Automating insight extraction from football data visualizations

Daniel Girela - Twenty3 Sport

Abstract

In this paper we define similarity metrics for two widely used football data visualizations: heatmaps and passing sonars. We illustrate how they can be used in automatically finding players that are particularly similar or dissimilar in terms of their moves or passing intentions. We define as well a visual tool to represent what the average player's passing intentions are given certain fixed patterns in his moves, and make use of the similarity metrics defined before to validate the tool and classify players in terms of their passing predictability. Finally, we introduce a framework designed to make use of these metrics in automating the extraction of insight from variations in a player's heatmap or passing sonar with time or game circumstances.

Keywords— Heatmap, area of influence, passing sonar, player comparison

1 Introduction

Football data visualizations are so widely used nowadays that discussing their importance seems naive. When executed well, they are able to convey an incredible amount of information very quickly. Moves, passing directions, shots, defensive actions and, virtually, every type of event that can be recorded in a football match is susceptible to be represented on a graph from which experienced analysts and casual fans will be able to easily extract insight[1].

One of the simplest methods to extract insight from a certain type of visualization is to execute it over two different datasets (think of events by two different players, of just by a player when playing in two different positions on the pitch) and compare the outputs. Uncountable pieces of content can be produced in this way, be it because two visualizations are too similar or too different to each other, and therefore it would be useful for writers to have a tool that tells them when these similarities or dissimilarities hold, instead of them needing to come up with the idea of a pair of visualizations to compare or, even worse, needing to exhaustively execute comparisons until one

[1]See this article by my colleague Mark Thompson for a more detailed discussion.

satisfies their *eye test*. Our aim in this paper is to define metrics that make this work easy: by converting the similarity between two instances of a certain visualization into a number, a user can automatically select pairs of instances for which this similarity satisfies certain conditions (be greater/less than a given threshold) and then interpret them.

Throughout the paper, we will fix our attention in two types of football data visualizations: heatmaps and passing sonars. These serve to visually represent where a certain player touches the ball, and what is the distribution of the directions of the passes he attempts. To build these, we will use event data provided by Wyscout. On the ball events in a match are sequentially recorded, and we can access information of the player in possession of the ball, the location on the pitch where the event happened, the type of event (shot, pass, tackle, clearance, etc.), the end location of the ball (when the event involves a ball movement), the instant in the match when the event happened, etc.. All examples will use Premier League data for the 2018/19 season, although in Section 5 we will need data for the 2017/18 as a training dataset as well.

The content of the paper is organized as follows:

- In Section 2, we mathematically describe heatmaps and define a measure to compare them, as well as a way to compute the area of influence of a player.
- In Section 3, we mathematically describe passing sonars and define a measure to compare them.
- Section 4 is devoted to present a way to compare players both in terms of their moves and passing intentions at the same time.
- Section 5 introduces expected passing sonars, as a way to quantify the predictability of a player's passing patterns given his moves.
- Section 6 discusses items of future work and, in particular, a framework to use all tools presented in the paper to automate the extraction of insight from visualization comparisons in practice.

2 Comparing heatmaps

A **heatmap** is a type of data visualization that serves to describe notions such as the space occupied by a player, or his area of influence. An image of the pitch is coloured by means of cold-to-hot scale, where hot areas represent parts of the pitch where the player has spent most of the time, and cold areas those where he spent the least. This type of visualization is ubiquitous in football analytics since its early days and, as such, can be found in most free access websites, such as WhoScored, but it is also shown in written media and, even, live on TV while games are broadcasted.

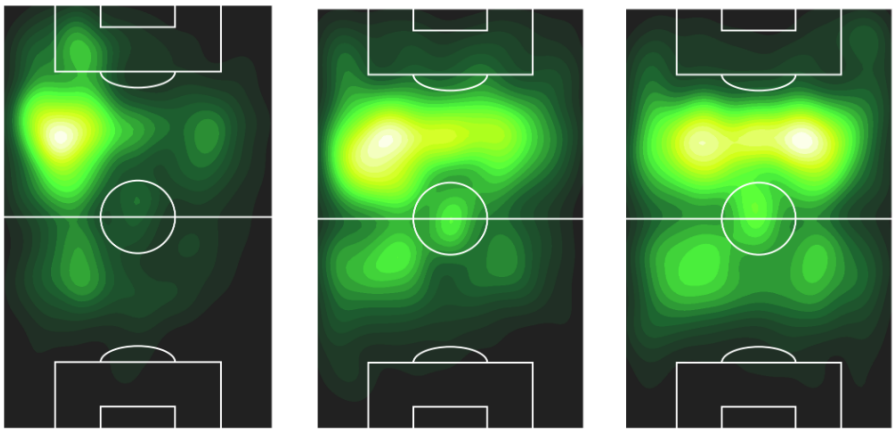

Figure 1: David Silva's, James Maddison's and Christian Eriksen's heatmaps of all their passes in EPL 2018/19

Formally speaking, they are generated by representing the contour plot of the probability density function for the location of the player on the pitch: if we identify the pitch with the rectangle $P = [0, 68] \times [0, 105] \subset \mathbb{R}^2$, this is simply a Borel measurable function $f \colon P \to [0, \infty)$ such that, for every Borel measurable set $E \subset P$, the probability of finding the player in E is

$$\int_E f(x, y) dx dy.$$

Unfortunately, one such density function is not known, so one has to estimate it from a sample of known locations of the player in the pitch. In this case, the sample we will be using is formed by the locations where the player touches the ball (passes, shots, carries, interceptions, clearances, etc.), which will give us insight on the player's movement patters when in possession of the ball. Naturally, using samples of locations of different types will yield different interpretations. The most common technique to estimate the density function from a known sample of locations is the *kernel density estimation* or, simply, KDE, which is implemented in the stats library from the scipy package in Python, and about which an interested reader can read on [2].

Then, if a heatmap is simply a visual representation of a density function, any distance in the space of density functions could, potentially, work to compare two heatmaps. However, the problem is not as simple as it appears to be: the most natural distances in this space, such as the L^2 distance[2]

$$d_{L^2}(f, g) = \left(\int_F |f - g|^2 dm \right)^{\frac{1}{2}}$$

or the Kullback-Leibler divergence

$$d_{KL}(f, g) = \int_F f \log \left(\frac{f}{g} \right) dm$$

[2]Here, and throughout the text, dm denotes the Lebesgue measure in $\mathbb{R}^d$, $d = 1, 2$.

have the significant drawback that they don't take into account the geometry of the space, a fact that, for simplicity and without loss of generality, we illustrate in the one-dimensional case. Assume f, g and h are the density functions plotted below:

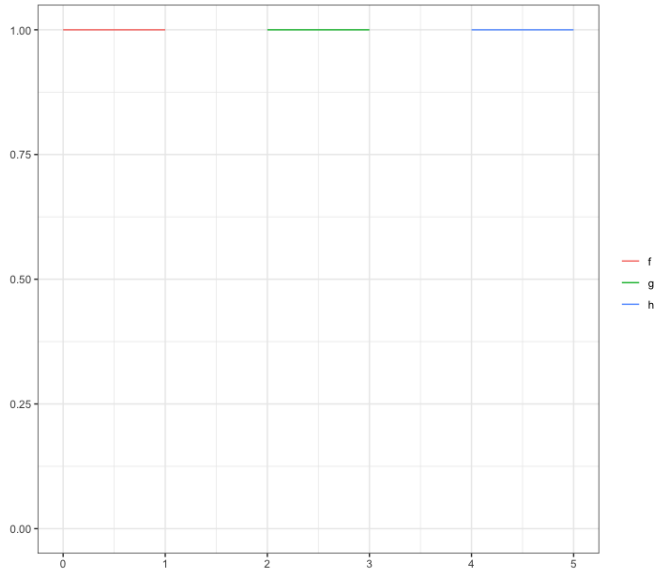

Figure 2: Density functions corresponding to uniform distributions on three disjoint intervals

Intuitively (or, at least, in the appropriate setting for the application we are interested in), one would want f to be closer to g than to h, since *transporting the mass* f requires less work if you want to turn it into g than if you want to turn it into h. However, it is straigtforward to check that $d_{L^2}(f,g) = d_{L^2}(f,h)$ and $d_{L^2}(f,h)$ and $d_{KL}(f,g) = d_{KL}(f,h)$.

More generally, and coming back to the bi-dimensional case, if f and g are two densities supported, respectively, in $F, G \subset P$ and F and G happen to be disjoint,

$$d_{L^2}(f,g) = \sqrt{2}$$

and

$$d_{KL}(f,g) = \infty,$$

and, in particular, these distances don't depend on how close to each other F and G are, which is, clearly, undesirable. Indeed, suppose that a right-centre back, a left-centre back and a striker touch the ball in areas of essentially the same shape and size, that are placed in natural locations for each other. Using the L^2 distance or the Kullback-Leibler divergence to measure the distances among their heatmaps would imply that the two centre-backs' heatmaps are close to each other as they are to the striker's.

Therefore, we need a definition of the distance between f and g that takes into account not only the point-by-point differences between them, but also *what effort is needed to transform f into g*. Fortunately, this is a well-known problem in probability theory, known as *Monge's optimal mass transportation problem*, which we state below for the sake of completeness.

Let μ be a Borel probability measure on P, and let $T\colon P \to P$ be a Borel measurable map. Define the push-forward of μ by T as a new probability measure $T_{\#}\mu$ on P defined by $T_{\#}\mu[B] = \mu[T^{-1}(B)]$ for all Borel measurable subsets B of P. Now, if $\nu = T_{\#}\mu$, we define the cost associated of transporting μ into ν by T as

$$I[T] = \int_P |x - T(x)| d\mu(x).$$

Intuitively, one can interpret this expression as follows: $|x - T(x)|d\mu(x)$ is the cost of transporting the mass located at x to $T(x)$, so summing (i.e., integrating) this on x represents the total cost associated to transporting all the μ-mass by the action of T.

Now, given two Borel probability measures μ, ν on P, Monge's optimal mass transportation problem consists in finding T such that $T_{\#}\mu = \nu$ and $I[T]$ is minimum. One such minimizer T is called an optimal transportation plan, and the associated cost $I[T]$ is called the **earth mover distance between** μ and ν, and will be denoted $emd(\mu, \nu)$. If $\mu = fdm$ and $\nu = gdm$, we will abuse notation and simply replace $emd(\mu, \nu)$ by $emd(f, g)$.[3]

Coming back to our initial example, we have:

- $emd(\text{Maddison}, \text{Silva}) = 0.0034$.
- $emd(\text{Silva}, \text{Eriksen}) = 0.0087$.
- $emd(\text{Maddison}, \text{Eriksen}) = 0.0019$.

In this case, the earth mover distance catches that Maddison and Eriksen are wider players than Silva, who spends most of his time closer to the left wing. The fact that Silva is closer to Maddison than to Eriksen follows from the fact that Eriksen's heatmap is hotter in the right half space, whereas Maddison's is hotter in the left half-space, where Silva concentrates his touches.

An interesting application of a heatmap comparison tool like this arises in scouting: assume a player is leaving your team and you need to find a replacement for him. Potentially, one of the attributes you would like to replicate is the way the player occupies space on the pitch, and this metric can help us to do so. As an example, the following table shows attacking midfielders with at least 1200 minutes on field over the last Premier league season, and is sorted by how close the player's heatmap is to that of Welsh midfielder Aaron Ramsey (who, by January 2019, was already known to be set to leave Arsenal by the end of the season).

[3]This optimization problem has solutions in the case that concerns us. We refer an interested reader to [3] for an excellent survey on the subject.

Name	Distance to Ramsey's heatmap
A. Ramsey	0.0000
M. Özil	0.0031
J. Lingard	0.0038
G. Sigurðsson	0.0039
B. Reid	0.0046
A. Pritchard	0.0058
R. Babel	0.0062
Juan Mata	0.0074
C. Paterson	0.0140
M. Antonio	0.0182

Table 1: Heatmap comparisons between Aaron Ramsey and other attacking midfielders

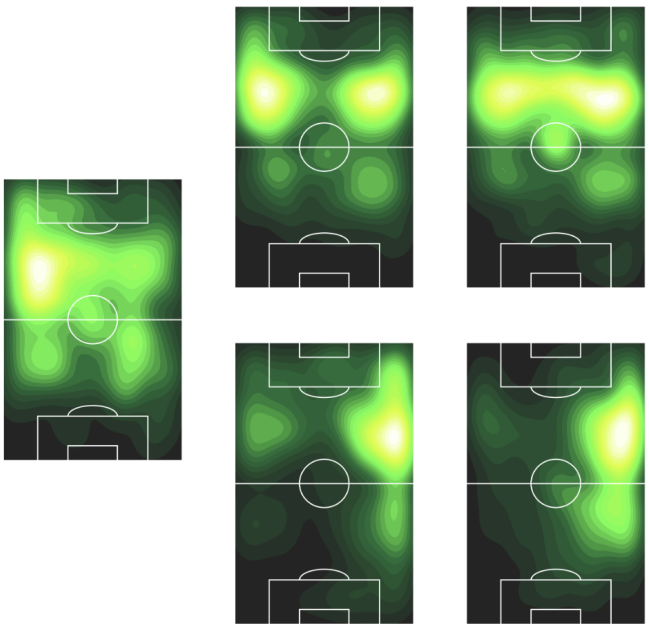

Figure 3: Ramsey's heatmap (on the left), compared to Özil's and Lingard's (at the top) and Antonio's and Groß's (at the bottom)

It is relatively clear that the similarity metric catches that Ramsey's moves concentrate around the three-quarter line, touching both wings (maybe with a bit more weight on the left) and that he is clearly different in his moves to players that mostly fall on the right wing.

2.1 The area of influence of a player

Having codified the moves of a player on the pitch as a probability density function enables us to extract quantitative information from that set of moves. A very simple application in this direction is the computation of the *area of influence of a player*, which can be understood as the area of the subset of the pitch where the player spends, say, 95% of the time. Naturally, there are infinite such subsets, but if one is precise when thinking about the notion of area of influence, quickly finds out that the right concept should be *the smallest subset of the pitch where the player spends* 95% *of the time*. It is easy to check that, if f is the probability density function for the location of the player on the pitch and we denote the level sets of f by

$$F_\lambda = \{(x, y) \in P \colon f(x, y) > \lambda\},\ \lambda > 0,$$

the set we are looking for is one of such level sets: precisely, the one with maximum possible λ and satisfying

$$\int_{F_\lambda} f(x, y) dxdx \geq 0.95.$$

Once the right F_λ is found, the area of influence of the player is, simply, the area of F_λ.

These sets and areas are easy to compute (or, at least, approximate) computationally thanks to the regularity properties of the probability density functions we are representing heatmaps by. Indeed, discretizing F into a sufficiently thin and uniform grid $\mathcal{G} = \{Q_i\}_{i=1}^N$ (i.e., $\mathcal{G}$ is a partition of $[0, 68) \times [0, 105)$ formed by half-open squares $[a, a + h) \times [b, b + h)$), inside each of the Q_i's we can approximate f by its value f_i at any point in Q_i, and so the integral of f over any of the Q_i's can be approximated by $h^2 \cdot f_i$. Thus, if we assume that the Q_i's are sorted in descending order by the f_i's, we just need to find the smallest M such that

$$h^2 \sum_{i=1}^{M} f_i \geq 0.95,$$

and, in that case, the level set we are looking for is, approximately,

$$\bigcup_{i=1}^{M} Q_i,$$

whose area is $M \cdot h^2$.

As an example, we compare the heatmaps of West Ham's Declan Rice when playing as a lone defensive midfielder and when playing as a left defensive midfielder in a double pivot. By applying the technique above, we obtain areas of influence of 4380 m^2 and 4200 m^2 respectively, which correlate with the fact that, visually, it seems that Rice needs to cover less ground when playing in a double pivot than when being the only defensive midfielder on the pitch.

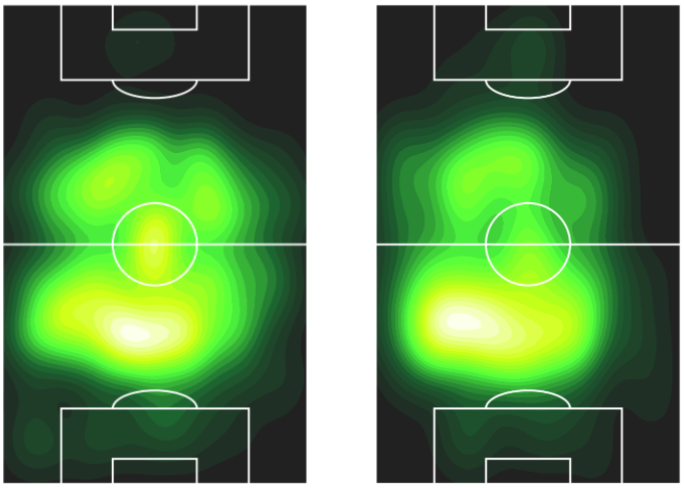

Figure 4: Declan Rice's heatmaps when playing as a lone defensive midfielder (left) and as a left defensive midfielder in a double pivot (right)

3 Comparing passing sonars

A **passing sonar** is a type of data visualization that summarises the distribution of directions of a set of passes by only looking at the angle with respect to a fixed direction, without taking into account the location of where each pass originates from. Formally speaking, they can be generated as the normalized histogram of the variable "signed angle between the segment that represents the pass and a vertical segment joining the pass starting location and the opponent's goal line", and representing the angle over a circle. First introduced by Elliot McKinley, with data from American Soccer Analysis [4], they were later enriched by the Football Whispers Data Science team [5] including a component that indicates pass success rates in each of the bins in the histogram, and are now widely used across the media. Throughout the paper, we will just look at the original version of the passing sonars as it is our aim to compare players (or sets of passes) in terms of their intentions, and not their accuracy or skill.

Figure 5: Sterling's, Hazard's and Mané's passing sonars of all their passes in EPL 2018/19

Mathematically, a passing sonar can be then represented as an s-dimensional vector

$$p = (p_1, p_2, \ldots, p_s)$$

where each p_i represents the proportion of total passes in the set such that their angle lies in the interval $\left[-\frac{\pi}{s} + (i-1)\frac{2\pi}{s}, -\frac{\pi}{s} + i\frac{2\pi}{s}\right]$, and s is the *resolution* of the passing sonar, i.e., the number of bins in the histogram[4]. Then, given two passing sonars p, q, a natural first attempt towards measuring the distance between them is, simply, the euclidean distance

$$d_E(p,q) = \sqrt{\sum_{i=1}^{s}(p_i - q_i)^2}.$$

However, a more careful look at the following example reveals that this is not the right choice for the same reason that the L^2 distance or the Kullback-Leibler divergence were not the right ones for heatmaps, as we saw in Section 2. Indeed, let p be a uniform passing sonar (i.e., all passes are uniformly distributed along all possible angles) and let us build two new passing sonars from p removing half the passes from one of the angular segments and:

- Adding them to one of the neighbouring angular segments (yielding q, left-hand side in the image below).
- Adding them to the opposite angular segment (yielding r, right-hand side in the image below).

[4]The reader is encouraged to simply put $s = 16$, as this is the parameter that will be used for all visualizations.

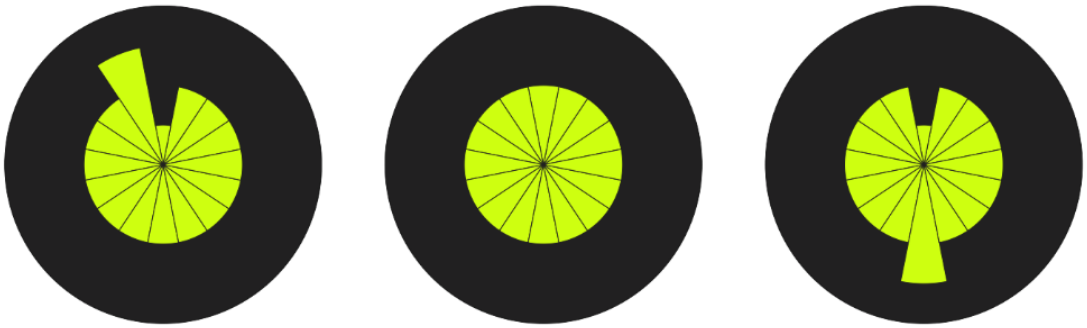

Figure 6: Three passing sonars built to illustrate the problems with using euclidean distance to measure their similarity. From left to right: q, p, r

Intuitively, one would want the distance from p to q to be smaller than the distance from p to r, as the *few* passes that make p and q different go in directions that are close to each other, whereas that does not happen between p and r. However, it is quite clear that $d_E(p,q) = d_E(p,r)$.

Seeing why that is a problem, let us now make our own attempt at defining a distance for passing sonars. Identifying $p_{-i} = p_{s-i}$ for the sake of simplifying notation, we set:

$$d(p,q) = \sum_{i=1}^{s} \left| (p_i - q_i) + \sum_{j=i-2}^{i+2} (p_j - q_j)\delta(i-j) \right|,$$

where

$$\delta(t) = (1+t)\chi_{\{t>0\}}(t).$$

The idea that guides the definition is that, if the difference between p_i and q_i is replicated by an opposite sign difference between p_j and q_j for j close to i, they should compensate each other when we compute the distance. One can then check that, for p, q and r as above (and $s = 16$), we have $d(p,q) = 0.1875$ and $d(p,r) = 0.53125$.

Coming back to the passing sonar examples we started with, we would expect Hazard and Mané's passing sonars to be closer to each other than to Sterling's, as the first two are more focused on passing towards their right (left winger to striker, say), whereas Sterling seems to pass more towards his left and, definitely, backwards. In fact, we have:

- $d(\text{Sterling}, \text{Hazard}) = 3.565$.
- $d(\text{Sterling}, \text{Mané}) = 2.746$.
- $d(\text{Mané}, \text{Hazard}) = 1.314$.

As we did for heatmaps, we can take this a bit further and find, for example, the strikers whose passing sonars are the most similar to Manchester City's Argentinian striker Sergio Agüero's. The

following table shows strikers with at least 2000 minutes on field over the last Premier League season, and it is sorted by how close the player's passing sonar is to Agüero's:

Name	Distance to Agüero's passing sonar
S. Agüero	0.0000
A. Mitrović	0.6710
N. Redmond	0.7092
S. Rondón	0.7276
Lucas Moura	0.7548
J. Vardy	0.8109
Heung-Min Son	0.8342
P. Aubameyang	0.8931
M. Rashford	0.8978
R. Jiménez	0.9193
C. Wood	0.9731
Gerard Deulofeu	0.9756
G. Murray	1.0047
M. Arnautović	1.2388
Roberto Firmino	1.2395
C. Wilson	1.3333
Diogo Jota	1.3604
A. Lacazette	1.6329
A. Barnes	1.7219
R. Lukaku	2.3574
H. Kane	2.5849
T. Deeney	2.8043

Table 2: Passing sonar comparisons between Agüero and other strikers

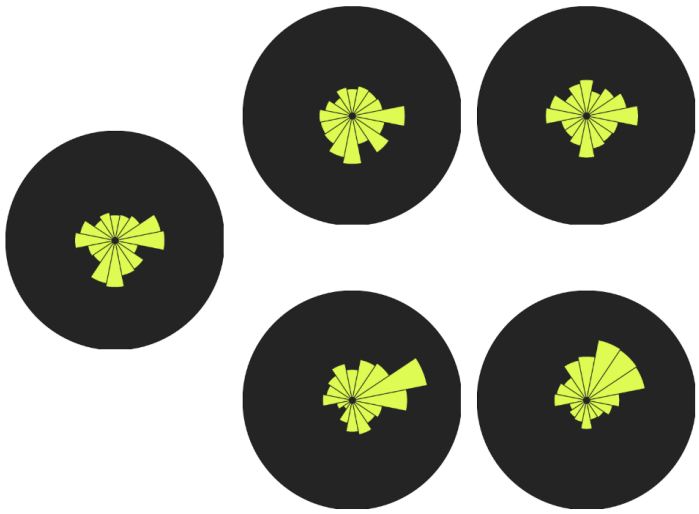

Figure 7: Agüero's passing sonar (on the left), compared with Mitrović's and Redmond's (at the top) and Kane's and Deeney's (at the bottom)

It is easy to see that the similarity metric catches that Agüero's passing sonar is similar to other relatively symmetric passing sonars with certain weight in passes that go backwards, and quite dissimilar to passing sonars with a clear assymmetry in one horizontal direction.

4 Comparing players in terms of their moves and passing intentions together

Having measures of similarity for both heatmaps and passing sonars enables us to compare two players both in terms of his moves and his passing intentions *at the same time*. Whereas it would seem natural to define one such combined distance as the sum of the heatmap distance and the passing sonar distance, the fact that these two have different scales advises against that. Therefore, a way to solve this issue, at least when we want to find the most similar player to a fixed player X, is to rank players in the list in terms of both distances to X, and define the *combined distance to* X as the sum of those two ranks.

Let us take a look at two examples of slightly different nature. The first one illustrates this for strikers with at least 2000 minutes on field in the last Premier League season against Liverpool's Brazilian striker Roberto Firmino (PS-Rank and H-Rank denote, respectively, the distance of the player's passing sonar or heatmap to Firmino's).

Name	PS-Rank	H-Rank	Combined distance
Roberto Firmino	1	1	2
S. Rondón	4	2	6
A. Barnes	5	6	11
Lucas Moura	7	5	12
N. Redmond	3	13	16
G. Murray	10	7	17
C. Wilson	2	16	18
J. Vardy	6	14	20
P. Aubameyang	9	11	20
T. Deeney	20	4	24
H. Kane	22	3	25
Heung-Min Son	16	9	25
C. Wood	18	8	26
Gerard Deulofeu	8	18	26
A. Mitrović	17	10	27
M. Arnautović	11	17	28
A. Lacazette	15	15	30
M. Rashford	19	12	31
S. Agüero	12	21	33
R. Jiménez	14	19	33
Diogo Jota	13	22	35
R. Lukaku	21	20	41

Table 3: Heatmap and passing sonar comparisons between Firmino and other strikers

We see that Rondón, who happens to be the closest to Firmino, has the most similar heatmap to Firmino's, while he ranks third in terms of passing sonar distance - their moves are quite wide, although they both tend to fall on the left wing, and are therefore very similar, and their passing distribution is very symmetric. Tottenham Spurs' Harry Kane moves quite similarly to Firmino as well, although his passing is more erratic, as he has a clear tendency to pass towards the right wing. On the other hand, Callum Wilson, whose passing is quite uniform as well, is the closest to Firmino in terms of passing sonar, although his moves look quite dissimilar, as he is more prone to fall on the right wing.

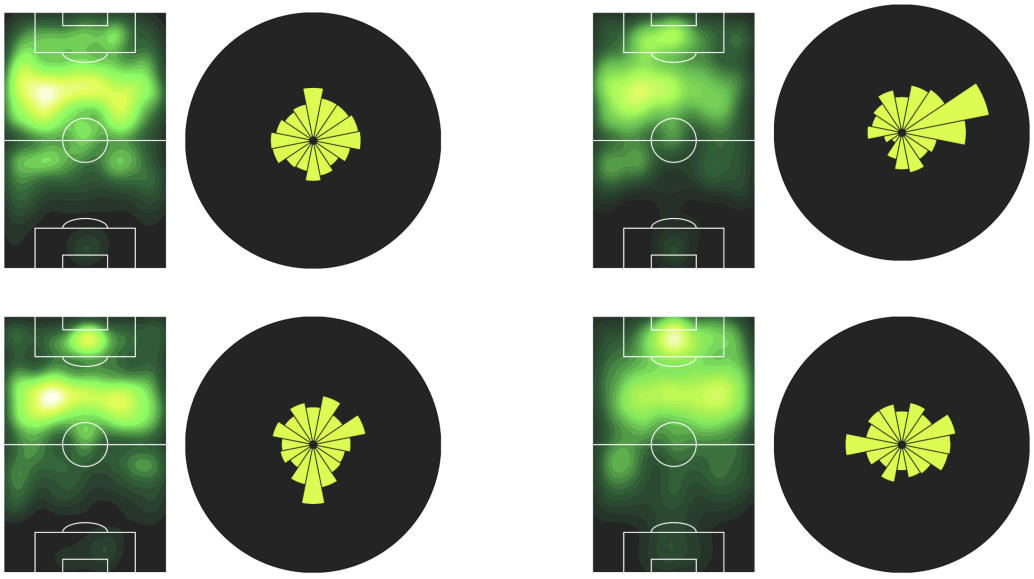

Figure 8: Firmino's (top-left), Kane's (top-right), Rondón's (bottom-left) and Wilson's (bottom-right) heatmaps and passing sonars, respectively

Running the same exercise for Leicester City's number 9 Jamie Vardy yields interesting results as well.

Name	PS-Rank	H-Rank	Combined distance
J. Vardy	1	1	2
Roberto Firmino	4	10	14
C. Wilson	6	8	14
P. Aubameyang	9	5	14
S. Rondón	5	11	16
C. Wood	14	2	16
A. Lacazette	13	6	19
N. Redmond	3	16	19
Gerard Deulofeu	2	19	21
M. Arnautović	8	13	21
T. Deeney	21	3	24
A. Mitrović	15	9	24
M. Rashford	18	7	25
S. Agüero	7	18	25
A. Barnes	12	14	26
H. Kane	22	4	26
Lucas Moura	11	15	26
Heung-Min Son	17	12	29
R. Jiménez	10	22	32
G. Murray	16	17	33
Diogo Jota	19	21	40
R. Lukaku	20	20	40

Table 4: Heatmap and passing sonar comparisons between Vardy and other strikers

We see that Firmino, Wilson and Aubameyang are all tied in terms of closeness to Vardy, although neither of them are too close to him in both moves and passing intentions at the same time. The player with the most similar moves to Vardy's is Burnley's striker Chris Wood, although he is thirteenth out of twenty-one in terms of passing sonar similarity. Likewise, Deulofeu is the closest player to Vardy with respect to passing intentions, but he is the eighteenth in terms of heatmap.

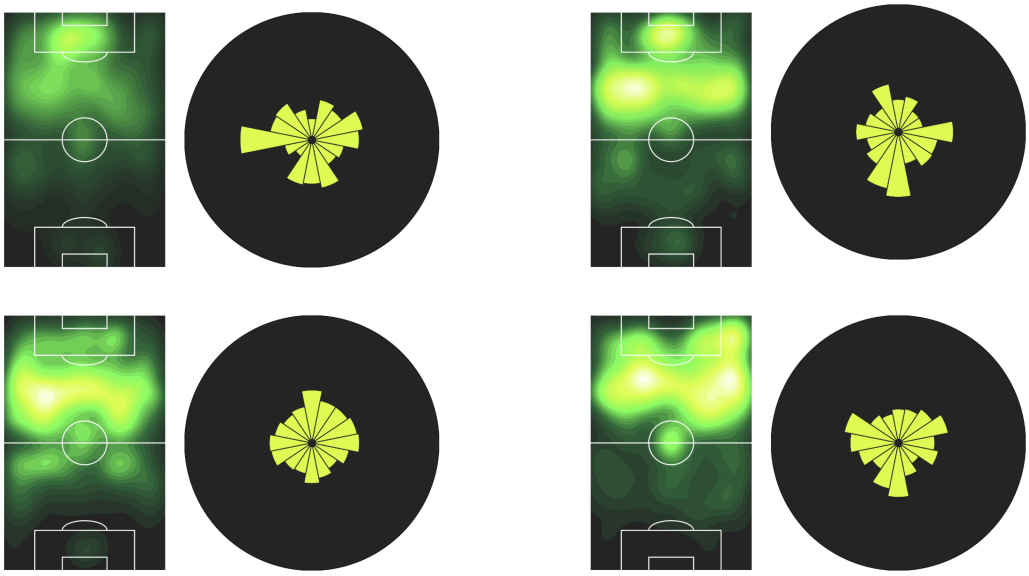

Figure 9: Vardy's (top-left), Wood's (top-right), Firmino's (bottom-left) and Deulofeu's (bottom-right) heatmaps and passing sonars, respectively

5 Expected passing sonars

As we have just seen, Vardy is a very interesting example because the players that are similar to him in terms of heatmap are not in terms of passing sonar, and viceversa. However, this is a quite unique situation, and most of the times there is a strong relationship between those. In general, the location of a player on the pitch determines a big part of his passing options, and we can exemplify this by means of a tool we call the **expected passing sonar of a player given his heatmap**, which we can concisely define as *the passing sonar an average player would have if his heatmap coincided with the heatmap of the player under study.* More precisely, the algorithm we execute to compute expected passing sonars is the following:

1. Divide the pitch into 16×12 rectangular cells (this is inspired in Karun Singh's work [1] on expected threat).

2. For each cell, estimate the probability for a player to pass the ball when being in that cell as the number of passes that originate from that cell divided by the number of events that happen inside that cell, and the average passing sonar for that cell as the passing sonar we obtain when we look at all passes that originate in that cell[5].

[5]These estimations are performed on an independent training dataset, which in our case is the set of all events in the 17/18 season of the Premier League.

3. For a player with a given heatmap, compute his expected passing sonar as the average of all the cells' passing sonars weighted by the time the player spends in each cell and the probability for that player to choose to pass when located in that cell.

Applying this technique to all Premier League field players with more than 2000 minutes on field during the 18/19 season yielded some interesting results:

- The distribution of distances between passing sonars and expected passing sonars looks like the following:

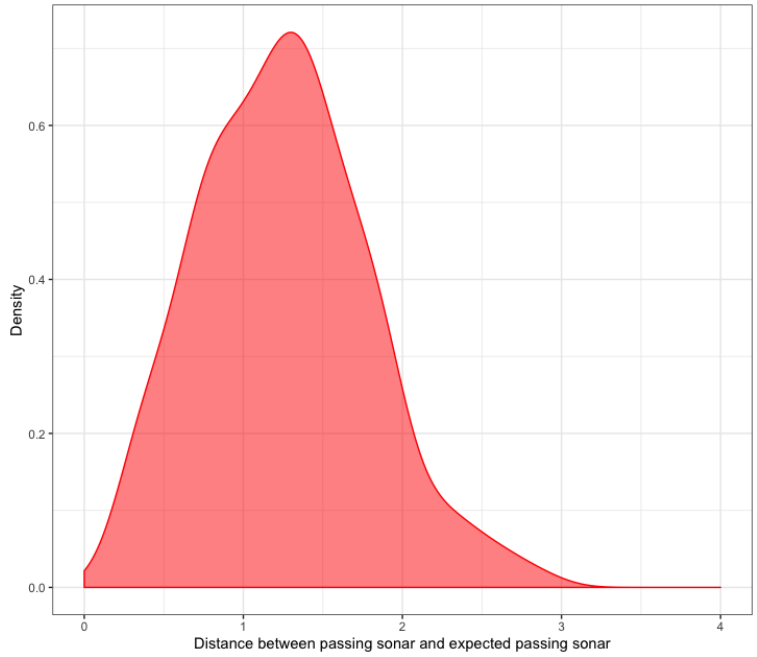

Figure 10: Density plot for the distribution of distances between passing sonar and expected passing sonar for all players

Comparing these with some examples such as the ones that we have shown in Section 3 should convince the reader of the fact that, in general, these distances are small, which *proves* that the location of the player on the pitch determines to a certain extent what his passing patterns are going to look like.

- Splitting the distribution of distances between passing sonars and expected passing sonars by positions (see the image below), we notice that, in general, central attacking midfielders are less predictable players than centre midfielders and defensive midfielders, as one could have expected. We can also use the quantiles of these distributions to classify players in terms of the predictability of their passing intentions given where they are located.

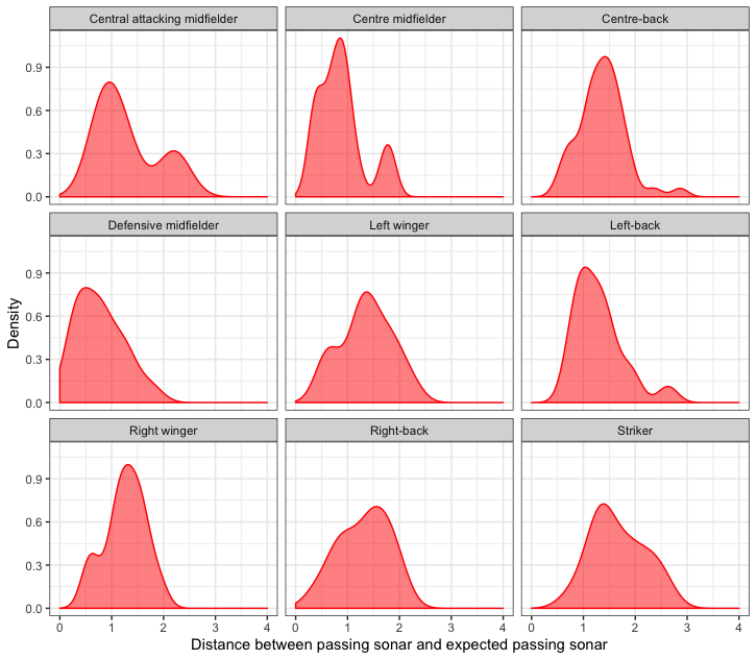

Figure 11: Density plot for the distribution of distances between passing sonar and expected passing sonar for all players, split by player position

For example, just fixing our attention on wingers, it is interesting to compare the differences between real (left) and expected passing sonars (right) for AFC Bournemouth's David Brooks (the most predictable winger, at the top, who plays prominently in the right wing and who tends to pass forwards and towards his left, i.e., possibly to the central attacking midfielder, the striker or a right-back who has overtaken him), Everton's Richarlison (who is around the median *predictability*, in the middle, and passes symmetrically towards both sides but passes backwards with more frequency than expected) and West Ham's Felipe Anderson (who is around the 90th percentile in the distribution of distances between passing sonar and expected passing sonars, at the bottom, who plays inwards more frequently than expected, and whose forwards and backwards passing proportions are inverted with respect to expectation):

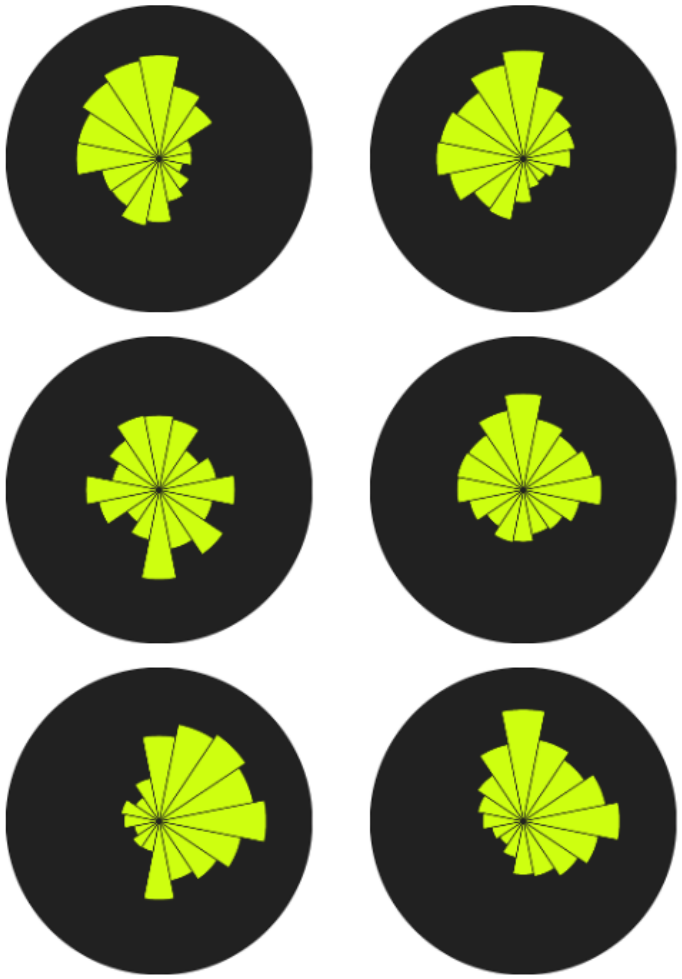

Figure 12: Comparisons between passing sonar and expected passing sonar for Brooks (top), Richarlison (middle) and Anderson (bottom)

6 Future work

While the approach taken in Section 4 enables us to compare players *from a list* in terms of their passing sonars and their heatmaps at the same time, this is far from ideal. We would like to have a closed formula to compute the distance between two players based on the two visualizations regardless of what other players we aim to compare them with. To be able to do that, we need to understand the scales of both distances to be able to normalize them before summing them up and combining them into a single distance.

On the other hand, all examples of visualization comparisons we have talked about so far are *static*, i.e., they rely on us having a significant and self-contained sample of events that describes the players moves/passes (say, a season). However, football is a dynamic subject and, obviously, writers need to be able to produce content every week. We can use the metrics defined above in order to automatically uncover stories such as the following[6]:

1. There is a significant difference between Pogba's heatmap over the last five games with respect to the previous ten.
2. Van Dijk's passing sonar is significantly different when he is paired with Matip to when he is paired with Joe Gomez.

For that, we would just need to understand the distribution $\mathcal{D}$ of distances between two instances of a visualization for a given player on different sets of games (varying, say, over all players in the competition and all pairs of sets of games in which the player has a significant amount of minutes played - this would be a heavy computation, but it only would need to be run once). Then,

1. We could compute the distance between Pogba's heatmap over the last 5 games against his heatmap over the previous 10 games and, if it is greater than the 90th percentile of $\mathcal{D}$, raise an alert if Pogba has played a significant amount of minutes in both timeframes.
2. If Van Dijk reaches, say, 300 minutes played with Joe Gomez and he had also played more than 300 minutes with Matip previously, that could trigger the computation of the distance between Van Dijk's passing sonar when playing alongside Matip and when playing alongside Gomez. If this happens to be greater than the 90th percentile of $\mathcal{D}$, we raise an alert.

References

[1] K. Singh, *Introducing Expected Threat (xT) - Modelling team behaviour in possession to gain a deeper understanding of buildup play*, available at `https://karun.in/blog/expected-threat.html` (2019).

[2] M. Rosenblatt, *Remarks on Some Nonparametric Estimates of a Density Function.* Ann. Math. Statist. 27 (1956), no. 3, 832–837.

[3] Villani, C. (2016). *Topics in optimal transportation.* Providence, Rhode Island: American mathematical society.

[4] `https://twitter.com/etmckinley/status/1046389278153068545`.

[5] `https://twitter.com/Mladen_Sormaz/status/1110110341864804352`.

[6]The reader should take these simply as examples: we are not claiming any of them to be true.

List of Figures

1 David Silva's, James Maddison's and Christian Eriksen's heatmaps of all their passes in EPL 2018/19 . . . 3
2 Density functions corresponding to uniform distributions on three disjoint intervals . 4
3 Ramsey's heatmap (on the left), compared to Özil's and Lingard's (at the top) and Antonio's and Groß's (at the bottom) . . . 6
4 Declan Rice's heatmaps when playing as a lone defensive midfielder (left) and as a left defensive midfielder in a double pivot (right) . . . 8
5 Sterling's, Hazard's and Mané's passing sonars of all their passes in EPL 2018/19 . . 9
6 Three passing sonars built to illustrate the problems with using euclidean distance to measure their similarity. From left to right: q, p, r . . . 10
7 Agüero's passing sonar (on the left), compared with Mitrović's and Redmond's (at the top) and Kane's and Deeney's (at the bottom) . . . 12
8 Firmino's (top-left), Kane's (top-right), Rondón's (bottom-left) and Wilson's (bottom-right) heatmaps and passing sonars, respectively . . . 14
9 Vardy's (top-left), Wood's (top-right), Firmino's (bottom-left) and Deulofeu's (bottom-right) heatmaps and passing sonars, respectively . . . 16
10 Density plot for the distribution of distances between passing sonar and expected passing sonar for all players . . . 17
11 Density plot for the distribution of distances between passing sonar and expected passing sonar for all players, split by player position . . . 18
12 Comparisons between passing sonar and expected passing sonar for Brooks (top), Richarlison (middle) and Anderson (bottom) . . . 19

List of Tables

1 Heatmap comparisons between Aaron Ramsey and other attacking midfielders . . . 6
2 Passing sonar comparisons between Agüero and other strikers . . . 11
3 Heatmap and passing sonar comparisons between Firmino and other strikers . . . 13
4 Heatmap and passing sonar comparisons between Vardy and other strikers . . . 15