What's a Midfielder Anyway?

Clinton Raye

Problem Statement

Association football, commonly known as soccer, is the most popular sport in the world. A team consists of eleven players: one goalkeeper and ten 'outfield' players. These outfield players are typically categorized into one of three positions: defense, midfield, and attack. However, with the evolution of soccer's tactics in the modern era there is an endless number of names that are used to describe a player's position. The popular video game FIFA's 'Ultimate Team' mode allows managers to assign their players one of seventeen positions [FIF]. This excludes popular labels commonly assigned by pundits: 'false nine', 'second striker', 'trequartista', 'inverted wingback', etc.

This analysis aims to evaluate several data driven approaches to classifying player 'positions' based on their in-game actions, not their qualitative descriptions. Previous research has analyzed the network associations between players [Wal], optimal team formation solutions [Sal07], and the performance characteristics of players [SIA19]. However all of these analyses rely on assumed player positions, which are the result of expert based labeling. No known research or analysis shows an attempt to define players roles and positions based on statistical approaches.

The goal of this analysis is to provide a model through which individual players, based upon their in-game statistics, can be meaningfully clustered into a position. This position will be more representative of a player's role and characteristics than the traditional position naming method is. Various clustering methods will be used and compared: K-Means, Gaussian Mixture Model (GMM), and Spectral Clustering. The resulting model selected as best performing could be used by fans and analysts alike to better understand a player. Players could be compared to similar players on other teams or in other leagues in a more meaningful manner. Additionally, the number of clusters in the model will reveal how many different 'positions' exist in the population.

Data Source

The data to be utilized in this analysis was sourced from Kaggle [RSK]. The data consists of observations for 2000+ players in Europe's "Top Five" soccer leagues. Each observation contains 400 different factors. The data set from the 2019-2020 season will be used for this analysis.

The factors will be reduced to only contain quantitative factors. This is justified based on the desire of the analysis to cluster players based only on their measured in-game performance. As a result, fields such as 'nationality', 'foot preference', etc. will be discarded.

Several of the existing factors are listed as 'per 90', or in other words the value is representative of the player's per game statistics. Additional factors will be treated to become per game values. This is in effect normalizing the data. Normalization is justified to account for the variation in the number of games each player has played in a given season. Furthermore, players who recorded less than 450 minutes, or the equivalent of 5 full games, will be discarded. This action is taken to reduce the impact of observations based on a small sample size.

The dataset contains numerous statistics that relate only to goalkeepers. The data will therefore be split into 3 sets: one consisting of all observations and all players, one consisting only of outfield players with no goalkeeper specific stats, and one consisting only of goalkeepers with goalkeeper specific stats. The aim of splitting the data into distinct sets is to produce models which are more informative given the large difference between the in-game statistics of outfield vs goalkeepers.

Methodology

Prior to training any clustering model on the data, Principal Component Analysis will be performed to reduce the data to two and three dimensions. Each model will then be trained on six data sets, each of the different player/statistics sets described above for each of the two and three component reductions. Although two or even three dimensions may not be sufficient to capture the variability of the data, the dimensions will be limited to a number that can be represented visually. This is due to the stated goal to produce a result that is digestible by the average soccer fan.

K Means will be utilized as the first of three clustering models trained on the dataset. K Means was selected as it is the most straight forward clustering model to understand and as a result provides a strong baseline for comparison with the other model. K Means may prove to be inefficient at clustering the observations in a meaningful manner since it relies on compactness of the data to be organized.

GMM will be utilized as the second of the three clustering models. GMM was selected as it has the ability to capture multiple, partial assignments to clusters for any given observation. Given the nature of the dataset, it is not unreasonable to assume that any given player may belong to multiple different positions or roles. Players are often deployed in different manners between games. GMM may assist in revealing the complexity involved in cleanly delineating players by position based on data.

The final of the three models to be utilized is the Spectral Clustering method. Spectral Clustering was selected because it provides a different method of assessing similarity than K Means and GMM. Specifically, it is a graph clustering method rather than a compactness clustering method. Spectral Clustering has the ability to describe the relationship between observations in a manner that is not the simple distance or probability of belonging to a distribution. Additionally, Spectral Clustering has built-in dimensionality reduction, which may be of use for the high dimensional dataset.

Each of the three models will be trained on the data using varying numbers of clusters. A final number of clusters for each combination of model and dataset will be selected using the Kneed algorithm [Rag], which finds the trade-off model performance and number of clusters.

Evaluation

The result of the Kneed algorithm is the nominal number of clusters in the dataset. This will answer the question posed regarding how many 'positions' truly exist in the population. For each model, however, the Kneed algorithm must be run on an appropriate statistic. The K-Means models will be evaluated using the sum of squared differences, based on euclidean distance, between each observation and it's assigned cluster center. Due to the underlying probabilistic nature of the GMM and it's lack of discrete assignments, GMM models will be evaluated using the Akaike Information Criteria (AIC). Finally, the Spectral Clustering models will be evaluated by examining the 'eigengap' between the eigenvalues of the Laplacian matrix.

The success and validity of each of the clustering methods for the full player and statistic dataset can be evaluated using known features of the dataset. The dataset includes goalkeepers, which operate under a different set of rules than outfield players. One would expect that each clustering model has a cluster that is dedicated to containing only goalkeepers. The GMM may contain goalkeepers that have some probability of belonging to another position cluster based on the players style.

The original dataset does include labels, that is positions, however since this analysis is aimed at evaluating players outside of their labels, labeling accuracy is not an appropriate method of evaluating each model. Instead, a measure comparing between cluster variability and within cluster variance must be used, namely the Calinksi-Harabasz Index (also known as the Variance Ratio Criterion) [Cal74]. This criterion will assist in evaluating whether the Gaussian Mixture Model performs better than the K Means model. However, it may be insufficient to describe the performance of the Spectral Clustering model since Spectral clustering is not reliant on simple distance metrics for assignment.

Finally, a survey of players assigned to each cluster will be analyzed to determine how the cluster model's assessment aligns with popular notions of player positions. This will speak to the usefulness of the model in regards to how well the common fan can understand the output.

Results

Table 1 below displays the explained variance ratio for each combination player and statistics dataset for PAC reductions to two and three components. Each dataset's variance can be explained with 2 components to 95% or greater. This ensures that visualizations in two or three dimensions will accurately capture the variability in the data.

Dataset	2 Component PCA	3 Component PCA
All	97.66%	99.54%
Outfield	98.61%	99.55%
Goalkeepers	96.38%	98.90%

Table 1: PCA Explained Variance

Figure 1 below illustrates the observations for the outfield player and outfield only statistics dataset once reduced to 2 components using PCA. The dataset shows a distinct compactness of the overall dataset without clear separation of groupings. From these images alone, one could assume that there is no non-linear behavior linking the observations and that Spectral Clustering will perform poorly as a result. Additionally, since there is not groupings, K-Means may struggle to delineate the clusters in a meaningful manner. GMM appears to be the most appropriate model as there is one primary grouping that likely could be split into smaller, overlapping clusters.

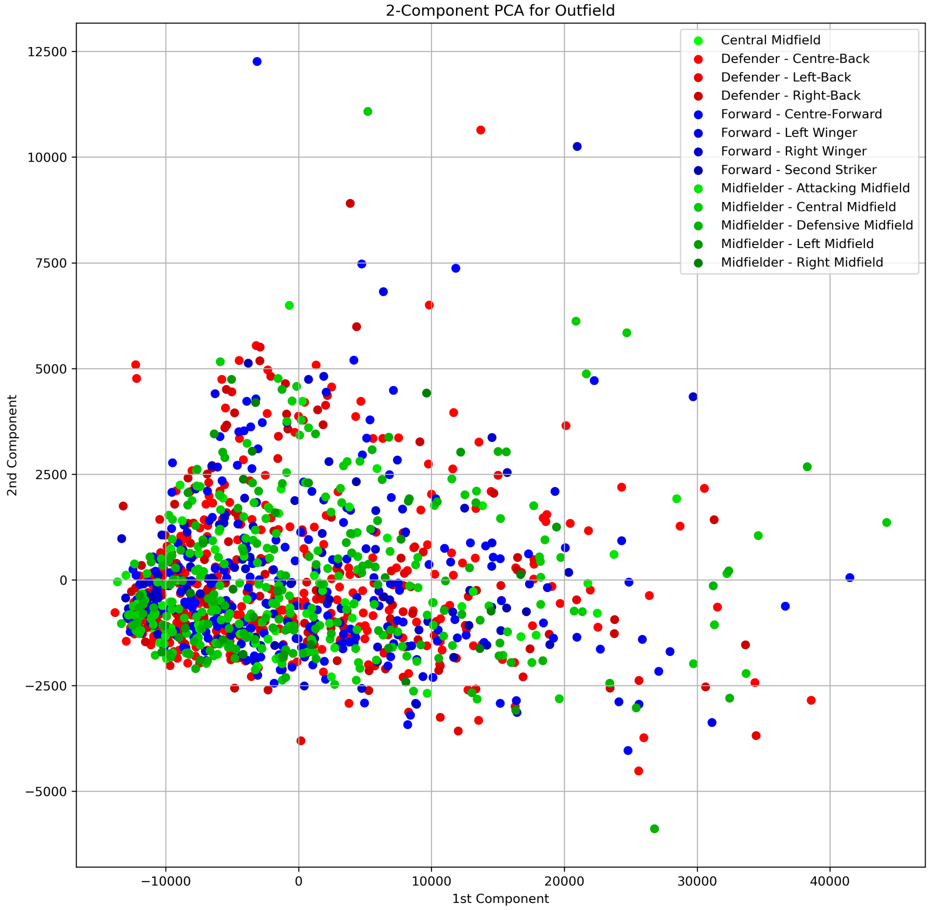

Figure 1: Outfield Players/Outfield Statistics 2 Component PCA

Table 2 below lists the number of clusters selected for each model, dataset, and number of components combination. Notably, Spectral Clustering yields the same number of clusters for all but one dataset. This can likely be traced again to how compact the observations are when transformed into the principal directions.

Model	All		Outfield		GKs	
	PCA - 2	PCA - 3	PCA - 2	PCA - 3	PCA - 2	PCA -3
K-Means	5	6	5	5	6	7
GMM	4	7	8	6	2	4
Spectral Clustering	3	3	3	3	2	3

Table 2: Number of Clusters Selected by Model

Figure 2 below illustrates the sum of squared differences for cluster numbers ranging from 2 to 20 for the outfield players and outfield statistics dataset when transformed into two principal components.

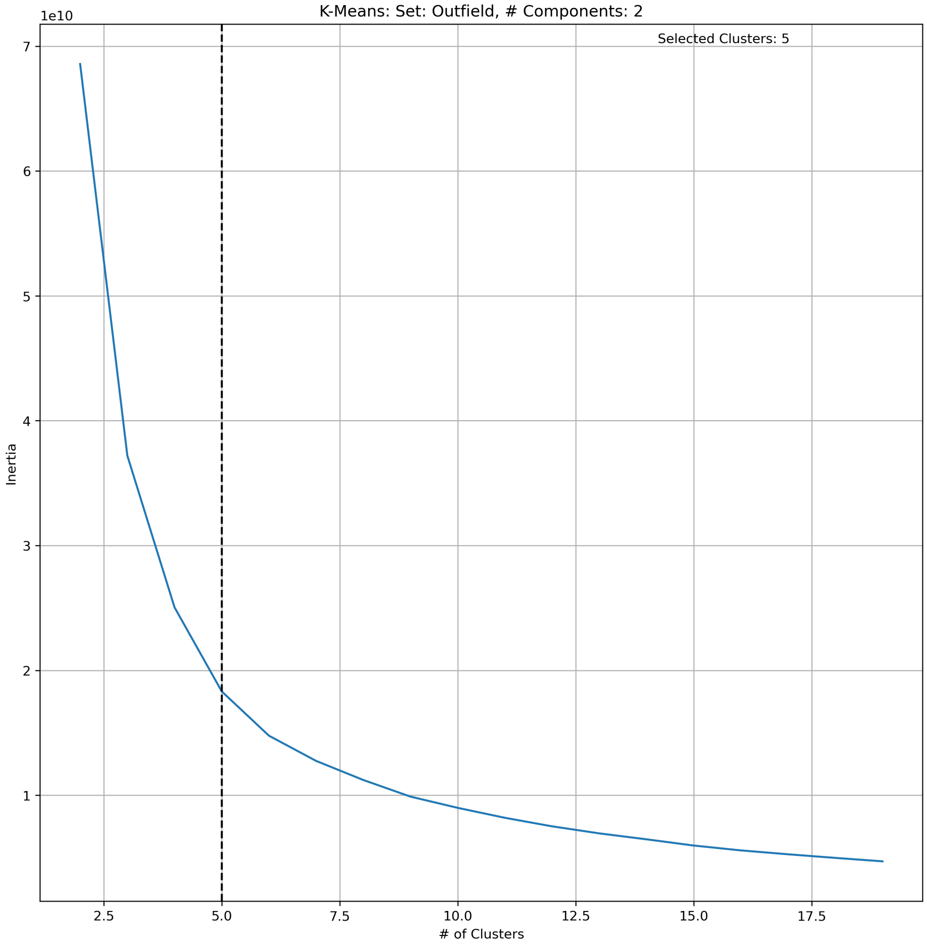

Figure 2: Outfield Players/Outfield Statistics 2 Component PCA - Cluster Selection

Figure 3 below displays a selected example of a trained model. The model in question is the K-Means as trained on the outfield only dataset in 2 principal directions. The figure shows that the majority of the variability exists in the first principal direction and as a result the clusters are effectively 'vertical slices' of the observations.

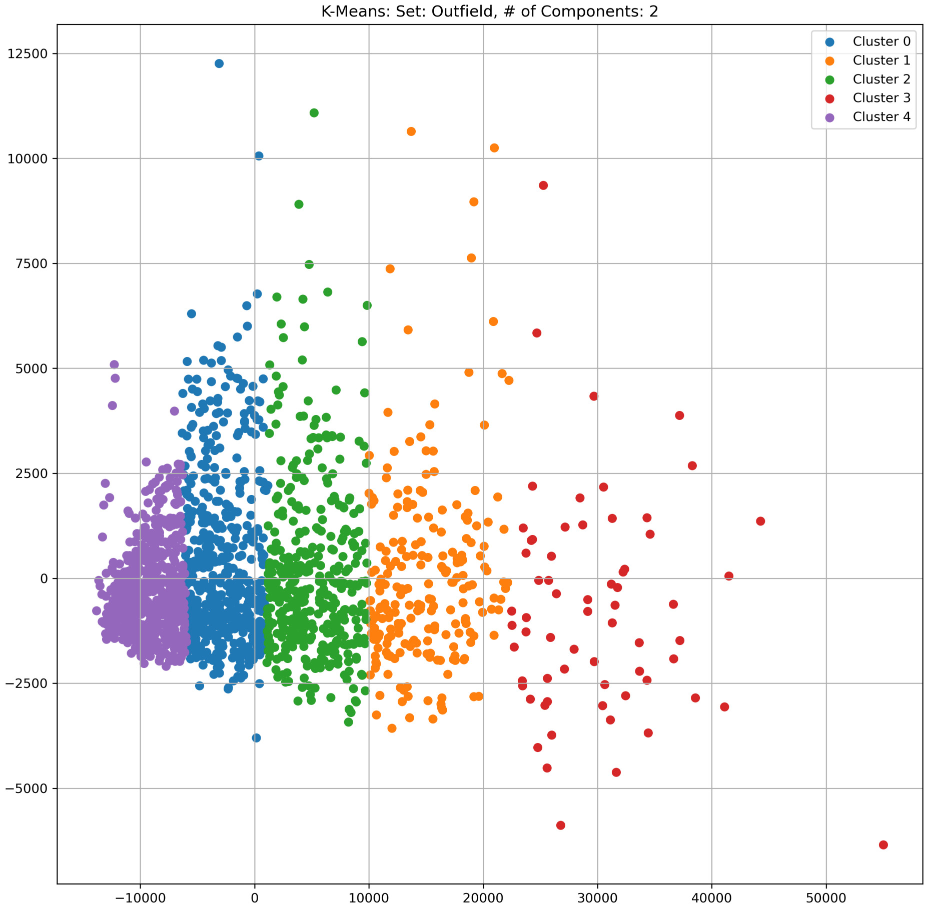

Figure 3: Outfield Players/Outfield Statistics 2 Component PCA - K-Means

Figure 4 below is another selected example with a GMM model trained on the outfield only dataset in 2 principal directions. Again the majority of the variability is in the first principal direction, however GMM captures variability in the second principal direction as well. Additionally, GMM captures some of the overlap in clusters that is to be expected with such a compact dataset.

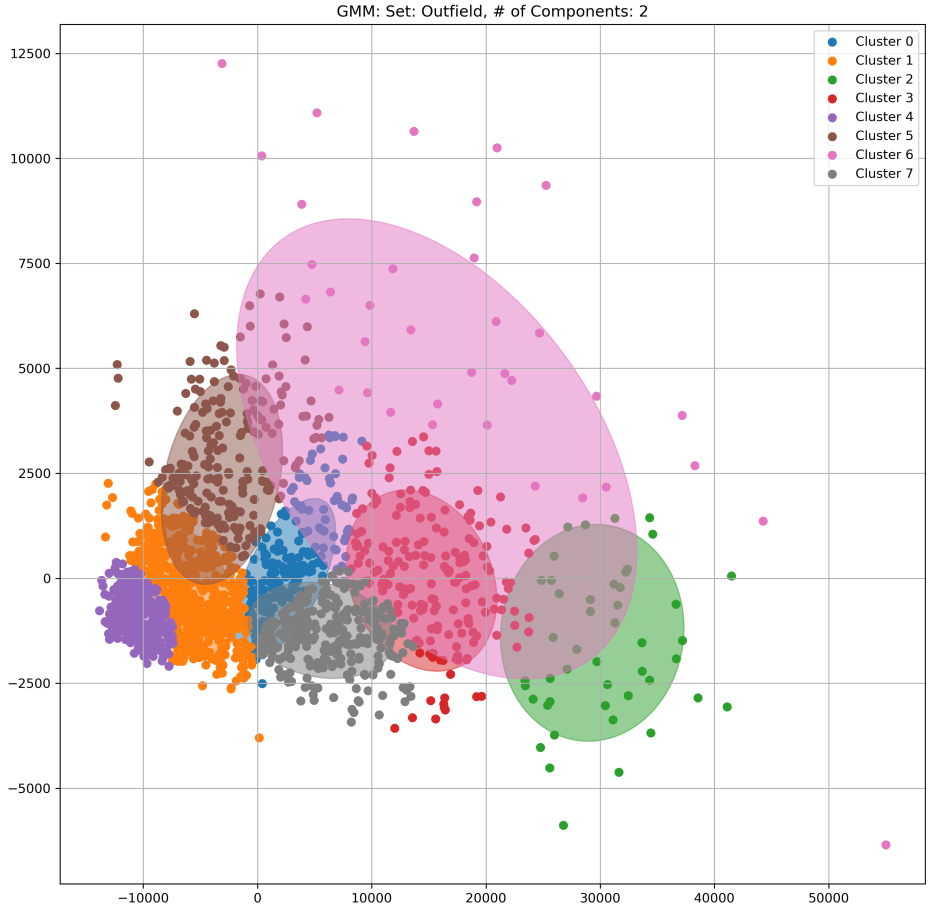

Figure 4: Outfield Players/Outfield Statistics 2 Component PCA - GMM

A visual analysis of the dataset containing all players and all statistics shows that goalkeepers are not distinct from outfield players in terms of recorded statistics despite quite literally playing by different rules. Figure 5 below illustrates this point.

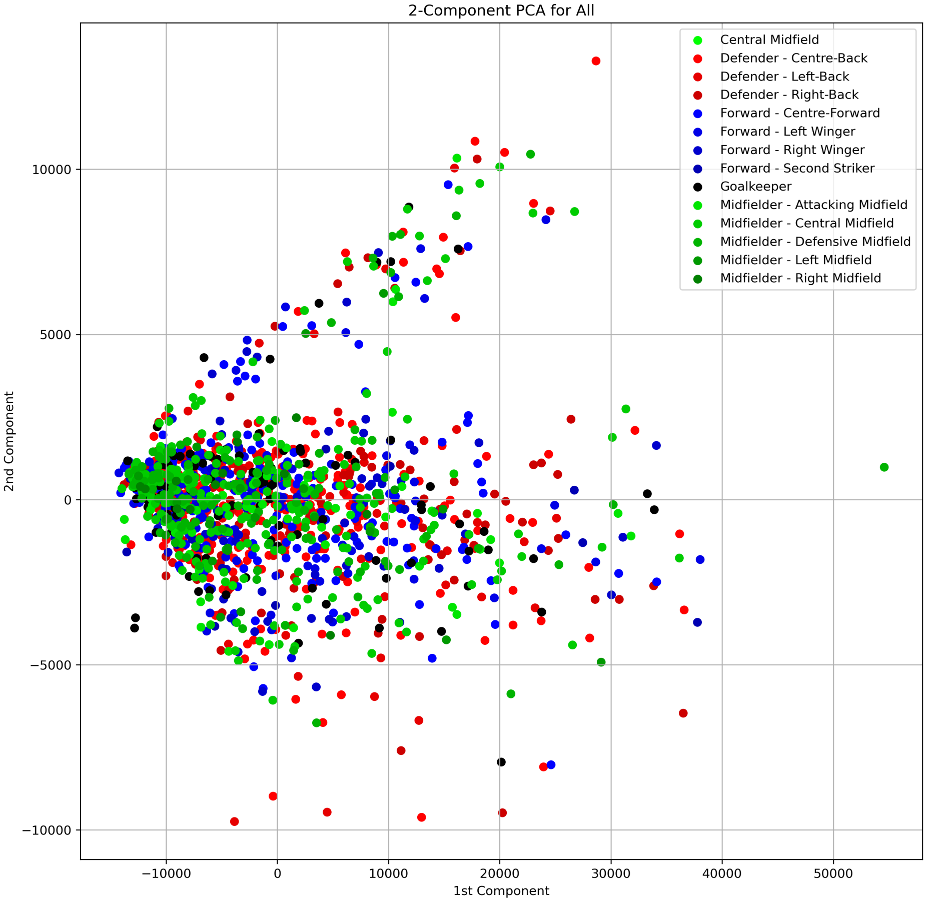

Figure 5: All Players/Statistics 2 Component PCA

The quality of each model for each dataset is captured in table 3 below. A higher value indicates better clustering, that is that the observations are more spread between clusters than within clusters. While K-Means outperforms GMM in all cases, both are orders of magnitude better than Spectral Clustering. Spectral Clustering receives the lowest possible scores, this is due to the compactness of the observations, all observations are interconnected and placed within the same cluster.

Model	All		Outfield		GKs	
	PCA - 2	PCA - 3	PCA - 2	PCA - 3	PCA - 2	PCA -3
K-Means	3896.21	3239.77	4336.26	3949.82	248.83	205.94
GMM	2850.93	2809.21	2555.64	3193.15	209.30	232.55
Spectral Clustering	0	0	1.28	2.41	0.15	0

Table 3: Calinksi-Harabasz Index

While K-Means performs the best per the Calinksi-Harabasz index, GMM will be used for a survey of players. This is rationalized by the desire for the selected model to capture the overlap in player positions, that is the casual user of this output is not beheld to discrete categories.

Figure 6 below lists the difference in percentage occurrence of each position label in the outfield only dataset for each GMM with two components cluster as compared to the population.

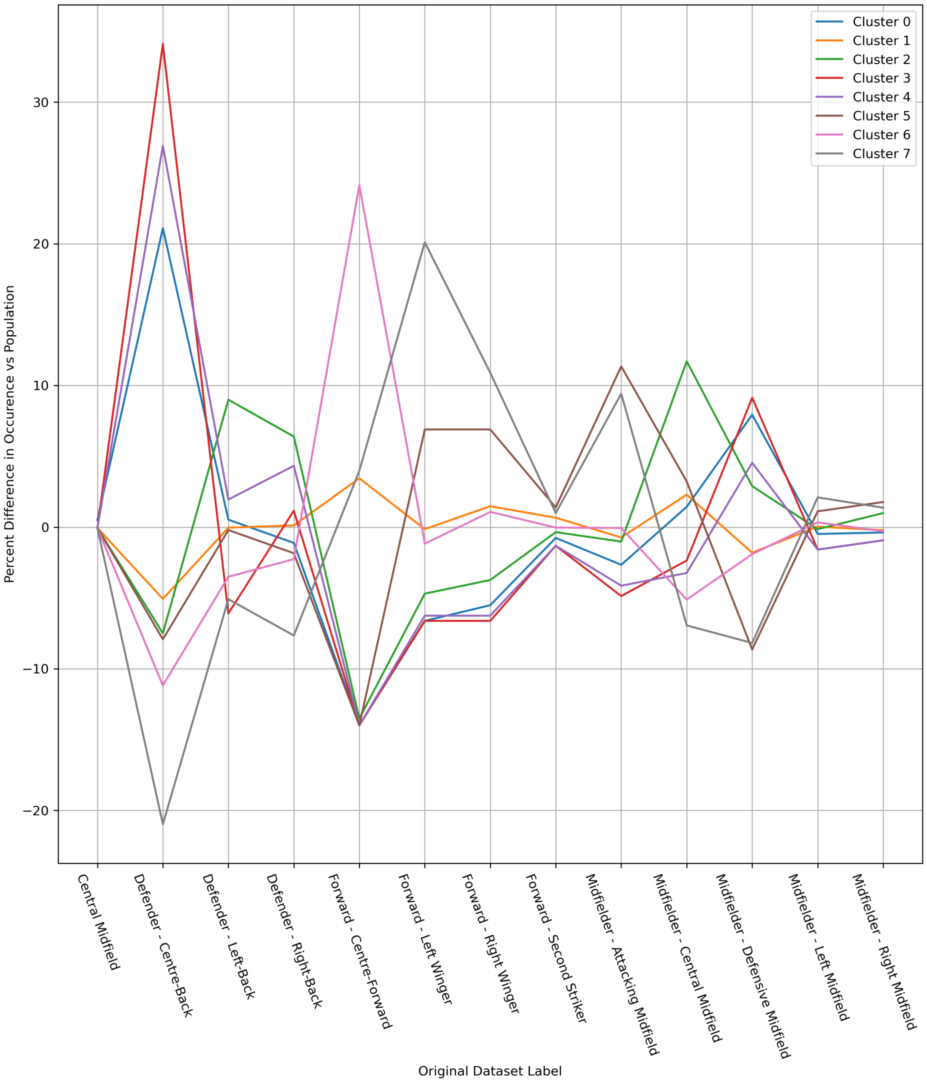

Figure 6: Occurrence of Position in Each Cluster for GMM, PCA 2, Outfield Only

Cluster 1 contains more center forward labels as compared to the population and other clusters. This can be interpreted as this cluster being representative of players who 'play like center forwards' regardless of their labeled position. An example of a notable player who is not labeled by experts as a center forward is Dani Olmo. Olmo, a Barcelona player, is listed as an attacking midfielder but based on this analysis has in-game characteristics more typical of center forwards. Olmo is likely labeled as such because of his literal position both on the field and in relation to his teammates. However this analysis shows he may be best considered a forward.

Another example is that cluster 3 contains significantly more center defender labels compared to the population and other clusters. Casemiro, the former Real Madrid and now Manchester United player, is labeled by the GMM model as having characteristics more akin to that of

Cluster	Assignment Percentage
0	0.0
1	5.3
2	0.0
3	81.0
4	0.0
5	0.0
6	13.7
7	0.0

Table 4: Casemiro's Assignment Percentage

a center defender, despite popularly being considered a defensive midfielder. GMM allows us to partially assign to multiple clusters, table 4 lists Casemiro's cluster assignments.

Casemiro primarily, about 81% so, belongs to the cluster mostly described by center defenders, but also shares characteristics with clusters 1 and 6, at 5.3% and 13% respectively, corresponding to groups of typically center forwards. This could be represntative of his usage at certain points of the game or in specific formations.

In conclusion, the GMM model yielded by the analysis can be used to understand a player's in-game statistics in the context of other players. This understanding goes beyond simple positional labels. While GMM, nor any of the other models in the analysis, yielded a strongly performing clustering model, the intent to provide a model through which fans may understand the game better was satisfied.

References

[Cal74] Jerzy Caliński Tadeusz; Harabasz. “A dendrite method for cluster analysis”. In: *Communications in Statistics* 3.1 (1974), pp. 1–27.

[Sal07] V. Di Salvo. “Performance Characteristics According to Playing Position in Elite Soccer”. In: *International Journal of Sports Medicine* 28.3 (2007), pp. 222–227.

[SIA19] Md. Tanzil Shahriar, Yashna Islam, and Md. Nur Amin. “Player Classification Technique Based on Performance for a Soccer Team Using Machine Learning Algorithms”. In: *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. 2019. DOI: 10.1109/ICECCT.2019.8868989.

[FIF] FIFA. *FIFA 21 Positions*. https://fifauteam.com/fifa-21-positions/. Accessed: 2024-10-20.

[Rag] Ville Satopaa;Jeannie Albrecht; David Irwin; Barath Raghavan. “Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior”. In: ().

[RSK] RSKriegs. *Soccer players values and their statistics*. https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics. Accessed: 2024-10-20.

[Wal] Luke Walsh. *Network Analysis of Soccer Player Positions*. https://medium.com/inst414-data-science-tech/network-analysis-of-soccer-player-positions-ed4c58e5419b. Accessed: 2024-10-20.