Turner Schwiebert

Seminar in Data Analytics

Dr. Sarah Supp

October 28, 2022

Understanding the Positional Roles of Professional Soccer Players: A Clustering Analysis

The modern landscape of professional soccer is constantly evolving. One important aspect of this evolution is the stylistic innovation of Europe's top teams and their respective coaching staffs. Much of this innovation comes in the form of positional use of players in new spaces and roles. Thus, it is increasingly important to understand the various roles that these players inhabit on the field. This research aims to profile all players in Europe's top five leagues[1] by their modern positional role.

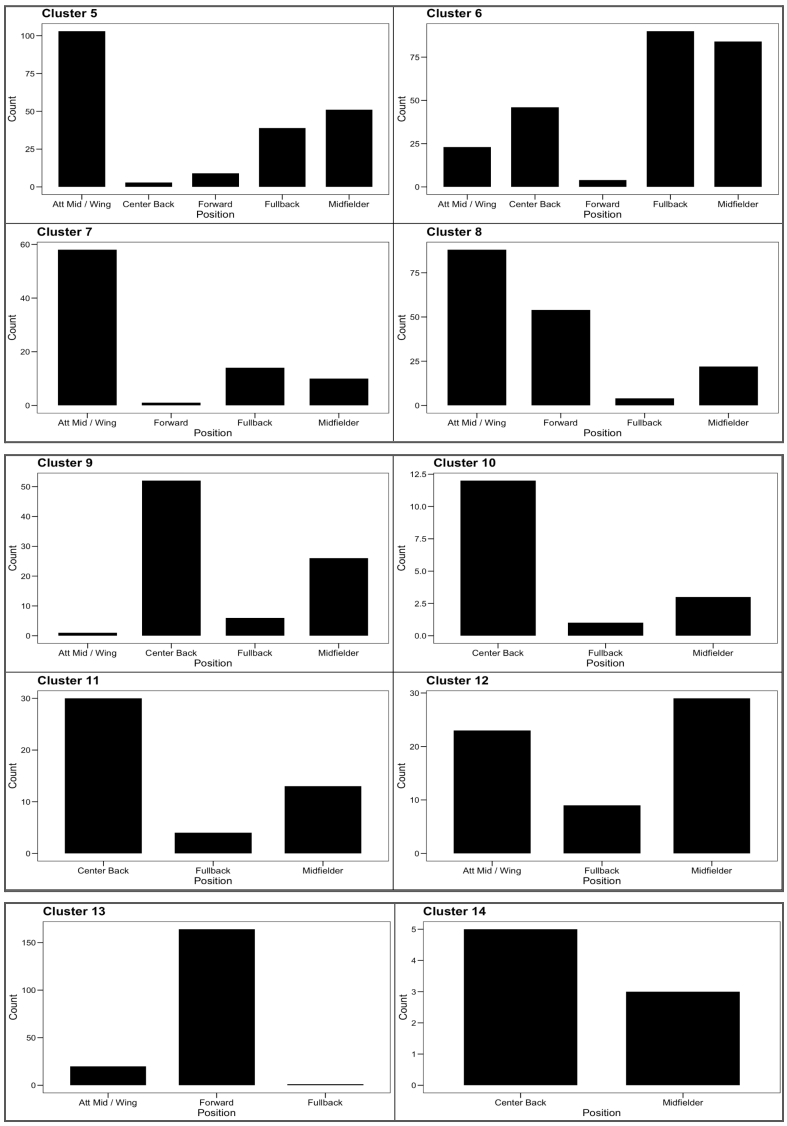
Before any methods were employed to provide insights into this research topic, data was first verified and collected. All data was web scraped from FBRef.com, the soccer website of the Sports Reference family of websites (FBRef, 2022). Sports Reference is dedicated to providing open-source sports data, and their soccer data is extensive and accurate. In fact, previously published research has used Sports Reference for research in various sports contexts (Terner, 2021). Thus, data on all players from top European leagues in Europe from FBRef is both trusted and suitable for this research topic as well. More specifically, the data collected from FBRef is "per 90-minute" performance data of all players in the top leagues. Approximately 130 of these variables were collected. These variables cover shooting, passing/pass types, goal/shot creation, defense, possession, and miscellaneous aspects of each player's performance for the past 365 days.

[1] 1685 players from the Premier League (Eng.), Bundesliga (Ger.), Serie A (Ita.), Ligue 1 (Fra.), La Liga (Spa.)

The methodology used to provide insights into the modern positional roles of professional soccer players consists of k-means and hierarchical clustering. These methods have been successfully employed in previously published studies with similar goals (Gyarmati, 2014, Shelly, 2020). However, this clustering analysis has and will continue to occur in two phases. First, all data will be run through both hierarchical and k-means clustering. For phase two of the research, all the data will be split into groups representing the players' traditional position on the field. It is important to note that this traditional position is significantly different from the profiles I am attempting to assign to each player. Player performance varies within each position, so this phase aims to capture this variance in a clustering analysis, revealing positional roles within each position. Thus, I expect actual insights into modern positional roles to be generated from this phase.

In accordance with my expectations for the first phase of my research, the results when all the data was inputted into clustering algorithms were not very interpretable. The prevalence of positions within 14 clusters outputted from hierarchical clustering can be found below in Figure 1.

Figure 1: Prevalence of Traditional Position within Hierarchical Clusters



Note: Unobserved positions in bar plots indicate no prevalence of those positions in that cluster.

Figure 1 shows that there is a large amount of overlap among the clusters. This overlap of traditional positions within clusters using all players in the data set yields little to no possible insights. There is some pattern in that very few center backs are being clustered with forwards.

However, everything in between is completely uninterpretable. For example, midfielders, fullbacks, and attacking mids/wingers are frequently clustered together. Thus, there is very little possibility to generate useful insights when clustering with all the players pooled together. This is because there is too small variability between these three positional groups, but there is enough variability between center backs and forwards. In general, clustering all the players together is simply too complex to profile the players by their distinct positional role. This complexity is also evident in the figures below.

Figure 2: Mean Dribblers Tackled (per 90 minutes) by Cluster

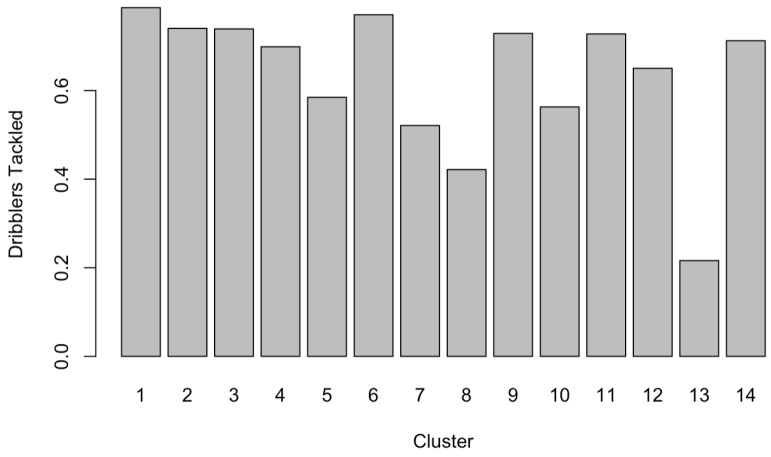
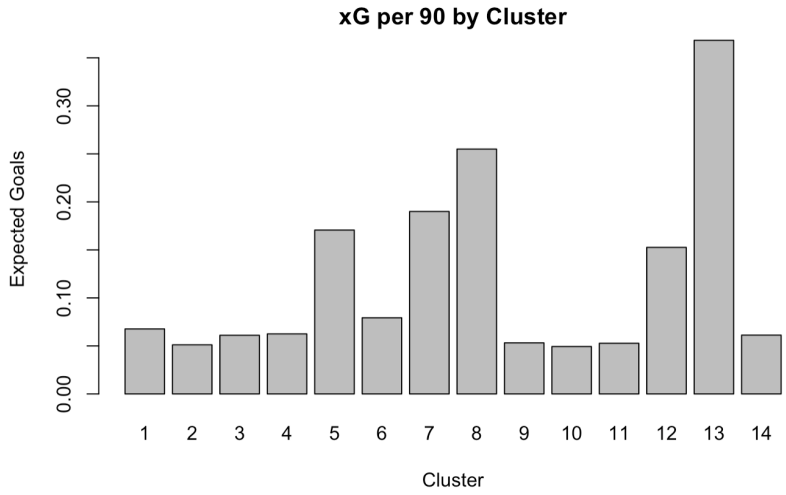


Figure 3: Mean Expected Goals (per 90 minutes) by Cluster



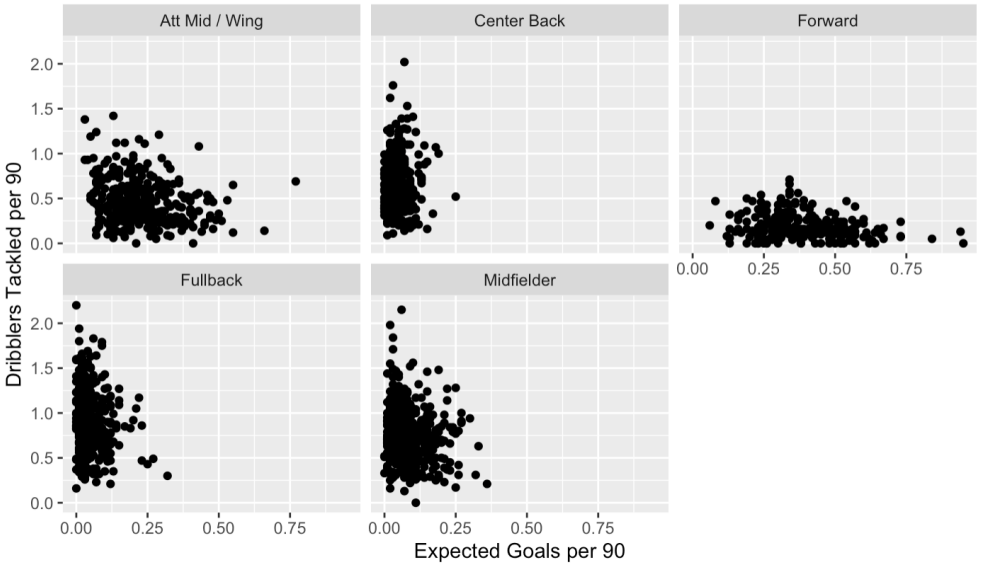
Before analyzing Figures 1 & 2, it is important to note that “expected goals” is a metric for offensive output, and “dribblers tackled” is a metric for defensive output. There are few patterns across clusters for these two variables. This further shows the difficulty when interpreting clusters generated using all players in a single dataset. It is impossible to distinctly profile these players into their positional roles, especially with 130 performance variables.

For phase two of the player profiling process, the dataset will be split by traditional position. This split is illustrated below.

Table 1: Distribution of Observations by Position

Position	Observations
Att. Mid./Winger	340
Center Back	364
Forward	232
Fullback	340
Midfielder	409
Total	1685

Figure 4: Scatterplot of Expected Goals and Dribblers Tackled Across Traditional Position



The plots in Figure 4 begin to show how phase two will be better at profiling players into positional roles. Clustering algorithms used on these positional groups will group them into clusters of players who perform similarly at each position, thus leading to current positional roles. However, more analysis and tuning are required to test this hypothesis.

In conclusion, the first phase of this analysis shows that variability amongst midfielders, fullbacks, and attacking midfielders/wingers is not sufficient to provide in-depth player profiling when all players are pooled together. However, variability between center backs and forwards is great enough that they were very rarely clustered together. This shows that the variability of players who only stay on the defense of a team (center backs) perform extremely differently than those that stay on the attack of a team (forwards), which makes sense. Thus, one main takeaway from this research thus far is that variability in the performance of players between these positions on the field has blended in the modern game and is not sufficient for profiling positional roles. I am optimistic about employing clustering methods for these positional groups. This is the next step in my project.

Works Cited:

FBRef. (2022). Retrieved From https://fbref.com/en/

Gyarmati, L., Kwak, H., & Rodriguez, P. (2014). Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*. https://doi.org/10.48550/arXiv.1409.0308

Shelly, Z., Burch, R. F., Tian, W., Strawderman, L., Piroli, A., & Bichey, C. (2020). Using K-means clustering to create training groups for elite American football student-athletes based on game demands. *International Journal of Kinesiology and Sports Science*, *8*(2), 47-63.

Terner, Z., & Franks, A. (2021). Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, *8*, 1-23.