Ready Player Run: Off-ball run identification and classification

Sam Gregory

Abstract

One of the major downfalls of tracking data in football is the lack of a common language to describe actions that take place off the ball, particularly patterns of player movement. This paper provides a method for identifying and classifying off-ball in possession runs into similar groups to allow for more generalisable analysis. The objective is to create a vocabulary of run types that can be used to better describe or analyse specific runs and be queried more easily than raw tracking data. These runs are identified by segmenting the raw tracking data using periods of high player speed and acceleration, then classifying them using a clustering method with functional cluster centres modelled as Bézier curves. These run clusters are used to analyse positional trends as well as player archetypes based on common runs. We also discuss a series of potential extensions using these run types as the basis for further analysis based on pre-existing work with tracking data.

Key Words: Off-ball runs, player tracking data, functional clustering, Bézier curves

1. Introduction

One of the main advantages of relatively standardised and ubiquitous event data in football is that the data provide a common vocabulary for analysing the game. Despite different definitions from different vendors, everyone roughly knows what it means for a player to have completed five final third passes in a game.

With tracking data there lacks a similar taxonomy of movement, especially for off-the-ball actions. Even the language itself can be ambiguous: asking a group of coaches or analysts to define what an overlapping run is will lead to a series of different definitions that may or may not be easily codified within the tracking data.

This paper uses an unsupervised approach to classify off-the-ball run types in an effort to translate raw tracking data into discrete actions that can be used in analysis in a consistent manner, as is currently done with event data.

Trajectory analysis [1] is a burgeoning area of sports analytics with work being done to classify both ball [3] and player trajectories [2,6,7,8,9]. Some of the seminal work in player trajectory analysis models player runs as Bézier curves [2,7]; this paper takes a similar approach.

Using player and ball tracking data, we identify periods of acceleration and deceleration to pinpoint when players are making "intentional" off-the-ball runs. Then building off the methodology in Miller and Bornn 2017 [7] - we employ Bézier curves to align individual runs in space and time, comparing similar runs and clustering them into distinct groups using a classical k-means approach but with functional cluster centers.

The distinct run types are then used as building blocks or a vocabulary to identify patterns and trends or to query large data sets of tracking data.

2. Data

The raw tracking data are a sample of 74 games from a top European league; the players and teams have been anonymised. Data are sampled at a rate of 25 frames per second with the x,y locations of each player and the ball relative to the origin (0,0) at the centre of the pitch. The coordinates are adjusted so both teams are attacking from left to right.

3. Methodology

3.1. Creating run segments

In other sports defining what constitutes a "run" is a simpler task. In American football a receiver's route begins at the snap and ends when the catch is made or play comes to an end [2]. In basketball, players tend to be either moving from one location to another

or standing still [7]. In football, this is not the case, as players are almost constantly moving - walking, jogging or sprinting. Consider the distribution of player speed in Figure 1. The bump at 2 m/s is roughly the difference between a player walking and jogging.

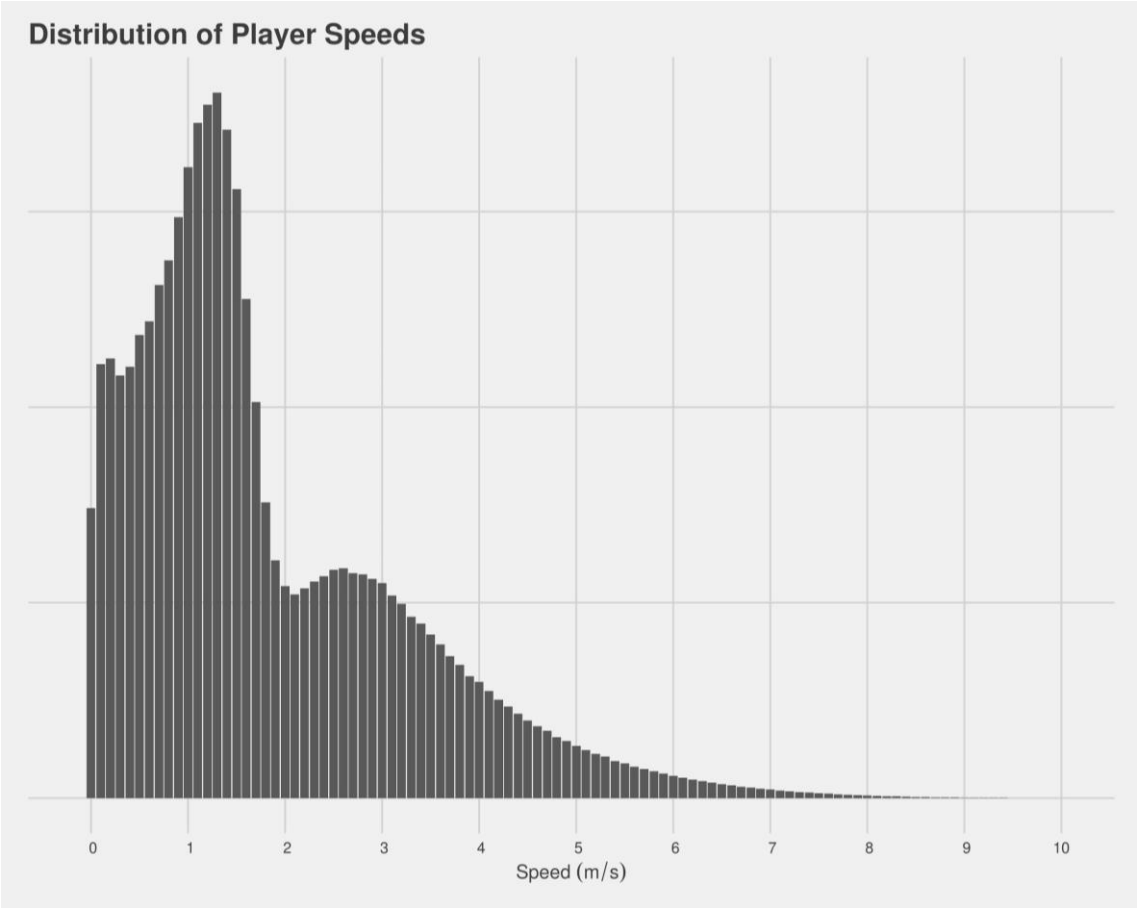

Figure 1 - Player speed distribution

For the purposes of classifying runs, the goal is to identify intentional runs from one location to another rather than just periods of walking or slow jogging. In order to identify these run segments we look not just at player speed but also acceleration (Figure 2).

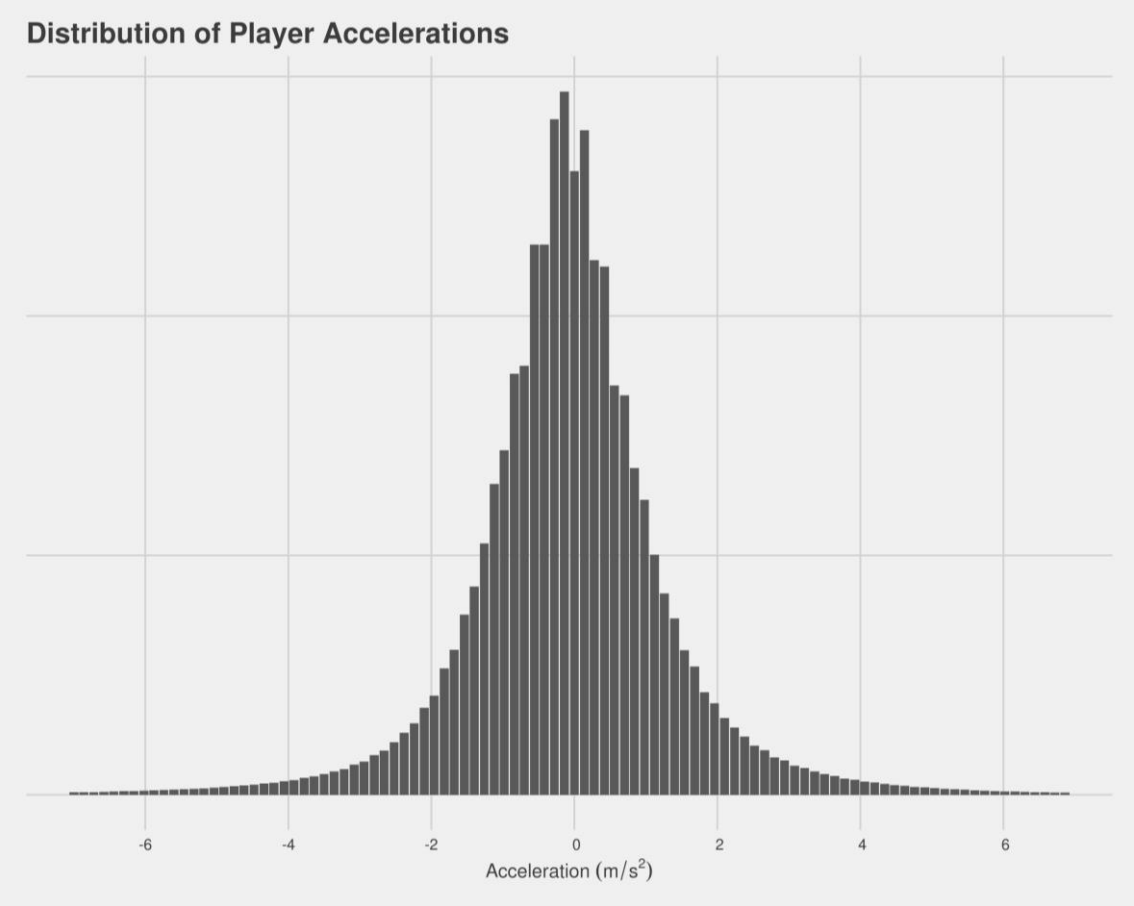

Figure 2 - Player acceleration distribution

To segment the raw player tracks into discrete runs we identify all periods where the ball was in play and a player reached a speed of at least 5 m/s with an acceleration of at least 2.5 m/s^2. These points are all classified as run starts. The run continues until one of four things happens:

- The player slows to a speed of 4 m/s or less
- The player acceleration drops to less than -2.5 m/s^2
- Team loses possession
- Stoppage in play

Using this methodology to create the run segments we then remove any runs which have a duration of less than 1 second.

Consider a player track segment illustrated using speed, acceleration (Figure 3) and trajectory (Figure 4). The orange sections are where the player is making a run by the definition outlined above.

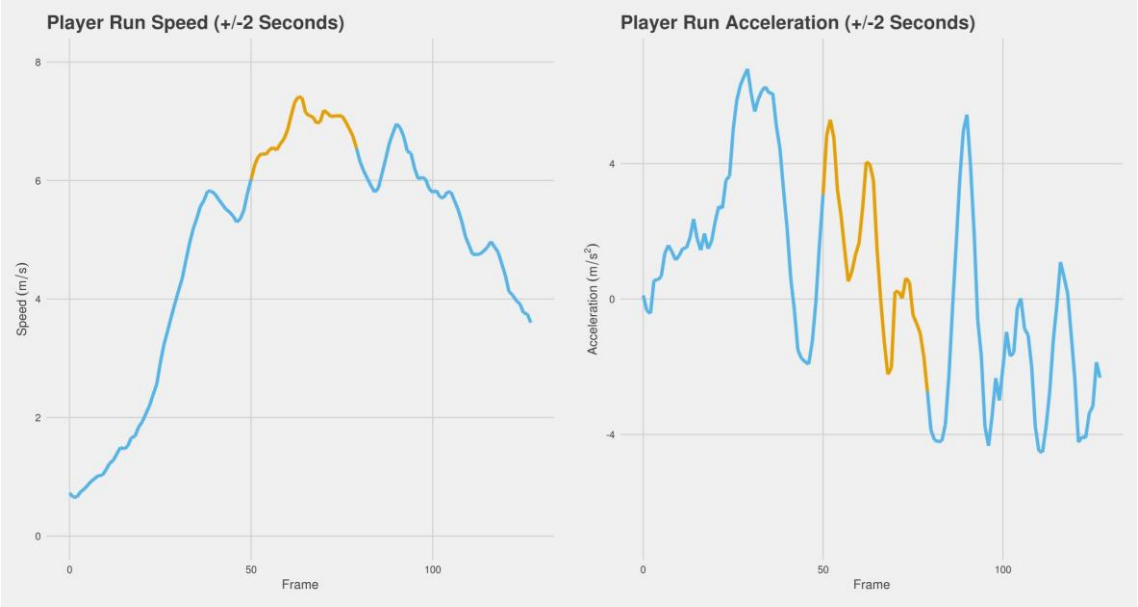

Figure 3 - Player Run Speed and Acceleration

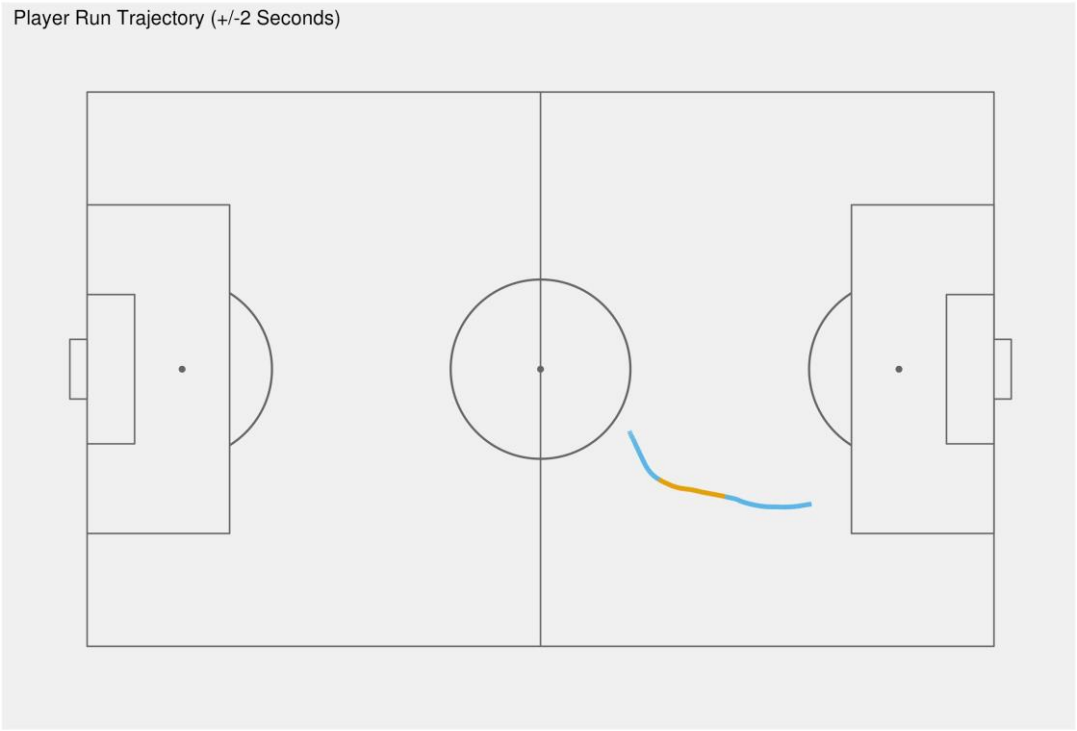

Figure 4 - Player Run Trajectory

This leaves a total of 80,710 intentional player runs for an average of 1,091 per match. Looking at only off-ball in possession runs reduces the number to 25,572 runs, or 346 per match.

3.2. Coordinate system adjustments - accounting for teammates

The first coordinate adjustment is a simple mirroring, assuming that runs on the left and right hand sides of the pitch should be treated the same. Mirroring the pitch so that every run originates from the left hand side of the pitch allows for more granularity in run type classification without increasing the number of clusters. This also allows for one to one comparisons between runs on the left and right hand sides of the pitch.

Note that even though this mirroring technique forces all runs to start on the left side of the pitch they may still move onto the right hand side.

The following equation illustrates how the mirroring is applied to the *y* coordinates of run trajectory $\square_{\square} \varepsilon \mathbb{R}^{\square_{\square} \square 2}$ which has duration $\square_{\square}$.

$$\square_{\square} = \square(\square_{\square 1} < 0)\square_{\square}[0,-1]^{\square} + \square(\square_{\square 1} \geq 0)\square_{\square}[0,1]^{\square} + \square_{\square}[1,0]^{\square} (1)$$

Another difficulty with modelling similar run trajectories in football is the relevance of other players' positions. In American football routes all start at a similar location (the line of scrimmage), whereas in football runs can begin anywhere and runs in similar locations may have very different intents depending on the location of teammates. To adjust for this we introduce a second coordinate adjustment making each run relative to the position of the team in possession at that time.

This adjustment treats the centroid of each team at the start of the run (excluding goalkeepers) as the coordinate system origin. So instead of (0,0) representing the centre of the pitch, (0,0) is now the centroid of the in possession team. A run trajectory $\square_{\square}$ by a player on team *a* that begins at time *t* is adjusted as follows. This equation assumes team *a* has 10 outfield players on the pitch at the time, which may be adjusted depending on injuries or red cards. Note the vector of centroids in (3) has length $\square_{\square}$ the duration of $\square_{\square}$.

$$\square\square\square\square\square\square\square\square_{\square\square} =$$
$$\frac{1}{10}\sum_{\square=1}^{10} [\square_{\square}, \square_{\square}] \forall \square\square\square\square\square\square\square\square\ \square\square\square\square\square\square\square\square\ \square\ \square\square\ \square\square\square\square\ \square (2)$$
$$\square_{\square} = \square_{\square} - [\square\square\square\square\square\square\square\square\square_{\square\square}, \ldots, \square\square\square\square\square\square\square\square\square_{\square\square}]^{\square} (3)$$

The top left plot in Figure 5 shows the location of all teammates and the run trajectory at the start of a run, the top right plot adds the team centroid. The bottom left plot mirrors the coordinates of all players so that the run starts on the left side of the pitch. The bottom right plot then shifts the run slightly towards the centre and slightly forward relative to the centroid instead of the centre dot.

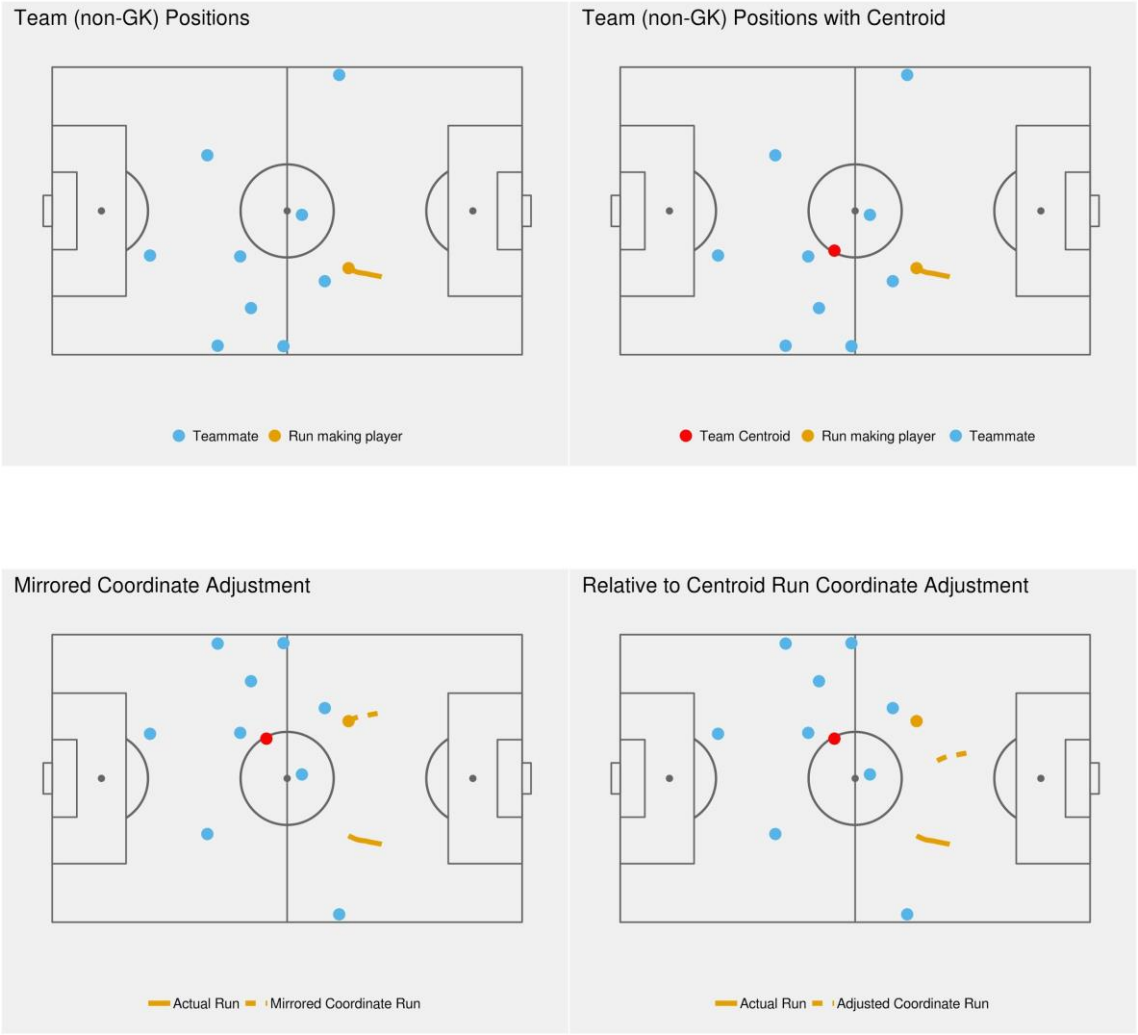

Figure 5 - Run coordinate adjustment. The top left quadrant is the initial player positions at the start frame of the run. The top right adds in the team centroid. The bottom left quadrant mirrors everyone's location so the run starts on the left hand side of the pitch. The bottom right adjusts the run coordinates relative to the team centroid.

3.3. Clustering run types

To cluster the runs we use a similar methodology to Miller and Bornn 2017 [7] and Chu et al. 2019 [2] exploiting the nature of Bézier curves.

Bézier curves provide an ideal functional form for measuring distances between trajectories that have different lengths because they can be evaluated at any arbitrary number of points along the curve. So for an observed run trajectory of length *d* the

distance from that trajectory to the cluster centre can be evaluated at *d* different points each an equal step along the Bézier curve.

The Bézier curves are defined by the following equations. *P* is the number of control points that defines the Bézier curve. The list of control points for each Bézier curve is defined as a matrix $\theta \in \mathbb{R}^{P \times 2}$ with each control point an *x,y* coordinate.

$$B(t, \theta) = \sum_{i=0}^{P} \theta_i B_i^P(t), t \in [0,1] \quad (4)$$

$$B_i^P(t) = \binom{P}{i} t^i (1-t)^{P-i} \text{ for } p = 0, \ldots, P\text{-}1 \quad (5)$$

For a given set of control points θ the Bézier curve can be evaluated at any arbitrary *t* between 0 and 1.

The clustering itself is a straightforward k-means approach using the distance from the cluster centre - defined as a Bézier curve - to the adjusted coordinate trajectories defined in 3.2. The first step is to fit Bézier curves to randomly initialised cluster centres. We set *k*=70 clusters randomly assigning 70 runs as cluster centres and fitting Bézier curves to these runs.

Fitting these Bézier curves is equivalent to defining the matrix of control points θ given a trajectory or list of trajectories. We define the observed run trajectory R_i as a T_i x 2 matrix where T_i is the duration of length of the run trajectory R_i. Then define the following equation as a classical linear least squares equation with an error term ε_i .

$$R_i = B_i \theta + \varepsilon_i \quad (6)$$

The matrix of regressors is made up of the polynomials along which the Bézier curve is evaluated.

$$B_i = \left[B_0^P(t_{i1}), B_1^P(t_{i2}), \ldots, B_P^P(t_{iT_i}) \right]^T \quad (7)$$

Solving the linear least squares problem gives an estimate for the matrix of control points θ.

After fitting the Bézier curves to the k randomly chosen cluster centres the next step is to calculate the distance from each trajectory to each cluster centre. Consider a trajectory *i* defined as R_i with length T_i and a cluster centre *j* defined as $B_j(t, \theta_j)$. The L1 distance between these is defined as:

$$D_{ij} = \frac{1}{T_i} \sum_{t=1}^{T_i} \sum_{l=1}^{2} \left| R_{it} - B_j\left(\frac{t}{T_i}, \theta_j\right) \right| \quad (8)$$

Each run trajectory□□ is assigned to the cluster *j* that minimises □□□□□□in (8). The objective function at iteration *m* is the mean of the distances of each of the *n* run trajectories to its cluster centre.

$$□□□□ = \frac{1}{□}\sum_{□=1}^{□} □□□(□□□□□_{□1}, □□□□□_{□2}, \ldots, □□□□□_{□□}) \quad (9)$$

The cluster centres are then updated by taking all of the run trajectories in each cluster to solve for a new set of control points $\theta_{□}$ for cluster *j* in equation (6). The distance calculations in (8) are repeated with the new centres and these iterations continue until a predetermined tolerance level is reached at which point the clusters are finalised.

$$□□□□ = □□□□_{□-1} - □□□□□ \quad (10)$$

4. Results

4.1. Cluster Centres

Looking at the final score objective function by number of clusters (Figure 6) there is no clear "elbow point" so we settled on 70 clusters based on the interpretability of the results. For different applications, different numbers of clusters may be more appropriate.

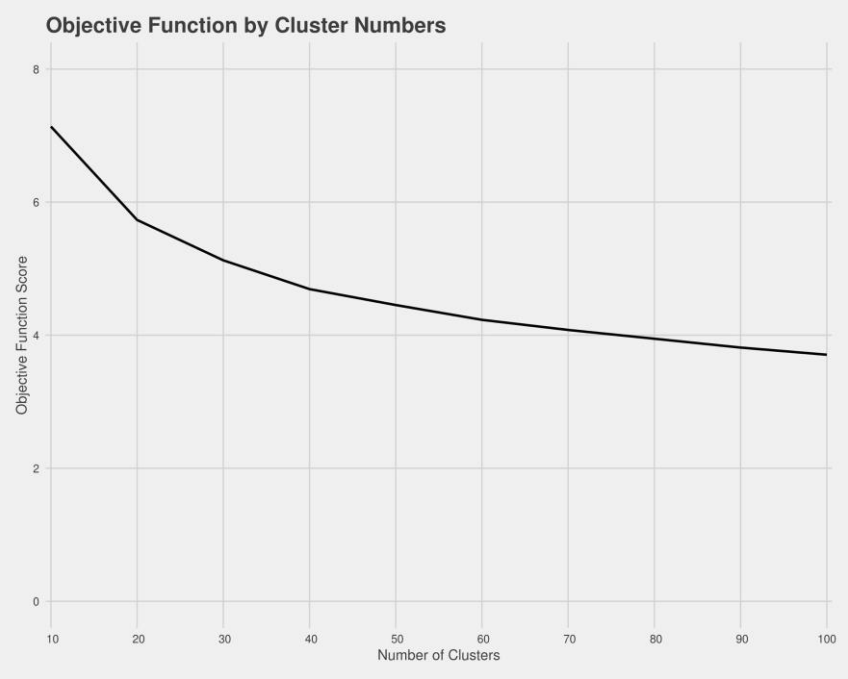

Figure 6 - Objective function by number of clusters

Figure 7 shows the Bézier curves which represent the cluster centres, the blue dots show where the trajectories start from. Note that these cluster centres are all relative to the team centroids so for the purpose of illustration Figure 7 assumes the centre dot to be

the team centroid (which is why some run trajectories extend beyond the dimensions of the pitch).

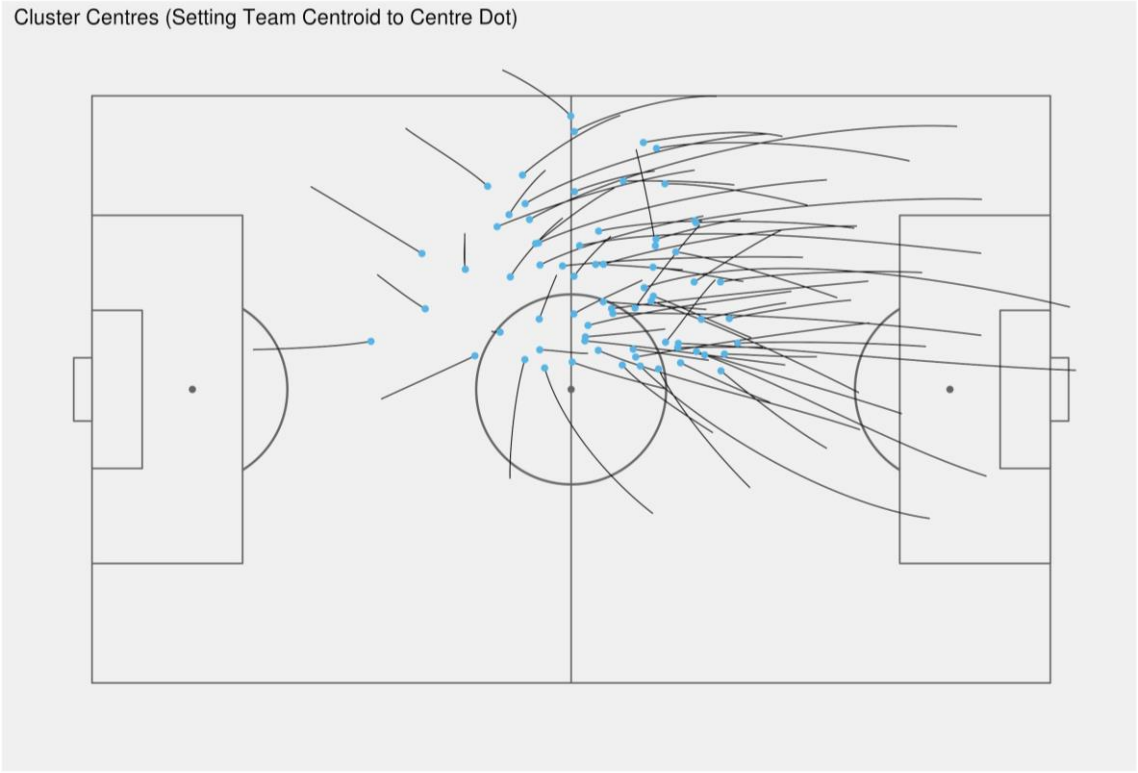

Figure 7 - Cluster Centres, blue dots are run start points

For the purpose of this analysis we treat the runs originating from the left and right as identical, but these 70 run types could easily be extended to 140 by considering the mirror set of runs.

Although this method was unsupervised some familiar run types already begin to jump out. Overlapping runs near the touchline, runs originating from behind the team centroid to offer a supporting option, direct runs into attacking position, horizontal runs to try and create space, attacking runs cutting into the centre etc. These trends become even more evident looking at which positions are making each run type.

In analysing these runs it is important to note the clustering is done on adjusted coordinates and in order to maintain the comparison all runs in the rest of this paper are mapped in adjusted coordinates with the centre dot as the centroid. To illustrate how this translates to actual coordinates consider all of the runs which fall into the following run type first plotted in adjusted coordinates then real world coordinates (Figure 8).

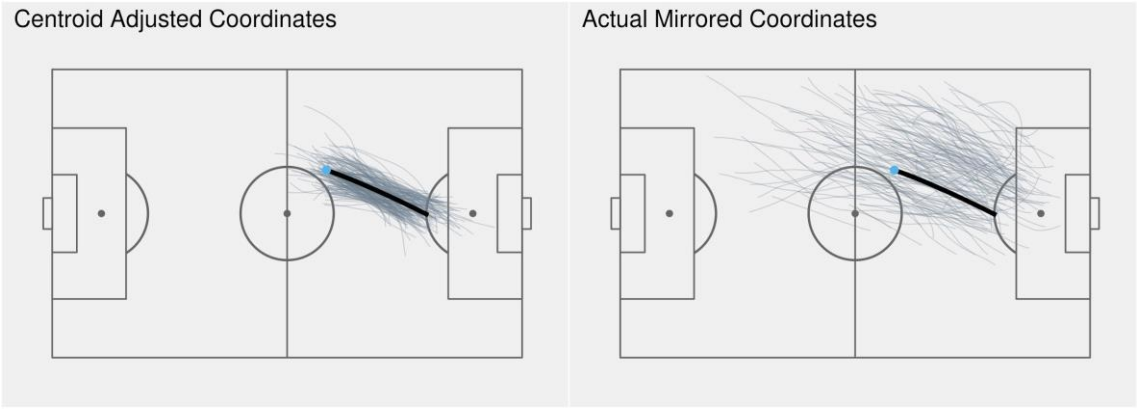

*Figure 8- Example run type in real and adjusted coordinates (*Bézier *curve centre relative to centre dot)*

In Figure 8, it is still evident that the purposes of these runs are similar - a player already ahead of the team centroid is making a direct attacking run - however they are not grouped as tightly together by location as they are when plotted with the same relative team centroid.

The fact runs with similar intent but in different parts of the pitch are being clustered together in the same group validates the approach of adjusting the run coordinates relative to the team centroid.

4.2. Easily identifiable run templates

4.2.1. Fullback runs

The motivating example given in the introduction was identifying overlapping runs typically made by fullbacks.

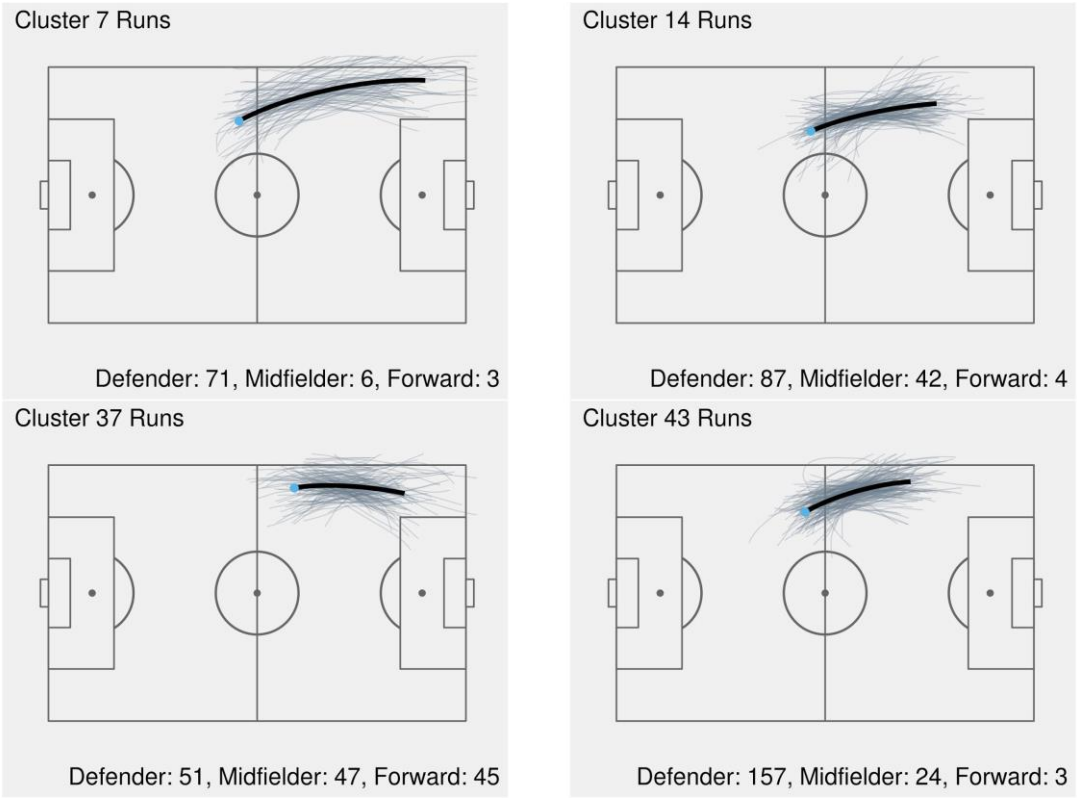

Figure 9 - Overlapping style runs

All of the runs in Figure 9 were made by defenders more than forwards or midfielders. Overlapping runs are usually defined as runs which begin from behind the ball and move beyond the ball, this information could easily be added to these run types to identify both the number of times a particular player made this run and the number of times they actually overlapped the player in possession. There is also some variation in the shape and length of these overlapping run types which may be of interest to differentiate between runs or players.

4.2.2. Supporting defender runs

This series of runs in Figure 10 highlights four different backward run types made by defenders (typically centre backs). Given this analysis focuses on in possession runs we can assume the intent in most of these cases is to support a teammate in possession.

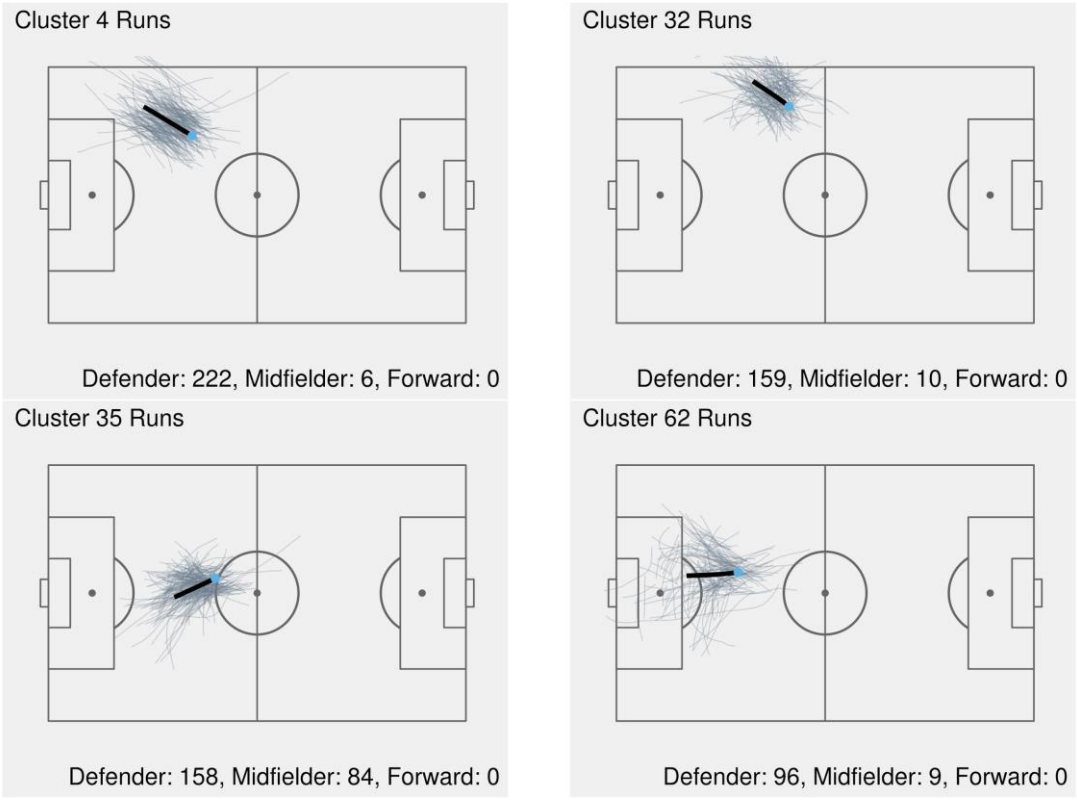

Figure 10 - Defender support runs

4.2.3. Midfielder runs

The run templates in Figure 11 are mostly made by midfielders. They mostly occur near the team centroid and appear to either be moving forward with the play (clusters 5 and 26) or stretching the field by moving wide (clusters 21 and 22).

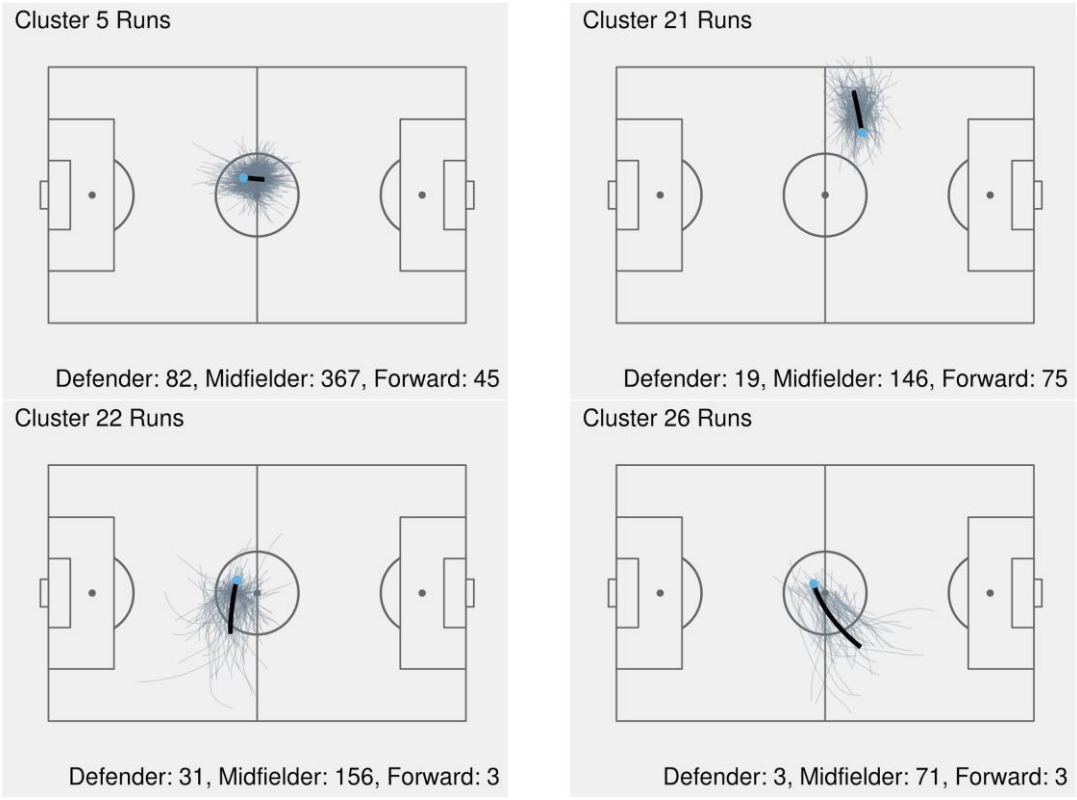

Figure 11 - Midfielder runs

4.2.4. Forward runs

Figures 12 and 13 highlight runs mostly made by forwards and attacking midfielders. The first set of runs in Figure 12 are all very direct attacking runs while the set of runs in Figure 13 are quite different attacking runs cutting from wide positions to more central ones. In general this group of runs shows how this approach can distinguish between different types of runs forwards make to find space and lose opposition defenders.

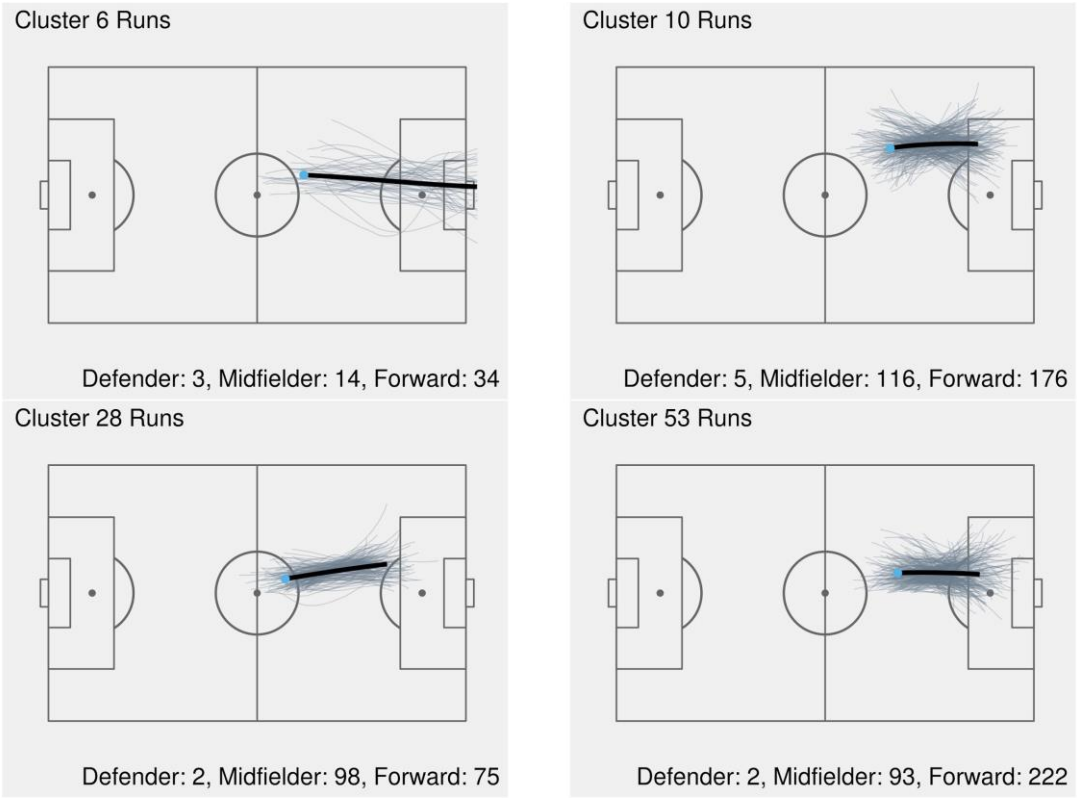

Figure 12 - Direct forward runs

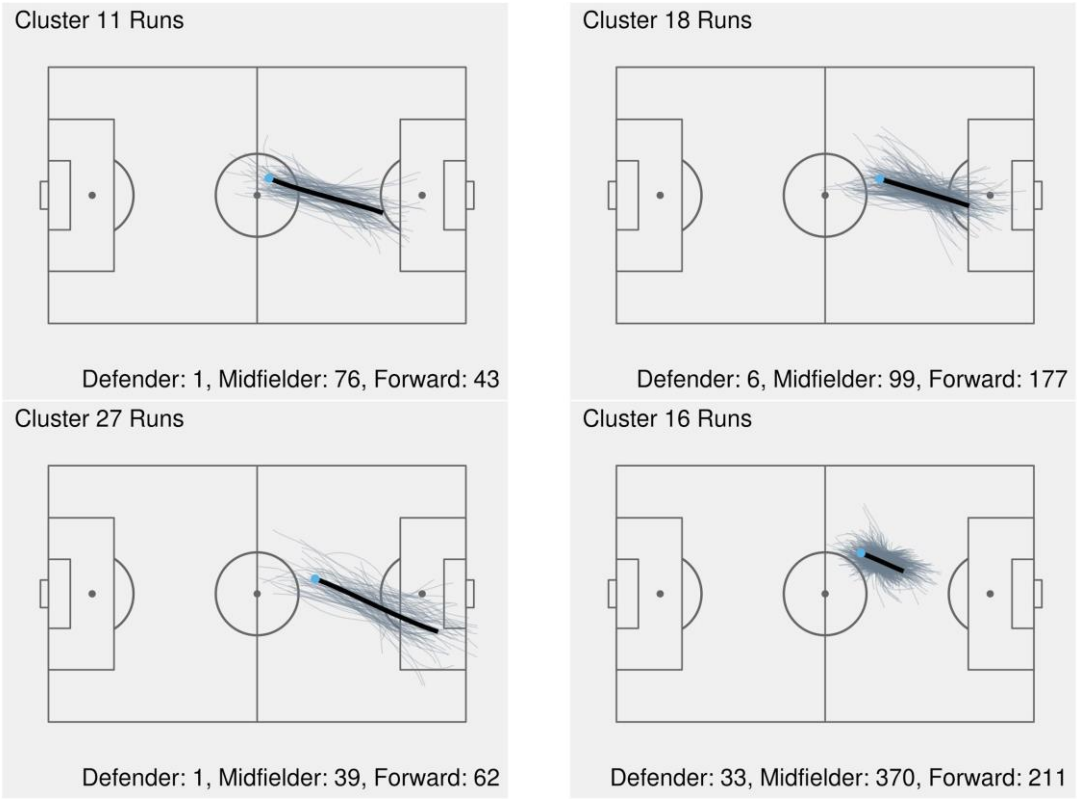

Figure 13 - Cuts inside

4.3. Player specific run types

The results of the run clustering can also be used to describe individual players. Consider the five most common run templates made by two different players in Figure 14.

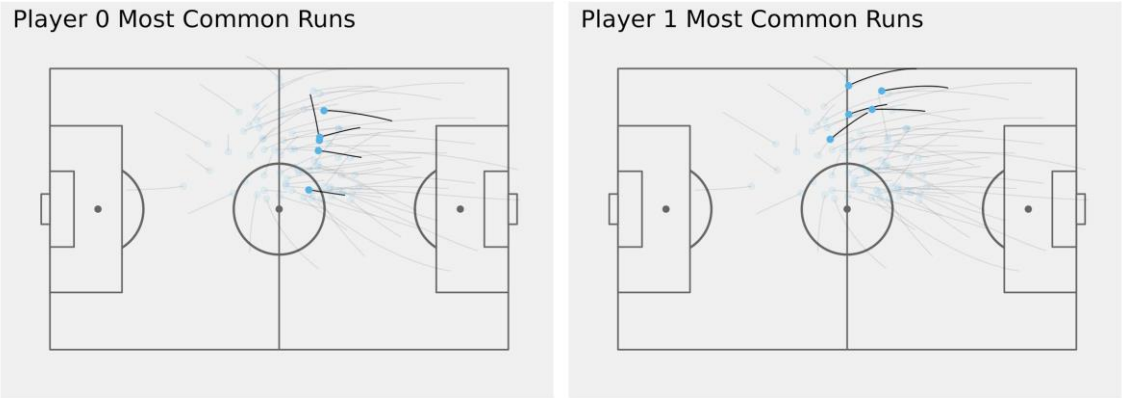

Figure 14 - Five most common runs by player

Only by looking at their most frequent in possession off-ball runs we already have an idea of the type of players these two are. Player 0 is an attacking player who gets wide to stretch the field and cuts in from these wide position, whereas Player 1 is a full back who makes lots of wide overlapping-style runs from deeper positions relative to the team centroid.

This can be extended to team level analysis as well seeing which runs - by which players - a team makes when in possession.

5. Practical Applications

5.1. Professional and media applications

Some of the initial applications of these run types have been demonstrated above, with player, team and position archetyping. These types of analyses have potential applications across recruitment, performance analysis and potentially even sport science. Being able to synthesise off-ball run information will allow teams to summarise how individual teams or players behave in possession using data in a much more complete manner than is possible with only event data.

There are also media use cases, the ability to query tracking data just to pull out run segments in general but to also get video of specific and similar run types could be useful for creating stories about what players do off the ball.

5.2. Modelling extensions

In terms of improving the run classification the next steps would be to incorporate more information beyond just the centroid of the in possession team. Taking into account the

position of opposition players or the ball may help better distinguish between runs with similar trajectories but different intents.

Segmenting and classifying run types to create a vocabulary or dictionary for different run types makes further analysis on off the ball movement much easier. There has been work valuing different off the ball movement [4], but much of this has focused on off the ball movement in a very general sense rather than the effect or value of an individual run. These run types could be assigned values in different situations based on how much space they generate [4], how they affect pass probabilities [10] or even expected possession value scores [5].

6. Conclusion

This paper outlines a relatively simple way of segmenting tracking data into individual runs and grouping them into easily understandable run templates. Using a rules based approach based on a track's duration, speed and acceleration we demonstrate how these tracks can be divided into intentional run segments. Based on their trajectories relative to the team's centroid they are classified into run templates building up a vocabulary of run types. The clustering approach exploits the nature of Bézier curves for mapping distances between trajectories of different lengths. The final output is a series of runs categorised into 70 different types which can be used to analyse player and team tendencies or simply as a convenient way of querying off the ball tracking data.

7. Acknowledgements

The data analytics and research teams at Sportlogiq provided extensive feedback and ideas at various stages of this paper, particularly Eimear O'Leary Barrett, Evin Keane, Daniel Daly-Grafstein, Michael Horton, David Yu and David Vallett. Luke Bornn also provided additional support discussing the approaches he took in Miller and Bornn 2017 [7].

8. References

[1] Bian, J., Tian, D., Tang, Y., & Tao, D. (2018). A survey on trajectory clustering analysis. Retrieved from https://arxiv.org/pdf/1802.06971.pdf

[2] Chu, D., Reyers, M., Thomson, J., & Wu, L. (2019). Route Identification in the National Football League. Retrieved from https://arxiv.org/pdf/1908.02423.pdf

[3] Daly-Grafstein, D., & Bornn, L. (2019). Rao-Blackwellizing field goal percentage. *Journal of Quantitative Analysis in Sports*, *15*(2), 85–95. Retrieved from https://www.degruyter.com/downloadpdf/j/jqas.2019.15.issue-2/jqas-2018-0064/jqas-2018-0064.pdf

[4] Fernandez, J., & Bornn, L. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference*. Retrieved from http://www.lukebornn.com/papers/fernandez_ssac_2018.pdf

[5] Fernandez, J., Bornn, L., & Cervone D. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference*. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2019/02/Decomposing-the-Immeasurable-Sport.pdf

[6] Hochstedler, J., & Gagnon, P. T. (2017). American Football Route Identification Using Supervised Machine Learning. *MIT Sloan Sports Analytics Conference*. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2017/02/1542.pdf

[7] Miller, A., & Bornn, L. (2017). Possession Sketches: Mapping NBA Strategies. *MIT Sloan Sports Analytics Conference*. Retrieved from http://andymiller.github.io/docs/acm-possession-sketches-final.pdf

[8] Nistala, A., & Guttag, J. (2019). Using Deep Learning to Understand Patterns of Player Movement in the NBA. *MIT Sloan Sports Analytics Conference*. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2019/02/Using-Deep-Learning-to-Understand-Patterns-of-Player-Movement-in-the-NBA.pdf

[9] Sha, L., Lucey, P., Yue, Y., Carr, P., Rohlf, C., & Matthews, I. (2016). Chalkboarding: A New Spatiotemporal Query Paradigm for Sports Play Retrieval. Retrieved from http://www.yisongyue.com/publications/iui2016_chalkboarding.pdf

[10] Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-Based Modeling of Pass Probabilities in Soccer. *MIT Sloan Sports Analytics Conference*. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2017/02/1621.pdf