Mohammad Mustafa Arif

mustafa.arif@mail.utoronto.ca

Thesis Interim Report

SOCCER PLAYING STYLES: A Data driven approach

Contents

[Introduction 2](#_Toc62389084)

[Literature Review 4](#_Toc62389085)

[Data 7](#_Toc62389086)

[Progress to Date 7](#_Toc62389087)

[Future Work 7](#_Toc62389088)

[References 7](#_Toc62389089)

[References 7](#_Toc62389090)

# Introduction

Soccer is the most popular sport in the world. Today, more than 240 million people play soccer regularly, and a combined 3.5 billion viewers tuned in to the 2018 FIFA World Cup broadcast. Despite its rich history, the first record of analytics in soccer is from 1996, when Opta began recording match statistics for the English Premier League. Since then, the field has rapidly evolved and teams have begun using data-driven approaches for tactical decision making. With increased data collection, methodologies in statistics, computer science and network theory have been employed to study teams and exploit strengths and weaknesses. Recently, Liverpool FC, a first division English soccer club, announced the development of “pitch control”, a probabilistic model used to quantify space creation during matches in real time.

Traditional methods in soccer analytics involve utilizing high level match statistics such as possession percentages, numbers of shots, passes, crosses, fouls, and offsides. However, these methods are prone to oversimplifying 90+ minute interactions between two teams to a set of global statistics. An increasingly popular approach to add nuance to these methods is the use of expected goals (xG), a metric to measure the quality of shots taken by their probability of resulting in a goal. In addition, modern approaches in literature have attempted to study team passing profiles through the distribution of passes amongst the players. However, these approaches are also prone to oversimplification; they do not distinguish between different types of passes or the distribution of passes across different regions of the pitch. Furthermore, they do not explore defensive activity with respect to team performance.

The first objective of this project is to expand on existing literature characterizing team passing distributions and fill in the aforementioned gaps by:

1. Augmenting the passing distributions with nuanced descriptive metrics measuring deeper passing statistics using network theory.
2. Incorporating defensive metrics: the distribution of tackles, interceptions, and fouls to represent team performances.

The appropriate features are derived from the Wyscout spatio-temporal soccer event logs, a public dataset with reports of soccer match events marked by time, spatial coordinates, players involved, and a standardized taxonomy of event tags. The dataset covers Europe’s top five domestic leagues for the 2017/2018 season, as well international fixtures in the 2016 Euros and the 2018 World Cup. Some features are engineered though direct manipulation of the logs, while others require a deeper analysis through the construction of passing networks across players and pitch-zones.

After engineering the features describing each performance, feature importance will be measured via calculating correlation coefficients and mutual information with respect to some measure of performance. The purpose of this exercise is to gauge which features contain the most descriptive power about a team’s performance.

Next, features for each team will be aggregated for the entire season, and teams will be categorized into distinct playing styles. For this step, unsupervised clustering methods such as k-means and hierarchical clustering, along with probabilistic soft-clustering algorithms such as gaussian mixture models will be used. Once cluster assignments have been established, they will be analysed from a fundamental perspective to attempt an intuitive interpretation of the established playing styles. Finally, performances will be analysed with respect to playing styles study if there are certain playing styles that tend to overperform/underperform in comparison to others.

Thus, the desired outcome of this project is to arrive at classification of teams in terms of a feature space spanned by custom-defined metrics. The project aims to study these playing styles from an intuitive perspective and analyze any discrepancies in effectiveness/performance across these styles.

# Literature Review

Increasingly available data has facilitated recent developments in the use of analytics in soccer. Soccer-logs capturing all events within a match have commonly been used to study many aspects of soccer [CITATION NEEDED]. In particular, there exists literature investigating certain squads and explaining unexpected performance patterns through a data-driven lens. There have also been notable research attempts to derive representations and insights via statistical methods for teams as well as individual players. While this project is framed at segmenting teams into characteristic playing styles and investigating performance discrepancies across these playing styles, the following literature survey explores all the aforementioned sub-domains as the methodology discussed can be reapplied in the context of this project.

One key question individual player analysis address is: What makes a player “good” or “bad”? A popular method to quantify player quality using data-driven approaches is to use information from in-game events for extracting information. For example, Pappalardo et al. use soccer-logs to design and implement PlayeRank, a role-aware, multi-dimensional evaluation framework for individual soccer players [CITE], tuned through team performance outcomes. They validate the performance by testing the algorithm against ground truth ratings from professional soccer scouts, achieving a significant level of agreement. They also demonstrate the use of PlayerRank to distinguish top players from others, and measure player versatility. As such, Pappalardo et al. addressed issues with the time’s simple data-driven player performance metrics by:

1. Designing a “role” aware system: The evaluation function is dynamic with respect to a player’s role on a team.
2. Evaluating performance across many dimensions instead of reporting a scalar score combining all aspects of player performance.

However, a major drawback of this method is its limited scope at individual player level. Analysing squads of 11 players (with possible substitutions/injuries) becomes an increasingly challenging task with this framework. In addition, the input features for PlayeRank are solely derived from individual player activity and fail to measure interactions between players or “team chemistry”. Nevertheless, the use of unsupervised clustering with the K-Means algorithm on spatial features for role detection is a robust method, and can be generalized, with the help of involved feature engineering, to segment teams into playing styles.

To address the intrinsic drawbacks of individual player analysis, attempts have been made to identify team performance metrics from a higher level of abstraction. For instance, Cintia et al. extract five separate team metrics summarizing passing behaviour amongst players and across zones of the pitch, eventually combining them via a weighted harmonic mean to a single indicator, the ‘H-Statistic’ [CITE THIS]. They report a significant correlation between the H-Statistic and performance measures and explain game outcomes based on the H-Statistic difference across the teams. Further, they demonstrate out-of-sample predictive power in the H-Statistic by utilizing various classifiers predict match outcomes through a season based exclusively on H-Statistic values and find high agreement between real and simulated rankings. The study successfully managed to quantify passing behaviour across teams, but one limitation is that only 5 simple statistics are used to describe passing: mean and variance of passes across players, mean and variance of passes across zones, and total passing volume. These shortcomings present a requirement for more expressive features to describe team passing, and other aspects of the game such as defense.

Another popular way to study teams and player-player interactions is through network science. Originally proposed by Gould and Gatrell [CITE] in 1979, the idea of constructing passing networks from team performances gained popularity after Duch et al. [CITE] demonstrated in 2010 that flow centrality in passing networks, a measure of the frequency of the player being involved in paths resulting in a shot, can be used to quantify the contribution of individual players in team performances. More recently, Clemente et al. [CITE] employed the use of network science to study team performances in the 2014 World Cup and demonstrated to a statistically significant degree that large connectivity between teammates is associated with better overall team performance. In their 2019 article, Buldu et al. [CITE] also utilized metrics engineered using network science to compare Pep Guardiola’s FC Barcelona team from the 2009/2010 season, considered one of history’s best soccer teams, with the rest of the teams in the Spanish First Division. They found a significant difference in the advance ratio, a measure of passing directness, between FC Barcelona and the rest of the teams in the league, indicating that FC Barcelona’s squad employed a relatively indirect approach to passing. Moreover, they ran graph algorithms to compute the clustering coefficient (a measure of a network’s local robustness), average topological shortest path, largest eigenvalue of the connectivity matrices, and algebraic connectivity (a quantification of the team’s division or fault tolerance). Across all these metrics, they found statistically significant differences between FC Barcelona and the rest of the teams in the Spanish league. Thus, Buldu et al. [CITE] defined a network science framework for studying the playing style for a specific team against its rivals. For this project, their methodology can certainly be augmented to incorporate features measuring other aspects of the game beyond passing and generalized to compute numeric representations for more teams across multiple leagues to find commonalities/differences and discern team playing styles.

Each of the works discussed approached soccer analysis through a data-driven method utilizing soccer-logs datasets. However, they do not apply their methods to general team playing style identification. Furthermore, relatively little exploration is done beyond passing, and especially for defensive attributes. This project will therefore attempt to expand on methodology described above to fill in these gaps.

Finally, note that the use of soccer-logs is just one of many different approaches for utilizing analytics in soccer. There have also been notable attempts to exploit video footage, GPS tracking, and player physiological signals [CITE ALL 3]. These approaches, however, are fundamentally different in scope to the problem addressed and the data used in this project. Thus, they are not discussed comprehensively here.

# Data

For this project, the data being used is sourced from a public, spatio-temporal dataset of soccer-logs spanning Europe’s top 5 domestic leagues for the 2017/2018 season: Spanish first division, Italian first division, English first division, German first division, and French first division [CITE]. In addition, international games from the 2018 World Cup and the 2016 European Cup are also covered. The data is provided by Wyscout, a leading company in the soccer industry, and the collection of data is segmented into a rigorous three phase process:

1. Expert video analysts set team formation at the beginning of each game. This includes mapping the on-field players to their positions as well as the list of available players on the bench.
2. For each touch on the ball, the analysts, using a propriety tagger software, select one player and create a corresponding event on the timeline. The event description involves specifying an event type and subtype, along with the spatial coordinates, and other special tags to specify additional attributes.
3. The logs are quality controlled, both algorithmically and through manual cross comparisons.

The exact taxonomy of the events, subevents, and event-tags can be found in the appendix. Altogether, the dataset covers 1941 games, 3,252,294 events, and 4,299 players.

# Progress to Date

# Future Work

# References

Clemente: <https://www.tandfonline.com/doi/abs/10.1080/24748668.2015.11868778>

# Appendix

## Event Taxonomy

|  |  |  |
| --- | --- | --- |
| **Event** | **Subevent** | **Tag** |
| pass | cross, simple pass | accurate, not accurate, key pass, opportunity, assist, goal |
| foul |  | no card, yellow, red, 2nd yellow |
| shot |  | accurate, not accurate, block, opportunity, assist, goal |
| duel | air duel, dribbles, tackles, ground loose ball | accurate, not accurate |
| free kick | corner, shot, goal kick, throw in, penalty, simple kick | accurate, not accurate, key pass, opportunity, assist, goal |
| offside touch | acceleration, clearance, simple touch | counterattack, dangerous ball lost, missed ball, interception, opportunity, assist, goal |

Table taken from Pappalardo et al. [CITE]