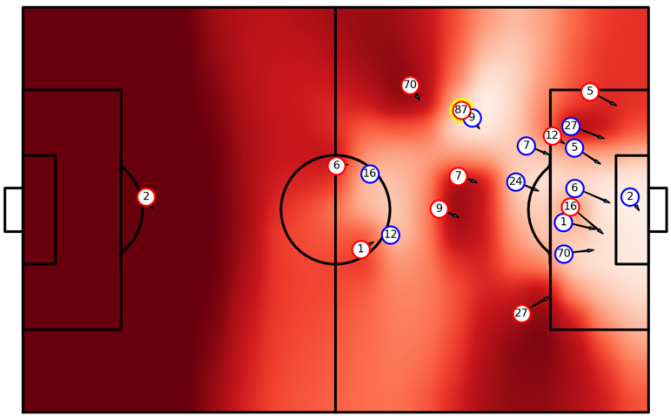

Figure 1. The example of the potential pitch control field (PPCF) is displayed above. Note that in reality, the PPCF is a three dimensional with the third dimension representing the player. In this example, it has been integrated over all players on the Red team. Red #87 controls the ball as highlighted in yellow. Dark red indicates strong control by the red attacking team (corresponding to a PPCF value of 1) while white indicates control by the blue team (corresponding to a PPCF value of 0). Players are represented by colored circles with the anonymized jersey number in the center. Player velocities are represented by vectors emanating from the corresponding player circle.

In Figure 1, we can identify a few important features of the *PPCF*. Notice how Red #9 and Red #7 are moving forward at speed. This opens up the space behind them which they no longer control due to the time it would take them to turn around. Red #5 is offside and thus is unable to legally influence play. Blue #27 is running towards the goal which opens up space left of the goal for Red #12 to control. The midfield region of the pitch is less dark indicating the potential for a counter attack by Blue #16 and Blue #12 if a turnover were to occur.

The differential equation used to compute the control probability for each player at a specified location, r, at time t is:

$$\frac{dPPCF_j}{dT}\left(t,\vec{r},T|s,\lambda_j\right) = \left(1-\sum_k PPCF_k\left(t,\vec{r},T|s,\lambda_j\right)\right) f_j(t,\vec{r},T|s)\lambda_j \qquad (3)$$

Where $f_j(t,\vec{r},T|s)$ represents the probability that player j at time t can reach location r within some time T. This term is addressed in more detail in Section 3.1.1. We incorporate time of flight of the ball by setting $PPCF_i(t,\vec{r},T|s,\lambda_i) = 0$ when T is less than the time of flight of the ball at location r. These flight times are discussed in more detail in Section 3.1.2. The parameter λ_i is the rate of control and represents the inverse of the mean time it would take a player to make a controlled touch on the ball. This parameter is discussed further in Section 3.1.3.

Integrating Equation 2 over T from 0 to ∞, we build a per-player probability of control. The resultant *PPCF* is three dimensions with two spatial dimensions on the pitch and a third dimension for the players. Generally, when visualizing the *PPCF*, we integrate over the players of the attacking team.

3.1.1. Time to Intercept

The probability that a player will be able to intercept the ball at a given location on the pitch within some time, T, is given by the term $f_j(t, \vec{r}, T|s)$ in Equation 3. To model this, we first compute the *expected* intercept time, $\tau_{exp}(t, \vec{r})$. This is done by finding the time it would take player j to reach location $\vec{r}$ from their start location $\vec{r}_j(t)$ with a starting velocity $\vec{v}_j(t)$ assuming they are able to accelerate with some constant acceleration a to a maximum speed v. The values of a and v should be chosen to correspond with the average sustainable speed and acceleration during a match. For this analysis, the values of 5 m/s and 7 m/s^2 are chosen respectively.[3] Numerous factors can lead to the *true* intercept time, $\tau_{true}(t, \vec{r})$, showing significant variance from this computed *expected* value, these include: tracking data inaccuracies, player facing, player awareness, tactical decision making, and other factors. To avoid the explicit modeling of these and other issues, we model the distribution of residuals, $\tau_{exp}(t, \vec{r}) - \tau_{true}(t, \vec{r})$, using the logistic function (we choose the logistic over the normal because of its heavier tails). This allows us to compute the probability that the player will be able intercept the ball using the cumulative distribution function of the logistic:

$$f_j(t, \vec{r}, T|s) = \left[1 + e^{-\pi \frac{T - \tau_{exp}(t,\vec{r})}{\sqrt{3}s}}\right]^{-1} \quad (4)$$

3.1.2. Time of Flight

To compute the time of flight of the ball, we simulate different trajectories using aerodynamic drag with a coefficient of drag that exhibits non-linear dependence on the speed of the ball as described in [9]. By varying the angle and speed of the ball, we can understand the minimum and maximum time it would take the ball to travel a certain distance.

When building the *PPCF*, we select the time of flight that most closely matches the arrival time of the nearest attacking player. This modeling choice advantages successful passing and helps to counteract the limitations of the decision model in the next section.

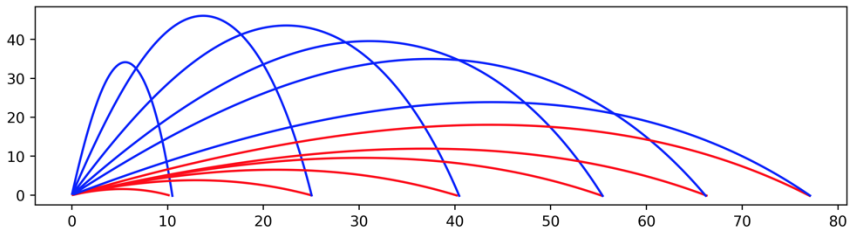

Figure 2. This figure illustrates possible ball trajectories (all units are in meters). Trajectories with a large flight time are shown in blue while trajectories with a smaller flight time are shown in red.

Although we do not explicitly model blocked passes, the control exerted by defenders through the *PPCF* will serve as a proxy for balls blocked near the receiver. Because long balls are likely to be lobbed, and therefore unable to be intercepted through the majority of their trajectory, this

[3] There is no reason that a and v cannot be set on a per-player basis allowing for more intelligent modeling of break-away situations with a fast attacking player, but for simplicity, we use fixed values of a and v.

simplification is largely justified. It bears mention that distant lobbed passes are more difficult to execute precisely than short ones. Although this is not modeled explicitly in the *PPCF*, it is treated by proxy in Section 3.2 where we utilize a normal distribution to disadvantage long-passing.

3.1.3. Control Rate

The parameter, λ_i is the control rate and has units of $1/\text{seconds}$. Higher values of λ_i indicate less time is required before the player is able to control the ball. In the context of computing opportunities, we would expect that this control rate might be different for attacking players and defending players. Attacking players want to make a precise controlled touch that results in a shot or continued possession while a defender will often be satisfied with heading the ball away or kicking it out of play. To account for this, we introduce a second parameter, κ, which scales the attacking control rate, λ. Thus, we can write the following expression for λ_i where A is the set of attacking players and B is the set of defending players.

$$\lambda_i = \begin{cases} \lambda & i \in A \\ \kappa\lambda & i \in B \end{cases} \tag{5}$$

Note that the parameter, λ_i is set to zero for attacking players when that attacking player is in an offside position.

3.2. Transition Model

The final term in in Equation 2 quantifies the likelihood that the next on-ball moment will occur at an arbitrary point $\vec{r}$. Figure 3 shows a histogram of the average displacement, $\Delta\vec{r} = \vec{r}_{i+1} - \vec{r}_i$, between subsequent on-ball events in data.

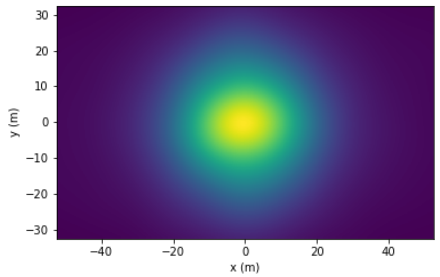

Figure 3. A 2D histogram of the relative location of subsequent ball touches, $\vec{r}_{i+1} - \vec{r}_i$. The standard deviation of the average displacement is 14 m.

Given that the ball is moved around the pitch through collisions with players (e.g. passes, headers, blocks, interceptions, etc.) it makes sense that the ball's motion will appear to be a form of two-dimensional Brownian motion and that the distribution of displacements between subsequent events will be normally distributed in aggregate.[4]

[4] Another way to conceptualize this is that players are more likely to attempt easy short passes and if they attempt a long pass, the resultant location will have higher variance due to the implicit angular variance present when passing.

Although, on average, the distribution of displacements for subsequent on-ball events may be normally distributed, we would expect there to be intelligent decision making on the part of passers. Passers are likely to select passes that are less likely to be intercepted. As we have already constructed a model that describes the probability that a pass to a given spatial location will be successful, we can superimpose these models and construct a decision probability density field using the following equation:

$$\mathrm{T}(t, \vec{r}|\sigma, \alpha) = \mathrm{N}(\vec{r}, \vec{r}_b(t), \sigma) \cdot \left[\sum_{k \in A} PPCF_k(t, \vec{r})\right]^{\alpha} \quad (6)$$

In this expression, A represents the set of all players from the team in possession, α is a model parameter used to scale the dependence of the decision conditional probability by the *PPCF* and N is a two-dimensional normal distribution.[5] The expression in Equation 6 is normalized to unity.

3.3. Score Model

The first conditional probability in Equation 2, $P(S_r|C_r, T_r, D)$, represents the likelihood of scoring from a location r assuming that the ball is successfully controlled at that location by the attacking team. To simplify our model, we ignore game state, D, for this term and base the model solely on distance to the goal. Our assumption is that defensive positioning will be proxied through use of the *PPCF* in the other conditional probability models.

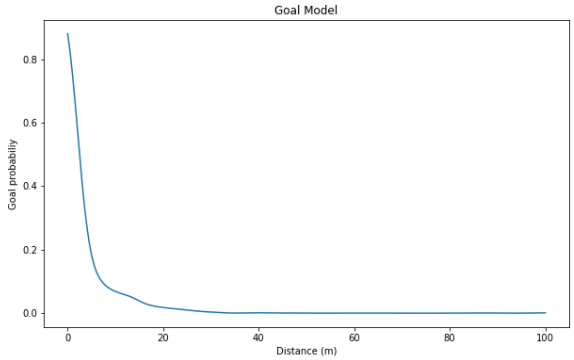

Figure 4. The probability of scoring plotted against the distance from goal. Gaussian kernel smoothing is used to ensure a continuous distribution.

Event data is used to compute the probability of scoring a goal conditioned on distance to the goal. The functional dependence seen in Figure 4 represents the average scoring chance given an on-ball event at a particular distance from goal. It is likely that this will not represent the true distribution due to an inherent shot selection bias. For example, we would expect that if unpressured, a player is more likely to shoot from and score at *20 m* than the average player making a ball-touch *20 m* from

[5] One term conspicuously absent from the next on-ball event probability density is a preference for events that move the ball closer to the target goal. In future variants of this analysis, we intend to incorporate such a term. For now, however, the failure of the model to incorporate this information means that opportunities that may appear as clear scoring chance to a domain expert may be under-estimated by the model.

the goal. To allow the data to account for this, we add a model parameter, β, that permits the shape of the above distribution to vary while maintaining the monotonic decreasing behavior we expect:

$$S(\vec{r}|\beta) = \left[S_d(|\vec{r} - \vec{r}_g|)\right]^{\beta} \quad (7)$$

In Equation 7, $\vec{r}_g$ is the location of the target goal and $S_d(x)$ represents the data-derived function displayed in Figure 4.

3.4. Combination

Using Equation 2, we combine the conditional probabilities to give a single, unified model which represents the posterior probability of scoring with the next on-ball event at a particular location. The constituent conditional probability and probability density maps are represented spatially in Figure 5. Note that Figure 5a and Figure 5b are probability maps while Figure 5c and Figure 5d are spatial probability densities which must be spatially integrated to be interpretable as a probability.

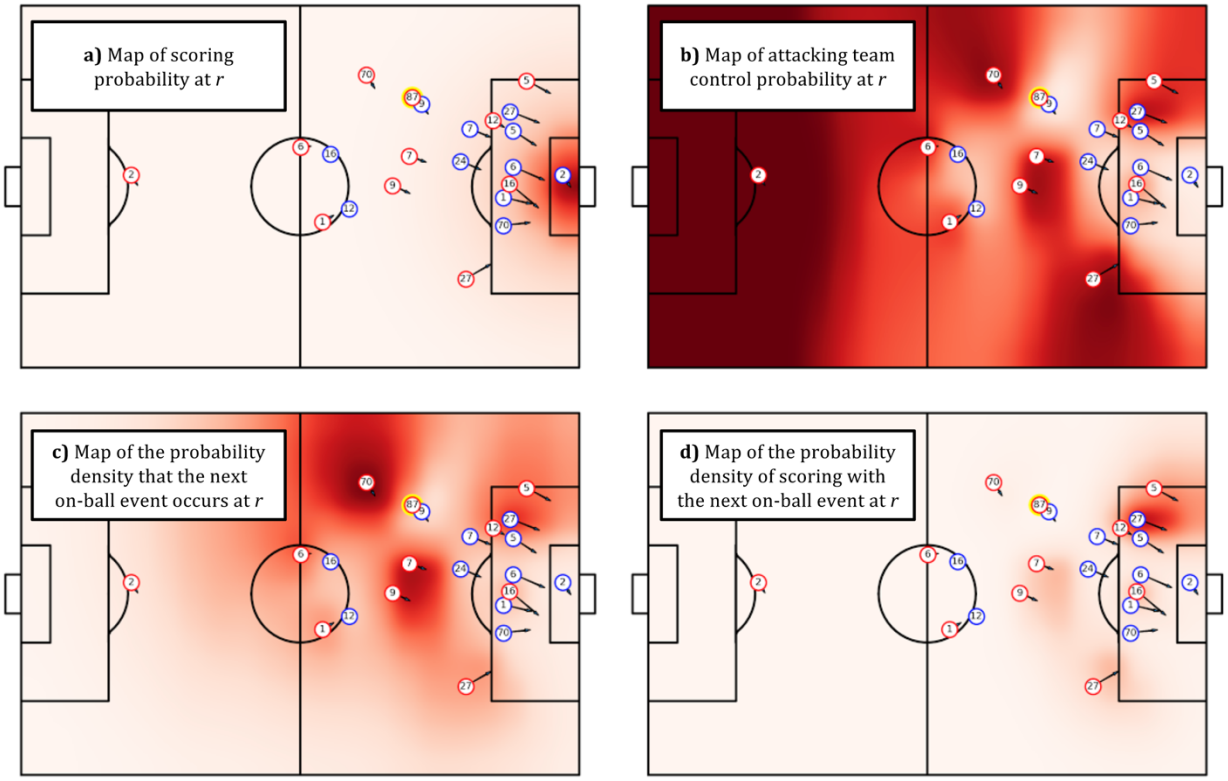

Figure 5. a) The scoring model (no spatial normalization, dark red corresponds to unity). b) The control model: probability of the attacking team controlling the ball at a given location assuming the next on-ball event occurs there (no spatial normalization, dark red corresponds to unity). c) Transition model: probability density of the location of the next on ball-event (normalized to unity) d) Off-ball scoring opportunity model: probability density of scoring with the next on-ball event at the specified location. Red #87 has the ball. Red #5 is offside and is not included in the computation. The integrated magnitude of the OBSO is 1.1%.