**Recruiting the most effective talents based on visual positional**

**features and model based zonal passing attributes**

**Submission category**:  Open submission: event and tracking data

**Are you submitting on behalf of a commercial organization**: No.

**Research question:** How to identify players for a specific position and role, who would complement the rest of the team based on the potential player’s **positional zone on the pitch and zonal passing network** that results in **chance creation events**.

**Objective of this research:** The objective of this research is manifold:

* Help in evolving a multidimensional gameplay that is well distributed over the key areas of the pitch and not skewed to a side/ zones.
* Tracking weaknesses in gameplay by identifying the areas of pitch and spaces that are not well covered by players.
* Identifying a player’s naturally preferred ‘zone of control’ and promoting them to play in those zones
* Recruiting players in a way such that they do not much overlap/interfere with existing players’ zone
* Developing a ML model to predict whether an incoming player’s ‘zone of control’ and zonal passing network can improve the chance creation for the team.

**Rationale for chosen topic:** Recruiting the next big star for a team is often the biggest challenge in modern football, especially at the highest level. Throwing money at big names does not always guarantee the best results as players often take a lot of time to ‘fit’ into a team and complement the team’s style of play. So how DO we find the player who brings out the best from your team and vice versa?

In modern football, even players with similar positions tend to perform slightly different roles for the team and more often than not end up covering different areas of pitch. For example, Lionel Messi and Arjen Robben were both right wingers, with quite different heatmaps. While Messi prefers to drift in more centrally in a playmaking role, Robben would hog the sidelines and cut in sharply to the left and finishing with a deadly pass or finish. Moreover, while the best of players is adaptable and can play in different roles and systems, they always tend to occupy or shift to the areas they are most comfortable on a pitch. **Identifying players whose preferred ‘zone-of-control’ matches the zone where the team is leaving a lot of spaces/ current players are not occupying could be a key method to recruit players**. This way the team doesn’t end with multiple ‘similar’ type of players without addressing the key areas of weaknesses. For example, **Fig6** shows that the heatmaps of Hakim Ziyech and Kai Havertz closely matches each other while its similar for Timo Werner and Christian Pulisic. All of these players currently play for Chelsea and it remains to be seen how Chelsea manages all these players.

The previous part solves only a part of the problem, that of players effectively occupying distinct zones with enough overlaps while not encroaching on each other and using the entire field. But what about the system that they are about to be part of or the role? For that, we could look at the zonal passing network of the team over the course of a season and applying machine learning techniques to predict how a prospective player’s style/ passing map from a zonal point of view could improve the chance creation of the team. The combination of the above two techniques could recruit a player who would almost seamlessly fit into a team and start contributing to positive results from get go.

**Your approach to answering this question:**  Lets’ look at first part of the problem, that is identifying players by positional zones. I have chosen Man Utd as the team to analyze and provide insights on. The data that was provided recorded the first four matches of every team in the 2018-19 season. To begin with, the touches of all the players who played any number of minutes in any of the matches were modified into a data frame. But for the sake of this particular topic and because there can be only 11 players on the pitch at a time, 11 players with the highest number of touches, i.e. players with than more than 145 touches spanning the four matches were selected. Before diving into analyzing the entire team, I focused on an individual player’s positional plots to define how to approach the problem. **Fig1** shows all the touches of Paul Pogba in Man Utd’s first match of the season, against Leicester City. We use [DBSCAN](https://en.wikipedia.org/wiki/DBSCAN) clustering method to remove the outliers and get the zone of the pitch with his highest number of touches. This can be his ‘zone-of-control’, usually the densest part of heat zones. To properly define the zone, a [convex hull](https://en.wikipedia.org/wiki/Convex_hull#:~:text=In%20geometry%2C%20the%20convex%20hull,convex%20set%20that%20contains%20it.) has been drawn and that has been shown in **Fig2**. Now we can see a pattern emerging and we can begin to use the above process for the entire team. As mentioned above, the clustering parameters for the 11 players with the highest number of touches across the four matches was optimized, their touch coordinates plotted, and the resulting convex hulls drawn. This is shown in **Fig3.** While we can see patterns emerging, due to overlapping plots it is difficult to draw much insights from that. So then we turn to [Voronoi](https://en.wikipedia.org/wiki/Voronoi_diagram#:~:text=In%20mathematics%2C%20a%20Voronoi%20diagram,%2C%20sites%2C%20or%20generators).) tessellation by taking the mean point of the clusters as the Voronoi points. That technique helps us to identify the zones easily. The transparency of the coloring of the zones is inversely proportional to the number of touches of the players, whose names and number of touches along with the Voronoi slices are shown in **Fig4.**

From the figure, it’s clear that the style of play is heavily skewed to the left, with larger and fainter Voronoi slices to the right. The mean position of 7 outfield players will fall to the left side of the pitch. The zone of the players overlap heavily towards the center with lack of number of touches in the final third. Lukaku, the striker’s zone is almost same as that of Pogba and Fred, which shows a lack of attacking threats. The right side also leaves open spaces that can be exploited by teams. Another interesting thing to note is that the mean of the clusters closely match the mean of the Voronoi slices for players with higher number of touches. So it can be inferred that with a larger dataset, the zones would start to look more accurate and close to the cluster shapes. To identify an ideal prospect, the prospect’s Voronoi slice/ convex hull over an entire season should closely resemble the shape of the zone he has to occupy and that basically becomes a computer vision problem (using image similarity algorithms).

**Fig5** shows the passing network from Man Utd vs Burnley match (only match where all goals were non-penalty for United) with every pass creating a chance for United. The linkage for the goals is shown in bolder lines and bigger points. To solve the second part of the problem, a team whose gameplay significantly improved after the addition of a player in the latter part of the season has to be identified. The eleven players with the highest number of touches (including the player who changed the team’s fortune) have to be identified and a zonal passing network for big chances created (chances in the 20 X 20 area in front of goal) that involves any of those 11 players has to be created. Because it’s a labeled dataset (we know the big chances created improved after the addition and which of those chances are goals/ shots on target/ post etc.) we could easily develop a model using supervised classification techniques like Logistic regression, Random forests, Decision trees and so on. The validity of the model can be evaluated by doing a train-test split on the dataset and it can be seen whether or not the positional features extracted from the dataset can make the model accurately predictive. The dataset would have to include the player in question features from both his previous and current teams.

**The applications of this research within a professional football environment and how this work will help teams win:**  As mentioned previously, this has an almost direct influence on how team recruitment strategy and gameplay formulation based on incoming players. This can be seen as a two-step filtering procedure. The first step involves identifying the players that will not have a conflict with other important players on the team and will strengthen the team from a purely mathematical space partitioning point of view. The players identified by this process can be further evaluated by the second step. The model that I talked about developing should be able to predict whether the player’s positional and passing attributes will actually improve the team’s ability to create more big chances. The results from the first part have already been shown to some extent here and that is a more visual analytics approach for recruitment. The second part is a more model-based statistical approach to predicting outcomes. The combination of the two could be an interesting and potent methodology for recruiting players.