few seconds. Defending players are in this zone if they are acting to intercept or defend a pass within the next few seconds.

3. **Cooperation zone**: Players in this zone are further away from the play and not expected to receive the ball during within next few seconds. In attack, these players aim to occupy dangerous areas of the pitch and control space. In defense, they aim to minimize the area used by the opposition.

An example for showing these zones is shown in Figure 1, which is a single frame of a match between FC Barcelona and Real Betis, played during the first match day of the 2017/2018 La Liga season.

Much of the art and science of soccer coaching is about giving instructions about how players should act in these different zones. Our intention in this paper is to find a starting point for an SPP model of soccer, which builds on these three zones.

The basic assumption of the model we now develop is that when a player's team has the ball and are attacking, the player attempts to optimally balance three different criteria, namely

Pass probability: A player moves to maximize the probability of pass success to either himself or to another player [5]. For example, by opening up a passing lane either to himself or a teammate. (Figure 1c)

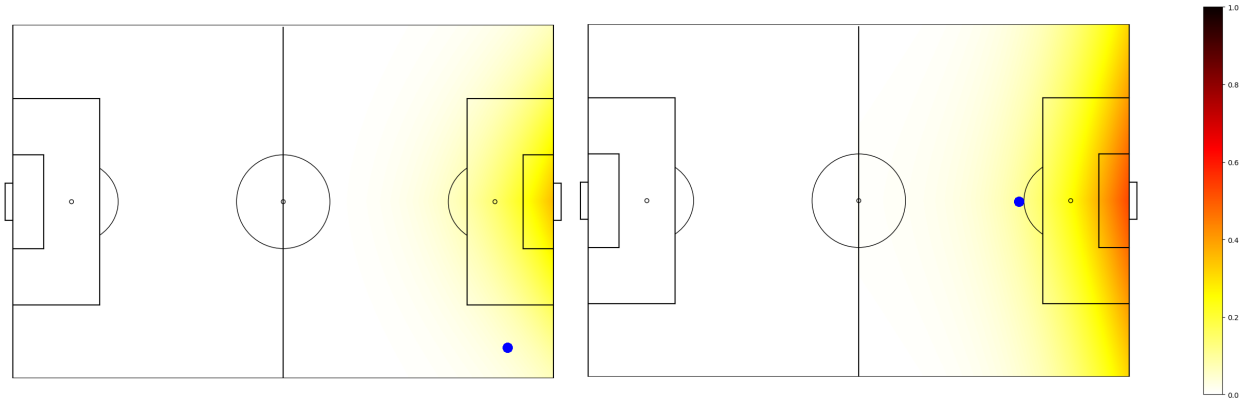


Figure 2: Pass impact for two different starting coordinates of a pass (blue dot). Heat gives probability that a pass ending at that point results in a goal. See Appendix A.2 for details of how impact is calculated.

Impact: Occupy point on the field which is maximally dangerous [3]. For example, a striker moves to a point directly in front of goal. (Figure 2)

Control: Maximize the amount of space controlled by the team [6]. For example, controlling the area in front of the penalty area, or the area between the defensive lines. (Figure 3a)

The mathematical details of the model are given in Appendix A.

With relation to the three zones defined above, we are going to (later in the article) test the hypothesis that the players in the mutual help zone primarily act to maximize a combination of pass probability and impact and that players in the co-operation zone attempt to maximize pitch control.

That the complex actions of a football player can be modelled in terms of these three criteria is a strong claim, and thus the first aim of this paper is to test this hypothesis: to what extent is it possible to model the movement patterns of players according to these three principles? To this end, we perform both a quantitative comparison to model predictions and a qualitative evaluation of the model by professional players and coaching staff.

To be useful in a coaching context and to improve their decision-making, the value of different actions should be straightforward to present to the players in post-match analysis. Thus, the second aim is to look at the insights that are gained from this model and to discuss these insights with players in a coaching intervention.

In order to simulate attacking situations our model should (given an initial state of players positions and velocities, and ball position) return an optimal position for all the attacking players based on a weighted combination of passing probability, impact and control. To do this, we make one further assumption: that all the attacking players can 'see' in to the very near future. In soccer (and other sports) players are continuously predicting the next movements of their opponents and the ball in order to gain advantage over them. Indeed, players often praise their teammates and players in other teams for their ability to see the game in the future. Moreover, there is evidence from studies of collective motion of pedestrians that humans make movement decisions based on the projected positions of others [27] and studies of cognition during sporting activities [28].

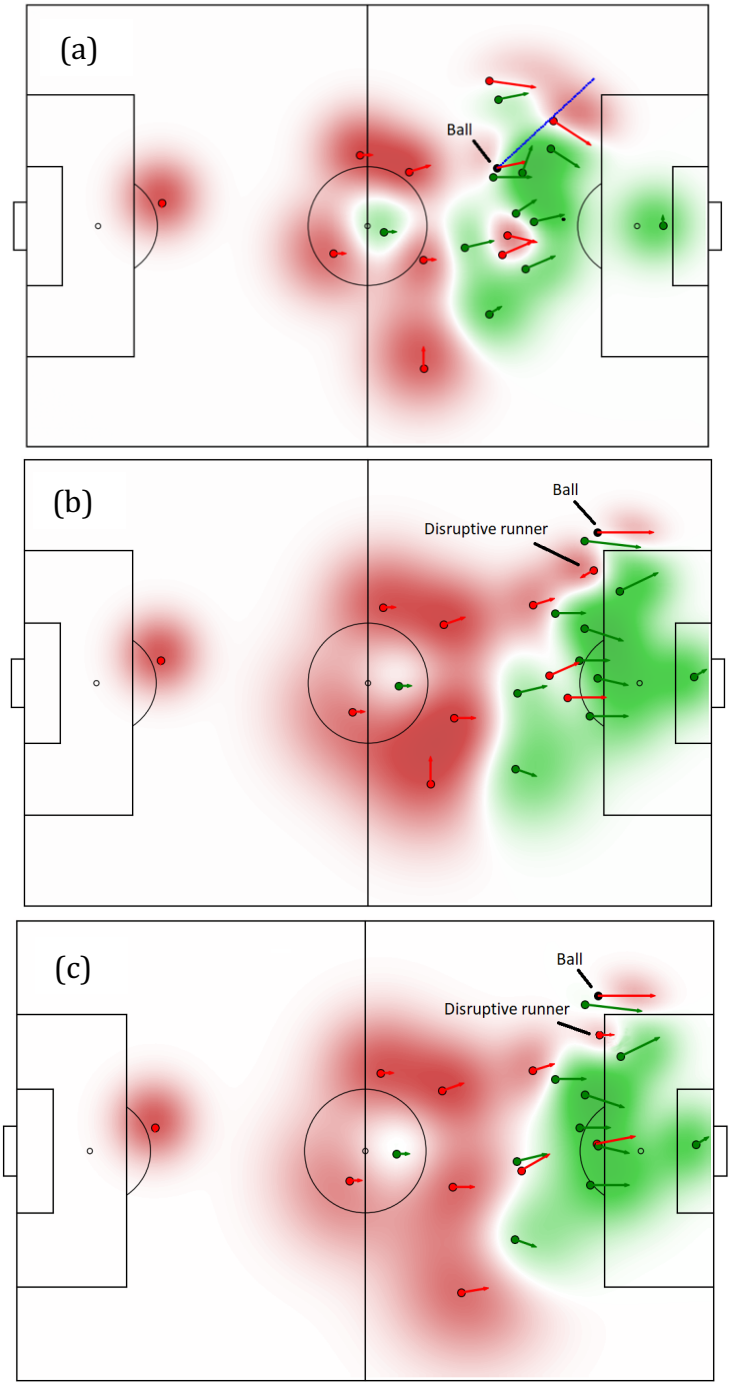


Figure 3: (a) Pitch control before the pass is played in the situation shown in figure 1 (b) Pitch control directly after the pass is received (c) As (b) but in this case the player labelled 'Disruptive runner' is positioned so as to maximize pitch control given the positions of the other players.

We thus base each player's optimization, not on the current position of all the players but instead on their future positions. For a focal player we ask, given that he knows the future positions of all his teammates and the opposition, where is the optimal position for him to stand to maximize a weighted of pass probability, impact and control. Figure 3b and 3c give the real and, respectively, simulated optimal positions for the disruptive runner in our example assuming he is trying to maximize pitch control. In this example, the optimal position is very close to the actual positioning, because in both cases the player occupies a small pocket of space behind the defense. For more details see the illustrative video at https://youtu.be/K-pP6W1FGIA.

In the next section we make a quantitative comparison between the simulation model. In section 4 we present visualizations and model simulations for players in order to make a qualitative comparison and to perform a coaching intervention.

3. Quantitative comparison of data and simulation

Our hypothesis was that the movements of players at different distances from the ball, i.e. different zones, would be predicted by different optimization criteria. We also expected players in different playing roles, i.e. attacking players, wingers, central midfielders, etc. to have different optimization criteria.

In order to evaluate the model quantitatively we identified all passes that started in the final third during two of Hammarby's home games (against Malmö FF and IFK Göteborg). For each player, at the time of each pass, we calculated the model's prediction according to seven different optimization criteria: pass probability (PP); pitch control (PC); pitch impact (PI); pass probability and pitch control (PP*PC); pass probability and pitch impact (PP*PI); pitch control and pitch impact (PC*PI): and all three criteria (PP*PC*PI). Notice that since PP, PC and PI are all probabilities, multiplying them together also gives a probability. For example, PP*PI is the probability of a pass being received and a goal resulting from that pass. As a control, we further defined an eighth optimization criteria of the player staying in the current position (CP). Each of these eight criteria was maximized over the duration of the pass, up to a maximum two second interval.

We computed the reachable areas for all the players that we want to simulate and sample 200 random points inside each of them. Equations (6) and (7) (in the Appendix) were used to compute the points on the pitch that the players can reach between the start and end frame. The only thing that we need to know, apart from the current player position and speed (which determines the center of the reachable area with Equation (6)) is the timespan between the start and end frame (which gives the radius of the area, as in Equation 7). For each reachable point we applied the eight optimization criteria and selected the point that optimized each of the criteria.

Figure 4a shows the distance in meters between the positions of the players predicted by three of the optimization criteria (PC, PP*PI and CR) and their actual position after a pass has been made. In this figure, the player's positions are ordered as a function of their distance from the target of the pass. So that the player who receives the pass has ordinal distance 1. The next nearest player to the pass target has ordinal distance 2 and so on, with the furthest away player assigned ordinal distance 10 (the player who makes the pass is excluded from the analysis).

For the Malmö match, the best predictor of the position of the player receiving the ball (ordinal distance 1) was PP*PI, outperforming both the PC model and the control model. Furthermore, for the 2 or 3 players closest to the ball, the PP*PI criteria best predicts the position of the players. For ordinal distances of 4 or greater the current position (CP) predicts positions best. For the Göteborg match, the PP*PI was only better than the CP model for the player receiving the ball, and only marginally so.

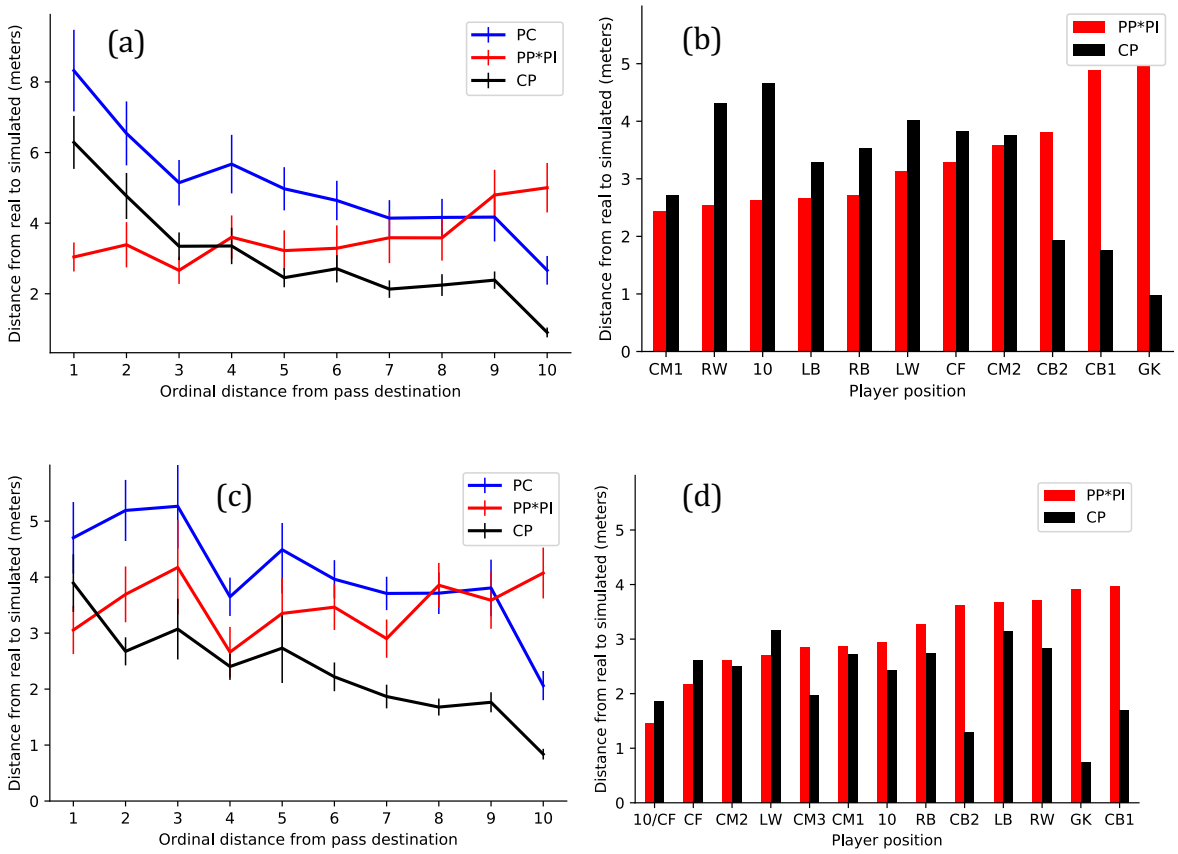


Figure 4: Comparison of positions predicted by model simulation and actual position. Error bars indicate standard errors. (a) Distance between model prediction for PC, PP*PI and CP (current position) models and actual position taken for 21 passes made by Hammarby in the final third against Malmö (b) Same data organized by player position in a 4231 formation (c) Distance between model prediction for PC, PP*PI and CP (current position) models and actual position taken for 41 passes made by Hammarby in the final third against Göteborg (d) Same data organized by player position.

In terms of our zonal model of football (Figure 1b) these results suggest that players in the mutual help zone tend to maximize PP*PI. That is, the players nearest the ball maximize the probability of both receiving a pass and of a goal resulting from that pass. The results also suggest that the mutual help zone typically might encompasses the nearest two or three players. The other models (not shown in the figure) give poorer predictions than PP*PI in most cases. We thus considered PP*PI the best model of the mutual help zone.

The attacking players (10, RW, LW) tend to be closest to the position that maximized PP*PI. Figure 4b, in the Malmö match it was the player playing the 10, Niko Djurdjic, and the right-winger (LW),

Vladimir Rodic, whose movements were best predicted by optimizing passes into high impact areas. In both matches (see also Figure 4d) the attacking players (for example, 10, RW, CF, LW) were better predicted by PP*PI than defending players (for example, CB1, CB2, GK).

Our hypothesis was that pitch control (PC) would provide a good model of players in the co-operation zone. This turned out not to be the case. The current position of a player was, at all ordinal distances, a better predictor of position than pitch control. Current position was also, in the majority of cases, a better predictor even when accounting for pitch impact (PC*PI). This result does not imply that standing still is the optimal model for players in the co-operation zone, instead it means that we have not yet identified a model that outperforms the control.

4. Qualitative comparison and coaching intervention

During the Allsvenskan season 2019 we used the visualization and simulations as part of coaching interventions within Hammarby's first team. Two of the co-authors met with the entire team, smaller groups or individual players and discussed various attacking situations. About ten structured meetings took place with a presentation made on a larger screen, along with about twenty unstructured meetings where one of the authors discussed performance, using visualizations on a laptop, with players over coffee or lunch. Further meetings took place with the head coach, Stefan Billborn, and assistant coach, Joachim Björklund, to discuss player performances.

We presented examples of match situations, analyzed using our model, to the players involved in them. These discussions were part of our first scientific aim, giving us an additional way of judging the realism of our model. Does it suggest passes and movements that real players make or consider reasonable? Such an assessment does involve a degree of subjectivity, but in modelling complex collective motion it is a very useful way of checking that a model makes sense [29,30]. Models that do not pass an 'eye test' are less likely to be useful in coaching interventions, simply because players and coaches will not give credence to their output.

The second scientific aim was to see if the approach provided a way of conducting post-match analysis. We asked both coaches and players at Hammarby involved in the simulated match situations to comment on what, if anything, they learned from the visualizations. The players discussed the output of the model between themselves and with two of the co-authors.

4.1. Passing probabilities

We started by presenting the pass possibility model for passes occurring within the mutual support zone. Figure 5 shows three pass situations in Hammarby's matches during the first half of the 2019 season, and three from later in the season. The first example (Figure 5a) is chosen because the optimal pass as identified in the model was different than that chosen by the player. In this example Kacaniklic (20), chose a pass with a low probability of success, to Tankovic (22), over a more likely to succeed pass to Kjartansson (17). Even accounting for pitch impact (PP*PI), the choice was suboptimal. The players, including Kacaniklic, and coaches who saw this situation were in unanimous agreement that he had made a poor decision and that the model was correct in its alternative suggestion.

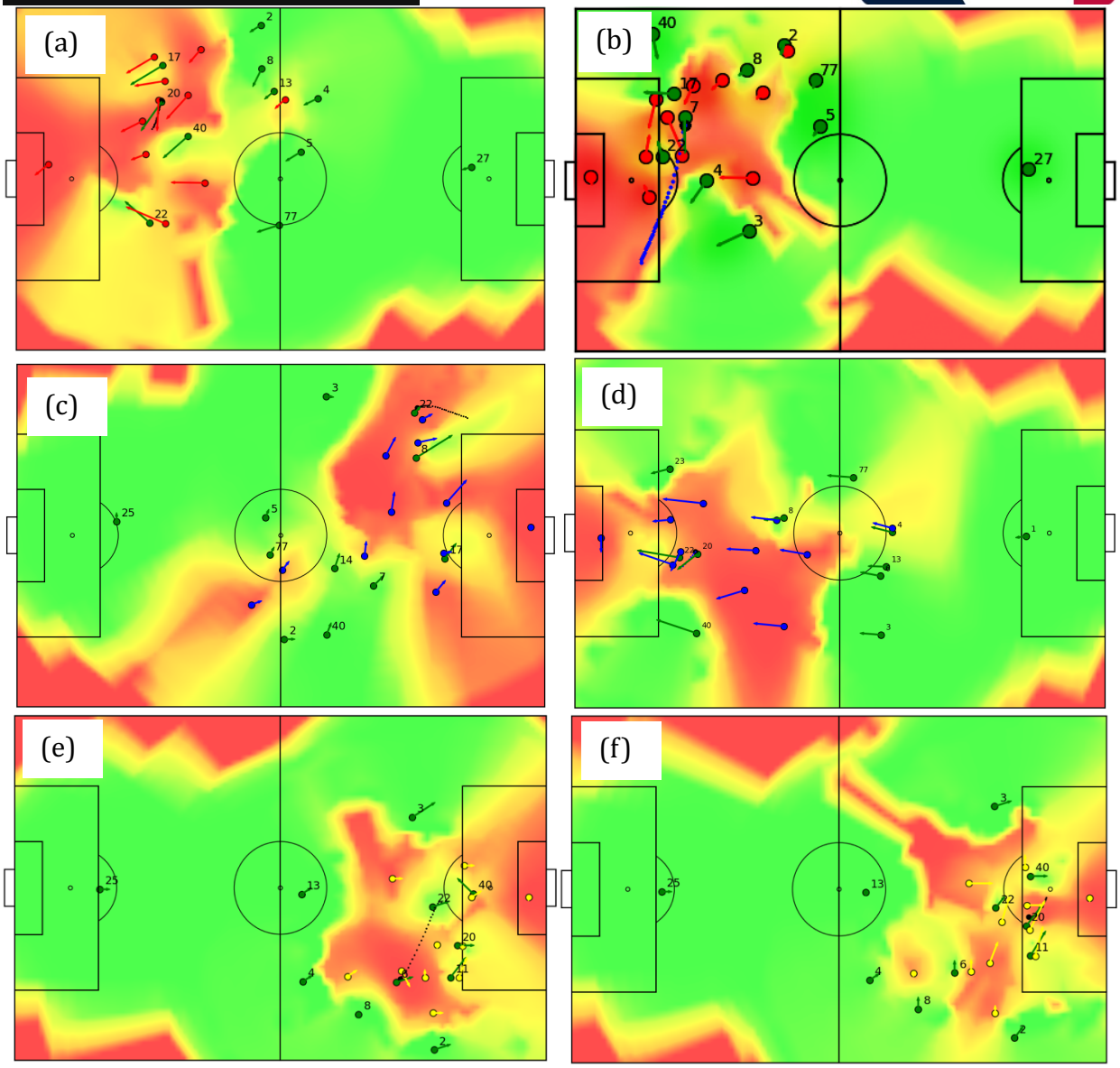


Figure 5: Pass decisions in four different situations for Hammarby during the 2019 season. Colur shows probability of a successful pass success probability, green 100% to red 0%, predicted by the model. Hammarby players represented with green circles, with player numbers. Dotted line traces balls trajectory one second in to the future. (a) Kacaniklic (20) failed pass to Tankovic (22) against Kalmar (red). (b) Khalili (7) failed pass to Widgren (3) against Östersunds. (c) Tankovic (22) successful pass to Andersen (8) against Djurgården. (d) Kacaniklic (20) successful pass to Djurdjic (40) against Sundsvall (e) Bojanic (8) successful pass to Tankovic (22) against Häcken (f) (22) Kacaniklic (20) assist to Djurdjic (40) against Häcken.

The pass made by Khalili (7) in figure 5b was analyzed on request of the player himself. In this case, Tankovic (22) had expressed frustration that Khalili should have passed him, instead of making the longer pass to the left back, Widgren (3). The model showed that if the actual pass Khalili made, had it reached its target, was the optimal choice in terms of pass success probability and impact. Talking

about this situation led to a more in-depth discussion about pass choice, in particular the higher impact of playing longer passes across the face of the box to change the direction of play.

The pass in figure 5c came late in the match between Hammarby and Djurgården. It is a more complex pass than those in 5a and 5b and was only made possible by the sudden acceleration of Andersen (8). Again, the players agreed that the model captured the opportunity which both players, passer and receiver, had created. In this case, the discussion around the pass focused on the optimal position of the players after it was played (see section 4.3).

In these examples and others, the players clearly understood plots of pass probabilities. Later in the season, when shown pass probability in figure 5d, Djurdjic (40), who received the pass, commented, "Alex (20) does that so well, holding up the ball and not playing the first, most obvious pass. This is what I keep telling the others. Be patient."

These comments and others, demonstrated that this particular group of attacking players could use images showing pass probabilities to discuss with coaches and with each other how they played. There was a feeling among the players, commentators and fans that their decision-making in these situations improved over the season. This is exemplified a combination of passes between Bojanic (6), Tankovic (22), Kacaniklic (20) and Djurdjic (40) in figure 5e and 5f, from the last match in the season.

4.2. Pitch control

Under our zonal model of soccer, the co-operation zone involves all players who don't have or cannot immediately receive the ball. The correct positions in this zone are, in theory at least, decided by the tactics set out by the manager. Working together with Hammarby head coach Stefan Billborn and assistant coach Joachim Björklund, we created a template for how we should defend in different situations, using a combination of pitch control and pitch impact as a guide.

Once this template was created we used it in discussions with the players. We cannot reveal the full template since it is an internal tactical document for the club, but we give one example where we could use pitch control to show how a player was out of position. Figure 6a shows the pitch control for both teams. The player of most interest in this situation is Vladimir Rodic (11), furthest from the ball. He is playing in a left-back position and although he controls a lot of area in this position, it is of very low impact. By showing Rodic figure 6b, in which Widgren (3) is positioned more centrally in a similar situation, we were able to explain how Rodic could improve his own positioning. In figure 6b, Widgren's more central position allows both Andersen (8) and Bojanic (6) to move closer to the ball and control areas with higher pitch impact. Rodic immediately saw the advantage of Widgren's positioning, and the difference between the two images, and said, "I didn't properly understand where I was meant to be in these situations, and now I do."

The left-back role was unusual for Rodic, partly explaining why he was incorrectly positioned. When the same comparison was shown to Widgren, the regular left back, he said, "Yes, of course, my role is to move in to midfield when we are in the final third."

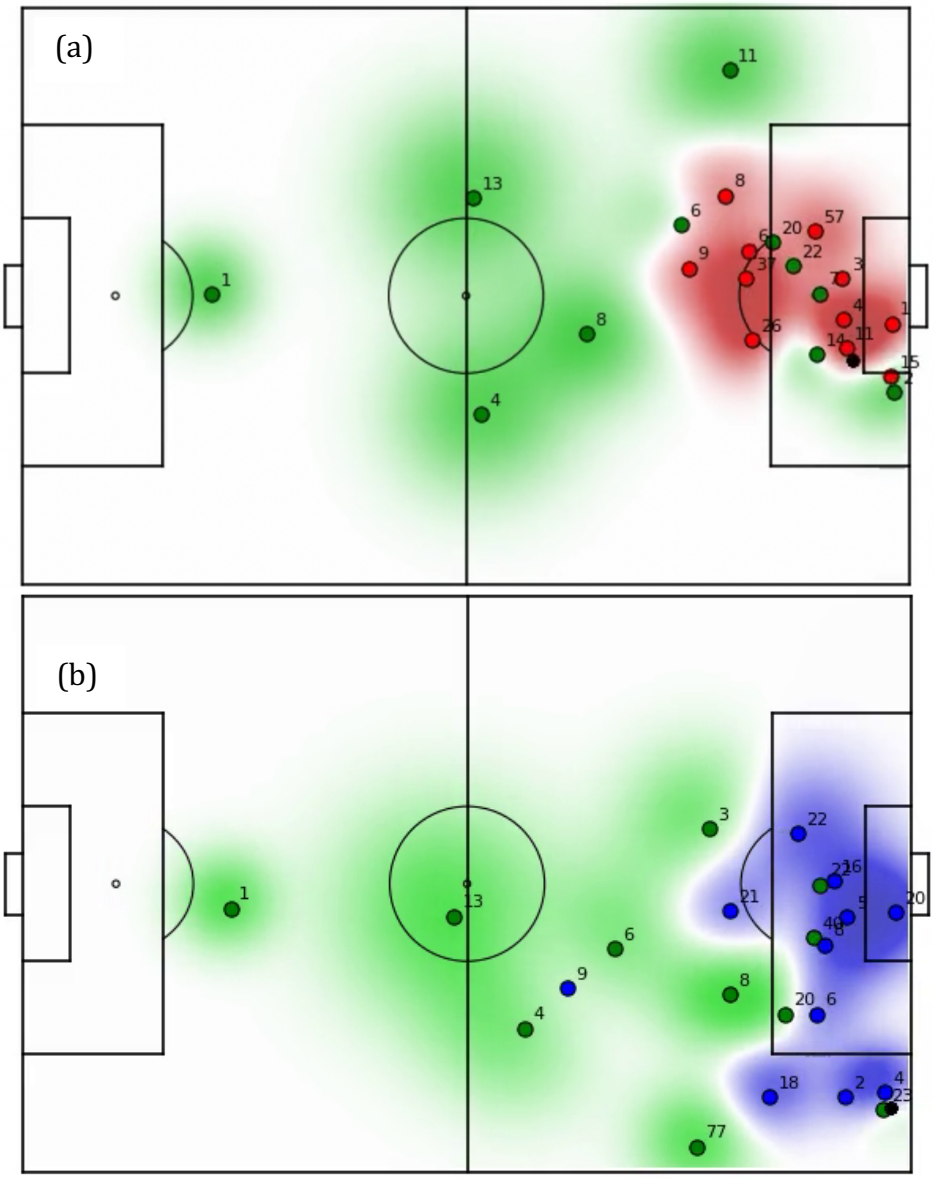


Figure 6: Pitch control in two different situations for Hammarby during the 2019 season. Colur shows pitch control, green 100% to blue/red 0%. Hammarby players represented with green circles, with player numbers. (a) Rodic (11) is poorly positioned in terms of controlling dangerous areas when playing against Helsingborg (b) Widgren (3) is better positioned in a similar situation against Sundsvall.

4.3. Simulating future positions

In the SPP simulations presented to the players the weightings (PP, PI and PC) were chosen manually, based on the zone each player is in. Automatically identifying the zone of a player is difficult, because it depends on context. Moreover, in section 3, we could not find any one simple optimization criteria which players tend to follow. Eventually, we envisage that player zones can be automatically

identified by an algorithm based on player positioning, but in order to get useable results in the current study we manually assigned the roles to the players.

Even though roles were assigned on a case-by-case basis we did develop some general rules of thumb by which we decided the criteria optimized in each simulation. Specifically, attacking players in the mutual help zone maximize pass probability multiplied by pitch impact (PP*PI) and players in the co-operation zone attempt to maximize either impact and control. In most cases, we would assign one or two players in the co-operation zone a striker role, and they maximize impact (i.e. find dangerous positions). The rest of the team either maximize control or do not maximize any criteria.

The simulations were particularly useful for allowing players to explore alternatives to the decisions they actually made when in the mutual support zone. Figure 7a shows an attacking situation, where the center forward, Kjartansson (17) has taken the ball in to the box. Figure 7b shows the same situation 1.5 seconds later, with the shaded player positions showing the optimal positioning according to the model. In this situation we made the following assumptions in modelling the players: Tankovic (22) and Kacaniklic (20) are in the mutual support zone and maximize pitch impact multiplied by passing probability; Djurdjic (40) is in the co-operation zone and optimizes pitch impact only; Söderström (13) and Andersen (8) are in the co-operation zone and aim to maximize pitch control; and Solheim (77) is assumed to minimize the opposition's pitch control.

For two players, Kacaniklic (20) and Djurdjic (40), the simulated positions are very close to the positions they actually adopted 1.5 seconds later. In the case of Kacaniklic, an opposition defender marked him and the simulation confirms that his run was the best he could do. The model's suggestion for Tankovic (22) differs from his actual run in to the box. The simulation indicates that a better choice would be for him to take a supporting position on the side of the box, where there are no opposition players and he could potentially receive a pass from Kjartansson (17). When presented to players and coaches this suggestion was considered reasonable, although it was by no means conclusive that Tankovic had made the incorrect decision.

The model also suggested that Söderström (13), Andersen (8) and Solheim (77) should all move further up the pitch than they did. The coaches agreed strongly with the model results in this case, and this became a focus for discussions with the full-backs and central midfielders during the season. They were encouraged to push up the pitch during attacking situations, even when they were not immediately involved in an attack.

Similar conclusions were drawn from other simulations of final third attacking simulations. One example is given in figure 7c and 7d. In this case, Tankovic (22), Khalili (7) and Kjartansson (17) are all considered to be in the mutual help zone, and thus maximize pitch impact multiplied by passing probability. Again, the model suggests that, Tankovic should run along the side of the box and that Khalili should move further forward. When these results were presented to Khalili he noted that there was only five minutes left in this match, a derby against rivals Djurgården, and Hammarby had a 2-1 lead. He decided in this case not to participate in the attack. Other players commented that, at this point, they were extremely tired, and pushing up the pitch was risky. In general, the simulations provided a way for the players to talk to each other and the coaches about their decisions in a more open and constructive way.

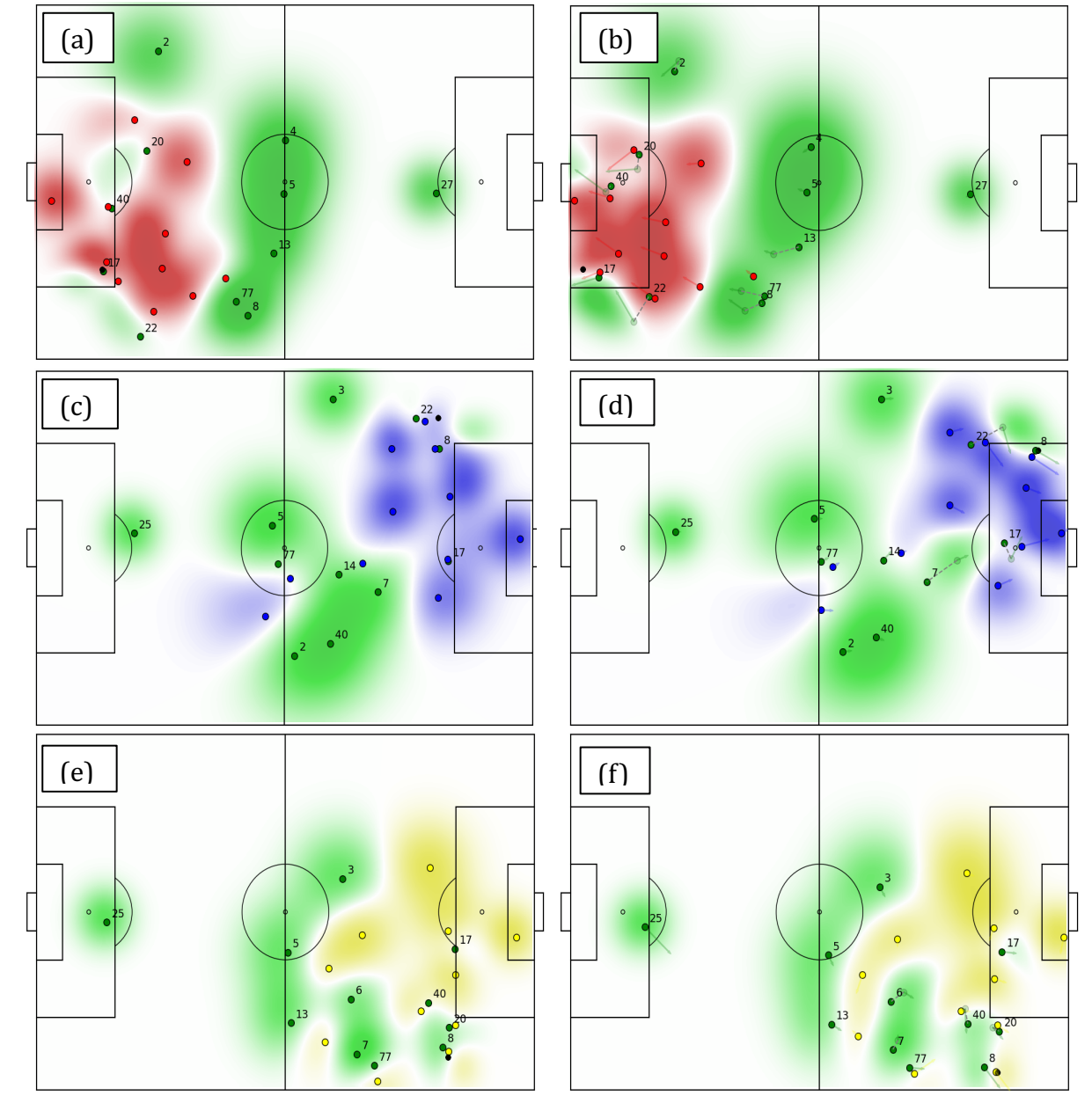


Figure 7: Simulations of three match situation for Hammarby during the 2019 season. Color shows pitch control predicted by the model before (a,c,e) and after (c,d,f) a pass is made. See main text for details of each of these situations. Optimal positions based on seeing 1.5 second in to the future are shown as 'shadows' while actual positions are filled in. Dotted line joins shadows to true position.

Another aspect of the game for which the simulations proved useful was in talking about press in the opponent's half. Figure 7e and 7f show the situation, respectively, before and after a 1 second simulation where opponents, Häcken, have the ball deep in their own half. Here, Khalili (7), Bojanic (6), Djurdjic (40) and Kacaniklic (20) are all assumed to optimize pitch control, to occupy as much of the field, and thus block passes, as much as possible. All four player positions are very close to the

optimal suggested by the model. As such, the model confirmed that a tactic the coaches knew to be effective was working as it should, and that all the players were adopting near optimal positions.

5. Effect of intervention

Coaching based on the output of a simulation model has never before been attempted in soccer. With that in mind, the main scientific aim of these coaching interventions was to judge the degree to which players could relate to, discuss and understand the visualizations and the outcome of the model. The experiment was not controlled in the sense that we could study the effect on players who did and didn't receive the intervention. Indeed, the environment of a soccer club meant a controlled experiment wasn't possible. Moreover, it was unavoidable that some players were more engaged than others, and thus received more feedback.

The visualizations and simulation, were further supplemented by visual aids during the coaching interventions. These included showing: attacking runs by Barcelona players; the passing patterns and positioning of Manchester City during sustained attacks; and shot maps showing the success of shooting (expected goals) from different positions around the penalty area.

Hammarby's play improved as the intervention season progressed, with the team scoring 33.9% more goals than any other team in Allsvenskan, and beating the goal scoring record for the league with an average of 2.5 goals per match. Niko Djurdjic was chosen as best forward in Allsvenskan. The intervention was not extensive enough, making up only a small fraction of the time the players spent training and talking about football, to be considered as a primary cause of this footballing success. The players and other aspects of coaching were much more important.

Our results do however point towards a relationship between a coaching staff that are open to assessing decision-making using data, players who are able to talk and reason about their own performance using visualizations and success in attacking play. At the end of the season, there was consensus throughout the club---from the chairman and the director of football, through the head coach and the players themselves---that the players could gain from the approach taken here. The first author of this article was offered an extended contract by the club.

6. Automated player feedback and other applications

Once we had familiarized the players with these types images and films, we created an automated system that shared videos of important moments within matches. The system automatically identifies situations where players performed actions which significantly decreased/increased their own team's pitch impact or reduced the opposition's pitch impact. For example, if they completed a pass with high impact (as illustrated in figure 2) then it would be added to a personalized highlight reel. After the match the highlight reel was shared with the players through our internal communication app, showing both their top 10 actions increasing pitch impact and bottom 3 decreasing pitch impact.

The highlights were shown first the clip in video form, then as a video of pitch control for the situation then finally a video of them both simultaneously. While the results of section 4 showed that players

do not typically appear to optimize pitch control, they did find videos of pitch control intuitive and useful visualizations of how they are controlling space.

This automated feedback is just one of many potential applications of this type of approach. Automatically identifying key moments in a match, then using a SPP model to find the optimal actions at those points in time, can give players a range of feedbacks on their positioning and decision-making. In particular, it can be used for "ghosting" or showing "what if" scenarios to players [33]. This can be particularly useful at an academy level or with younger, developing players. The approach can be made interactive, so players can see the effect of changing their positions. For example, asking "what would have happened if I had made this run instead?". In currently available 'coach paint' tools or in TV commentary, the user can 'move' a player and say what he or she could have done better, but this does not properly account for how the other players would have moved had the player moved in proposed manner. The SPP techniques described here can solve this type 'what if' problem by simulating all the other players according to the one-second-rule.

7. Conclusions

This article documents a first step in predicting and modelling the 'self-propelled' movement of soccer players. This approach is complementary to a purely data-driven one, where player movements are predicted by, for example, feeding trajectories in to a neural network [12]. The advantage of our approach is that it builds on existing coaching knowledge and the assumptions can easily be explained to coaches and players. Building all of this knowledge in to a model, validated against data, for a wide variety of match situations, is a large project of which this paper is a first step.

Here we have established that the zonal model, previously used informally by coaches, can offer a starting point for a simulation model. Specifically, the mutual help zone corresponds to finding the position that maximizes the probability of both receiving the pass with the maximum pitch impact. Both quantitative comparison of simulation and data and qualitative discussions with players confirms that the two or three players closest to the point to which a pass is delivered can be considered in the mutual help zone.

In the discussions with players and coaches, pitch control proved a useful concept for discussing tactics and positioning in the co-operation zone (Figure 6). It was not however an accurate model for movement of players in the co-operation zone when their teammates were passing the ball in the final third. Indeed, none of our models captured player movement in these situations better than simply standing still. The discussions with the players showed that pitch control could potentially prove a more useful model of co-operative movement when pressing (Figure 7). We envisage an approach to soccer

Although the approach taken here emphasizes simulation, the ultimate aim is not to build a complete soccer simulation. Rather, the aim of this research should be to build up a set of coaching tools that explain what players do and suggest alternative in certain situations. Creating these tools will involve a modelling cycle, very similar to that used in collective animal behavior, in which models of different situations are iteratively improved, both through quantitative and qualitative comparisons of data and model [20,21]. This will happen most efficiently if, as has been the case here, the research is conducted within professional football clubs.

References

[1] Pollard, Richard, and Charles Reep. Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)* 46, no. 4 (1997): 541-550.

[2] Decroos, Tom, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1851-1861. ACM, 2019.

[3] Rudd, Sarah. A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains." In *New England Symposium on Statistics in Sports*. http://nessis.org/nessis11/rudd.pdf

[4] Fernández, J., Bornn, L., & Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In *13th MIT Sloan Sports Analytics Conference*.

[5] 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*.

[6] Fernandez, J., & Bornn, L. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. In *Sloan Sports Analytics Conference* (Vol. 2018).

[7] Sumpter, D. (2016). *Soccermatics: mathematical adventures in the beautiful game*. Bloomsbury Publishing.

[8] https://github.com/statsbomb/open-data

[9] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific data, 6(1), 1-15.

[10] Spearman, W. (2018). Beyond expected goals. In Proceedings of the 12th MIT sloan sports analytics conference (pp. 1-17).

[11] Felsen, P., & Lucey, P. (2017). Body Shots: Analyzing Shooting Styles in the NBA using Body Pose. In MIT Sloan, Sports Analytics Conference.

[12] Le, H. M., Carr, P., Yue, Y., & Lucey, P. (2017). Data-driven ghosting using deep imitation learning. In MIT Sloan, Sports Analytics Conference.

[13] Mehrasa, N., Zhong, Y., Tung, F., Bornn, L., & Mori, G. (2018). Deep learning of player trajectory representations for team activity analysis. In *11th MIT Sloan Sports Analytics Conference*.

[14] Dick, U., & Brefeld, U. (2019). Learning to Rate Player Positioning in Soccer. Big data, 7(1), 71-82.

[15] Hobbs, J., Power, P., Sha, L., & Lucey, P. (2018). Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering. MIT Sloan Sports Analytics Conference.

[16] Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., & Shochet, O. (1995). Novel type of phase transition in a system of self-driven particles. *Physical review letters*, *75*(6), 1226.

[17] Couzin, I. D., Krause, J., James, R., Ruxton, G. D., & Franks, N. R. (2002). Collective memory and spatial sorting in animal groups. *Journal of theoretical biology*, *218*(1), 1-11.

[18] Sumpter, D. J. T. (2010). *Collective animal behavior*. Princeton University Press.

[19] Tunstrøm, Kolbjørn, Yael Katz, Christos C. Ioannou, Cristián Huepe, Matthew J. Lutz, and Iain D. Couzin. "Collective states, multistability and transitional behavior in schooling fish." *PLoS computational biology* 9, no. 2 (2013): e1002915.

[20] Sumpter, D. J., Mann, R. P., & Perna, A. (2012). The modelling cycle for collective animal behaviour. *Interface focus*, *2*(6), 764-773.

[21] King, A. J., Fehlmann, G., Biro, D., Ward, A. J., & Fürtbauer, I. (2018). Re-wilding collective behaviour: an ecological perspective. *Trends in ecology & evolution*, *33*(5), 347-357.

[22] Herbert-Read, J. E., Perna, A., Mann, R. P., Schaerf, T. M., Sumpter, D. J., & Ward, A. J. (2011). Inferring the rules of interaction of shoaling fish. *Proceedings of the National Academy of Sciences*, *108*(46), 18726-18731.

[23] Mann, R. P., Perna, A., Strömbom, D., Garnett, R., Herbert-Read, J. E., Sumpter, D. J., & Ward, A. J. (2013). Multi-scale inference of interaction rules in animal groups using Bayesian model selection. *PLoS computational biology*, *9*(3), e1002961.

[24] Strandburg-Peshkin, A., Twomey, C. R., Bode, N. W., Kao, A. B., Katz, Y., Ioannou, C. C., ... & Couzin, I. D. (2013). Visual sensory networks and effective information transfer in animal groups. *Current Biology*, *23*(17), R709-R711.

[25] Buldú, J. M., Busquets, J., Echegoyen, I. & Seirul-lo, F. (2019). Defining a historic football team: Using Network Science to analyze Guardiola's FC Barcelona. *Scientific reports*, *9*(1), 1-14.

[26] Seirul·lo, F. (2010). Estructura sociafectiva. Documento INEFC – Barcelona. http://www.motricidadhumana.com/estructura_socioafectiva_doc_seirul_lo_Outline_drn.pdf

[27] Moussaïd, M., Helbing, D., & Theraulaz, G. (2011). How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences*, *108*(17), 6884-6888.

[28] Körding, Konrad P., and Daniel M. Wolpert. Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences* 10, no. 7 (2006): 319-326.

[29] Herbert-Read, J. E., Romenskyy, M., & Sumpter, D. J. (2015). A Turing test for collective motion. *Biology letters*, *11*(12), 20150674.

[30] Webster, J., & Amos, M. (2019). A Turing Test for Crowds. *arXiv preprint arXiv:1911.06783*.

[31] Cavcar, M. (2000). The international standard atmosphere (ISA). *Anadolu University, Turkey*, *30*, 9.

[32] Z. Lowe. "Lights, Cameras, Revolution". *Grantland*. 19 Mar 2013.

Appendix A: Mathematical model

Here we describe the mathematical details underlying each of the three maximization criteria outlined above.

A.1 Pass probability

Spearman et al. proposed a probabilistic approach to model the probability of success of a pass [5]. We base our work on an implementation of their model, with a few changes that improve the ball dynamics. In Spearman et al. the ball movement is modelled using a pure ballistic approach, i.e., only aerodynamic drag is considered to be responsible of the natural deceleration of the ball during a pass. The general formula of aerodynamic drag is used with no Magnus force giving an equation of motion for the ball of

$$\ddot{\vec{r}}_{aero} = -\frac{1}{2m}\rho C_D A \dot{r}\dot{\vec{r}} \tag{1}$$

where $m = 0.42\ kg$ is the mass of the ball, $\rho = 1.225\ kg/m^3$ is the density of the air [31], $C_D = 0.25$ is the so-called drag coefficient and $A = 0.038\ m^2$ is the cross section area of the ball. The omission of the Magnus force, which alters the trajectory as a result of differences in pressure on opposite sides of the object, is made primarily because of the lack of data about the spinning movement of the ball.

Since many of the important passes in our study are ground passes, during which the ball is almost all the time in close contact with the grass, friction needs to be added to our model. The equation of motion here is

$$\ddot{\vec{r}} = -\mu g\, \hat{r} \tag{2}$$

Due to the changing conditions of the pitch and the unavailability of data, we settled on an estimate of a value of $\mu = 0.55$, which is a value in the middle of the interval that FIFA recommends for high quality artificial grass surfaces.

To determine how the forces act along the whole trajectory of the ball we performed some trial and error experiments with passes in which the ball did not get too high over the pitch (at most 10 or 20 cm), so that they can be considered to be ground passes. Comparing the real trajectory of the ball to simulations, indicated that the first two thirds of its trajectory was mainly caused by the aerodynamic drag and that ball-grass friction was the main force acting during the last third of the pass. Putting together both forces with their respective time intervals, the final equation of motion for the ball we used is

$$\ddot{\vec{r}} = \begin{cases} -\frac{1}{2m}\rho C_D A \dot{r}\dot{\vec{r}}, & t \leq \frac{2t_{max}}{3} \\ -\mu g\, \hat{r}, & t > \frac{2t_{max}}{3} \end{cases} \tag{3}$$

As should be clear from our discussion above, there is certainly scope for improving a model of ball dynamics.

To model time taken for a player to intercept the ball in a certain pass, Spearman et al. solved an equation of motion for all the players, with two constraints which set limits to the maximum players' speed and acceleration. For our purposes however to solve a minimization problem for all the possible points along the trajectory of the pass would be too computationally expensive. So, we adopted the model of Fujimura and Sugihara in 2005 in which they consider the players as objects whose movements are described by an equation of motion with a driving force (which represents the force exerted by the players' legs) and a drag force (which bounds their maximum possible speed). This gives the following equation

$$m\frac{d}{dt}\vec{v} = \vec{F} - k\vec{v} \tag{4}$$

whose solution is given by:

$$\vec{x} - \vec{x}_0 = V_{max}\left(t - \frac{1 - e^{-\alpha t}}{\alpha}\right)\vec{e} + \frac{1 - e^{-\alpha t}}{\alpha}\vec{v}_0 \tag{5}$$

Where $V_{max} = F/k = 7.8\, m/s$ is the maximum velocity that a player can reach, $\alpha = k/m = 1.3$ is the magnitude of the resistance force and $\vec{e}$ is the unit vector that denotes the direction of the acceleration of the player. The values for these constants are the same ones as Fujimura and Sugihara used in their paper and were obtained by performing a study with several field hockey players.

With this result, it is possible to see that all the points that a player with starting position $\vec{x}_0$ and initial velocity $\vec{v}_0$ can reach are enclosed inside the circle with center

$$\vec{x}_0 + \frac{1 - e^{-\alpha t}}{\alpha}\vec{v}_0 \tag{6}$$

and radius

$$V_{max}\left(t - \frac{1 - e^{-\alpha t}}{\alpha}\right) \tag{7}$$

This makes the finding of the interception times for the players easier than with the minimization problem that Spearman proposed. In order to obtain them the time must be discretized (we do that in steps of 0.04 seconds) and, for each time step, we check the already computed position of the ball and calculate the reachable area of the player. If the current ball position falls outside the circle, we advance to the next time step and repeat the process until the ball is in inside the player's reachable area; that moment determines the interception time.

Once the physical models behind the ball and the players have been established, the probabilistic model proposed by Spearman et al. can be used. Its main feature is the usage of a logistic distribution to determine the probability of a player getting the ball at time T knowing his arrival time t.

$$P_{int} = \frac{1}{1 + e^{\frac{T - t_{int}}{\sqrt{3}\sigma/\pi}}} \tag{8}$$

Note that this function does not compute the probability for a certain player to get the ball during a pass, but the probability of him being able to intercept the ball after T seconds (without considering the rest of the players). Furthermore, another consideration that is made is that a player has to be in the vicinity of the ball for a certain time in order to have control over it, this is modelled with the term:

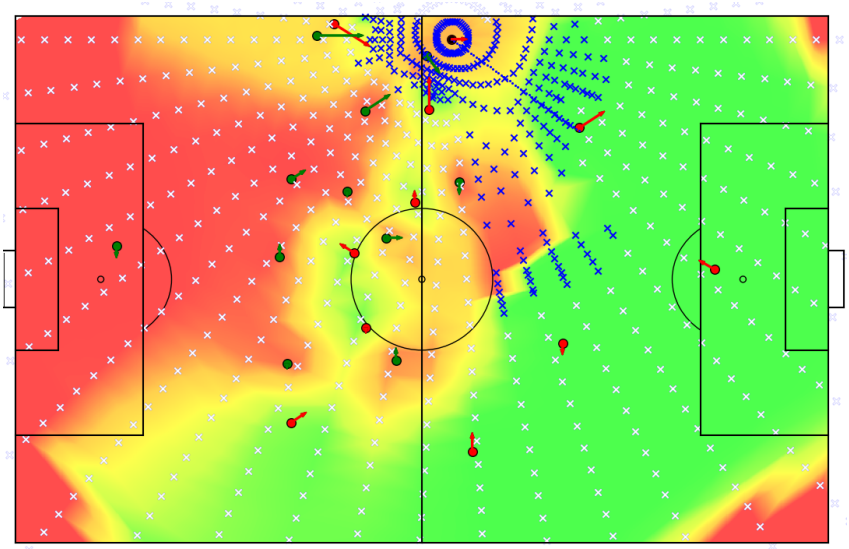


Figure A1: Pass probabilities and interception points inside (with blue markers) and outside (white markers) the possible ground-pass area around the ball.

$$P(t) = 1 - e^{-\lambda t} \tag{9}$$

With the combination of these two, the final system of differential equations that gives the probability for each player to receive the pass is built as follows,

$$\frac{dP_j}{dT}(T) = \left(1 - \sum_k P_k(T)\right) P_{int,j}(T)\lambda \tag{10}$$

In order to analyse the pass probabilities for any given match moment a, total of 750 potential passes are Calculated, distributed over 50 angles all around the ball and 15 pass speeds between 1 and 20 m/s. For each of these passes the probabilities of pass success together with their respective interception point, i.e., the location on the pitch where it is most likely to be received, are computed and a heatmap is generated showing the probability that a teammate of the player passing the ball will receive it. Figure 1c shows the points calculated in one example in a match between FC Barcelona and Real Betis.

Since the simulations that we perform are based on ground passes, it is clear that there will be some areas on the pitch (mainly all of the points that lay behind a player) that cannot be reached with one

of these passes, either because the ball is always intercepted before it gets there or because a really strong pass is needed for the ball to get there and there is no possibility of interception due to its speed. The points that bound the reachable area with a ground pass are what we call the "last interception points".

A.2 Pitch control

Pitch control was proposed by Javier Fernández and Luke Bornn in [6]. In their work they follow a different approach than Spearman's, basing it on what they call "player influence areas" instead of arrival times. The player influence at a certain point on the pitch p at time t is determined by the position and speed of the player and defined by:

$$I_i(p,t) = \frac{f_i(p,\ t)}{f_i(p_i(t),\ t)} \tag{11}$$

Where,

$$f_i(p,t) = \frac{1}{\sqrt{(2\pi)^2 det[COV_i(t)]}} exp\left(-\frac{1}{2}\left(p-\mu_i(\vec{s}_i(t))\right)^T COV_i(t)^{-1}(p-\mu(t))\right) \tag{12}$$

Pitch control is also a way of mimicking long passes in an easier and, possibly, more trustful way than simulating the trajectory of a long ball in the air for two main reasons: the first one is that factors that are unknown with the datasets that we use like wind speed and direction or ball spin play an important role in these passes and, even if we could perfectly model the flight of the ball, not all the players have the same skills when sending high passes and modelling this "player accuracy factor" properly would be almost impossible.

We thus also use pitch control to extrapolate the pass probability model in section A.1. A grid of points is created outside the zone that is already covered by possible ground passes and, for each of them, we calculate pitch control. Figure A1 shows an example of this extrapolation for a frame of the same game between FC Barcelona and Real Betis.

A.3 Pitch impact

Attacking players, in particular strikers playing CF or 10 roles, are often instructed by coaches to take specific dangerous positions even though these are not reachable by a pass from the position the ball is now. These positions are often directly in front of goal, or at one of the posts, where a rebounding ball might provide a good-quality shot opportunity, but they don't necessarily need to be positions which have high pitch control or where a pass is currently possible. The idea is to be well placed for the second ball.

To measure impact, we used a model developed by the company Twelve using event data from three historical seasons of the Premier League, La Liga and Champions League (output of this model can be found at twelve.football/analytics). All the matches used to train the model are broken down into sequences of possession, i.e., fragments of the game during which one of the teams holds the

possession of the ball without losing the ball and without any stops in the play (due to fouls, throw-ins, offsides, etc.). A chain was considered broken and another chain begun whenever the opposition team made two consecutive touches of the ball.

Once all actions are allocated to a possession chain then two logistic regressions are fitted in order to assign a value to each pass. The first regression is obtained by assigning each possession chain a value between 0 (if the play ends without a shot) and 1 (if the sequence finishes with a goal). This gives the probability of a pass (defined by its starting and ending co-ordinates on the pitch) leading to a shot. A second regression is then used to compute the probability of a shot leading to a goal (i.e. to obtain the expected goal value of the shot). Multiplying these two probabilities for every shot gives the probability that a pass of with certain starting and ending co-ordinates and qualifiers is likely to result in a goal. It is this value which we call the pass impact.

Figure 3 shows two examples of the pass impact for two different starting coordinates of the pass. A limitation of this method is that the most valuable point on the pitch is always the point nearest to goal. In practice, however, these positions tend to be offside and thus not of maximum impact.