Toward a better understanding on changes of soccer team tactics over time.

Patton Yin

**Abstract**

This paper aims at understanding team tactics with the use of data and qualitative analysis. The goal is to produce good visualization in getting people more insights understanding teams. The Principal Component Analysis and K-means Clustering techniques will be applied. The results would be the graphs of analysis on teams with a favor on attacking, and how there are doing for different standards. Such analysis provides lots of insights into what could be done next.

**Introduction and background**

Soccer has developed for over hundreds of years, but the need for updating playing tactics never change. It is easily observed that, tactic change is common and takes place frequently in modern soccer competitions. There would be a lot of reasons for why a team needs to change its tactics: there might be players got injured, it might be a need to adjust tactics towards opponents of different level… In short, if a team cannot advance its tactics, they easily fall behind because they’re not optimizing toward the best outcome. Pep Guardiola’s coaching in Manchester City is extremely successful for over 7 years, and soccer fans observed that this team’s tactics always evolve, and that could be a reason explaining why this team dominates the Premier League. There are a lot of videos online explaining how this coach manages to overcome the difficulty brought by injuries of certain players by making use of certain traits of other players. Such analysis is amazing. It worth mentioning that capturing the details of every other team’s tactics change which are not under the spotlight over time is hard: the substitution of 1 player out of the 11s can lead to big structural changes: as players have different technical characteristics, to best exploit their strong suit, team’s tactics must be updated. Human beings are having a hard time to capture every subtle changes. In this case, beginning from the perspective of a soccer fan, this paper aims at making us of data and computing power of machine to establish an analyzing methodology on how a team’s tactics could change across time basing on basic stats.

According to the tradition of soccer, the main phases of a play can be divided into 2: attack and defense, and the attacking strategies can be divided into 2 categories: established attack and counterattack. According to a paper, there are a lot of other parameters that can feature or indicate the team’s implementation of these strategies (e.g. possession level, total number of passes) (2016, Hewitt). If these data are made use of and applied to data analysis, the teams’ tactics could be observed. With the idea of establishing a methodology first, this paper focuses on building an analysis on established attack related tactics.

**Methods**

The dataset is from the website <https://fbref.com/en/>. The data are collected from 4 tabs “Miscellaneous”, “Possession”, “Pass Types”, “Passing” of every team in the Premier League. There are many variables under each tab, each variable would be one parameter of a team’s stats in one match. Correspondingly, the observations will be different matches. Then, with the use of domain knowledge, variables are selected. The variables picked are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Data Tab Name | Data Sample | Special treatment | Explanation (if needed) |
| Date | Miscellaneous | “2023-02-14” |  |  |
| Result | Miscellaneous | “W”, “L” | “W” is converted to integer 3, “D” is converted to integer 1, “L” is converted to integer 0. | This is to change the Win/Lose/Draw into the points got by the team. |
| GF (Goals For) | Miscellaneous | 3 |  | Goals scored by this team |
| GA (Goals Against) | Miscellaneous | 2 |  | Goals scored by the oppoenent team |
| Crosses | Miscellaneous | 17 | Cross/Att (Passes Attempted) | Better to understand the allocation of tactic resources |
| Att (Passes Attempted) | Pass Types | 869 |  |  |
| TB (Through Ball) | Pass Types | 3 | TB/Att (Passes Attempted) | Better to understand the allocation of tactic resources |
| Touches | Possession | 936 |  |  |
| Att.3rd (Touches in attacking 1/3 area) | Possession | 234 | Att.3rd/Touches |  |
| Mid.3rd (Touches in midfield 1/3 area) | Possession | 484 | Mid.3rd/Touches |  |
| Def.3rd (Touches in defending 1/3 area) | Possession | 227 | Def.3rd/Touches |  |
| Carries | Possession | 580 |  |  |
| PrgC (Progressive Carries) | Possession | 31 | PrgC/Carries | Carries that move the ball towards the opponent’s goal line at least 10 yards. Better to understand the allocation of tactic resources |
| PrgP (Progressive Passes) | Passing | 51 | PrgP/Att (Passes Attempted) | Better to understand the allocation of tactic resources |
| Possession.R (Possession Rate) | Possession | 75 |  |  |
| P3rd (Passes to attacking 1/3) | Passing | 52 | P3rd/Att (Passes Attempted) | Better to understand the allocation of tactic resources |
| PPA (Passes to Penalty Area) | Passing | 9 | PPA/Att (Passes Attempted) | Better to understand the allocation of tactic resources |
| Take-on (Attempts to take on defenders) | Possession |  |  |  |

When doing the analysis, a lot of the parameters were made into ratios: For example, number of Crosses are divided by the total number of passes attempted. The reason is, to measure the tactics of a team, it would be the best to focus on which aspect a team decides to lean forward the resources to. In this case, the higher the ratio of Crosses indicates a higher tendency of shifting the resources of passes on the “Cross” tactic. Majority of the data are treated in such way.

Principal Component Analysis, and K-means Clustering are tried out in this analysis. Principal Component Analysis and K-means Clustering worked to generate a meaningful / interpretable outcome in terms of classifying the team with different traits. For Principal Component Analysis and k-means clustering, the variables to be used will be the variables selected and treated above.

When performing analysis, all values are scaled, as different parameters have different scales. Such differences in magnitudes would affect the persuasiveness of each variable. To begin simple first, values are scaled.

**Results**

Figure 1

Principal Component Analysis result using the selected variables.

PCA

As mentioned before, only the variables indicating a team’s preference on attacks are selected purposefully. If a team gets a higher value on such parameter that means this team favor doing established attacks. In this case, imagining a vertical line from PC1 at coordinate 0 dividing the teams into two groups, it is reasonable to see that generally teams are divided into 2 groups, one is the group with a strong desire to do the established attacks, the other is the group without (Figure 1). However, it is observed that some vectors are pointing to the right. For the vector “Touches in defending region” makes sense by the name. If a team has higher touches of ball in their own defending region, that is a strong indication of the team’s lack of ability to execute the established attack. For the vector “Cross”, it might not make sense to audiences who are not familiar with soccer. In modern days, performing crosses in higher ratio is usually an indication of team’s lack of ability to perform established attacks, as such way of attack doesn’t require much skills as the other ways do. In this case, it is reasonable that there are only two vectors pointing to the right accounting for teams who do not prefer to perform established attacks.

Figure 2

Elbow method graph for k-means clustering.

A picture containing text, receipt, font, algebra

Description automatically generated

Figure 3

Results of k-means clustering.

K-means Clustering

After implementing the k-means clustering, a consistent result was derived. According to the elbow method, cluster of 2 is picked (Figure 2). Then, the cluster composition is derived: Arsenal, Brighton, Chelsea, Liverpool, Manchester City, Manchester United, Newcastle United. The only exception is that Leeds United are not counted here. This is reasonable considering it’s falling right on the edge between the two clusters.

Radar graph for teams and analysis on the dynamic

Based on the Principal Component Analysis, and k-means Clustering, a few vectors are selected to represent team’s attacking attribute. Then, for each team, the mean value of their stats on these variables in matches across the whole season was picked. By taking the maximum mean value among all teams as the maximum point of the edge and minimum value as the interior point, radar graphs for different teams are got. (Figure. 4.1 4.2 4.3) This is a visualization of how teams are allocating their resources in their attempts to win.

Figure .3

Figure 4.2

Figure 4.1

**Discussion**

PCA and k-means clustering

The initial attempt with such method is successful in terms of the goal. Basing on that, more other parameters could be added to the Principal Component Analysis for deeper analysis. However, such initial attempt was not how traditionally PCA analysis and k-means clustering are used in gaining insights. As for the next step, more clusters could be developed, and thus more categories of teams might be developed. This would be a great step moving toward abstractly understanding a problem.

Visualization of the graph

This is found to be consistent with human observations throughout the season: Manchester City is excellent and is exceptionally willing at holding the ball, which made them the highest possession rate and most touches in the midfield region; Arsenal is ferocious in attacking, which made them the most touches in attacking region, most passes to attacking region, and most number of progressive passes; Chelsea is poor at doing attacks over this season, and it is indeed occupying much less space compared with two teams before.

Although the graphs did tell some aspects about the teams’ allocation of resources on tactics, they are not enough. Most importantly, these variables are not parallel in terms of their representation on the team tactics. Some might be an implication of the other, some variables might be influenced by the same parent factor. Ideally, each parameter should have been given a weight which could comprehensively tell audience about teams’ tactic variability. To improve on such aspect, more domain knowledge is needed. Plus, team’s tactics can change over time, which means that using the mean value is not enough in representing team’s tactic variability across the whole season. What is expected to be done next step is to plot out the data for each team across the season, and to find a way which captures the trend. Finally, it would be interesting to see how teams’ ability to adapt to different tactics related to the chance of winning.