Soccer analytics

Unravelling the complexity of "the beautiful game"

While national soccer teams compete this June to win the World Cup, analysts will be crunching data behind the scenes to pursue an advantage. **Luke Bornn**, **Dan Cervone** and **Javier Fernandez** explore the evolution of soccer analytics

With the soccer World Cup due to kick off in June, many fans will be scrutinising the schedules to assess their country's chances. By analysing squad strengths and weaknesses, formations, injuries, and other aspects of the tournament, fans and media alike will be engaging with the competition long before the first kickoff. How does one country's "tiki-taka" style match up to another's strong defensive pressure? How will one country's young goalkeeper fare against another's world-class striker?

This sort of analytical engagement with soccer is quite common among fans, commentators and pundits – it is part and parcel of the game. But soccer's team management and strategy is far from being recognised as analytics-driven. This is not the case for sports such as baseball and basketball, where analytics is a well-established discipline. It is 15 years since Michael Lewis's *Moneyball*[1] and Dean Oliver's *Basketball on Paper*[2] were published. However, the first mainstream book-length treatment of soccer analytics – *The Numbers Game*[3] by Chris Anderson and David Sally – arrived only five years ago.

While the delay in the soccer community's acceptance of quantitative metrics might be attributed to its traditional and well-entrenched culture, or its semantic and geographical distance from baseball's analytical roots, the data are also to blame.

The problem for soccer is this: data across all major team sports have focused on what is happening to the ball throughout the game. In baseball, for instance, tracking ball events alone captures nearly every impactful moment in the game. In basketball even, while off-ball actions such as defensive positioning do affect game play, on-ball events provide the data that comprise modern metrics (such as player efficiency ratings). In other words, all these metrics rely on is aggregate counts of how a particular play began and ended, with basic on-ball information such as who took the shot and whether it was made or missed.

In soccer, however, on-ball actions provide less insight into strategy and player evaluation. Indeed, soccer games are often won and lost away from the ball. As soccer legend Johan Cruyff explained: "When you play a match, it is statistically proven that players actually have the ball three minutes on average … So, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you're a good player or not."

vverve/Bigstock.com

ABOVE Player-tracking camera installation. Courtesy of Metrica Sports.

This might explain why soccer analytics was slow to get going. Early research focused on game-level events such as home advantage, the prevalence of drawn games, and the relative strength of individual teams and leagues. Later, the collection of data for on-ball events led to analyses of possession ratios and the use of long-ball tactics.[4] Now teams have the ability to track player movements over the whole pitch, throughout the course of a game. This has enabled analysts and coaches to take their eyes off the ball, quite literally – leading to breakthroughs in soccer analytics for team management and strategising. This year, for the first time, FIFA will be providing all 2018 World Cup teams with real-time player-tracking data, allowing them to make tactical or personnel changes informed by data (bit.ly/2HGesf8).

Where are they now?

In all dynamic sports, player positioning, movement, and off-ball actions are important for creating scoring chances. In basketball, for example, attacking players "screen" or "pick" to create space for teammates away from the ball by blocking the movement of defensive players. One of the most common techniques in soccer, known as an "overload", involves using the ball to draw defenders to one side of the pitch with the primary aim of creating space and opportunity on the other side. To be effective, off-ball teammates must be prepared to capitalise on this newly created space; hence the interplay between spatial creation and spatial exploitation comprises much of modern soccer tactics.

Player-tracking data helps inform these tactics by providing time series of the locations of each player, and the ball, throughout the game. These data can be – and have been – used in multiple ways, including measuring passing skill by looking at the geometry of players within passing lanes[5] and measuring how players create space for themselves and teammates.[6] The photo to the left shows a player-tracking camera installation; after stitching together the video from the two cameras to create a panoramic video covering the whole field, standard computer vision techniques track the position of players and the ball.

One idea that continually arises in the use of tracking data is that of pitch control, where the underlying estimand, or variable of interest, is the space, or fraction of space, owned by a particular player. This definition has varied based on application, from measuring points in space where a player is closest, to measuring the area that the player is most likely to reach first given their position, velocity, and physical abilities.

To explore what it is possible to learn from these data, we will review several variants of pitch control models and compare them visually (in Figure 1, page 28), using a single frame of data from a Spanish La Liga match.

Pitch control models

The first model, called a *Voronoi model*, allows an analyst to partition space by assigning every location to the closest player, producing what is known as a Voronoi tessellation or diagram (Figure 1a). Provided a distance function $d_m^i(t)$ between

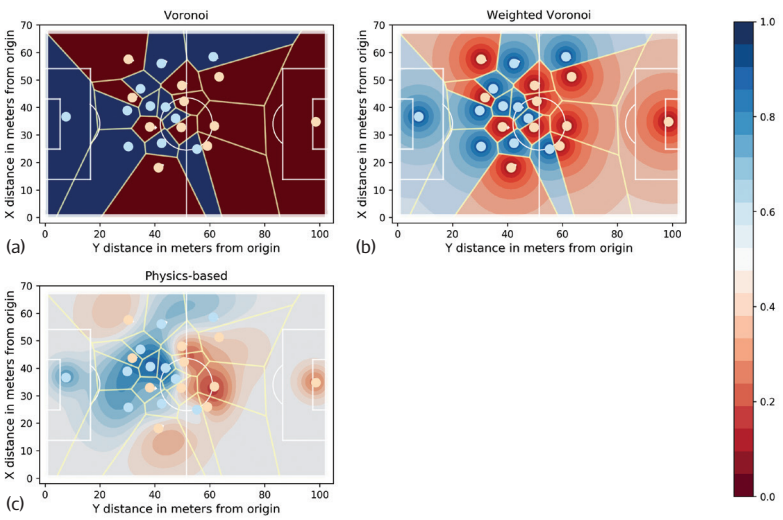

FIGURE 1 Comparison of pitch control models: (a) Voronoi; (b) weighted Voronoi; and (c) physics-based (with Voronoi lines included for comparison).

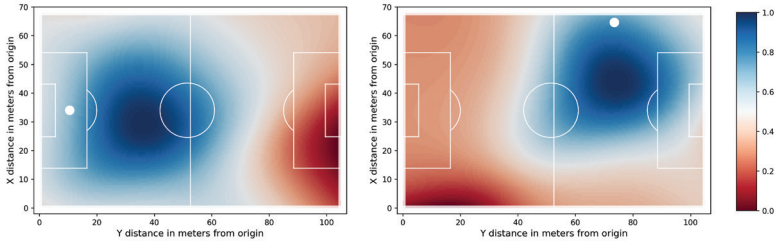

FIGURE 2 Distribution of defender positions as a function of the ball's location (white dot) over 20 Spanish La Liga games. Defending team's goal is on the right of the pitch.

A physics-based model of pitch control

A player's influence area I is modelled through a simple function, in this case $I_i = \pi(m; x_i, \Sigma)$ where π is the density of a bivariate normal evaluated at m, with mean x_i the location of player i. The covariance matrix Σ can be decomposed through a singular value decomposition to be expressed in terms of a rotation matrix R and a scale matrix S. We can then adjust this matrix dynamically to account for the orientation of a player's velocity vector, and the scale matrix can be adjusted to account for player's speed, scaled according to the distance to the ball. Once the distribution is [0,1] normalised from x_i, player i's current position, we obtain a degree of pitch influence that follows a multivariate normal. A pitch control surface for the team can be obtained by aggregating each player's influence degree at each given location, as expressed by the following equation:

$$C_m^i(t) = \alpha\sigma(\beta_1 \sum_r I_r(m,t) - \beta_2 \sum_j I_j(m,t))$$

Note that this equation expresses pitch control by team to keep consistency with previous equations, but the same result will be obtained for any player since aggregation is performed over both teams. Here σ corresponds to the logistic function which transforms the aggregated influence into a degree of control within the [0,1] range. Constants β_1 and β_2 allow one to account for different impact of the aggregated influence of one team depending on contextual factors (e.g., the attacking team might be thought to have higher influence on the pitch) or individual player skills, while α allows one to shrink or expand the logistic function.

player i's location and a given location m on the field at time t, the level of control of player i over any given location can be defined as

$$C_m^i(t) = \begin{cases} 1 & i = \text{argmin}_j d_m^j(t) \\ 0 & \text{otherwise} \end{cases}$$

The resulting partitions are often referred to as "Voronoi dominant regions". Through this expression we can assign dominance of cells to individual players or to one or both teams, noticing that regions are set to be fully controlled by one player (or team) at any given time.

A variant on the strict region dominance of classical Voronoi tessellation is that of *weighted Voronoi*,[7] where a weighting function $w_m^i(t)$ is used to account for the relative level of influence a player i has over location m (see Figure 1b for illustration). We can express this pitch control model as

$$C_m^i(t) = \begin{cases} \dfrac{1}{1+w_m^i(t)} & i = \text{argmin}_j w_m^j(t) \\ 0 & \text{otherwise} \end{cases}$$

A typical weighting function would be $w_m^i = \beta d_m^i$, where the distance to any given location is controlled by a constant factor.

While Voronoi tessellation is simple to understand, explain, and visualise, it misses various characteristics intrinsic to sport. First, it ignores the physical characteristics of the players, including their current velocities. Second, the hard thresholds separating players' space do not reflect space occupation in actual play, where balls falling between players become battlegrounds to gain control.

Fundamentally, deciding on a model depends on the variable of interest. To find the space closest to a player, then Voronoi will be suitable. However, in soccer it is more strategically useful to know the *probability* that a given player controls the ball if it is passed to a given location. This has led to more physically motivated models, which give probabilistic outputs.

The literature presents at least two approaches to these *physics-based* models of pitch control. One defines the degree of control of any given location according to players' velocity and the distance to the ball, in order to define regions that are theoretically easier for a certain player to reach.[4,5] This approach also assumes that any location can be controlled by a given player if that player is closest to that location, regardless of his distance to that point. Another approach models a continuous degree of influence a player has on the field while accounting for his velocity and distance to the ball. Then the influence of each team's players is aggregated to obtain a surface of control (see box for a detailed explanation of this model).

Figure 1c shows this method. We see from this single frame that the Voronoi model creates strong edges between players from the two teams. The weighted Voronoi also has strong edges, but down-weights areas away from the players. In stark contrast, the physics-based model results in a smooth surface of control, allowing for increased control when multiple teammates cluster together.

Measuring spatial value

While pitch control models can help an analyst or coach understand how a team or individual player is controlling space on a frame-by-frame basis, they can also be averaged over many moments to understand where an individual, team, or group of teams spend their time, perhaps as a function of another variable, such as ball location. This conditioning is especially important in soccer, where the player positioning desired by the coach can vary drastically as a function of game state, such as where the ball is or the current pace of attack.

Figure 2 provides an example. It shows the average space owned by defenders of a Spanish La Liga team as a function of where the ball is, over 20 games played during 2017. Here we see that defender positioning changes drastically as a function of the ball's location. Because of the nature of conditioning on a continuous variable such as ball location, using limited data, it is important to leverage spatial information to reduce variability. Here we use a machine-learning program – a feed-forward neural network with one hidden layer[6] – to get a model-based estimate for defensive team control as a function of ball location.[5] This plot can be used, for example, as one estimator for how valuable space is as a function of the ball, with the underlying idea being that defensive players will, on average, position themselves in areas of importance.

Pressing forward

In soccer, there exists a culture of teams searching for the next high-profile coach to avoid league relegation, while those same coaches spend hundreds of millions of euros to attract elite-level talent in order to compete for trophies. But throwing money or sporting talent at the challenge requires more. Smaller teams have in the recent past triumphed over their bigger rivals, whether it is Leicester City winning the English Premier League in 2016, Deportivo de La Coruña winning the Spanish La Liga or Champions League classification in multiple consecutive years, or Portugal and Greece winning two of the four previous UEFA European Championships.

This ability of countries with small populations (or clubs with small budgets) to win championships suggests that it is possible to overcome the favourites even with highly constrained resources. While small clubs cannot compete on budget, they can make smart, data-informed decisions to help close the gap on their heavy-spending competitors, levelling the playing field – as demonstrated by the Oakland Athletics baseball team, whose success through analytics was documented in *Moneyball*.

With this year's World Cup marking the first time that teams will have real-time access to player-tracking data throughout the sport's biggest tournament, it is clear that paltry data collection is becoming a thing of the past. Now that these data are in place, it is time for the statistics community to lean in and develop metrics which leverage the data to measure the game in ways more closely aligned with winning strategies. Armed with this information, teams will be battling not just on the field, but also behind the scenes – using data to better inform their strategies and attack opponents' weaknesses. ■

vwerve/Bigstock.com

Luke Bornn is vice president of strategy and analytics for the Sacramento Kings basketball team.

Dan Cervone is a senior analyst in research and development with the Los Angeles Dodgers baseball team.

Javier Fernandez is head of sports analytics at FC Barcelona.

References

1. Lewis, M. (2003) *Moneyball: The Art of Winning an Unfair Game*. New York: Norton.

2. Oliver, D. (2004) *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington, DC: Brassey's.

3. Anderson, C. and Sally, D. (2013) *The Numbers Game: Why Everything You Know about Soccer is Wrong*. New York: Penguin.

4. Gudmundsson, J. and Horton, M. (2017) Spatio-temporal analysis of team sports. *ACM Computing Surveys*, **50**(2), 22.

5. Spearman, W., Basye, A., Dick, G., Hotovy, R. and Pop, P. (2017) Physics-based modeling of pass probabilities in soccer. Paper presented to Sloan Sports Analytics Conference 2017.

6. Fernandez, J. and Bornn, L. (2018) Wide open spaces: A statistical technique for measuring space creation in professional soccer. Paper presented to Sloan Sports Analytics Conference 2018.

7. Cervone, D., Bornn, L. and Goldsberry, K. (2016) NBA Court Realty. Paper presented to Sloan Sports Analytics Conference 2016.