Analyzing Football Tactics through Finishing Sequences Classification

Alexandre Llinàs Martínez

Grau en Ciència de Dades Aplicada
Analítica Deportiva

Tutor/a de TF
Teresa Divorra Vallhonrat
Professor/a responsable de l'assignatura
Susana Acedo

Gener 2025

Universitat Oberta de Catalunya

FITXA DEL TREBALL FINAL

Títol del treball:	*Analyzing Football Tactics through Finishing Sequences Classification*
Nom de l'autor:	*Alexandre Llinàs Martínez*
Nom del consultor/a:	*Teresa Divorra Vallhonrat*
Nom del PRA:	*Susana Acedo Nadal*
Data de lliurament (mm/aaaa):	*01/2025*
Titulació o programa:	*Grau en Ciència de Dades Aplicada*
Àrea del Treball Final:	*Analitica deportiva*
Idioma del treball:	*anglès*
Paraules clau	*Football analytics, Machine learning, play styles*

Resum del Treball

Aquest estudi introdueix un marc sòlid per analitzar i comparar els comportaments tàctics en el futbol mitjançant l'agrupació de seqüències de possessió i l'ús de dades tàctiques. Utilitzant tècniques d'aprenentatge automàtic no supervisat, s'han classificat les seqüències de possessió que acaben en tir en diferentes tipus, capturant contextos tàctics diferenciats, com ara contraatacs, inicis de joc i recuperacions a prop de l'àrea rival. A més, s'ha aplicat l'anàlisi de components principals (PCA) per identificar tendències clau dins d'aquests grups, permetent caracteritzar els estils de joc dels equips i facilitar les comparacions entre ells en diversos contextos.

L'anàlisi ha demostrat la capacitat de l'agrupació per destacar diferències tàctiques subtils, revelant que els equips tendeixen a adaptar els seus comportaments segons el context, més que mantenir-se fidels a un perfil tàctic rígid. També s'ha entrenat una xarxa neuronal amb dades tàctiques per predir aquests contextos tàctics, aconseguint resultats moderadament satisfactoris tot i les limitacions en les dades disponibles. Els resultats mostren que les dades tàctiques, encara que no són perfectes, són suficientment representatives per discernir els comportaments tàctics, especialment quan els grups estan ben definits i arrelats en la semàntica del futbol.

Aquest marc ofereix aplicacions pràctiques per al cos tècnic, permetent avaluar i comparar estratègies d'equip, tant entre rivals com dins del propi equip al llarg del temps. Això facilita la comprensió de l'evolució tàctica i els ajustos necessaris. Tot i així, l'estudi destaca diverses limitacions, com la mida reduïda del conjunt de dades, restringit a una temporada, i el fet de centrar-se només en seqüències que acaben en tir. Les investigacions futures haurien

d'abordar aquestes limitacions incorporant dades més diverses, perfeccionant les característiques tàctiques i explorant metodologies alternatives.

Abstract

This study introduces a robust framework for analyzing and comparing tactical behaviors in football through the clustering of possession sequences and the use of tactical data. Leveraging unsupervised machine learning techniques, possession sequences ending in shots were classified into meaningful clusters, capturing distinct tactical contexts such as counterattacks, long build-ups, and high-pressure recoveries. Principal Component Analysis (PCA) was applied to identify key trends within these clusters, enabling the characterization of team playing styles and facilitating comparisons between teams in various contexts.

The analysis demonstrated the capacity of clustering to highlight nuanced tactical differences, revealing that teams tend to adapt their behaviors within specific contexts rather than adhering to rigid tactical profiles. Furthermore, a neural network was trained using tactical data to predict these tactical contexts, achieving moderate success despite data limitations. Results showed that tactical data, while not perfect, are representative enough to discern tactical behaviors, especially when clusters are well-defined and grounded in football semantics.

The framework provides practical applications for coaching staff, allowing them to evaluate and compare team strategies, not only between opponents but also within their own team across different periods. This enables insights into tactical evolution and adjustments over time. However, the study also highlights several limitations, including the small dataset size, restricted to one season, and the focus solely on sequences ending in shots. Future research should address these limitations by incorporating more diverse data, refining tactical features, and exploring alternative methodologies.

Contents

Figures

Tables

1. Introduction

1.1. Context

In recent years, sport analytics (particularly in football) has advanced significantly, driven by innovations in data collection and processing technologies. Tools such as GPS-enabled devices and high-resolution cameras now capture every movement and action during matches, producing vast datasets. When combined with on-ball events, such as passes and shots, this data holds the potential to deliver actionable insights into team and player behaviors [8][10]. However, despite the proliferation of sports data providers and access to raw data through platforms or APIs, the metrics and models available often require extensive manipulation to contextualize the information effectively. As a result, most football analyses continue to focus on individual performance metrics or broad team statistics, leaving a gap in understanding collective behaviors and tactical patterns [12].

Football is inherently complex, involving 22 players, a dynamically moving ball, and an array of contextual factors, including unique rules (e.g. offside), environmental conditions, and psychological pressures. This complexity is further heightened by the diversity of tactical approaches teams adopt. These tactical variations, collectively referred to as a team's *playing style*, represent a critical aspect of match planning and analysis. Coaches and analysts dedicate significant resources to understanding playing styles, which serve as the foundation for crafting strategies that exploit an opponent's weaknesses while maximizing their team's strengths. A well-defined playing style can be the difference between winning and losing a match, as it influences all phases of the game, from defense to attack.

Given this complexity, this study narrows its focus to finishing sequences (Figure 1), a piece of game where a team keeps possession of the ball, ending with a shot. These sequences, also known as possession chains [5], offer a lens through which a team's tactical identity can be analyzed. By examining the actions that lead to goal-scoring opportunities (from ball recovery to the final shot) this research aims to capture essential elements of a team's playing style. Understanding these sequences is vital, as they reflect the decisions and strategies employed by both players and coaches to create scoring chances.

Despite their importance, there is no standardized method for identifying and classifying tactics among those sequences or possession chains [14]. This gap underscores the need for a robust classification framework to better understand how teams construct these critical plays. Accordingly, this study poses the

following questions: Are the tactics employed by teams distinguishable through their finishing sequences? How can these sequences be classified to reveal patterns that define a team's style of play?

The goal of this research is twofold: to deepen the understanding of team tactics in football and to provide coaches and analysts with tools to enhance decision-making, match preparation, and strategy development. Leveraging detailed data provided by a leading analytics company such as Kognia, this work seeks to explore tactical diversity with higher precision.

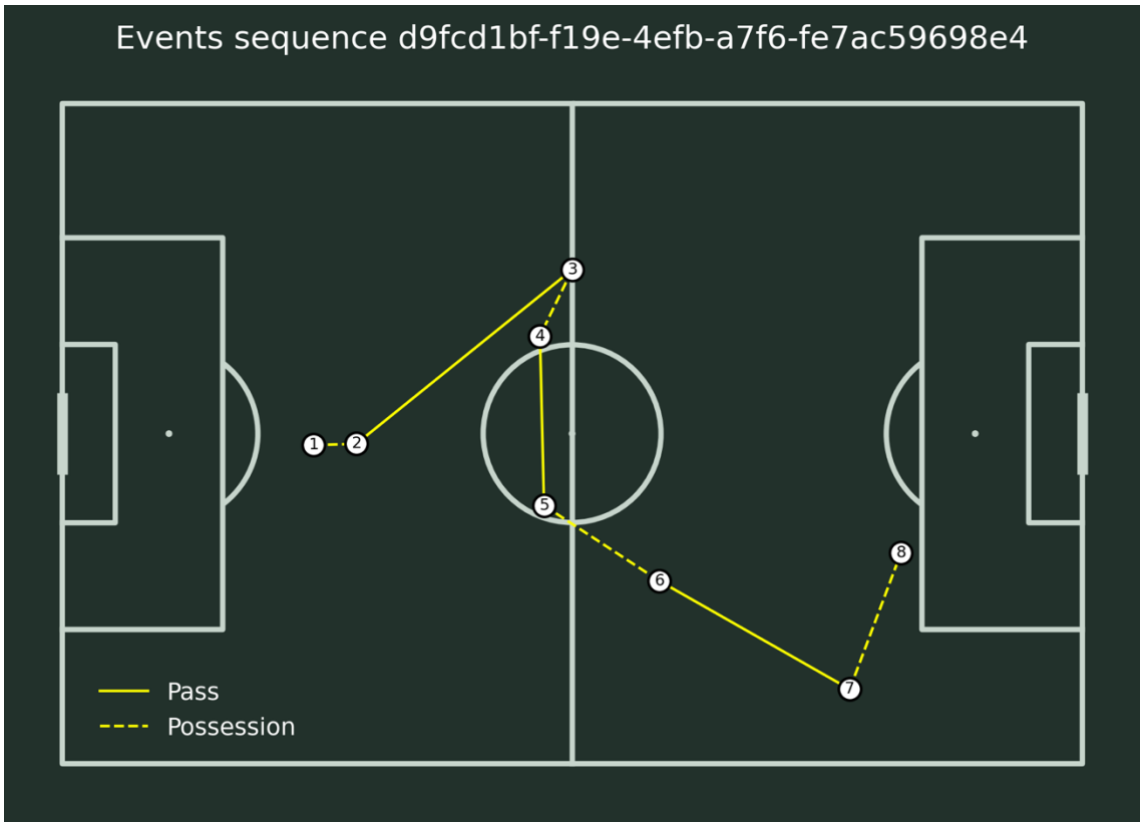

Figure 1. Finishing sequence. Attack direction from left to right

1.2. Goals

Main goal

- To create a methodology for classifying finishing sequences in football games and analyzing whether the underlying tactics can effectively identify playing styles and tactical differences between teams.

Specific goals

- Prepare and clean the dataset containing tracking and event information from football games, ensuring it is of high quality for analysis.
- Implement machine learning algorithms to classify finishing sequences using objective data.
- Analyze the tactics used by teams in these sequences and compare them to uncover differences in playing styles across teams.
- Assess the effectiveness of Kognia's tactics in identifying distinct tactical patterns within finishing plays.

1.3. Sustainability, diversity and ethical/social impact

Based on the *Guide on Ethical and Global Competence* provided by the program management, the following impacts are outlined for the project design:

On the sustainability dimension:

The project aligns with SDG 9 – Industry, Innovation, and Infrastructure, specifically with target 9.4 – Upgrade all industries and infrastructures for sustainability [15]. The processing and analysis of large football datasets may have varying energy costs depending on the machine learning models used. To minimize the environmental footprint, efforts will focus on selecting efficient models during the development phase, prioritizing those with lower computational

demands while maintaining high accuracy. This approach seeks to balance innovation with environmental responsibility.

On the ethical-social dimension:

This project underscores the ethical and transparent use of data in sports analytics. Handling extensive player behavior datasets poses potential privacy risks. To address these, the project adheres to strict compliance with regulations like the General Data Protection Regulation (GDPR).

Key considerations include:

- Ensuring that data collection and processing are supported by clear legal bases, such as player consent via contracts. Attention will also be given to potential power imbalances (e.g., employer-employee relationships) that may affect freely given consent) [1].
- Recognizing the sensitivity of certain data types, such as biometric or health data, which require additional explicit consent under the GDPR [7].
- Incorporating Data Protection Impact Assessments (DPIAs) to identify and mitigate privacy risks, especially when introducing new technologies or

handling sensitive data [6]. These assessments will be periodically reviewed to ensure compliance and the protection of individual rights .

It is worth noting that, for this specific project, these legal and ethical aspects are already addressed, as the provider company (Kognia) manages the data in full compliance with existing regulations.

On the diversity dimension:

This project primarily focuses on data from men's soccer due to the unavailability of equivalent datasets for women's soccer. While this represents a limitation in diversity, it highlights a broader issue of data inequality in sports analytics. The absence of comprehensive data on women's matches underscores the need for greater investment in data collection for women's sports, aligning with SDG 10 – Reduced Inequalities, specifically target 10.2 – Promote universal, social, economic, and political inclusion [16]. Future iterations of the project aim to address this gap by incorporating women's soccer data as it becomes available, promoting inclusivity in tactical analysis.

By addressing these dimensions, the project strives to balance innovation, responsibility, and inclusivity in the domain of football analytics.

1.4. Approach and methodology

The approach for this project integrates data analysis techniques, advanced statistical methods, and machine learning algorithms to classify and analyze football teams' finishing sequences based on tracking, event, and tactical data.

The first step involves a detailed study of the available variables in the tracking and event datasets. Key characteristics that influence the classification of finishing sequences will be identified. Subsequently, relevant features will be selected, and additional metrics will be engineered to enhance the representation and interpretability of the sequences.

Unsupervised clustering techniques will be applied to group finishing sequences into meaningful categories from a football perspective. These clusters aim to capture distinct tactical profiles of sequences, such as build-up plays, counterattacks, or high-pressure recoveries.

An analysis will be conducted to evaluate if teams use varying tactics across different sequence types. This step explores the diversity of tactical preferences and provides insights into team-specific strategies.

A simple deep learning algorithm will be developed to predict the sequence type based on the Kognia-defined tactics (refer to Appendix 7.2 for their definitions) employed in each sequence. This predictive model will validate the robustness of the identified clusters and tactics in explaining the finishing sequences.

The project will follow a structured yet adaptable management approach:

- PMBOK (Project Management Body of Knowledge) will be employed for overall project organization, as it provides clear definitions of project phases: initiation, planning, staged execution, monitoring, and closure. Its iterative nature allows for adjustments as the project evolves. A well-defined list of deliverables will serve as a foundation for the project plan.

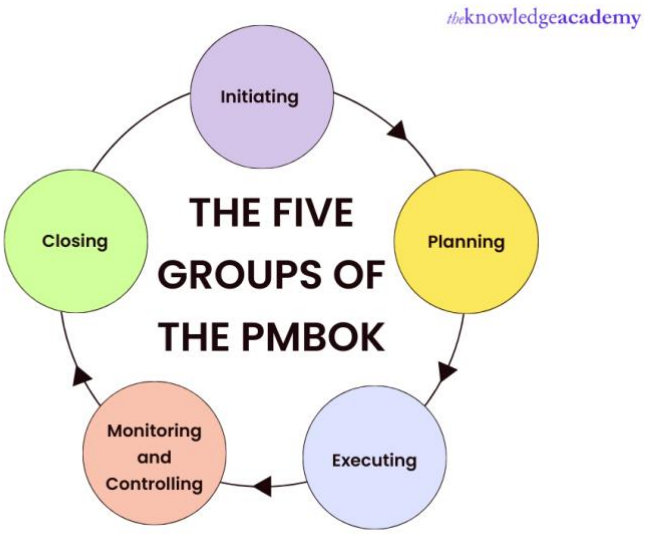

Figure 2. PMBOK circle. Source TheKnowledgeAcademy

- CRISP-DM (Cross-Industry Standard Process for Data Mining) will be adopted for data mining and model development. This methodology ensures a systematic approach, with steps including understanding the problem, data preparation, modeling, evaluation, and deployment.

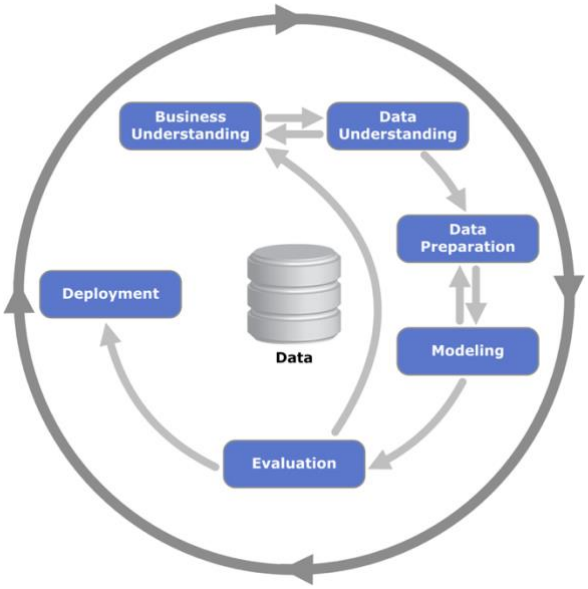

Figure 3. CRISP-DM. Source Wikipedia

Python will be used as the primary programming language. Python is an open-source language widely utilized in data analysis due to its extensive library ecosystem (e.g., scikit-learn, PyTorch, numpy, pandas, matplotlib, xarray, mplsoccer) and its active online community for support and resources.

The infrastructure provided by Kognia ensures the availability of servers equipped with multiple CPUs and GPUs for intensive data processing and model training.

This high-performance hardware enables efficient execution of computationally demanding tasks and ensures high availability.

1.5. Schedule

The planning is based on the following tasks, which are deployed from the general and specific objectives:

- Data research and collection: search and collection of all data necessary for the project, including tracking data, events and proprietary tactics of Kognia.
- Data preparation and cleaning: initial processing of the collected data to ensure its quality and integrity, including data cleaning and normalization.
- EDA: data exploration to identify patterns, trends and key relationships, using various visualization techniques.
- Initial clustering: application of unsupervised methods to obtain a first classification of the sequences using the objective data (tracking, eventing).

- Labeling of sequences: collaboration with football expert human annotators to manually label the sequences, creating a "ground truth" to validate and train the supervised models.
- Classification model training: implementation and training of supervised models, to classify finishing sequences.
- Validation and refinement of models: validation of trained models using the “ground truth” and refinement based on the results obtained and feedback from experts.
- Preparation of the final report: drafting of the final report of the project, which will include all the findings, methodologies and results obtained during the development of the project.
- Final review and delivery: final review of the report and the platform, and delivery of the final products to the corresponding stakeholders.

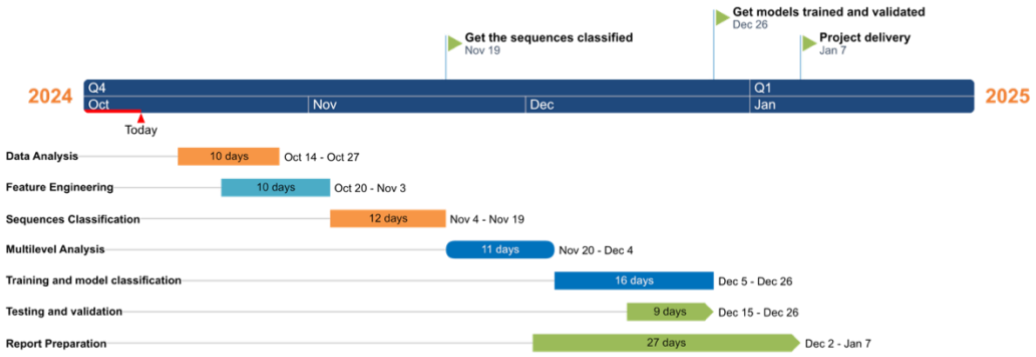

Figure 4. Gantt Diagram. Own elaboration

1.6. Product summary

The primary product of this project is this document, which presents a validated methodological approach for analyzing and comparing the behavior of football teams under varying tactical contexts. By leveraging robust data analysis techniques, clustering, and predictive models, this study provides a systematic framework for understanding team dynamics, offering valuable insights to enhance tactical analysis and decision-making in football analytics.

1.7. Document structure

The remainder of this document is structured as follows:

- **Chapter 2** describes the materials and methods employed in this project, including data collection, preprocessing, feature engineering, clustering techniques, and predictive modeling approaches such as PCA decomposition and deep learning.

- **Chapter 3** presents the results of the study, covering the classification of possession sequences, an assessment of tactical variability, analysis of tactical differences between teams, identification of playing styles, and predictions of sequence types using machine learning models.
- **Chapter 4** concludes the study by summarizing the findings and analyzing their alignment with the project's objectives. It also outlines potential avenues for future research in tactical football analysis.
- **Chapter 5** provides a glossary of key terms, including acronyms, football-related terminology, and technical machine learning concepts, ensuring clarity and accessibility for the reader.
- **Chapter 6 and 7** include the references and appendices, which provide supplementary information and additional resources related to the study.

2. Materials and methods

This chapter outlines the design, development, and methodology employed throughout the project. It integrates data analysis techniques, machine learning, and advanced statistical methods to achieve the dual objectives of identifying relevant tactical patterns and classifying them within specific football contexts.

The methodological approach centers on the processing and analysis of large volumes of tracking and event data extracted from football matches. This data is systematically transformed to generate actionable, interpretable, and tactically useful insights. The choice of tools and techniques was based on criteria such as the ability to capture complex tactical relationships, interpretability of results and computational efficiency.

As previously discussed, there is no standardized method for tactical analysis in football. Each analyst or researcher must select an approach tailored to their specific needs and objectives. Given the inherent complexity and heterogeneity of the game, these methods can vary widely, reflecting the diverse nature of football tactics. Some of the methods explored in the literature include:

- Modeling collective behavior with statistical tools: Hidden Markov Models (HMMs) have been used to analyze tracking data, focusing on spatial dynamics such as effective playing space and pitch control [8][13]. These models offer a robust framework for understanding team tactics at each time point by estimating states such as space occupation and passing networks.
- Deep learning for off-ball movement and event prediction: Neural network architectures have been applied to analyze event data, such as passes, to predict outcomes and understand player contributions, including off-ball movements [10]. This provides insights into how individual player decisions influence team dynamics and match outcomes.
- Network analysis to study team coordination: Network-based approaches quantify connections between players, such as passing networks, revealing tactical structures like centralization, density, and communication between teammates [2].

This project adopts a combination of techniques tailored to analyze finishing sequences and evaluate team tactics within these contexts. The main methods include:

Unsupervised clustering of sequences

Finishing sequences are classified using unsupervised clustering techniques, allowing the identification of distinct tactical profiles and categorizing sequences based on their characteristics.

Leverage Kognia's tactical data

Kognia's dataset provides rich, semantically meaningful tactical annotations for each sequence. These annotations are utilized to:

1. Detect differences in tactical approaches between teams.
2. Identify patterns and trends in the use of specific tactics across teams and contexts.

Trend identification and comparative analysis

The methodology explores trends in both the types of sequences used by teams and the tactics applied within these sequences. Comparative analyses evaluate how teams adapt their strategies to different match situations, providing actionable insights into playing styles and tactical diversity.

2.1 Data collection, exploration and preprocessing

The project used data from 379 matches of the 2022/2023 Spanish *La Liga* season (one match excluded due to a third-party company data collection issue). Three primary data sources were available for each match:

- CSV of events: This dataset contains detailed information about every event in the match, including the type of action (e.g., passes, shots, recoveries), the corresponding frame in the match video, the team and player executing the action, and other contextual variables.
- .h5 files with trajectories (tracking data): These files capture continuous positional data for players and the ball throughout the match. They include additional variables, though not all were utilized in this project.
- JSON with tactics: Provided by Kognia, these files offer detailed tactical insights for each match, identifying applied tactics and their defining characteristics.

To prepare the data for analysis, Python scripts and Jupyter Notebooks were developed. These tools facilitated the integration of event, tactics, and tracking data into a unified structured dataset. Key steps included:

- Data Cleaning: Filtering and cleaning the raw data to remove redundancies, inconsistencies, and irrelevant sequences.
- Sequence Selection: Ensuring that only relevant sequences (those culminating in a shot on goal) were included in the dataset.

An exploratory analysis provided valuable insights into the treatment of variables and data processing. A total of 6292 finishing sequences were identified, during which 89,818 events occurred. This analysis informed the development of feature engineering strategies and revealed early patterns relevant to the project objectives. Table 1 shows a basic statistical summary of the identified finishing sequences.

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
sequences per match	16.60	4.57	4.00	13.0	17.00	20.00	33.00
time in seconds per seq	41.17	18.47	0.08	11.35	19.32	32.64	147.48
events per sequence	14.27	10.99	1.00	7.0	11.00	19.00	98.00

Table 1. Finishing sequences. Basic stats

The number of events per finishing sequence shows considerable variability, as evidenced by the standard deviation of 10.99, which constitutes 77% of the mean (14.27). This indicates that finishing sequences exhibit a wide range of complexity, from simple sequences consisting of a single event to highly elaborate ones involving up to 98 events. Such diversity reflects the dynamic nature of football plays, emphasizing the importance of categorizing and analyzing these sequences in detail.

Regarding the duration of the sequences, while 75% of the sequences last less than 32.64 seconds, the maximum recorded duration is 147.48 seconds, highlighting the presence of outliers. These longer sequences may represent exceptional plays where the team retains possession for an extended period before attempting a shot. These could occur in situations where teams prioritize control or strategic build-up. However, they could also arise from errors in the dataset, which underscores the need for careful handling of such anomalies during analysis.

As figure 5 shows, the distribution of sequences per match appears to follow a normal distribution. This suggests consistency in the frequency of sequences across matches. However, the long right tail in the duration distribution indicates the occasional appearance of atypical sequences, likely due to the aforementioned long possessions or specific match situations.

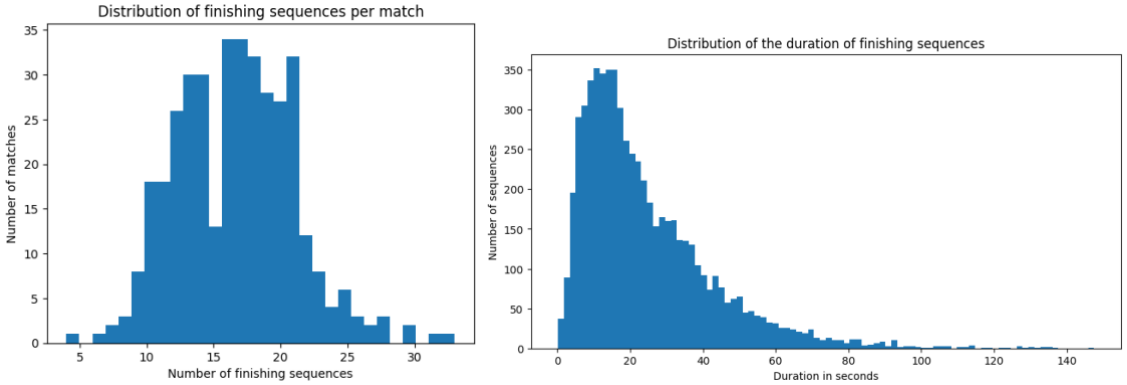

Figure 5. Basic distributions

Figure 6 illustrates the different events that trigger a finishing sequence. The most common events are throw-ins and interceptions, highlighting how offensive opportunities often start in ball-recovery situations. The ratio of starting sequences with the ball in play to those that start after a stoppage in play is approximately 60:40. This ratio is indicative of the fact that, although the continuous flow of play is fundamental for the creation of opportunities, plays that start after stoppages (such as throw-ins) also play a critical role.

It is particularly relevant that around 20% of the sequences start with a throw-in, highlighting its relevance, which is often underestimated in tactical analysis. Set-piece situations, such as throw-ins, emerge as key opportunities for teams to develop effective offensive plays, reinforcing their relevance in tactical analyses.

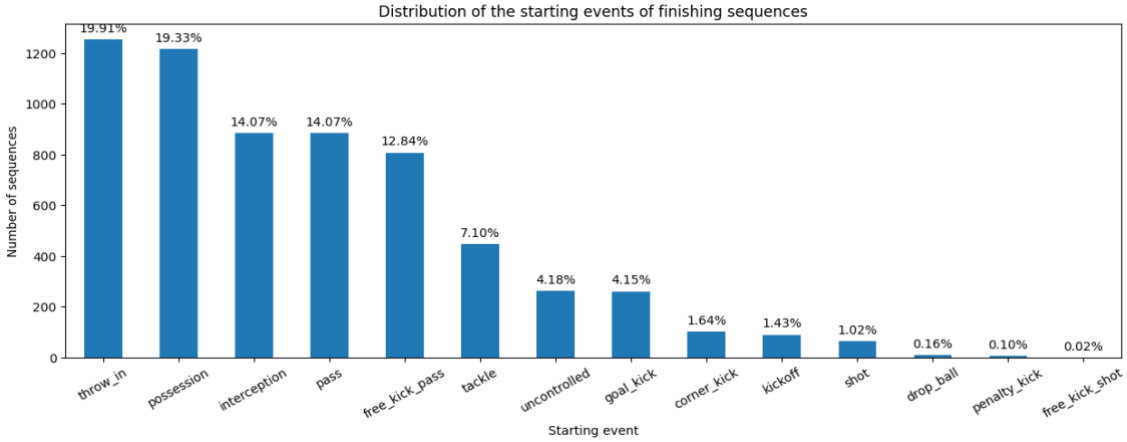

Figure 6. Starting event distribution

The starting zones heatmap (Figure 7) illustrates the areas of the field where finishing sequences most frequently begin (refer to Appendix 7.1 for zone definitions). Most sequences originate in defensive areas or the sides of the field. As expected, finishing sequences most frequently culminate in attacking areas (Zone D), given their nature as sequences ending in a shot.

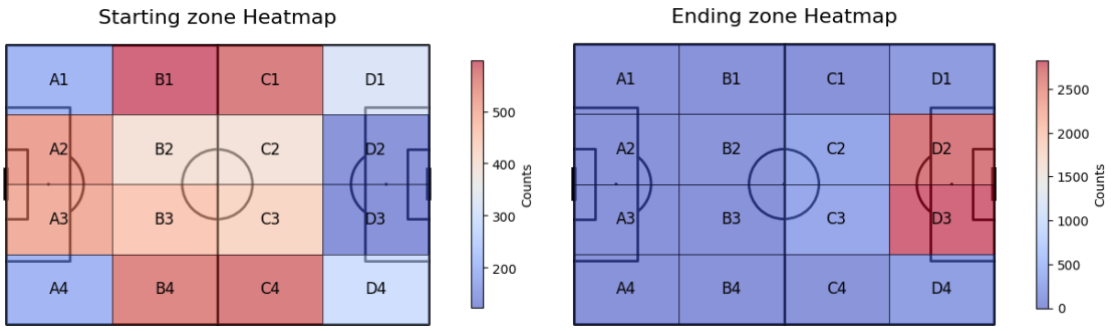

Figure 7. Starting and ending zones. Attack direction from left to right

Focusing on sequences starting with the ball in play (figure 8), the hottest zones are found in the areas close to the middle of the field (zones B and C), and in the center zones close to the own goal (A2, A3). This it means that most of the sequences begin with the team getting the ball when the opponent either is progressing (slightly tendency recover in the center zones) or it's already in the finishing zone (strongly tendency to recover in the center zones since the goal is there).

For sequences that begin after a stoppage in play, the starting concentrate on sides of zoned B and C, offensive corners, and the own goal area. This is consistent with findings already explained about starting events (most important starting event is the throw-in and goal-kicks are executed from the own goal).

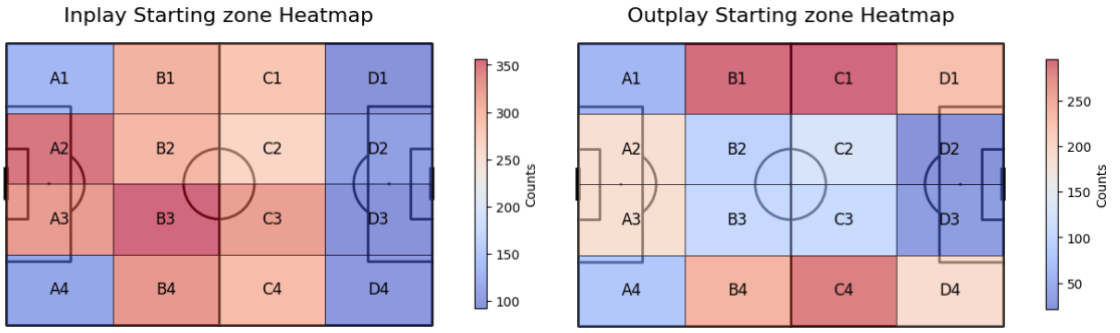

Figure 8. Starting zones. Inplay vs Outplay. Attack direction from left to right

2.2 Feature engineering. Indices creation

To enhance the analysis of game sequences and improve their representation from a tactical perspective, a series of specific indices have been developed. These indices are designed to capture key aspects of the tactical and dynamic behavior of teams on the field [13, 14]. The goal is to enrich each sequence with relevant information, allowing a deeper understanding of how teams progress, use space and structure their actions during a match. By leveraging these indices is possible to obtain a more detailed and meaningful characterization of

sequences, which facilitates the identification of fundamental patterns for advanced tactical analysis.

Verticality Index

The verticality index is a metric that quantifies the average vertical movement per event during a game sequence. It is defined as:

$$Verticality\ Index = \frac{\sum_{i=1}^{n}(end\ zone_i - start\ zone_i)}{n}$$

End zone and start zone represent the zones where each event i in the sequence starts and ends, respectively and n are the total number of events in the sequence.

This index offers a measure of the vertical change within a sequence, providing insights into whether the team is advancing forward (positive value), maintaining its position, or moving backward (zero or negative value) in tactical terms.

To compute the Verticality Index, the zones (A, B, C, and D) are assigned numerical values: 1, 2, 3, and 4, respectively. This conversion ensures that the metric reflects a logical progression along the field. A positive value indicates offensive progression, where the team's actions are directed towards more advanced zones on the field. A negative value suggests defensive retreat, reflecting scenarios where the team is blocking an opponent's attack or repositioning to seek better opportunities to attack.

Laterality index

The laterality index is a metric designed to evaluate the distribution of events on the field with respect to their lateral or central positioning. This index provides insights into whether a team tends to play more through the center or utilize the wings to develop its plays. It is defined as:

$$f(x) = \begin{cases} central\ if\ x\ \in \{2,3\} \\ wide\ if\ x\ \{1,4\} \end{cases}$$

x represents the lane in which an event ends, and $f(x)$ categorizes the lane as central (if it is in zones 2 or 3) or wide (if it is in zones 1 or 4). Then it can be calculated as:

The total number of events in the central zones.

$$E_c = \sum_{i=1}^{n} I(f(x_i) = central)$$

The total number of events in the lateral zones.

$$E_l = \sum_{i=1}^{n} I(f(x_i) = wide)$$

The total number of events.

$$E_t = E_c + E_l$$

And finally:

$$Laterality\ Index = \frac{E_c + E_l}{E_t}$$

A positive value indicates greater number of wide events than central events, suggesting more lateralized play. A negative value indicates greater number of central events than wide events, suggesting more centralized play. A value closes to zero suggests a balanced distribution between wide and central events.

Progression Speed Index

The progression speed index is a metric designed to measure the rate of advancement or retreat of a team on the field during a sequence of play. This index quantifies the dynamics of the sequence, offering insights into the pace at which a team progresses towards the opponent's goal or retreats towards their own goal. It is calculated as:

$$Progression\ Speed\ Index = \frac{Vertical\ distance}{Time}$$

Vertical distance represents the distance between the starting point and the ending point of the sequence along the x-axis of the field. The x-axis measures advancement towards the opponent's goal (positive) or retreat towards the team's own goal (negative). Time is the total duration of the sequence in seconds.

This index provides valuable insights about the dynamics of the sequences. It allows to identify if a team is being aggressive in its advance towards the opponent's goal or if it is retreating or advancing slower. A high value indicates that the team advances quickly towards the opponent's goal, which is typical of counterattacking situations or fast transitions. A low or negative value suggests that the team advances slowly, stays in position, or even retreats towards its own goal.

To better understand the relationships between the calculated indices and other features, a correlation matrix was generated (Figure 9). This analysis provides insights into the redundancy or complementarity of the variables, helping to refine the dataset by identifying which variables contribute most effectively to the representation and interpretation of finishing sequences.

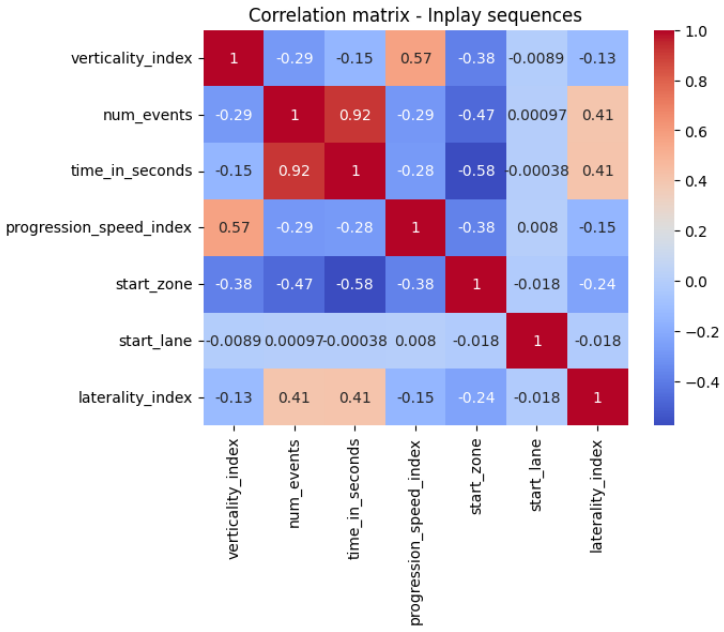

Figure 9. Variables correlation matrix

Strong correlation between *num_events* and *time_in_seconds* (0.92) suggests a significant overlap in the information they provide. As a result, *num_events* was discarded to avoid redundancy in the dataset.

The lack of significant correlations between *start_lane* and other variables indicates that it does not directly contribute to the patterns identified in the sequences. While its removal could be considered, it may still serve as a useful feature for specific exploratory analyses or context-specific interpretations.

Verticality_index and *progression_speed_index* exhibit a moderate correlation (0.57), which highlights that they capture related yet complementary aspects of the sequences. This reinforces their inclusion as they provide distinct tactical insights into team behaviors during sequences.

2.3 Sequences clustering

Since the sequences were not labeled and the resources to manually label thousands of sequences were not available, an unsupervised machine learning approach [9] was chosen. Clustering, an unsupervised learning technique, groups data points into clusters based on their similarity, allowing patterns and

structures to emerge without requiring predefined labels. This approach is ideal for categorizing sequences into tactically meaningful groups, enabling the identification of recurring behaviors and trends in team play.

In this project, *Hierarchical Clustering* was employed to segment sequences. This method provides flexibility in exploring clustering structures and does not require specifying the number of clusters beforehand, making it particularly suitable for the dynamic and heterogeneous nature of football data.

Methodology

- Distance metric: The similarity between sequences was quantified using Euclidean distance, a standard measure in multidimensional spaces.
- Linkage criterion: Average linkage was applied, which calculates the average distance between all points in two clusters. This approach minimizes the impact of outliers and is well-suited for noisy data.
- Dendrogram construction: A dendrogram was generated to visualize hierarchical relationships between sequences. This allowed for exploring clustering levels and provided insights into how sequences group together.
- Cluster validation:
    - Silhouette Score: Used to evaluate the quality of clustering by comparing the cohesion within clusters to the separation between clusters.
    - Visual inspection: The dendrogram was reviewed to identify optimal clustering cut-off points and ensure tactical interpretability.

Compared to other methods such as K-Means and DBSCAN, Hierarchical Clustering was chosen for its ability to:

- Explore relationships at multiple levels of similarity without requiring a predefined number of clusters.
- Capture hierarchical structures, which are particularly useful for understanding the layers of tactical complexity in football.
- Handle the non-linear relationships common in dynamic game data.

By implementing Hierarchical Clustering, this study established robust and semantically meaningful sequence groupings, providing a foundation for advanced tactical analyses in the subsequent results chapter.

2.4 Analyzing tactic differences between teams: Z-score

The Z-score [18] is a statistical measure that indicates how many standard deviations a data point is above or below the mean of a distribution. This metric

allows data to be standardized and compared uniformly, even when originating from different distributions Is calculated as:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X is a data point.
- μ is the population mean.
- σ is the population standard deviation.

A positive Z-score indicates that the data point is above the mean, while a negative Z-score indicates that it is below the mean. By standardizing data, Z-scores make it easier to identify significant deviations and compare values across distributions with differing scales.

In this study, Z-scores are used to standardize and compare tactical differences across teams. This approach is essential because the distributions of Kognia tactics usage and sequence counts are often imbalanced, with some teams using certain tactics more frequently or engaging in more sequences overall. By measuring how many standard deviations a proportion (e.g., tactic usage) is above or below the mean, Z-scores highlight statistically significant differences in team tactics while accounting for these imbalances.

A difference is considered statistically significant if the corresponding p-value falls below a predefined threshold, typically 0.05. The p-value represents the probability of observing a difference as extreme as the one calculated, assuming the null hypothesis (no difference between groups) is true. In this context, a lower p-value indicates stronger evidence against the null hypothesis, suggesting that the observed difference in tactic usage is unlikely to be due to random chance and is instead attributable to real tactical variability between teams or contexts.

When analyzing proportions, a variant of the Z-score formula is applied to account for differences in sample sizes between teams:

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Where p_1 and p_2 are the observed proportions of a specific tactic's usage for two different team and n_1 and n_2 are the total sample sizes (e.g., total number of sequences) for each team.

As explained in the results section, both the number of Kognia tactics utilized, and the number of sequences vary significantly across teams. This imbalance makes a direct comparison unreliable. Using proportions allows the analysis to:

1. Standardize tactic usage across teams with different sequence counts.
2. Account for how tactics are employed within specific sequence clusters, enabling a detailed and context-aware analysis.
3. Highlight meaningful differences in how teams approach similar situations, providing actionable insights for tactical evaluation.

2.5 PCA decomposition to define play styles

After clustering possession sequences into groups based on similar tactical patterns, additional dimensionality reduction is applied through Principal Component Analysis (PCA) approach [9]. This approach allows for a deeper understanding of how scoring opportunities are generated and offers both theoretical and practical advantages for tactical analysis.

PCA reduces the dataset to two or three principal components while preserving most of the variability. This simplification enables easier visualization and interpretation of trends and differences between teams in a graphical format. By projecting clusters into the PCA-reduced space, it becomes possible to observe groupings of teams that share similar playing styles or trends in generating scoring opportunities. This provides a visual framework for comparing teams within a tactical context. The principal components represent linear combinations of variables that explain the largest portions of variance in the data. These components help identify which tactical aspects are the most relevant in differentiating teams' styles of play.

PCA relies on the eigenvalues and eigenvectors of the covariance matrix of the data. Eigenvalues represent the amount of variance explained by each principal component, with higher eigenvalues indicating components that capture more variability in the data. Eigenvectors, on the other hand, define the direction of the principal components and are essential for interpreting their semantic meaning. By examining the eigenvectors, it becomes possible to label the principal components semantically, connecting them to specific tactical aspects.

To ensure that the selected number of principal components captures sufficient variability without oversimplifying the data, Mean Squared Error (MSE) is used to measure the reconstruction error. MSE quantifies the difference between the original data and its reconstruction from the reduced dimensions. The goal is to minimize the reconstruction error while maintaining interpretability. A lower MSE indicates that the chosen number of components effectively captures the underlying structure of the data.

2.6 Deep learning for sequence type prediction.

Classifying the sequences using Kognia tactical data allows for an assessment of whether the tactics generated by Kognia are sufficiently representative and discriminative to identify tactical contexts previously defined from objective data, such as eventing and tracking. While objective data provides a solid foundation for classifying sequences in a structured and reproducible way, Kognia's tactics capture qualitative nuances of tactical behavior that could enhance the characterization of sequences.

By training a neural network with these tactics as input features and the previously identified clusters as labels, the model directly evaluates the ability of tactical data to reflect the specific dynamics of each tactical context. This approach not only validates the practical utility of the tactics but also explores how they complement objective data in analyzing game patterns.

Input data

As input data each sample in the input tensor represents the proportions of tactics observed in a sequence. These proportions provide a "tactical fingerprint" of each sequence, summarized in a compact and relevant representation.

The labels (ground truth) correspond to the tactical context assigned to each sequence, such as the previously defined clusters. The model's task is to learn the relationship between tactical proportions and contexts.

Since the dataset is imbalanced, with certain clusters containing significantly fewer examples than others, oversampling was applied to the training dataset. This technique was used to increase the representation of underrepresented clusters, ensuring that the model does not become biased towards more frequent categories.

Model architecture

The simple neural network has an input layer size equal to the number of tactics considered, since each tactic has an associated proportion. Hidden layers use densely connected neurons to learn non-linear relationships between tactics. As output it produces probabilities for each possible context label (e.g. 7 classes if there are 7 contexts). A ReLU activation function is employed in the hidden layers, and the model is tailored for multiclass classification using a cross-entropy loss function. The dataset is split into an 80:20 ratio for training and testing purposes, ensuring a robust evaluation framework.

Metrics for Evaluation

To evaluate the performance of the neural network, the following metrics [17] were used:

1. Confusion matrix: A visualization of the performance across all classes, showing true positives, false positives, true negatives, and false negatives for each class.

2. Accuracy: Measures the overall correctness of the model's predictions.

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions}$$

3. Precision: Assesses how many of the predicted positives are truly positive for each class.

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

4. Recall: Measures how many of the actual positives are correctly identified.

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

5. F1-Score: The harmonic mean of precision and recall, providing a balanced measure.

$$F1 - Score = 2 \times \frac{Precision\ \times\ Recall}{TPrecision + Recall}$$

3. Results

3.1 Finishing sequences classification

After experimenting with various combinations of explanatory variables discussed in the previous chapter, the best results were achieved when segmenting sequences into seven clusters. This segmentation was based on the verticality index, sequence duration and initial possession zone. The chosen variables provided both quantitative robustness and qualitative interpretability.

As shown in Figure 10, the PCA visualization demonstrates a reasonably clear separation between clusters along the first two principal components. However, some overlapping regions such as between *opp-half-recovery-and-play* and *long-build-ups* or between *short-build-ups* and *long-build-ups* indicate areas of similarity among sequences. This is expected, as football tactics and game phases often do not exhibit strict boundaries. The Silhouette Score of 0.422 supports the clustering's moderate quality [11].

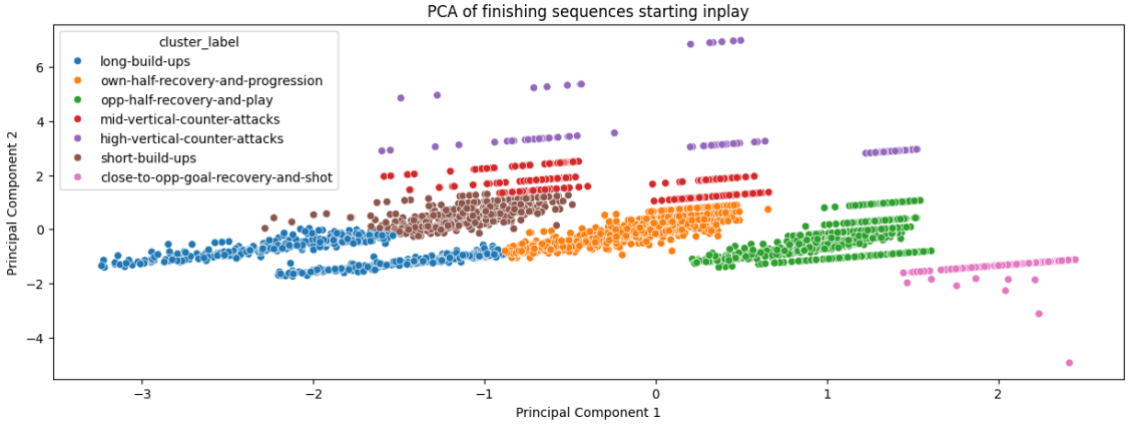

Figure 10. PCA visualization of the final clustering

To label the clusters semantically, the descriptive statistics of the sequences were analyzed (Table 2).

cluster label	verticality index		time (s)		start zone (1-4)		
	mean	**std**	**mean**	**std**	**mean**	**std**	**count**
long-build-ups	0.124	0.049	43.492	7.767	1.445	0.497	542
short-build-ups	0.308	0.084	22.894	5.383	1.000	0.000	412
own-half-recovery-and-progression	0.235	0.095	19.030	6.760	2.000	0.000	771
opp-half-recovery-and-play	0.195	0.139	11.650	5.908	2.995	0.073	937
high-vertical-counter-attacks	1.082	0.246	8.652	5.791	2.216	0.832	116
mid-vertical-counter-attacks	0.585	0.095	12.782	4.503	1.551	0.498	303
close-to-opp-goal-recovery-and-shot	-0.010	0.076	6.908	4.218	4.000	0.000	245

Table 2. Sequence types. Descriptive statistics

As can be seen, the generated clusters capture a wide range of tactical styles and contexts:

- *Long build-up:* This cluster represents sequences with a focus on maintaining possession starting from defensive zones, slowly progressing towards the opponent's area, prioritizing control and careful construction.
- *Short build-up:* These sequences reflect a more compact build-up, generally from defensive zones, with higher tempo and less prolonged possession compared to long build-ups.
- *Own-half recovery and progression:* These sequences start with a ball recovered in defensive midfield but close to opponent's midfield. They are recoveries that evolve towards the opponent's half at a moderate pace. However, from a footballing perspective, this cluster possesses less pronounced semantic meaning, serving as a "catch-all" category that likely includes a wide variety of sequences with differing tactical intentions and styles.
- *Opp-half recovery and play:* Sequences generated after recoveries, generally in the opponent's half. They represent a quick attempt to establish control and generate plays in favorable positions. Like the previous cluster, this category also lacks a sharply defined tactical identity, likely aggregating diverse sequences. This makes it a broad grouping where varying styles and contexts converge, emphasizing the need for cautious interpretation.
- *High-vertical counterattacks:* This cluster captures fast and aggressive counterattacks, where vertical transitions predominate. These plays prioritize speed over possession. They start generally around the halfway

line so probably there are not a lot of players between the ball and the goal.

- *Mid-vertical counterattacks:* These sequences represent more balanced counterattacks that start from defensive zones, with a combination of pace and control probably due there are more distance and more players between the ball and the goal.
- *Close-to-opp-goal recovery and shot:* This cluster represents recoveries very close to the opponent's area that quickly end in a shot attempt. They reflect high pressure or recovery in contexts close to the goal.

Tactical implications

The cluster categorization captures a diverse range of tactical contexts. By identifying these clusters, the analysis provides a robust framework to evaluate and distinguish team strategies in various game situations. This segmentation lays the foundation for deeper tactical insights in subsequent analyses.

It is important to note that these contexts are the result of the specific methodology employed and the interpretation of the data within this study. While the method is replicable, the resulting contexts could vary if applied to different datasets (e.g., other leagues, multiple seasons) or with alternative variables. Additionally, interpretations by other analysts might lead to different categorizations, reflecting the subjective nature of tactical analysis. The domain expertise of the analyst (in this case, knowledge of football) is critical in shaping how the data is read and how specific tactical contexts are identified.

3.2 Assessing tactical variability between teams and sequence types

Assessing tactical variability between teams and sequence types is a key point for identifying the factors that most influence tactical decisions and behaviors in football. By analyzing whether variability in the use of Kognia tactics is greater among sequence types or among teams, a fundamental question can be answered: are tactics driven more by the game context or by the tactical identities and preferences of each team?

This analysis is key to understanding whether teams adapt their tactics dynamically based on game situations and context or they maintain a consistent and rigid tactical identity regardless of situational changes. Moreover, the analysis assesses the significance of sequence clusters as a framework fo tactical evaluation, offering insights into how they enhance our understanding of football dynamics.

Figure 11 (b) reveals a large variability in the number of in-play finishing sequences per team. For example, teams like FC Barcelona and Real Madrid have significantly more sequences than teams like Getafe CF or RCD Mallorca. This highlights the need for proportional analysis to ensure fair comparisons across teams with different sequence counts.

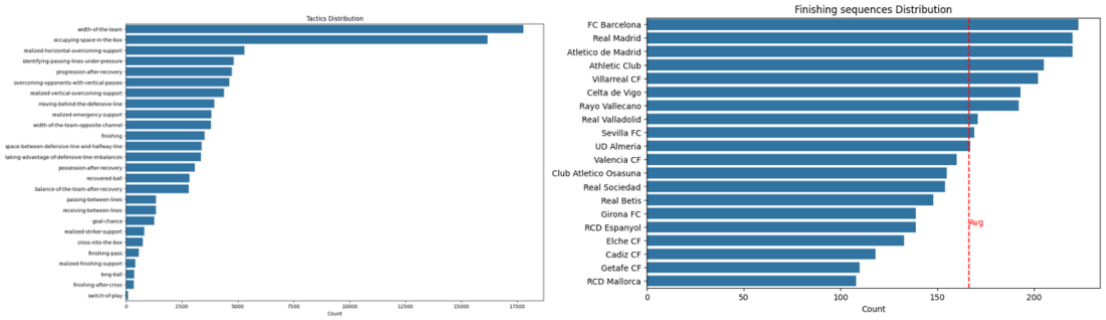

Figure 11. Tactics a) and in play finishing sequences distributions b)

Figure 11 (a) demonstrates an imbalanced distribution of tactics, with some tactics appearing much more frequently than others. For instance, tactics like *occupying-spaces-into-the-box* are dominant. This aligns with the nature of the dataset, which focuses on sequences culminating in a shot, typically occurring near the box, as shown in earlier analyses.

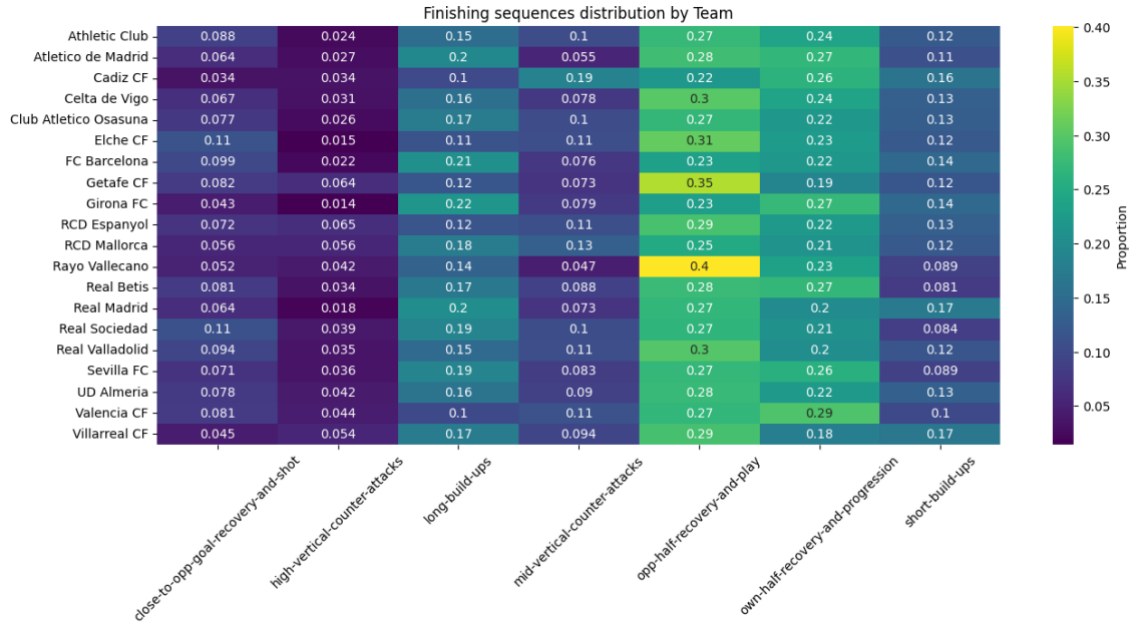

Figure 12. Finishing sequence types per team

Focusing on sequence types per team (Figure 12), teams like Barcelona, Real Madrid or Cadiz exhibit a more balanced distribution across multiple sequence types. This suggests their ability to adapt their tactical approaches to different game contexts. In contrast, teams such as Rayo Vallecano or Getafe CF show a

stronger concentration in specific categories, such as *opp-half-recovery-and-play*, reflecting a more rigid and less adaptable tactical focus.

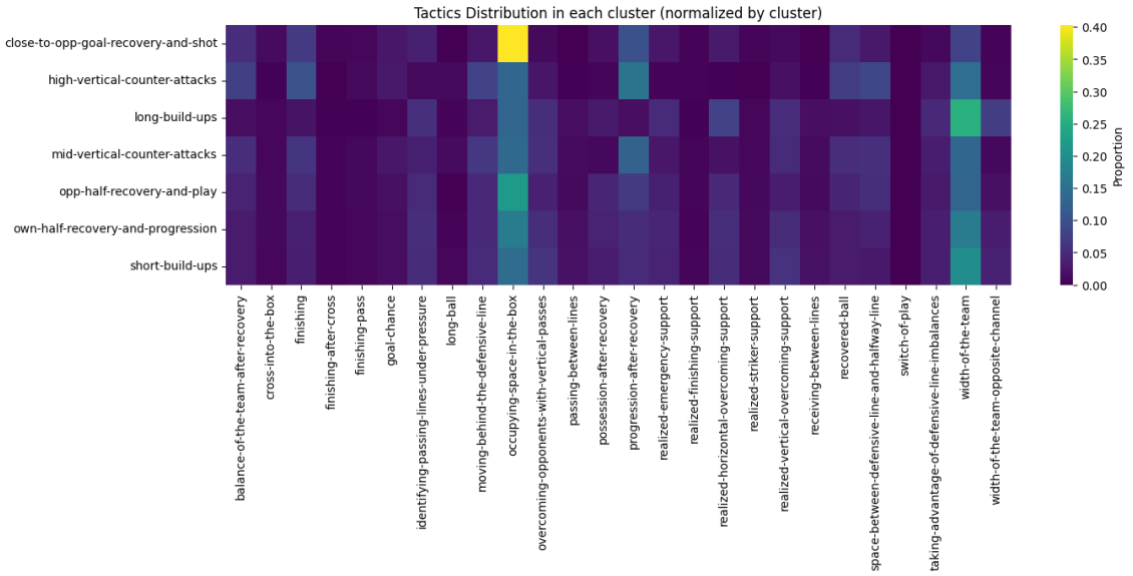

Figure 13. Tactics distribution per sequence type

As illustrated in Figure 13, there is notable tactical variability between clusters. For instance, *high-vertical-counter-attacks* cluster emphasizes tactics like *progression-after-recovery* and *moving-behind-the-defensive-line* which associated with fast transitions. On the other hand, *close-to-opp-goal-recovery-and-shot* sequences rely heavily on tactics such as *occupying-spaces-into-the-box*, reflecting high-pressure scenarios. Meanwhile *long-build-ups* are dominated by possession-based tactics like *width-of-the-team*, highlighting a controlled and patient style.

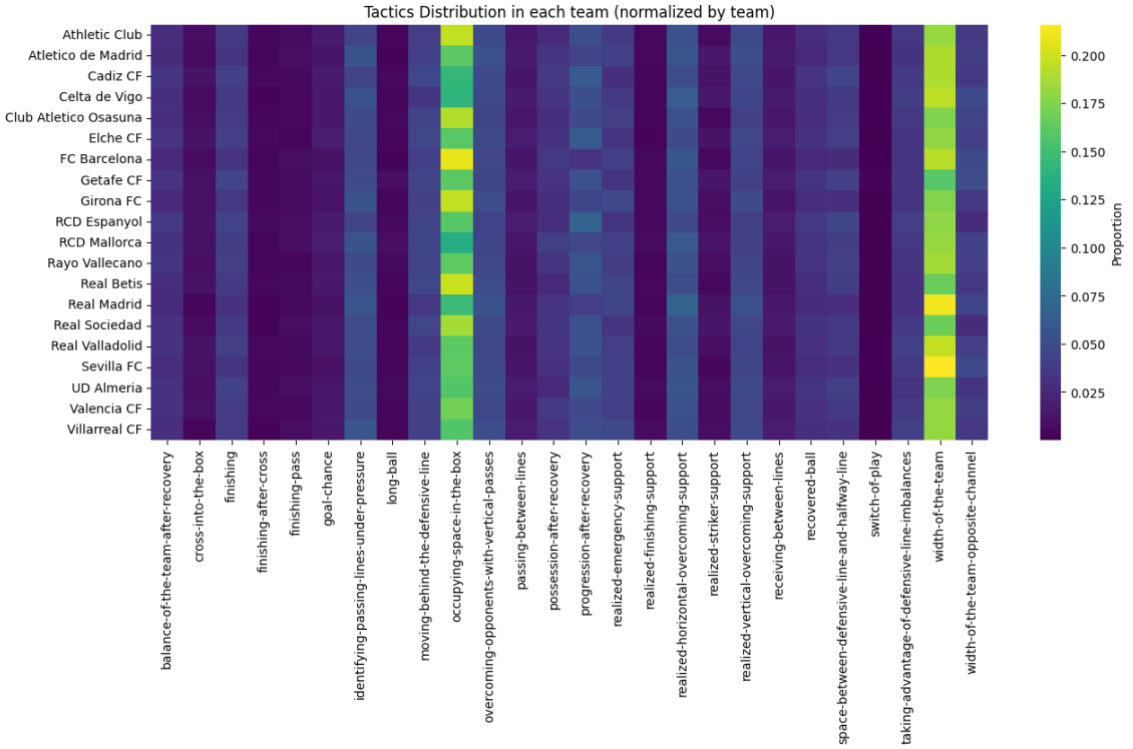

Figure 14. Tactics distribution per team

As expected, figure 14 reveals that *occupying-spaces-into-the-box* and *width-of-the-team* are the most frequently detected tactic across all teams. However, tactical differences between teams are relatively small, with many tactics being deployed in similar proportions.

The analysis of the covariance matrices provides additional quantitative evidence. The means of the covariance matrices for team tactics (8.677×10^{-23}) and tactics clusters (-2.321×10^{-21}) are extremely small, effectively close to zero. This suggests that, on average, the tactics are largely uncorrelated when examined across either teams or clusters.

The nearly zero mean for team tactics indicates that tactical proportions across teams are not strongly interrelated. This supports the observation that tactical behaviors are relatively consistent between teams, with limited variability in the way individual teams deploy their tactics. These findings align with the thesis that teams tend to share similar tactical profiles, making it challenging to differentiate them based solely on tactical proportions.

Similarly, the small mean for clusters implies that the tactics used in different sequence types have weak correlations on average. However, specific clusters, such as *high-vertical-counter-attacks*, show a stronger emphasis on certain tactics (e.g., fast-transition tactics). This highlights that while some clusters are tactically distinct, the overall relationship between tactics within clusters remains

weak. This reinforces the importance of contextual factors in shaping tactical behavior.

These results provide evidence that:

- Tactical differences between teams are relatively small, aligning with the finding that many tactics are shared across teams in similar proportions (as seen in heatmaps and distributions).

- Tactical differentiation across clusters is more pronounced in specific cases but does not imply strong interdependencies between tactics within clusters.

3.3 Analyzing tactic differences between teams

Understanding tactical differences between teams is a cornerstone of advanced football analysis. By comparing the tactical behaviors of teams, we can highlight the nuances in their playing styles and how these differences manifest in specific game scenarios. This section demonstrates the approach by analyzing the tactical differences between FC Barcelona and Atlético de Madrid. The comparison is performed both in general terms and within specific tactical contexts, showcasing the added depth that contextual analysis provides.

Figure 15 illustrates the tactics where statistically significant differences (p-value < 0.05) were identified between FC Barcelona and Atlético de Madrid. The Z-score quantifies the magnitude of these differences, with higher bars indicating greater deviations.

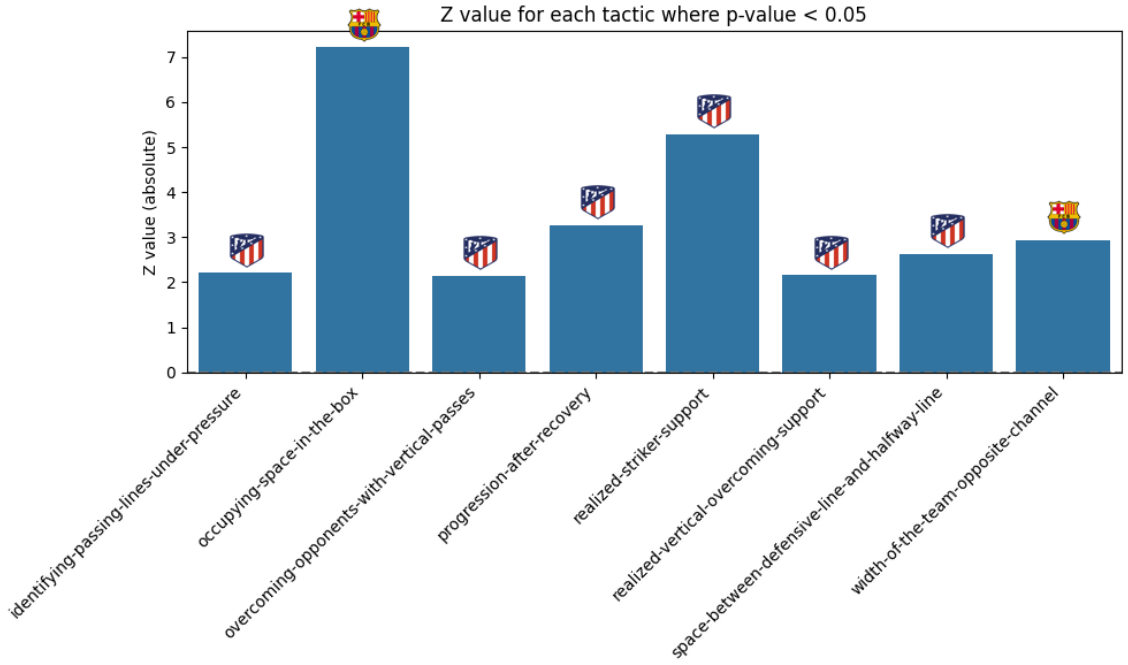

Figure 15. FCB vs ATM overall tactical differences

FC Barcelona: Excels with a particularly high Z-score in tactics such as *occupying-space-in-the-box* and *width-of-the-team-in-opposite-channel*. This suggests that they tend to crowd the opponent's box more and to have greater width in the opposite lane to where the ball is than Atletico de Madrid.

Atlético de Madrid: Predominates in tactics more related to verticality towards the opposing goal as *overcoming-opponents-with-vertical-passes*, *progression-after-recovery* or *realized-striker-support*.

When tactical differences are analyzed within specific sequence types, a greater number of significant tactics is detected, as shown in Figure 16. This approach not only increases the number of tactics detected, but also amplifies the semantic differences between both teams.

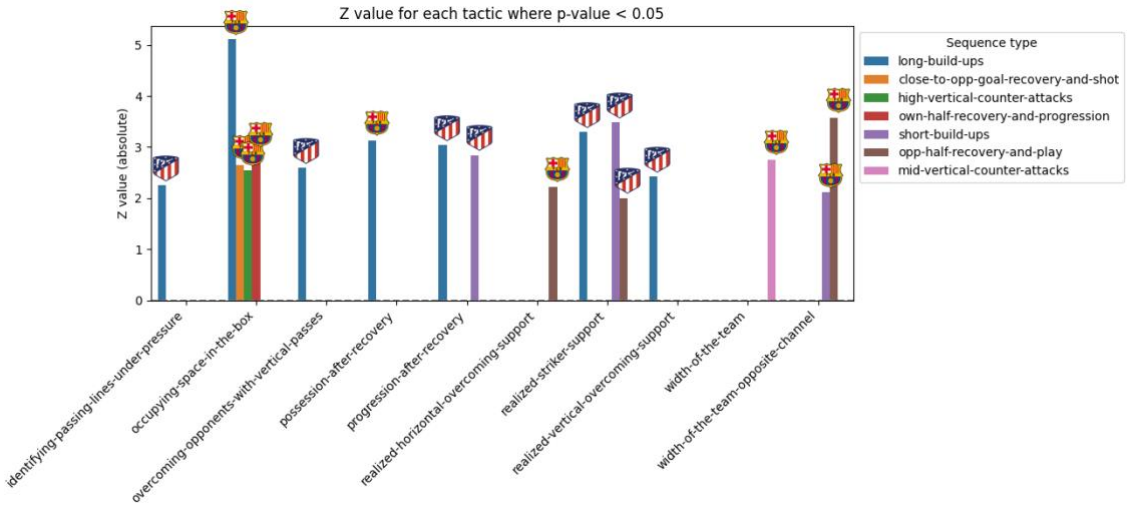

Figure 16. FCB vs ATM tactical differences considering sequence types

FC Barcelona continues to dominate tactics related to possession, field width or box penetration. Granularity in tactic differences has increased, revealing new insights. For instance, *possession-after-recovery* is a key tactic for Barcelona within the *long-build-up* cluster, highlighting a strategic focus on controlled build-ups stronger than Atletico de Madrid. Similarly, *realized-horizontal-overcoming-support* emerges as a distinguishing tactic for Barcelona specifically in the *opp-half-recovery-and-play* cluster, indicating a preference for lateral support in advanced recoveries.

On the other hand, Atlético de Madrid emphasizes tactics associated with quick transitions and direct play. *Progression-after-recovery* and *overcoming-opponents-with-vertical-passes* are particularly significant in both short and *long-build-ups*, emphasizing their vertical approach in possession sequences.

These results illustrate how clustering sequences allows to identify not only which tactics differentiate teams, but also in which specific tactical contexts those differences are most pronounced. This brings a richer and more contextualized perspective to tactical analysis.

Incorporating clusters into tactical analysis not only increases the number of significantly different tactics detected between, but also amplifies semantic and contextual differences. The inclusion of clusters allows tactics to be analyzed based on tactical context, providing an additional level of granularity. This allows differences to be detected that are not visible in a global analysis, where tactics can be “averaged” and lose their specificity.

The analysis of overall tactical differences between teams reveals an average of 5.48 significant tactics, with a standard deviation of 3.39. However, when the analysis incorporates sequence-specific contexts using clusters, the average increases to 8.15 significant tactics, with a slightly lower standard deviation of 3.32. This indicates that generalized analysis reveals fewer and less variable tactical differences between teams, while context-specific analysis using clusters uncovers a higher number of differences, accompanied by greater tactical variability and richness.

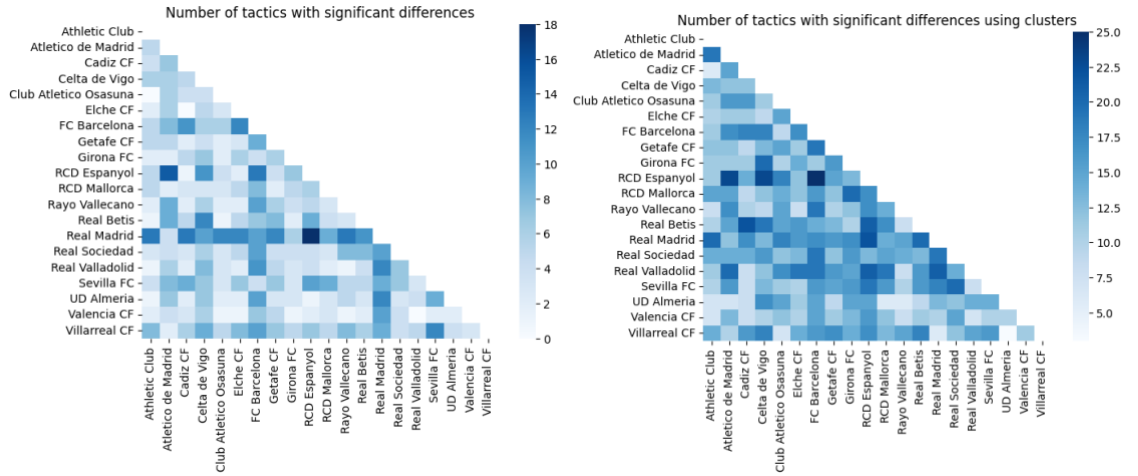

Figure 17. Overall a) and clusters tactic differences b)

From Figure 17 (a), it is evident that, overall, most teams show few significant differences from one another. However, Real Madrid stands out as having the largest tactical differences compared to other teams, suggesting a more distinctive tactical profile when analyzed at a general level.

Looking into Figure 17 (b), the increased intensity of the matrix color highlights how tactical differences become more pronounced and numerous when specific game contexts are considered. This emphasizes the value of using clusters to explore tactical variability, as clusters effectively reveal nuanced, context-dependent tactical behaviors. For example, certain tactics that might not stand

out in a general analysis become prominent in specific clusters, validating the hypothesis that clusters represent unique tactical scenarios where teams adapt their strategies more distinctly.

The results confirm that tactical variability between teams increases when analyzed through the lens of sequence-specific contexts provided by clusters. This demonstrates that tactics are not fixed or global, but are highly context-sensitive, varying according to the type of game sequence. Incorporating clusters into the analysis provides a finer level of detail and precision in tactical assessments, enabling the identification of differences that would remain hidden in a generalized approach.

3.4 Identifying playing styles

The application of PCA provides a structured approach to reduce data dimensionality while simultaneously identifying key trends in how teams construct and finish their possession sequences. This methodology enables the definition of pseudo *playing styles* grounded in the principal components derived from the data.

The selection of the number of components is guided by their associated eigenvalues, as shown in Table 3:

Component	Eigenvalue
1	2.132
2	1.811
3	1.505
4	1.088
5	0.638
6	0.194
7	0.000

Table 3. PCA eigenvalues

The eigenvalues reflect the amount of variance explained by each component. The first three components capture the majority of variance, with Component 1 holding the most explanatory power. Beyond Component 3, the eigenvalues show a marked decline, with Component 7 contributing negligible variance (eigenvalue = 0). Consequently, the first three components are retained for the analysis, as they provide a concise yet comprehensive summary of tactical trends.

Semantic interpretation of components

The eigenvectors (Table 4) provide key insights into the relative importance of each sequence type for defining the principal components. This facilitates the semantic labeling of components based on their tactical significance.

Sequence type	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
close-to-opp-goal-recovery-and-shot	0.280	-0.264	0.186	0.794	-0.005	-0.336	0.273
high-vertical-counter-attacks	0.356	0.492	-0.161	-0.150	0.629	-0.390	0.189
long-build-ups	-0.292	-0.550	-0.319	-0.066	0.497	0.222	0.459
mid-vertical-counter-attacks	-0.305	0.561	0.283	0.316	0.087	0.514	0.375
opp-half-recovery-and-play	0.624	-0.015	-0.160	-0.237	-0.404	0.309	0.520
own-half-recovery-and-progression	-0.177	-0.130	0.695	-0.429	-0.061	-0.356	0.392
short-build-ups	-0.444	0.235	-0.499	0.044	-0.428	-0.446	0.338

Table 4. PCA eigenvectors

The first principal component captures an offensive dimension that is strongly associated with aggressive recovery and play in the opponent's half. This is reflected in the high positive loading of *opp-half-recovery-and-play* (0.624), indicating that sequences where teams recover the ball in advanced areas and continue to play offensively are significant contributors to this component.

In contrast, *short-build-ups* (-0.444) has a negative loading in component 1, suggesting that this component also represents an opposing dimension to *short build-ups*. Teams with low scores in component 1 might favor slower, more deliberate sequences starting from deeper areas of the field, rather than emphasizing immediate recovery and offensive play.

Additionally, *high-vertical-counter-attacks* (0.356) has a moderate positive loading, highlighting that component 1 also aligns with teams that frequently employ fast, vertical transitions to capitalize on opportunities in advanced areas.

In summary, component 1 seems to encapsulate a tactical style centered on high pressing, quick recoveries, and offensive transitions in the opponent's half. Teams with high component 1 values are likely characterized by an aggressive, high-intensity approach, while teams with low component 1 values may exhibit more methodical and possession-oriented styles starting deeper in their own half.

The second principal component reflects a spectrum between fast and direct counterattacks and slower, controlled build-up play. This is evident in the high positive loading of *mid-vertical-counter-attacks* (0.561), which highlights that teams with high component 2 values tend to favor fast, vertical counterattacks initiated from intermediate field positions. Similarly, *high-vertical-counter-attacks* (0.492) also shows a strong positive contribution, further reinforcing the association of component 2 with a direct, counter-attacking approach.

On the other hand, *long-build-ups* (-0.550) has a strong negative loading, indicating that teams with low component 2 scores are more likely to engage in

possession-based, controlled build-up sequences. This suggests a tactical preference for maintaining possession and gradually advancing the ball, as opposed to relying on quick transitions.

In essence, component 2 captures the tactical dichotomy between a fast, transition-based approach (high values) and a slower, possession-oriented style (low values). Teams with high component 2 values prioritize speed and directness, while those with low values focus on control and gradual progression.

The third principal component captures a distinct tactical dimension related to the way teams progress the ball from deeper areas of the field versus their preference for more compact and shorter build-up plays. This component primarily differentiates between sequences originating from the defensive half and those characterized by tighter positional play.

The *own-half-recovery-and-progression* (0.695) sequence type has the highest positive loading. This indicates that teams with high scores in this component tend to favor recovering the ball in their own half and progressing it towards more advanced areas. Such sequences often reflect a focus on structured play originating from deeper positions, where teams aim to gradually build attacks from their defensive lines.

In contrast, *short-build-ups* (-0.499) contributes negatively. This suggests that teams with low scores in this component tend to engage in shorter, more compact sequences. These sequences might involve a higher tempo or an emphasis on maintaining possession within a smaller spatial area, rather than progressing the ball over longer distances.

To evaluate the effectiveness of the dimensionality reduction performed through PCA, a reconstruction of the original data was carried out using the selected components.

With two components, the Mean Squared Error (MSE) is 0.465, while incorporating a third component reduces the MSE significantly to 0.261. This drop indicates that the addition of a third component captures more variance in the original data, improving the reconstruction accuracy.

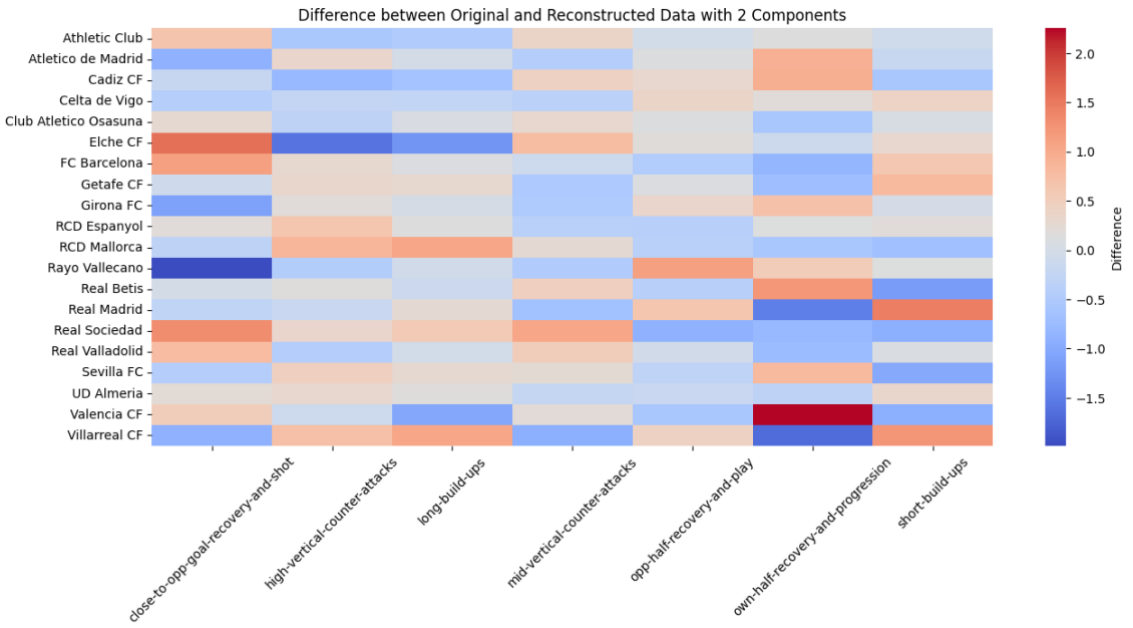

Figure 18. Reconstruction evaluation of error with 2 components

However, despite the higher average error with two components, the error is primarily localized to specific teams, such as Elche, and certain clusters, notably *close-to-opp-goal-recovery-and-shot* and *own-half-recovery-and-progression* (Figure 18). In contrast, teams like Celta, Osasuna, Getafe, Espanyol, and Almería exhibit consistently low errors across all clusters, suggesting that their tactical behavior may be more uniform or less complex, allowing it to be adequately represented with just two components.

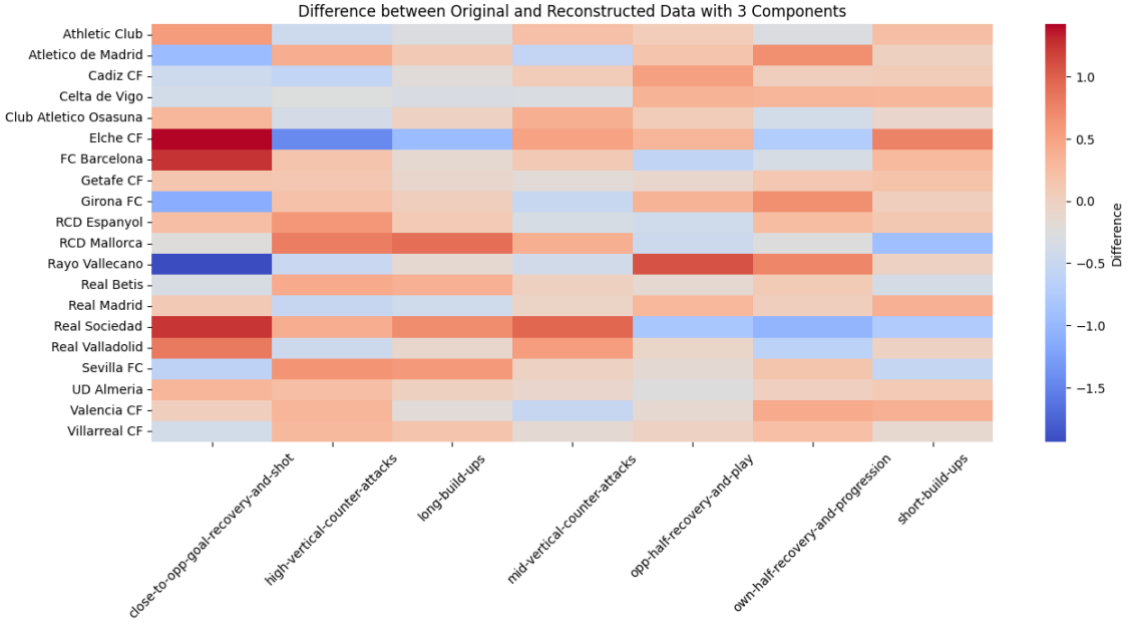

Figure 19. Reconstruction evaluation of error with 3 components

With three components (Figure 19), the overall error decreases substantially, but significant concentrations of error persist in the *close-to-opp-goal-recovery-and-*

shot cluster. This observation implies that this type of sequence may demand additional dimensions to fully capture its variability. Furthermore, certain teams, such as Elche and Rayo Vallecano, do not experience a notable reduction in their errors with the third component. This suggests that their playing styles may exhibit complex patterns that remain challenging to model effectively even with three components.

Tagging trends

Considering these results, the 2D (Figure 20) and 3D (Figure 25) PCA visualizations provide an insightful representation of team distributions based on the identified playing styles derived from the principal components.

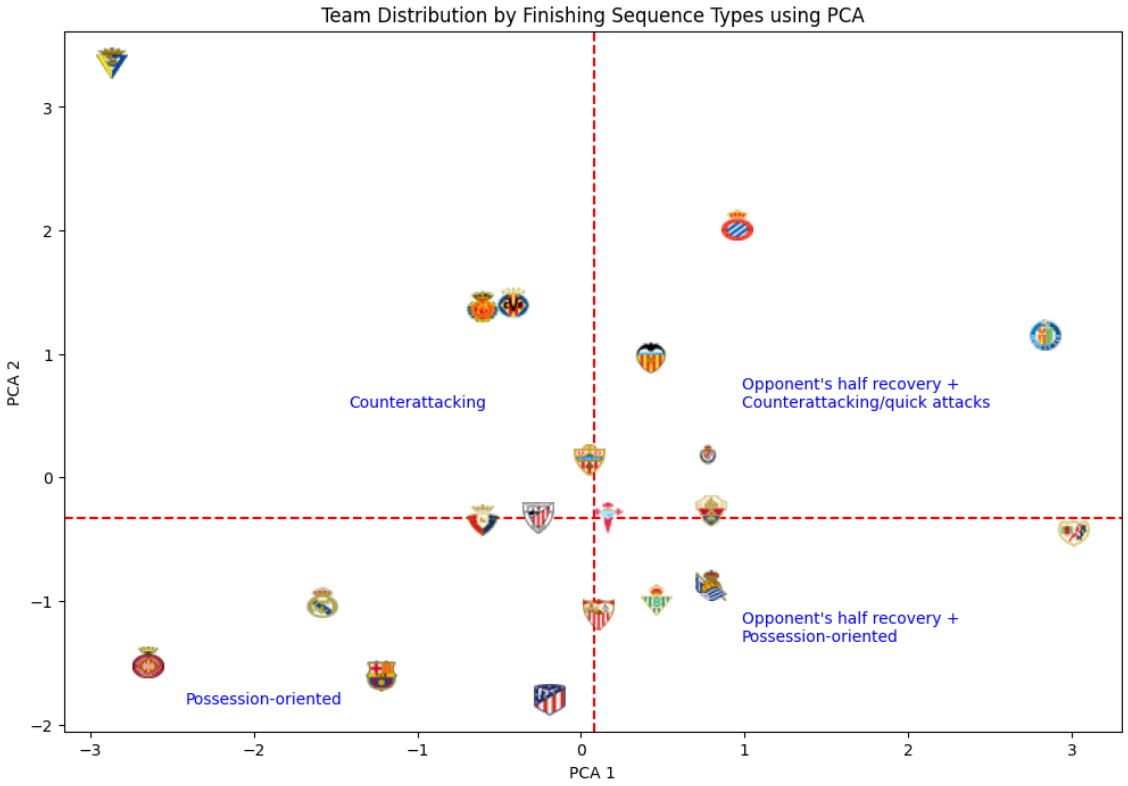

Figure 20. Team distribution by finishing sequence types using PCA

Figure 20 illustrates the positioning of teams along the first two principal components, revealing distinct groupings that suggest clusters of teams sharing similar tactical tendencies. The red cross represents the median values of the components, providing a reference for balanced styles.

Teams like Cadiz, Mallorca or Villarreal, located in the upper-left quadrant are likely those that heavily rely on counterattacking tactics, as indicated by their higher values on component 2 (vertical counterattacks) and lower component 1 values (indicative of reduced reliance on possession-oriented play).

On the other hand, teams like Girona, Real Madrid, Barcelona or Atletico de Madrid, located in the lower-left quadrant display negative values on both component 1 and component 2. This positioning suggests a reliance on long build-ups and possession-focused styles, possibly emphasizing slower and more controlled sequences.

In contrast, the upper-right quadrant contains teams like Getafe, Espanyol, Valencia or Valladolid, characterized by higher component 1 and component 2 values. This positioning indicates a tactical inclination towards recovering the ball in the opponent's half and executing quick, aggressive transitions towards goal.

Finally, teams positioned near the intersection of the red median lines display balanced tendencies, integrating elements of possession, counterattacks, and opponent's half recoveries. This balance reflects a more versatile and adaptable tactical approach.

The analysis of trends within the clusters can be extended to study the tactical tendencies employed by teams in specific contexts. By examining the use of specific tactics within each cluster, it becomes possible to visualize how, even within the same tactical context, teams adopt distinct strategies to reach a finishing opportunity. This deeper exploration highlights the diversity of tactical approaches and provides valuable insights into the adaptability and preferences of teams in varied situations. A detailed breakdown of these tactical trends, including the weights of each tactic within the PCA components, is included in Appendix 7.3.

3.5 Predicting sequence types

Predicting all sequence types

The results of the classification task provide an overview of the model's ability to predict tactical contexts using Kognia's tactical data. The overall accuracy of 0.471 indicates moderate success, highlighting both the potential and limitations of the input data in discriminating between sequence types.

The performance of the neural network in predicting tactical clusters, as reflected in Table 5 and the confusion matrix (Figure 21), reveals a nuanced evaluation of each sequence type.

Sequence type	Precision	Recall	F1-score	N
Close-to-opp-goal-recovery-and-shot	0.47	0.67	0.55	49
High-vertical-counter-attack	0.35	0.70	0.46	23
Long-build-up	0.66	0.83	0.73	109
Mid-vertical-counter-attack	0.34	0.57	0.42	61
Opp-half-recovery-and-play	0.51	0.37	0.43	188
Own-half-recovery-and-progression	0.41	0.19	0.26	154
Short-build-up	0.68	0.74	0.71	82

Table 5. Evaluation metrics for 7 clusters

Close-to-opp-goal-recovery-and-shot: This category shows relatively high recall (0.67), indicating the model's strength in correctly identifying sequences within this context. However, the lower precision (0.47) suggests a tendency to misclassify sequences from other clusters into this category. This is evident in the confusion matrix, where some sequences from clusters such as *opp-half-recovery-and-play* are incorrectly classified here. The overlap could reflect tactical ambiguities in high-pressure scenarios near the opponent's goal.

High-vertical-counter-attack: While the recall for this category is strong (0.70), indicating that most actual sequences in this cluster are identified, the low precision (0.35) highlights significant misclassifications. This is consistent with the confusion matrix, where a portion of *opp-half-recovery-and-play* and *mid-vertical-counter-attacks* sequences are misclassified here. This may be due to overlapping tactical features in rapid transitions, complicating the distinction between these contexts.

Long-build-up: This is the best-performing category, with high precision (0.66) and recall (0.83), demonstrating that the model effectively captures the distinctive features of this sequence type. As seen in the confusion matrix, the majority of true *long-build-ups* sequences are correctly classified, with only minor spillover into clusters such as *short-build-ups*. This reinforces the hypothesis that long, possession-oriented plays are well captured by Kognia's tactical data.

Mid-vertical-counter-attack: The moderate recall (0.57) and lower precision (0.34) indicate challenges in distinguishing this sequence type. The confusion matrix reveals frequent misclassifications into *own-half-recovery-and-progression* and *opp-half-recovery-and-play*, reflecting overlaps in tactical patterns related to transitions.

Opp-half-recovery-and-play: This sequence type presents significant challenges, with a low recall of 0.37 and moderate precision (0.51). The confusion matrix highlights a major source of error: frequent misclassifications into *own-half-recovery-and-progression, mid-vertical-counter-attacks* and *close-to-opp-goal-recovery-and-shot*. This reflects the semantic ambiguities discussed earlier,

where these clusters are less clearly defined and may encompass a wider variety of sequences.

Own-half-recovery-and-progression: This is the weakest-performing cluster, with very low recall (0.19) and precision (0.41). The model struggles significantly to identify sequences in this category, as seen in the confusion matrix, where a substantial number of these sequences are misclassified into *opp-half-recovery-and-play, mid-vertical-counter-attacks, long-build-ups* and *short-build-ups*. This supports the idea that the semantic definition of this context is weaker and potentially overlaps with other clusters.

Short-build-up: Alongside *long-build-ups*, this sequence type shows strong performance, with high precision (0.68) and recall (0.74). The confusion matrix supports this, with most sequences correctly classified and only minimal overlap with *long-build-ups* and *own-half-recovery-and-progression*. This suggests that short, compact build-up plays are well-defined in tactical data, making them easier for the model to distinguish.

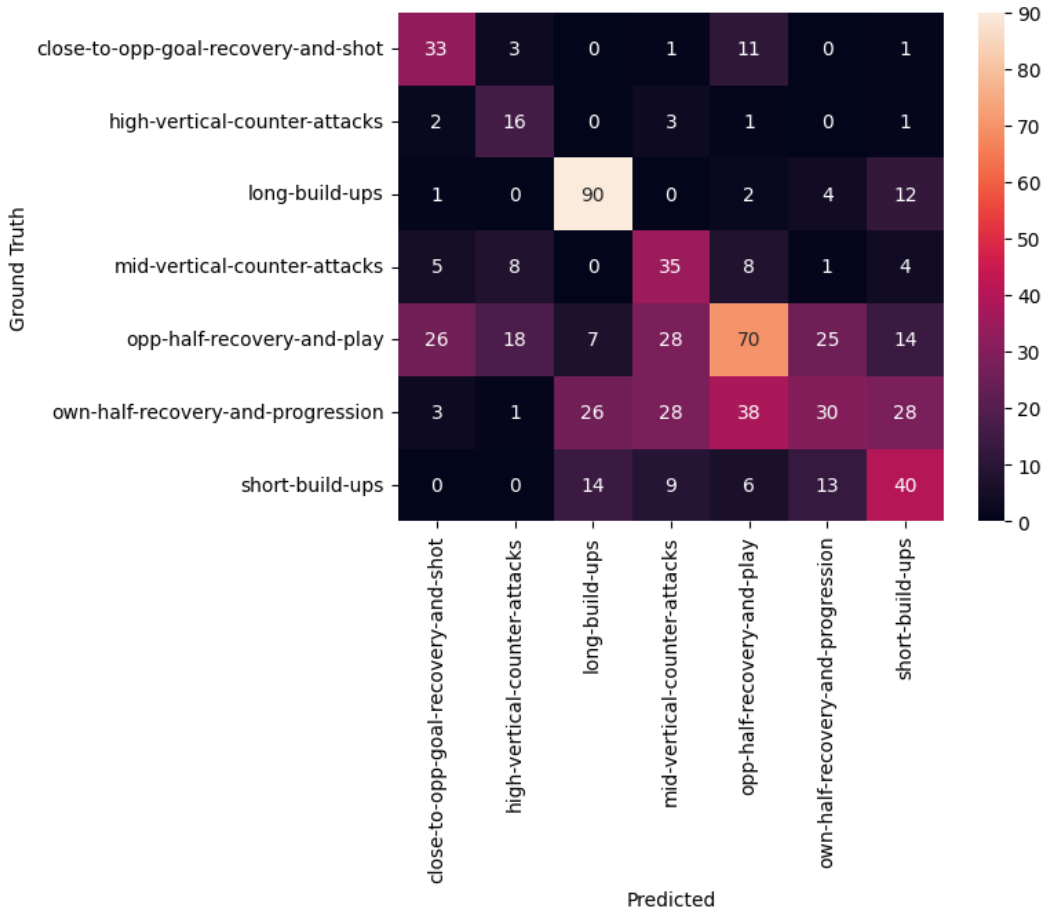

Figure 21. 7 clusters confusion matrix

In summary, the model's performance underscores the variability in how well Kognia's tactical features capture the nuances of different sequence types. The

strong performance for *long-build-ups* and *short-build-ups* highlights that possession-oriented contexts are better represented, while the weaker results for *own-half-recovery-and-progression* and *opp-half-recovery-and-play* reflect challenges in defining these contexts semantically. These findings emphasize the importance of refining tactical definitions and features to improve the model's capacity to distinguish between overlapping and complex sequence types.

Predicting most accurate contexts

Following the previous analysis, a refinement was made by removing the two most generic clusters, which, as previously explained, were the least semantically defined and introduced significant misclassification errors in the model. By removing these clusters, the system is tested under a more refined framework where contexts are better defined from a footballing perspective. This adjustment aims to evaluate whether better-defined tactical contexts lead to improved prediction results using Kognia's tactics. Below, the results of this refined analysis are presented.

Sequence type	**Precision**	**Recall**	**F1-score**	**N**
Close-to-opp-goal-recovery-and-shot	0.80	0.71	0.75	49
High-vertical-counter-attack	0.51	0.87	0.65	23
Long-build-up	0.88	0.81	0.84	109
Mid-vertical-counter-attack	0.75	0.62	0.68	61
Short-build-up	0.68	0.74	0.71	82

Table 6. Evaluation metrics for 5 clusters

The overall accuracy of the model increased significantly, reaching 0.747, which is a notable improvement compared to the earlier results with all clusters included. This improvement aligns with the hypothesis that better-defined tactical contexts enhance the system's predictive capabilities.

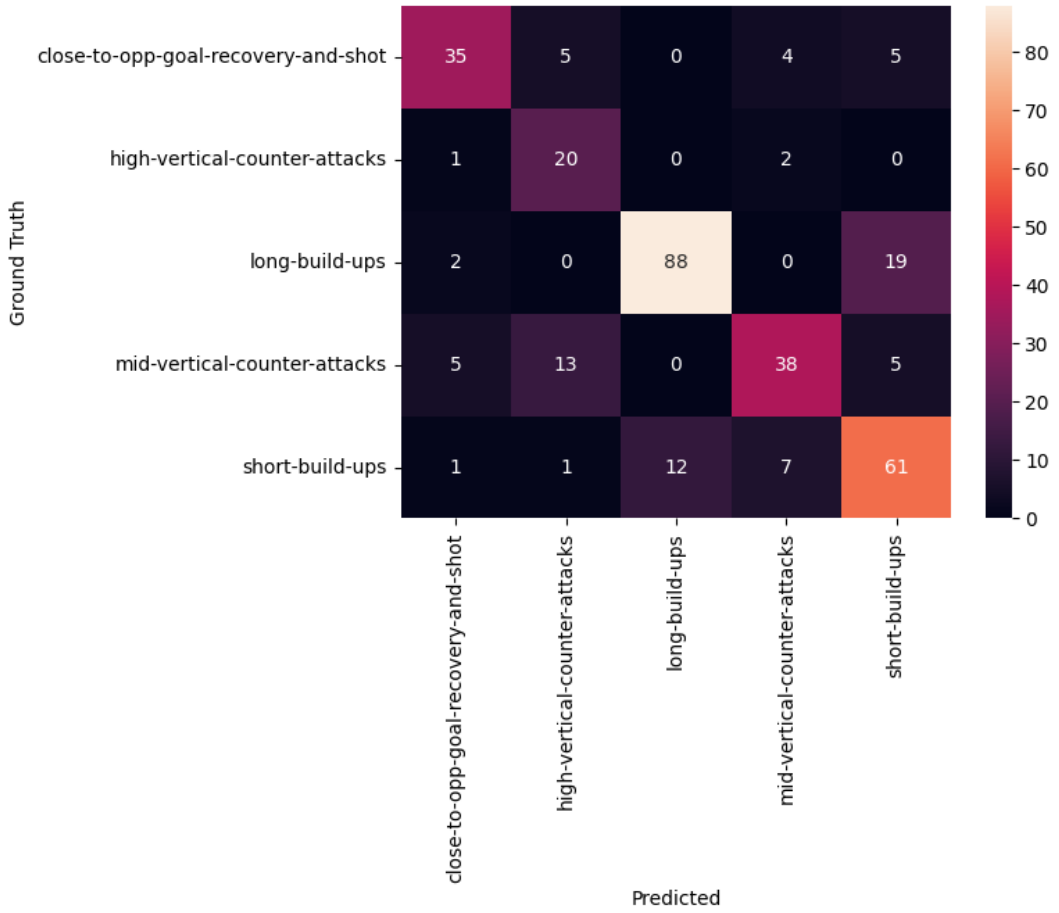

Figure 22. 5 clusters confusion matrix

Close-to-opp-goal-recovery-and-shot: With a precision of 0.80 and recall of 0.71, this sequence type shows robust performance. The F1-score of 0.75 indicates a strong balance between precision and recall, meaning that the model can effectively identify these sequences without significant false positives or negatives.

High-vertical-counter-attack: This category saw substantial improvement in recall (0.87), suggesting the model's ability to detect most instances of this sequence type. However, the precision remains moderate (0.51), indicating that some misclassifications persist, likely with other counterattacking-related clusters. The F1-score of 0.65 reflects these dynamics, showing good performance overall but with room for refinement in reducing false positives.

Long-build-up: This sequence type continues to perform exceptionally well, with a precision of 0.88 and recall of 0.81, resulting in the highest F1-score (0.84) among all clusters. The strong results for this cluster confirm that it is well-defined and clearly distinguishable in the tactical data.

Mid-vertical-counter-attack: With a precision of 0.75 and recall of 0.62, this sequence type shows improved performance compared to the earlier analysis. The F1-score of 0.68 highlights its relatively balanced performance, though some overlap with high-vertical-counter-attack sequences may still exist, as observed in the confusion matrix.

Short-build-up: Similar to long-build-ups, this sequence type performs well with precision (0.68) and recall (0.74). The F1-score of 0.71 suggests that the tactical fingerprint of this cluster remains well-defined and distinguishable in this refined analysis.

The significant improvement in overall accuracy and F1-scores across categories confirms that better-defined tactical contexts lead to enhanced predictive performance. The confusion matrix reveals reduced overlap between clusters, particularly between those that previously shared semantically similar characteristics (e.g., *opp-half-recovery-and-play* and *own-half-recovery-and-progression* in the earlier analysis).

Implications

These results underline the importance of clearly defined tactical contexts in improving the predictive power of models based on qualitative tactical data. This refined analysis demonstrates that Kognia's tactical data can effectively capture and represent the dynamics of well-defined football sequences, providing valuable insights for tactical analysis.

As part of the exploration of tactical behavior, an additional approach was attempted to classify teams within specific tactical contexts. The aim of this approach was to determine whether teams could be distinguished based on their unique use of tactics within the well-defined tactical clusters previously identified. This approach intended to leverage the team-specific tactical patterns to train a classifier capable of differentiating between teams within the same tactical context.

However, due to the lower variability in the use of tactics between teams within the same tactical context, combined with the limited amount of data available to train a classifier to differentiate teams based on these contexts, the results of this approach were extremely poor, achieving only a 2% accuracy rate. This outcome highlights the challenges of capturing subtle team-specific patterns within well-defined tactical contexts using the available dataset. As such, this limitation has been identified as an area for future research. Expanding the dataset to include more sequences and teams, as well as exploring additional features or alternative modeling approaches, could help address these challenges and improve the classifier's performance in distinguishing teams within tactical contexts.

4. Conclusions and future research

This study represents a step forward in understanding tactical behavior in football through the integration of data-driven clustering methods and qualitative tactical insights. By combining objective event and tracking data with Kognia's rich tactical annotations, the analysis provides a nuanced understanding of how sequences of play can be grouped into tactical contexts, how teams adapt within these contexts, and how these adaptations vary across teams.

The key findings of the study include:

1. **Clustering effectiveness**: The use of unsupervised learning to identify tactical contexts proved to be effective. The clusters obtained captured a wide range of different game contexts, from possession-oriented sequences to high-vertical counterattacks. The semantic validation of these clusters demonstrated that they provide meaningful insights into the game.

2. **Role of tactical data**: Kognia's tactical annotations showed great potential in predicting tactical contexts. However, the success of this prediction largely depended on the quality and granularity of the defined clusters. Contexts that were better defined from a footballing perspective yielded better classification results, highlighting the importance of semantic clarity in clustering.

3. **Team differentiation**: While tactical variability between teams is relatively small overall, the introduction of clusters as a context allowed for greater differentiation between teams, showing how tactics are adapted to specific contexts.

Implications and impact

This study introduces a robust framework that offers a valuable tool for coaching staff and analysts within the football industry. The framework enables teams to perform detailed comparative analyses of tactical behavior, both between teams and over time.

One of the most significant contributions of this work is the ability to answer questions about tactical tendencies in a structured, data-driven way. A coaching staff can use this framework to compare two teams within specific tactical contexts, identifying key differences in how they approach various game situations. This level of insight provides a clear picture of the strengths and

tendencies of different teams, helping to tailor match preparations and counterstrategies.

Additionally, the framework allows for internal evaluations within the same team over a defined time. It enables a coaching staff to track tactical evolution throughout the season, answering questions like: *At the beginning of the season, our team generated more shots through long-build-up sequences and possession-oriented tactics, whereas now we rely more on high-vertical counterattacks or opponent-half recoveries.* Such insights not only aid in identifying progress and areas for improvement but also help align tactical approaches with desired outcomes or match scenarios.

Furthermore, by visualizing how tactics evolve and adapt in different contexts, this framework empowers teams to assess the impact of strategic decisions, such as the integration of new players, shifts in formation, or adjustments in playing style. For example, a coaching team could evaluate whether a tactical adjustment made mid-season has had the intended effect of increasing efficiency in certain game phases or sequences.

Finally, this framework provides a replicable and transparent approach that bridges the gap between objective data and subjective football expertise. It offers a shared language and methodology for understanding and communicating tactical behavior, fostering collaboration between data analysts and football professionals. By doing so, it enhances the analytical capabilities of clubs, improves the quality of tactical insights, and ultimately contributes to more informed decision-making in the ever-evolving landscape of football analytics.

Limitations

Despite the promising results and practical implications of this study, there are several limitations that must be considered to contextualize the findings and guide future research efforts.

The analysis is based on data from a single season and 20 teams, which, while representative of a specific context, imposes constraints on the generalizability of the findings. After filtering and classifying possession sequences, some clusters contain relatively small numbers of examples, with just over 100 sequences in certain cases. This limited sample size requires caution when interpreting the results, as small datasets can affect the robustness and reliability of the conclusions.

Because the analysis is conducted on a specific league and set of teams, the findings may not generalize to other leagues, playing styles, or competitive contexts. Tactical behaviors could be influenced by factors unique to each league

or competition, so replication of this framework with other datasets is necessary to validate its broader applicability.

The study exclusively focuses on possession sequences that end in a shot, which, although strategically important, represent only a small and highly specialized subset of all possible possession sequences a team might execute. This narrow focus means that the conclusions drawn are specific to scoring opportunities and do not necessarily reflect broader tactical behaviors across all phases of play. Future research could expand the scope to include a wider variety of possession types to provide a more holistic understanding of team behavior.

The study employs all offensive tactics currently available from Kognia, without applying any filtering or selection criteria. While this approach ensures comprehensive coverage, it also introduces potential noise into the analysis, as some tactics may have low tactical relevance despite being labeled as such. The inclusion of such tactics might dilute the overall quality of the tactical fingerprints used for classification. Future work could involve curating a more selective set of tactics, prioritizing those with richer and more meaningful tactical content, or even designing new, contextually relevant tactics to enhance the framework's discriminative power.

Achievements assessment

The results of this study have been highly satisfactory, addressing key doubts that initially surrounded the research. One of the primary uncertainties was whether it would be possible to derive a meaningful and actionable characterization of possession sequences. Another concern was whether Kognia's offensive tactics would be sufficiently robust and informative to uncover and validate this characterization. Both doubts have been successfully dispelled. The clusters identified not only provide valuable tactical insights but also reveal distinct patterns of play that are reflective of real-world football dynamics.

At the outset of this study, alternative methodologies such as Hidden Markov Models (HMMs) were explored as potential approaches for analyzing possession sequences. However, they were ultimately discarded due to the inherent complexity of adapting these models to our specific dataset. The primary challenge lay in the non-sequential nature of our data, where tactics often occur simultaneously rather than in a strictly temporal sequence. Unlike traditional event-driven models, where one action follows another, the tactics in our dataset can overlap in time, making it difficult to apply HMMs effectively without significant modifications.

Ethical, sustainability, and diversity impact

The project effectively aligns with SDG 9 (Industry, Innovation, and Infrastructure), particularly target 9.4, which emphasizes upgrading infrastructure for sustainability. By selecting efficient machine learning models with lower computational demands, the project demonstrates a commitment to reducing the environmental footprint without compromising performance. This balanced approach highlights the importance of integrating environmental responsibility into innovative processes in sports analytics.

The reliance on Kognia's fully compliant data management reinforces the project's ethical foundation, ensuring transparency and accountability.

The current focus on men's soccer exposes an existing gap in the availability of comprehensive data for women's soccer, reflecting data inequality in sports analytics. This limitation underscores the need for future investment in data collection for women's sports. By addressing this, the project could contribute to SDG 10 (Reduced Inequalities), specifically target 10.2, which promotes inclusion across societal domains.

Future research

One of the main limitations of this study was the restricted dataset, which consisted of one season and 20 teams. Future research should aim to include data from multiple seasons, leagues, and competitions to ensure broader applicability and robustness of the framework. Expanding the dataset would also help address the issue of clusters with limited samples, providing a more balanced and reliable training environment for predictive models.

The current analysis focused solely on sequences ending in shots, which, while relevant, represents only a small portion of a team's overall possession patterns. Future studies could broaden the scope to include also those ending in turnovers or beginning with goal-kick actions. This would provide a more holistic view of team behavior and allow a deeper understanding of tactical diversity across different phases of play.

The tactics used in this study included all offensive tactics provided by Kognia without any filtering. While this comprehensive approach allowed for exploratory analysis, future research could focus on refining the selection of tactics by filtering out those with low tactical relevance or designing new tactics with richer semantic content. This refinement could enhance the discriminatory power of the models and improve the overall interpretability of the results.

The results obtained by using Kognia's tactics to predict tactical contexts showed promise, but there is room for improvement. Future work could explore the integration of complementary data types, such as player tracking, enhancing the contextual understanding of tactical behaviors.

While this study focused on football, the principles of tactical clustering and prediction could be adapted to other team sports, such as basketball or hockey, where tactical patterns and contextual behaviors are similarly critical.

5. Glossary

Acronym

EDA: Exploratory Data Analysis
CRISP-DM: Cross Industry Standard Process for Data Mining
GDPR: General Data Protection Regulation
CPU: Central Processing Unit
GPU: Graphics Processing Unit
DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Football terms

Game Style | Style of Play: Refers to the characteristic way in which a football team plays during matches. This can include the preference for possession of the ball, quick counterattacks, direct play, among others. The playing styles are influenced by the philosophy of the coach, the skills of the players and the general strategy of the team.

Patterns of Play | Tactical Patterns: Sequences of actions and movements in a football match that reflect the tactics and strategies used by a team. The set of playing patterns of a team define its game style.

Eventing Data: Football data that captures discrete events during a match (e.g., passes, shots, tackles), typically tagged with time, player, and location on the pitch.

Tracking Data: Continuous data that tracks player and ball movements in real-time, providing detailed spatiotemporal information (e.g., player positions, speed, and distance covered) throughout the game.

Tactic: Minimum unit on which a team's game model or style of play is based. The tactics can be individual, group or collective:

- **Individual Tactics:** Built on technical actions, such as a pass with specific intent.
- **Group Tactics**: Interaction of several players to achieve a common objective, such as keeping the defensive line high to cause an offside.
- **Collective Tactics:** Complex behavior of all the players on the team, such as pressing in a high block to prevent the opposing team from playing.

Finishing Sequence: Any possession that ends in a shot on goal. It involves a series of consecutive actions taken by the team that culminate in an attempt to score a goal.

Possession: Refers to the period during which a team maintains control of the ball. This includes passing, dribbling, and other actions until control of the ball is lost due to an interception, an error, or a shot on goal.

Technical terms of machine learning

Machine Learning: Branch of artificial intelligence that focuses on the development of algorithms that allow machines to learn from data and make predictions or take decisions based on them. It includes techniques such as supervised, unsupervised and reinforcement learning.

Feature Engineering: Process of selecting, modifying or creating variables from raw data to improve the performance of predictive models. It includes techniques such as normalization, codification and the creation of new characteristics based on combinations of existing variables.

Transformers: Deep learning architecture based on self-attention mechanisms, which has revolutionized the processing of sequences in natural language tasks and is being adapted for other applications, including the analysis of tactical sequences in sports. Transformers can capture long-term dependencies and complex relationships within the data.

Markov chains: Mathematical model that represents systems where the future state depends only on the current state and not on the previous states (property of Markov). In the context of football, they can be used to model transitions between different tactical game states.

6. References

[1] Barnard, T., Powell, T., & Kitchen, D. (2022). Data protection in sport – tackling GDPR issues. Irwin Mitchell. Available at https://www.irwinmitchell.com/news-and-insights/expert-comment/post/102hi9n/data-protection-in-sport-tackling-gdpr-issues

[2] Caicedo-Parada, S., Lago-Peñas, C., & Ortega-Toro, E. (2020). Passing Networks and Tactical Action in Football: A Systematic Review. Int. J. Environ. Res. Public Health, 17(18), 6649. https://doi.org/10.3390/ijerph17186649

[3] C3 AI. (n.d.). LIME: Local Interpretable Model-Agnostic Explanations. Available at https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

[4] Data Science Process Alliance. (n.d.). CRISP-DM 2.0. Available at https://www.datascience-pm.com/crisp-dm-2/

[5] Ghezzi, P., & Sotudeh, M. (2024). *Match phases in practice*. StatsBomb. https://statsbomb.com/wp-content/uploads/2024/10/Match-Phases-In-Practice-Ghezzi-and-Sotudeh.pdf

[6] Hassall, W. (2018). Guide To the GDPR For Sports Clubs - Privacy Protection - UK. Mondaq. Available at https://www.mondaq.com/uk/privacy-protection/731116/guide-to-the-gdpr-for-sports-clubs

[7] Martínez, G. (2018). The effect of GDPR in sports performance analysis. Sports Performance Analysis. Available at https://www.sportperformanceanalysis.com/article/2018/6/8/how-does-gdpr-affect-performance-analysis-in-sport

[8] Ötting, M., & Karlis, D. (2023). Football tracking data: a copula-based hidden Markov model for classification of tactics in football. Annals of Operations Research, 325(1), 167–183. https://doi.org/10.1007/s10479-022-04660-0

[9] Pate, A. (2019). *Hands-On Unsupervised Learning Using Python*. O'Reilly Media. https://learning.oreilly.com/library/view/hands-on-unsupervised-learning/9781492035633/copyright-page01.html

[10] Pei, Y., De Silva, V., Caine, M. (2024). Passing Heatmap Prediction Based on Transformer Model Using Tracking Data for Football Analytics. In: Bennour, A., Bouridane, A., Chaari, L. (eds) Intelligent Systems and Pattern Recognition. ISPR 2023. Communications in Computer and Information Science, vol 1940. Springer, Cham. https://doi.org/10.1007/978-3-031-46335-8_13

[11] Ul Islam, Q. (2023). Threshold silhouette score for cluster analysis. ResearchGate. https://www.researchgate.net/post/Threshold_silhouette_score_for_cluster_analysis

[12] Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite football: future challenges and opportunities for sports science. SpringerPlus, 5, 1410. https://doi.org/10.1186/s40064-016-3108-2

[13] Soares, M. (2019). Learning state representations and Markov models in football analytics. Master in Intelligent Interactive Systems, Universitat Pompeu Fabra.

[14] Stein, M., et al. (2015). Visual Football Analytics: Understanding the Characteristics of Collective Team Movement Based on Feature-Driven Analysis and Abstraction. ISPRS Int. J. Geo-Inf., 4(4), 2159-2184. https://doi.org/10.3390/ijgi4042159

[15] THE GLOBAL GOALS. (n.d.). Goal 9: Industry, innovation and infrastructure. The Global Goals. Retrieved September 18, 2024, from https://www.globalgoals.org/goals/9-industry-innovation-and-infrastructure/

[16] THE GLOBAL GOALS. (n.d.). Goal 10: Reduced inequalities. The Global Goals. Retrieved September 18, 2024, from https://www.globalgoals.org/goals/10-reduced-inequalities/

[17] VanderPlas, J. T. (2022). *Python Data Science Handbook*. O'Reilly Media.

[18] VanderPlas, J. T. (2020). *Math for the Non-Math Lovers: Chapter 8. Hypothesis Testing: Z and t Tests*. O'Reilly Media.

[19] Zader, N., et al. (2024). Engineering Features to Improve Pass Prediction in Football Simulation 2D Games. arXiv preprint. https://arxiv.org/abs/2401.03410

7. Appendices

7.1. Pitch zones and lanes

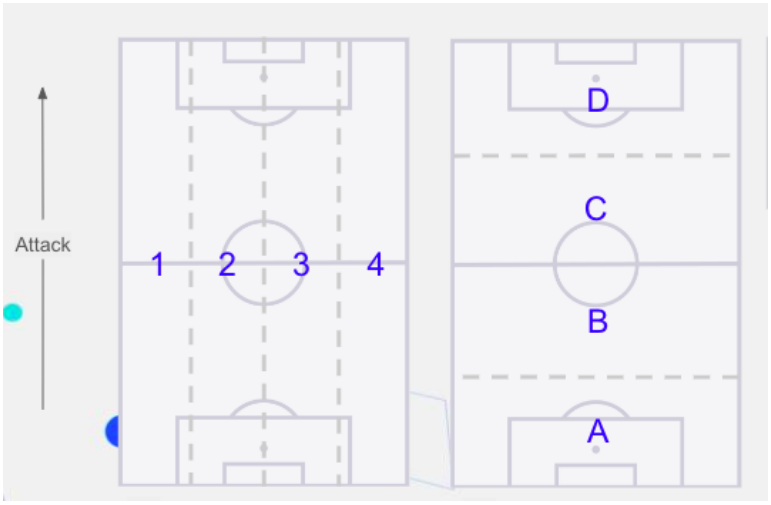

Figure 23. Pitch zones and lanes

Zones

Zone A: Closest to own goal (initiation zone).
Zone B: Progression in own half.
Zone C: Progression in opponent's half.
Zone D: Closest to opponent's goal (finishing zone).

These zones are not only based on distance, but also reflect the team's tactical approach in each phase of the game:

Zone A is used to initiate plays from defensive areas.
Zone B and Zone C cover transition and progression in own half and opponent's half, respectively.
Zone D is associated with the finishing phase of plays, where goal-scoring opportunities are attempted.

Lanes

Wide lanes: Zones 1 and 4, which correspond to the left and right flanks of the field. These zones are associated with plays down the wing, crosses and quick transitions on the outer lines of the field.

Central lanes: Zones 2 and 3, which cover the central part of the field. These zones are frequently used for quick vertical transitions, attacking combinations and possession plays in the center of the field.

7.2. Kognia tactics

- Long-ball: Tactic that illustrates a play that is developed using a long pass to progress vertically.
- Realized-emergency-support: Tactic attributed to a player who receives a pass from a teammate who is under pressure.
- Finishing: Tactic attributed to the player performing any type of action that finishes a play (except striking a cross).
- Finshing-after-cross: Tactic attributed to the player that finishes a play by striking a cross inside one of the 3 striking zones in the box.
- Finishing-pass: Tactic attributed to a player who makes a pass that at the moment of reception puts his teammate in a position ready to finish on goal.
- Realized-finishing-suppport: Tactic that shows the movement of a pass receiver to a pass that originates in front of the defensive line and is received behind the defensive line.
- Realized-horizontal-overcoming-support: Tactic that illustrates the movement of a player who receives a pass that horizontally overcomes at least one player from the opposing team.
- Passing-between-lines: This tactic is attributed to the player who passes the ball to a teammate in a free space between lines within an interior area delimited by several opponents around him.
- Identifying-lines-under-pressure: Tactic that shows the ball possessor being under pressure able to maintain possession of the ball by finding a pass to a teammate.
- Moving-behind-the-defensive-line: Tactic that describes the movements of offensive players without the ball who make runs starting in front of the defensive line and ending behind the defensive line.
- Occupying-space-in-the-box: Tactic attributed to offensive players without the ball who are inside the penalty area and are unmarked, thus able to receive and finish the play
- Overcoming-opponents-with-vertical-passes: The tactic that shows the ball possessor executing a pass to a teammate who is in more vertically advanced position and therefore overcomes opponents, generating progression towards the opposition goal.
- Possession-after-recovery: Common team behavior to keep ball in the immediate space that was recovered. The objective is to keep possession with the intention of organizing attack in a later period.
- Progression-after-recovery: Common team behavior to move fast forward from the area that the ball was recovered through the rest of the field using short/mid-range passes with the intention of finishing.

- Space-between-defensive-line-and-halfway-line: Tactic attributed to the rearmost line of the attacking team, illustrating their movement forward towards their attacking goal, reducing distance in between them and their teammates.
- Realized-striker-support: Tactic that illustrates the behavior of teammates giving support to a striker after he receives the ball.
- Switch-of-play: Tactic that shows a pass that crosses at least one of the four channels with the intention of changing the origin of the attack from one side of the field to the other.
- Taking-advantage-of-defensive-line-imbalances: Tactic attributed to offensive players without the ball who are positioned inside the imbalance area of the defensive line.
- Realizedd-vertical-overcoming-support: Tactic that shows a player who receives a pass that vertically overcome at least one player from the opposing team.
- Recovered-ball: Action that shows the moment in which ball possession is recovered by a team.

7.3. Identifying playing styles

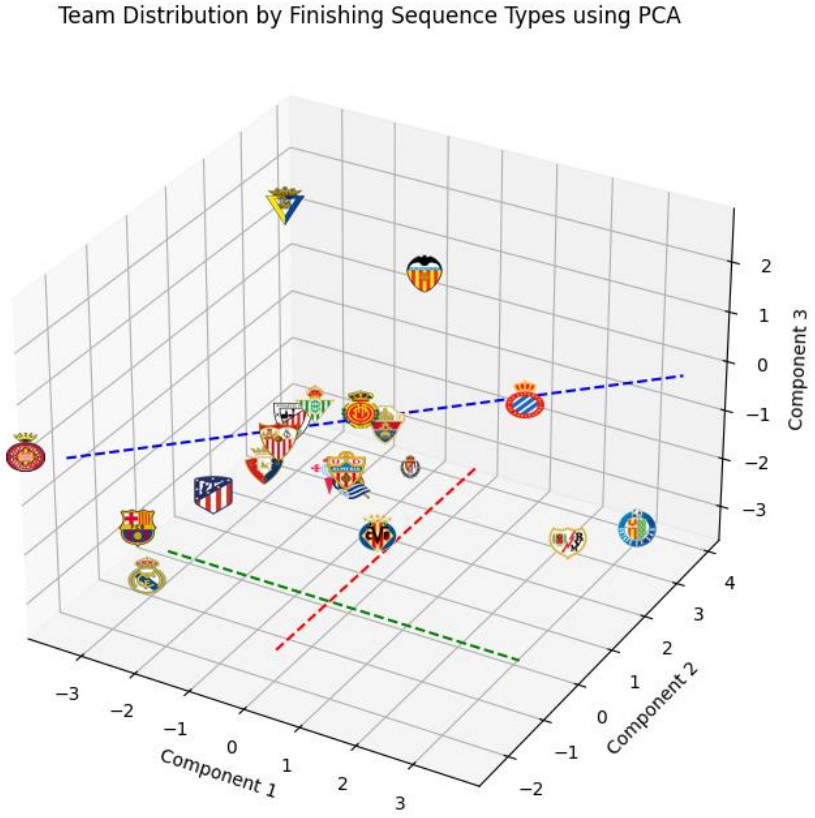

Figure 24. Team distribution by finishing sequence types using PCA (3D)

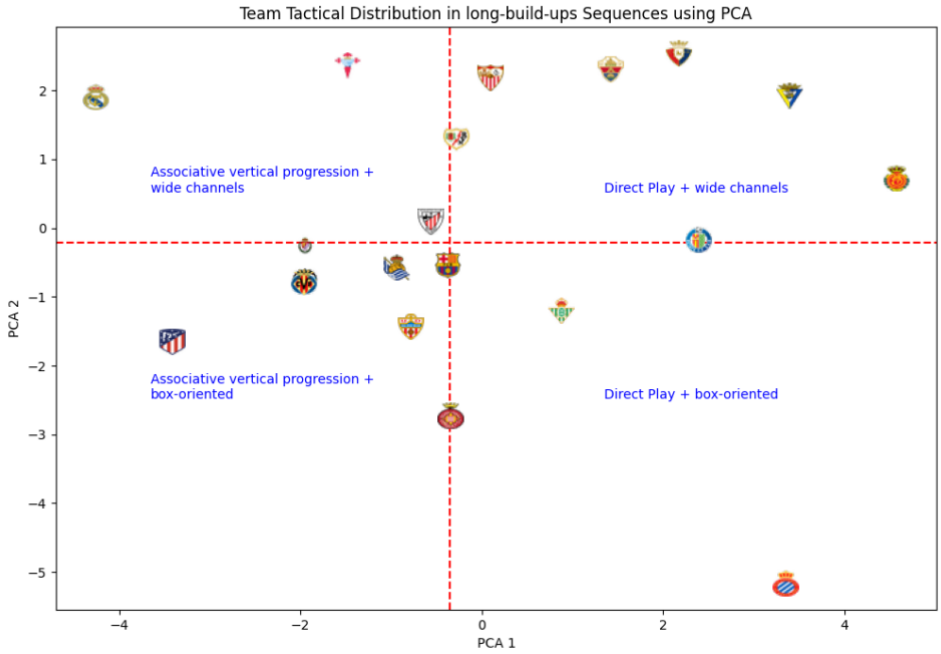

Figure 25. Team distribution by tactic's usage during long-build-up sequences

Tactic	Comp 1	Comp 2
balance-of-the-team-after-recovery	0.301712	0.126469
cross-into-the-box	0.248418	-0.184033
finishing	0.261697	-0.21766
finishing-after-cross	-0.022901	-0.264364
finishing-pass	0.181033	-0.2818
goal-chance	0.157974	0.086828
identifying-passing-lines-under-pressure	-0.248484	-0.083473
long-ball	0.268576	-0.097602
moving-behind-the-defensive-line	0.355362	-0.119059
occupying-space-in-the-box	-0.011178	-0.295123
overcoming-opponents-with-vertical-passes	-0.318586	-0.131113
passing-between-lines	-0.082119	0.056275
possession-after-recovery	0.142831	0.258306
progression-after-recovery	-0.03596	-0.182923
realized-emergency-support	-0.239322	-0.09539
realized-finishing-support	0.193071	-0.301061
realized-horizontal-overcoming-support	0.013331	0.265602
realized-striker-support	-0.122107	-0.038017
realized-vertical-overcoming-support	-0.316618	-0.112289
receiving-between-lines	-0.082119	0.056275
recovered-ball	0.210298	0.15949
space-between-defensive-line-and-halfway-line	0.158051	0.143787
switch-of-play	0.030579	0.21165
taking-advantage-of-defensive-line-imbalances	0.208183	-0.018166
width-of-the-team	-0.013712	0.275204
width-of-the-team-opposite-channel	0.0016	0.391329

Table 7. Long build-ups tactics eigenvectors