City Research Online

City, University of London Institutional Repository

Citation: Moussavi, L., Andrienko, G., Andrienko, N. & Slingsby, A. (2025). Visually-supported topic modeling for understanding behavioral patterns from spatio–temporal events. Computers & Graphics, doi: 10.1016/j.cag.2025.104245

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/35224/

Link to published version: https://doi.org/10.1016/j.cag.2025.104245

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

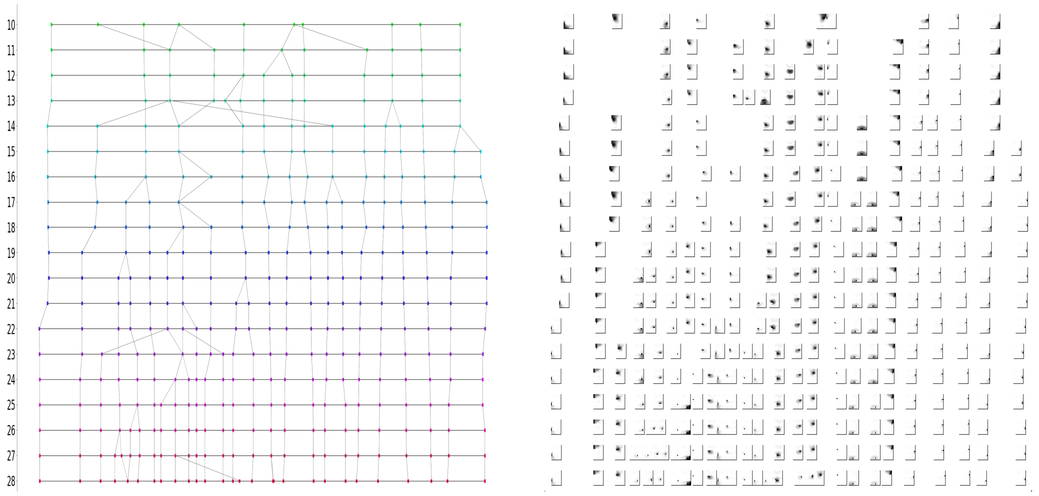
Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Graphical Abstract

Visually-supported Topic Modeling for Understanding Behavioral Patterns from Spatio-temporal Events

Laleh Moussavi, Gennady Andrienko, Natalia Andrienko, Aidan Slingsby



1D projections of topics obtained from topic modeling are shown for topic numbers ranging from 10 to 28. Each row represents the projection for a specific topic number. The topics are used to identify behavioral patterns.

Left: A visual analytics technique for selecting the number of topics with the most suitable result from multiple runs of topic modeling on spatial event sequences.

Right: A demonstration of how our technique can be used to discover spatially consistent and easily distinguishable topics from a football dataset. Topics are replaced by their spatial footprint.

Highlights

Visually-supported Topic Modeling for Understanding Behavioral Patterns from Spatio-temporal Events

Laleh Moussavi, Gennady Andrienko, Natalia Andrienko, Aidan Slingsby

- An end-to-end workflow for transforming raw spatio-temporal events into meaningful insights.
- A visual analytics technique for selecting the number of topics with the most suitable result from multiple runs of topic modeling on spatial event sequences.
- A demonstration of how our technique can be used to discover spatially consistent and easily distinguishable topics from a football dataset.
- A demonstration of how the obtained topics can be used to summarize a football game's dynamics and reveal insights into the games.
- Methods for using the topics to summarize and compare various matches and teams' playing styles.
- An exploration of how different teams deviate from their typical play style in different conditions.

Visually-supported Topic Modeling for Understanding Behavioral Patterns from Spatio-temporal Events

Laleh Moussavi[a], Gennady Andrienko[a,b], Natalia Andrienko[a,b], Aidan Slingsby[a]

[a] *Department of Computer Science; City, University of London, College Building, 280 St John St, London, EC1V 4PB, United Kingdom*
[b] *Fraunhofer Institute IAIS, Schloss Birlinghoven, Sankt Augustin, 53757, Germany*

Abstract

Spatio-temporal event sequences consist of activities or occurrences involving various interconnected elements in space and time. We show how topic modeling — typically used in text analysis — can be adapted to abstract and conceptualize such data. We propose an overall analytical workflow that combines computational and visual analytics methods to support some tasks, enabling the transformation of raw event data into meaningful insights. We apply our workflow to football matches as an example of important yet under-explored spatio-temporal event data. A key step in topic modeling is determining the appropriate number of topics; to address this, we introduce a visual method that organizes multiple modeling runs into a similarity-based layout, helping analysts identify patterns that balance interpretability and granularity.

We demonstrate how our workflow, which integrates visual analytics, supports five core analysis tasks: identifying common behavioral patterns, tracking their distribution across individuals or groups, observing progression at different temporal scales, comparing behavior under varied conditions, and detecting deviations from typical behavior.

Using real-world football data, we illustrate how our end-to-end process enables deeper insights into both tactical details and broader trends — from single match analyses to season wide perspectives. While our case study focuses on football, the proposed workflow is domain-agnostic and can be readily applied to other spatio-temporal event datasets, offering a flexible foundation for extracting and interpreting complex behavioral patterns.

Keywords: Visual Analytics, Topic Modeling, Spatio-temporal Events

1. Introduction

Spatio-temporal event sequences are chains of activities or occurrences that take place in different locations and at different times, with each event influencing or connecting to others. Because they span both space and time, these sequences can be especially complex to analyze and challenging to extract meaningful insights. Gaining insight into these types of data can help in controlling patterns or forecasting future trends based on historical data. However, spatio-temporal event sequences often contain a wealth of detailed, short-term patterns that may not be representative of the overall dataset. Extracting meaningful patterns that capture the essence of these sequences requires advanced analytical techniques.

Analyzing spatio-temporal events has extensive applications across various fields [6, 16]. For example, in environmental conservation and management, this research can unveil patterns of environmental change, such as deforestation, urban expansion, or shifts in land use [5]. In transportation and urban planning [2], spatio-temporal analysis can help understand traffic flows, optimize public transport routes, and plan urban infrastructure more effectively. It is also important in the domain of public health for tracking and analyzing epidemiological trends, helping to predict disease spread and manage health crises such as epidemics and pandemics. These diverse applications show the importance of research on spatio-temporal event sequences [8, 5].

Research into spatio-temporal data has grown significantly with the increased availability of geo-referenced and temporal datasets [25]. Many research studies have prominently used visual analytics to explore spatio-temporal data. Visual analytics leverages the collaboration between human intuition and creativity and the data-processing capabilities of computers [30, 5]. This field is notably vibrant in the study of spatio-temporal events, combining sophisticated analytical techniques with interactive visualizations to enhance the understanding of patterns, trends, and anomalies across both space and time. For example, Andrienko and Andrienko [4], as well as Andrienko *et al.* [3], provide systematic approaches to the exploratory analysis of spatial and temporal data, offering methodologies that include visual analytics to help uncover hidden patterns. In addition, Krüger *et al.* [14] introduce a technique designed to filter and explore long-term trajectory data using visual analytics, highlighting the potential for interactive exploration of spatio-temporal datasets.

Topic modeling [26] is an unsupervised learning method, originally developed for text mining, that facilitates moving from the granular level of individual words and documents to a more abstract understanding of the themes underlying large collections of text. This is achieved by transforming complex textual data into manageable numerical representations. The process begins with the creation of a document-term matrix, where each document is represented by word frequencies, disregarding word order. The model then analyzes patterns of word co-occurrence across the corpus to identify clusters of terms that frequently appear together, distilling the data into a set of topics.

These topics are essentially groups of related terms that provide a simplified, numerical summary of the text's content. Abstraction becomes meaningful when these groups are interpreted as themes or ideas that encapsulate the essence of the documents. By examining the most representative words within each topic, one can label and define the topics in a way that reflects the underlying themes in the corpus. In this way, topic modeling bridges the gap between raw textual data and higher-level insights, offering a structured approach to uncovering meaningful patterns and ideas.

While topic modeling techniques were originally developed in text mining to uncover hidden semantics within textual data, their underlying principles are generalizable and can be applied to other domains. This adaptability has enabled their use in various applications, including image data [28] and bioinformatics [13]. More recently, there have been efforts to leverage topic modeling for analyzing event sequences and spatio-temporal data. For instance, Chen *et al.* [11] examined sequences of user actions during sessions to understand behaviors in security management systems. They also applied their methodology to spatio-temporal visiting events in an amusement park, providing insights into participants' visiting behaviors. Andrienko *et al.* [2] explored road traffic movement data to reveal common patterns of space utilization.

Despite these promising applications, the use of topic modeling for spatio-temporal event sequences remains a relatively unexplored and evolving research area. Topic modeling is a process with many design choices, such as defining terms and documents, selecting the number of topics, and interpreting the obtained topics in the context of the data. These choices can significantly affect the outcome. Incorporating visual analytics techniques with topic modeling process enhances these design decisions. Specifically, visual analytics can be utilized not only in the interpretation of the topics but also during the earlier stages of model selection and data pre-processing, offering a more comprehensive understanding of the data.

In this paper, we propose an end-to-end workflow that combines computational methods (including topic modeling) with visual analytics to extract meaningful insights from raw spatio-temporal event sequences. We demonstrate the workflow using spatially-referenced event data from football matches [23]. This dataset is rich with various complex behaviors exhibited by teams and players, along with their potential influence on match outcomes - which may be possible to capture and model as topics.

Specifically, our analytical workflow consists of three main components:

1. **Data transformations**, which convert raw spatio-temporal events into structured segments (e.g., episodes of ball possession; Section 3) that serve as "documents" for topic modeling.

2. **Computational methods**, primarily topic modeling supported by dimensionality reduction to select the optimal number of topics (Sections 3, 4, and 5).

3. **Visual analytics**, which employs static and interactive visualizations to support exploration, interpretation, and comparison of the extracted patterns (Subsections 5.1, 7.1, 8.1, 8.2, 8.3, and 8.4).

This end-to-end workflow is modular and domain-independent, serving as a set of building blocks that transform raw data into insight. Each block can be independently fine-tuned to match the characteristics of a particular dataset, enabling both generalizability and adaptability.

Using the proposed workflow, we aim to support the following generic analysis tasks:

- **T1:** Identify common behavioral patterns using the representative units of behavior (e.g., zones most frequently activated within topics, as shown in Section 5.1).

- **T2:** Track how the occurrences of behavioral patterns are distributed across agents or groups (e.g., different teams, leagues, or over time).

- **T3:** Observe the progression of behavioral patterns at different temporal scales (e.g., shifts across halves, mid-season transitions, or coaching changes).
- **T4:** Compare how contextual conditions (e.g., playing at home vs. away, match outcomes) influence the occurrence of behavioral patterns.
- **T5:** Detect and quantify deviations from a typical behavior, helping to identify anomalies or shifts in behavior.

Throughout the paper, we demonstrate how the proposed workflow supports each of these tasks using real-world football data, providing both detailed and overall insight. Our approach uses topics to represent behavioral patterns.

One of the main challenges in topic modeling is determining the optimal number of topics to extract, as too few topics can oversimplify the data, while too many can fragment coherent themes. To address this, we propose a visual analytics technique (Section 5) that iteratively applies topic modeling across a range of topic numbers and visualizes the results. By arranging topics in a hierarchical graph based on their similarities and representing them with spatial heatmaps, our approach enables the identification of stable and significant topics, which aids in selecting a compact and representative set.

While our proposed visual analytics technique can work for any spatial event data (as long as we can represent individual topics by interpretable compact images such as spatial heatmaps or glyphs), we demonstrate our approach by applying this technique to a football dataset [23]. We begin with selecting a representative set of topics (Section 5) and then use the selected topics to summarize game dynamics and provide insights into individual games or series of games (Section 6).

This paper extends our earlier conference paper [20] by adding a suite of visualizations featuring five distinct visualization strategies that follow the same design principles and rely on topics derived from our modeling to analyze football games at varying levels of granularity. The first offers a detailed exploration of a single match, while the other four provide concise match summaries and reveal deviations from the "average" behaviors, facilitating comparisons across multiple games and teams.

In summary, our contributions are A) an overall analytical workflow to progress from raw event data to meaningful analytical insights; B) a visual analytics technique for selecting the number of topics with the most suitable result from multiple runs of topic modeling on spatial event sequences; C) a demonstration of how our technique can be used to discover spatially consistent and easily distinguishable topics from a football dataset; D) a demonstration of how the obtained topics can be used to summarize a football game dynamics and reveal insights into the games; E) methods for using the topics to summarize and compare various matches and teams' playing styles; and F) an exploration of how different teams deviate from their typical play style in different conditions.

High-resolution versions of all figures presented in this paper, along with additional examples, are available on the accompanying webpage: `https://lalehmoussavi.github.io/topic-modeling-for-behavioral-patterns/`.

2. Related Work

Topic Modeling.

Topic modeling is a powerful technique used to identify hidden themes from a collection of documents [26]. It is a tool commonly employed in text mining to discover the concealed semantics within text data. Topic modeling aims to discover combinations of co-occurring words that represent topics, allowing for the categorization of documents based on these topics. For example, in a collection of academic papers spanning subjects such as history, biology, and mathematics, topic modeling can identify distinct sets of terms commonly associated with each subject area, such as “empire”, “medieval”, and “revolution” for history; “DNA”, “evolution”, and “species” for biology; and “equation”, “calculus”, and “theorem” for mathematics.

Two popular methods used in topic modeling are Latent Dirichlet Allocation (LDA) [9] and Non-negative Matrix Factorization (NMF) [17]. They have their distinct mathematical foundations but share the same goal, use the same inputs, and produce similar outputs.

El-Assady *et al.* [12] proposed a visual analytics framework for performing and optimizing topic modeling on text data. It enhances the traditional topic modeling process by incorporating user interactivity and speculative execution, allowing for a more flexible and interpretable approach to text data analysis. In contrast to their approach which focuses on text data, our visual analytics approach is designed for spatio-temporal event sequences.

While topic modeling techniques were originally developed for text data, the underlying principles can be adapted to other domains. This adaptability has allowed topic modeling to be embraced for a broad spectrum of applications such as image data [28] and bioinformatics [13]. Topic modeling on spatio-temporal event sequences involves applying algorithms to uncover hidden topics within a dataset, which helps to understand how these topics/behaviors change across space and time. By analyzing the frequency of topics and their co-occurrence with other situations in various locations and over time, topic modeling can reveal insights into how certain activities or occurrences are distributed and evolve over space and time and in different situations.

Extensions and adaptations of topic modeling for spatial, temporal, or spatio-temporal data have been proposed in various studies. Chen *et al.* [11] applied a topic modeling approach to analyze non-spatial event sequences. They defined behavior as a sequence of actions an agent performs over time and aimed to uncover latent topics that represent distinct categories of behaviors using LDA. Their approach was demonstrated through two case studies: one analyzing operational behaviors in a real-world security management system, and the other examining visitor behaviors in an amusement park using a data set of IEEE VAST Challenge 2015 [29]. To address LDA's sensitivity to the number of topics, the authors introduced the concept of LDA ensembles, which involve running multiple LDA models with varying topic counts and projecting them to a 2D space.

Their model is performed on non-spatial event sequences, while we perform topic modeling on spatio-temporal terms and documents and visualize the results both spatially and through time. Additionally, one of their datasets, derived from the VAST Challenge, is synthetic, which may not fully capture real-world behaviors. In contrast, we have worked with the football dataset from Pappalardo *et al.* [22], which is based on real-world data. Although working with real-world data introduces additional complexity, it will offer more meaningful insights. While they used a 2D projection visualization to determine the number of topics, our current approach employs a different method based on 1D projection. We show that 1D projections make it easier to compare and analyze across different number of topics (Section 5).

Andrienko *et al.* [2] employed topic modeling to analyze road traffic trajectories, representing movement as sequences of place visits and transitions between them. Their work focused on uncovering patterns of space utilization and movement behavior but did not delve into interpreting the semantic meaning of the extracted topics. Our research differs significantly in data, objectives, and methodology. While their study analyzed continuous vehicle trajectories, we focus on discrete football events within matches, as detailed in Section 3. Crucially, our work emphasizes topic interpretation through visualization, which was not explored in their approach. Additionally, we introduce a novel 1D trapezium-like visualization for determining the optimal number of topics, in contrast to their 2D approach.

Overall, our study extends the prior work by addressing a different type of spatio-temporal data and incorporating methods that enhance both interpretability and topic selection.

Topic Modeling on Football Data.

The following works have used topic models on football data. Wang *et al.* [27] proposed the Team Tactic Topic Model (T3M) to analyze football data. T3M employs a generative process that considers players' positions and passing relationships to identify team strategies (topics). The model assumes that players' positions follow a Gaussian distribution within each identified topic. These Gaussian distributions ideally cover all different parts of the pitch. We note that this could be sometimes a wrong assumption since a forward like Messi might not be involved in any pass for a defending pattern.

Our work addresses multiple issues in [27] concerning data usage (considered events), episode definition, and methodology. First, while they only focus on passes and their receivers' positions, we include all event types (e.g., shots, fouls, tackles) and consider both the sender's and receiver's positions (Section 3). Second, their approach defines episodes based on stoppages (when the game is paused) and turnovers (when the other team gets the ball), whereas we define them more continuously tolerating short interruptions (Section 3).

Third, our approach is more scalable and allows a compact representation of the resulting topics. Fourth, while they did not employ a specific method for selecting the number of topics, as a core part of our work on Football data, we propose a visual analytics methodology to determine the optimal number of topics (Section 5). Finally, our main focus is not on the development of a topic model as a final product, but on using visual analytics at different stages of the topic modeling process to enhance the overall analysis process.

Andrienko *et al.* [7] introduced an approach for analyzing multivariate temporal data using topic modeling techniques. They represented variations in each attribute's value within an episode as 'words', combining these variations to create 'texts' for each episode and then applying topic modeling to these texts to uncover patterns and recurring themes. The authors demonstrated their methodology by analyzing two football matches from the German Bundesliga, and a mobility dataset for different countries.

Our work differs from [7] on used data, the ways to define terms, and overall methodology. First, the data used in the two studies are different types of movement data: continuous ball and player trajectories in [7], whereas we focus on discrete events during a match. Ball and player trajectory data are rarely accessible for scientific research, as they are expensive to obtain, hold significant commercial value, and is often considered proprietary. However, match event data, which describe on-ball actions during a football match, can be collected from broadcast footage through manual processing [24]. Consequently, event data are becoming available for an increasing number of competitions. Second, our dataset is significantly larger, encompassing 1,941 games from 7 different leagues and tournaments. This allows us to identify common or universal patterns across a wider range of teams and playing styles. Third, we define terms using an alphabet of grid cells on the pitch, while they used specific attributes, such as "Pressure on the ball" and "Pressure on attackers", to represent match situations, which were then discretized into bins. While their visual representations are limited to analyzing the temporal evolution of topics, our approach incorporates both spatial (visualizing the resulting topics, as discussed in Section 5.1) and temporal (tracking topic progression over time, detailed in Section 6) dimensions. Fourth, we extend their approach by introducing a visual analytics technique (Section 5) to determine the most suitable number of topics as a core part of our approach. Finally, in contrast to their proposed workflow for employing topic modeling, our aim is to provide a general framework for utilizing visual analytics at various stages of the topic modeling process.

Selecting the Number of Topics for Topic Modeling on Spatio-Temporal Event Sequences.

Chen *et al.* [11] and Andrienko *et al.* [2] employed 2D embeddings to determine the optimal number of topics by visually identifying clusters of similar topics across multiple modeling runs. In contrast, our approach uses 1D embeddings, which provide several advantages. A 1D representation simplifies interpretation by mapping each data point to a single axis, making it easier to match and compare topics across multiple runs. Additionally, this method enables more detailed and focused visualizations of topics' footprints (Figure 2).

3. Extracting Episodes from Football Dataset

This section addresses the first step of our analytical workflow (data transformations) by detailing how raw event logs are converted into episodes of ball movement.

The football dataset used in our paper [22] contains events from the 2017/2018 season of five top-tier European football leagues (Spain, Italy, England, Germany, and France), and two major international tournaments, namely the 2018 World Cup and the 2016 European Championship. Overall, this collection has information on seven prominent football competitions.

The dataset covers details on events, teams, matches, players, referees, and coaches [22], including 142 teams, $1,941$ matches with a total of $3,719,995$ recorded events ($1,917$ events per game on average). The events in the dataset include the following actions or happenings: *duel*, *foul*, *free kick*, *goalkeeper leaving line*, *interruption*, *others on the ball*, *pass*, *save attempt*, *shot*, and *offside*. These events are detailed with information about their sub-type (for example a *pass* could be a *hand pass*, a *cross*, etc.), timestamps, the involved player(s), and the positions on the field from where an event is made (and received). Previously, this dataset underwent statistical analysis to assess players' performance metrics [21].

We apply topic modeling to episodes of ball movement that contain sequences of events. Using episodes rather than single events makes it possible to extract meaningful information from the interaction between multiple players on different parts of the football pitch and during a period. We preprocess the dataset to extract episodes in a way similar to previous works [1, 7]. We define each episode as the duration during which a team possesses the ball until it loses possession. We ignore brief interruptions when the opposing team briefly gains possession for less than 10 seconds. For instance, if team A is moving the ball through a series of events and team B momentarily disrupts possession with a touch, team A's episode will still continue to incorporate subsequent events if the interruption lasts no more than 10 seconds. An episode for team B will be recorded from the moment of their interruption only if they manage to keep the possession for at least 10 seconds. With this pre-processing, the average number of episodes for each match is 316 (158 per team), and the average number of events per episode is 6.07.

Table 1 presents an example episode extracted from football data, showcasing a sequence of events involving players from Barcelona and Real Madrid match. The episode consists of six events, including actions by Barcelona players and also by Real Madrid players.

#	Time	Event	Player, Role	Team	X1	Y1	X2	Y2
0	2H 06:45	Duel - Ground defending duel	Suárez, FW	Barcelona	73.50	58.48	82.95	48.96
1	2H 06:47	Pass - Simple pass	Suárez, FW	Barcelona	82.95	48.96	90.30	34.68
2	2H 06:50	Duel - Ground attacking duel	Messi, FW	Barcelona	90.30	34.68	91.35	37.40
3	2H 06:50	Duel - Ground defending duel	Casimiro, MD	Real Madrid	90.30	34.68	91.35	37.40
4	2H 06:51	Shot	Messi, FW	Barcelona	91.35	37.40	0.00	68.00
5	2H 06:53	Save attempt - Reflexes	Navas, GK	Real Madrid	0.00	68.00	91.35	37.40

Table 1: An example episode extracted from football data, with events from the second half (denoted as 2H) of a Barcelona and Real Madrid match. Rows 0, 1, 2, and 4 represent actions by Barcelona players, while rows 3 and 5 involve Real Madrid players. The columns include the Time (half, minutes, and seconds), Event type (including event and sub-event), Player name and Role, Team name, and the starting (X1, Y1) and ending (X2, Y2) coordinates of each event.

4. Topic Modeling on all Matches of the Football Dataset

In this section and the next (Section 5), we discuss the second step of our workflow, i.e., the computational methods applied to the football dataset.

Topic modeling employs matrices to represent the relationships between documents, terms, and topics. The input to topic modeling is a document-term matrix X storing the distribution of terms (columns) across documents (rows). When applying NMF to the document-term matrix, it decomposes the matrix into two lower-dimensional matrices, W (document-topic matrix) and H (topic-term Matrix). Each element in the W matrix represents the weight of a topic within a document, while each element in the H matrix represents the weight of a term within a topic. The weights quantify the relevance of the topics and terms, respectively. Vayansky and Kumar [26] suggest that NMF is more suitable for short texts such as social media messages. Similar findings were reported by Andrienko *et al.* [7] when applying NMF to game episodes characterized by multivariate time series. These observations motivated us to utilize NMF in our study.

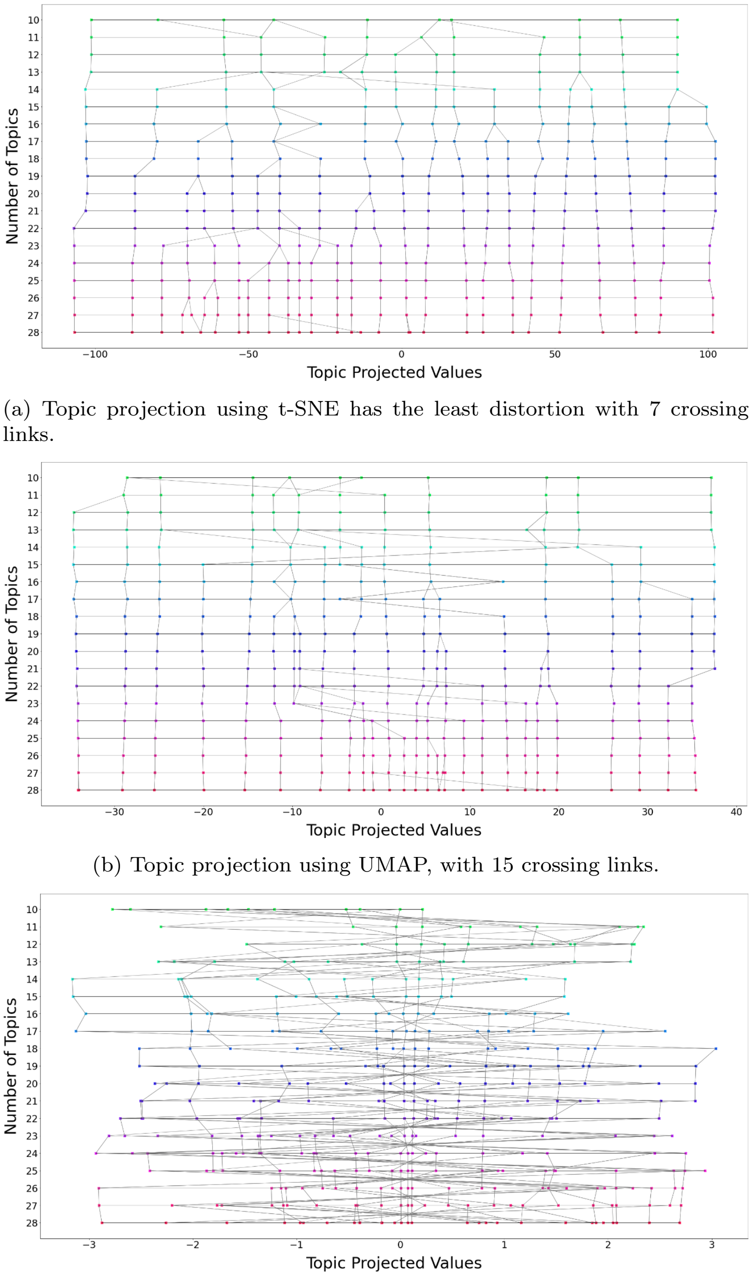
Next, we describe how we applied topic modeling to the matches in Pappalardo *et al.*'s football dataset [22]. In our approach, we define our documents as the episodes of ball possessions by teams during matches (Section 3). For each team, we consider a pitch to be oriented upwards in the direction of the team attack. We divide the pitch into grid cells with 15 rows and 12 columns, consisting of 180 grid cells in total. These grid cells will be treated as terms. For each team in a match, we generate a matrix $X_{M\times N}$ that represents the ball positions when controlled by that team, where M is the number of episodes (documents), N is the number of grid cells (terms), and each element X_{ij} equals the number of times that in episode i the ball-related event was recorded within grid cell j.

Our objective is to identify common patterns across all teams and matches, enabling us to compare different teams' behaviors during a match or an entire season. Applying topic modeling separately to each team would produce disparate sets of discovered patterns for different teams, making it challenging to compare the behaviors of the teams. Therefore, we create an integrated matrix including all episodes from all matches.

We apply NMF to this combined matrix and thus, our approach finds a document-topic matrix $\mathbf{W}_{\mathbf{M}\times K}$ and a topic-term matrix $\mathbf{H}_{K\times N}$, where $\mathbf{X} \approx \mathbf{W} \times \mathbf{H}$. These matrices are supposed to reflect common tactical patterns across all teams in all leagues.

5. Visual Analytics Technique for Selecting the Number of Topics

In this section, we introduce a visual analytics technique designed to identify the number of topics with the most suitable result from multiple runs of topic modeling on spatial event sequences. The objective is to determine the smallest set of topics that effectively captures all significant information, ensuring that the topics are clearly interpretable, distinct, and free of redundancy. Consistent topics across different runs are prioritized, while overlapping patterns should be consolidated into single topics for clarity. Our approach is adaptable to any dataset where topics can be represented as compact, interpretable images or glyphs. For demonstration purposes, we apply this technique to a football dataset.



(a) Topic projection using t-SNE has the least distortion with 7 crossing links.

(b) Topic projection using UMAP, with 15 crossing links.

(c) Topic projection using MDS with many crossing links.

Figure 1: 1D projections of topics obtained from topic modeling are shown for topic numbers ranging from 10 to 28. Each row represents the projection for a specific topic number. The dimensionality reduction techniques applied in these projections are: (a) t-SNE, (b) UMAP, and (c) MDS. Each topic is connected to its most similar topics in the projections above and below.

We iteratively apply Non-negative Matrix Factorization (NMF) to the data, varying the number of topics K over the range $\{K_{\min}, \ldots, K_{\max}\}$. For each iteration with a given K, the process yields a topic-term matrix $\mathbf{H}^{K}_{N \times K}$, where each row contains the term weights for one topic.

Previous work has applied dimensionality reduction techniques to project the topic-term matrices into a 2D space [2]. While this is useful in comparing topics obtained from different values of K, we argue that our approach enables a deeper analysis by projecting into a more compact 1D space allowing a hierarchical arrangement of topics across multiple runs.

For all values of K, we project the topics into a shared 1D space. This allows us to represent each topic solution as a line with K points corresponding to the positions of its K topics in the common projection. These lines are then segmented vertically in ascending order of K, with the top line showing topics for $K_{\min}$ and the bottom line for $K_{\max}$. Figure 1a illustrates the application of 1D projection for K ranging from 10 to 28, generated using the t-SNE dimensionality reduction technique [18].

To highlight relationships across different topic solutions, we connect each topic in the K-topics solution to its most similar topics in $K-1$- and $K+1$-topics solutions. These connections reveal how topics branch or split as the number of topics increases. The similarity between topics is determined by identifying nearest neighbors in the original N-dimensional space, where N is the number of terms, i.e., grid cells in our case.

Inevitably, dimensionality reduction introduces distortions, i.e., the distance between the points in the projection space (1D) might not always correlate with their distance in the original space (N-dimensional). In an ideal scenario, with the absence of distortions, comparable topics in each row should maintain their content and position. In this situation, each point will be connected to one of the closest points on both the top and bottom lines (either to its left or right). Without any distortions, the connecting links remain uncrossed forming a trapezium-like shape. We measure the distortion introduced by dimensionality reduction as the number of crossing links, i.e., links that connect a point to a point other than the closest one in the image space (7 crossing links in Figure 1a).

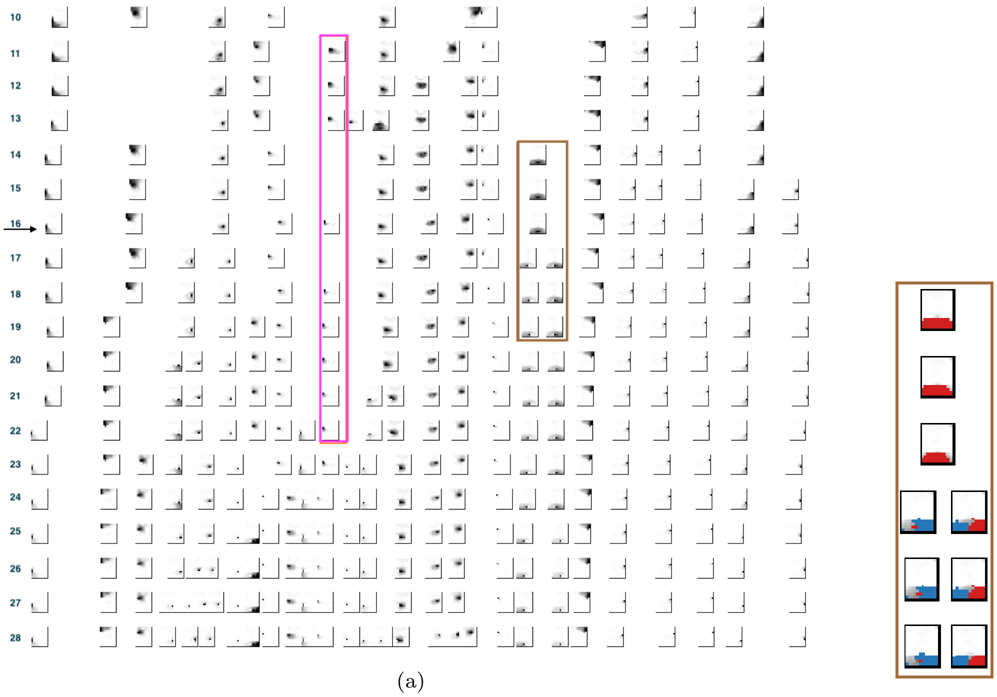
Dimensionality reduction can be performed with different methods. We explored three prominent dimensionality reduction techniques: t-SNE [18], UMAP [19], and MDS [15]. For each technique, we carefully adjusted the parameters to minimize distortion. Specifically, we selected a neighborhood size of 15 for UMAP and a perplexity value of 10 for t-SNE. All three methods produced consistent results, but t-SNE demonstrated the least distortion, yielding the clearest outcomes. Figure 1a presents the results for t-SNE that has the least distortion with 7 crossing links. Figure 1b shows the results for UMAP with 15 crossings that has more distortion than t-SNE, but less than MDS. Figure 1c shows the results for MDS with many crossing links.

5.1. Spatial Visualization supporting Topics Interpretation

To make an informed decision about the optimal number of topics, it is crucial to interpret the composition of topics within each set from a single run, understand their patterns, and compare results across different runs. To facilitate this, we created the visualization shown in Figure 2a which addresses these criteria. This visualization adopts the layout presented in Figure 1a but replaces the dots with small heatmaps showing topics' spatial footprints. The heatmaps' rectangle shape represents a football pitch with the attacking direction upward. The gray-scale shading reflects the weight of each grid cell ("term") in the topic. In other words, these heatmaps highlight the specific areas of the football pitch that each topic covers. Each topic's vector (180 values for one heatmap) has been normalized so that the values of its grid cells sum to 1. The connecting links are omitted, as they are no longer necessary, and retaining them would only clutter the figure. Now, the heatmaps allow us to visually interpret the topics in each row and compare their content with those above and below. Using Figure 2a, we can compare topics within the same row (for instance, moving from left to right in the scenario with 11 topics) and their resemblance to topics in adjacent rows (for example, comparing scenarios with 10 and 12 topics).

We observe gradual transitions from one topic to another in both vertical and horizontal directions. When looking vertically, it is interesting to note that the patterns in the rows with a smaller number of topics appear to be combinations of the patterns from the rows with a larger number of topics below them. When looking horizontally, we must remember that some level of distortion is inevitable, so we need to be careful when making sense of positions along the 1D axis. However, despite this, we can still find consistent patterns through different experiments with varying numbers of topics, showing how topics evolve across several rows.

This representation enables finding the most suitable set of topics for further analysis. In particular, we are looking for the smallest number of topics that satisfy two constraints: 1) its topics are *spatially consistent* across a wide



(a)

(b)

Figure 2: ***Figure (a)*** presents a comprehensive visualization of topic spatial footprints (spatial heatmaps), derived from topic modeling across a range of topic numbers (10 to 28, as labeled at each line). Each heatmap represents a football pitch, with the darkness of its grid cells proportional to the respective weights in the topic-term matrix (the attacking direction is from bottom to top). The final selected topic number is 16 (indicated by an arrow). The reason for this selection is that topics in this row are both spatially consistent and easily distinguishable. The group in the pink rectangle illustrates what spatial consistency means, while the group in the brown rectangle demonstrates redundancy.
Figure (b) provides a detailed view of the topics selected within the brown rectangle. The interactive tool is used to select the topics and color their cells red/blue when they are higher/lower than the average of the selection. The presence of red cells in the first three rows and a mix of red and blue cells in the last three rows suggests that the cell values have been distributed between two similar topics in the branches.

range of number of topics, i.e., they appear consistently, with minor fluctuations, within a range of rows, and 2) its topics are *easily distinguishable*, i.e., they do not have redundant patterns that can be replaced by one.

To assist the analyst in finding such number of topics that satisfy the above constraints, we added interactivity to this visualization. Our tool enables the analyst to pick a group of topics for a detailed exploration, in which the analyst can compare each of the selected topics against their average. For each topic, grid cells that fall below the average will be colored blue, while those exceeding the average will turn red, making it easier to perceive and understand the differences. As such, the tool color codes the cells to accurately reflect the comparison of heatmap pairs or groups within the entire range of different number of topics. Figure 2b is a screenshot of a selection of heatmaps from Figure 2a by our tool. It shows the dominance of cells' red colors for topics in the upper rows, contrasted with the mostly blue cells in the topics of the branching lower rows, and illustrates how a single topic at the top has been divided into multiple topics below. The selection is shown by a brown rectangle in both 2a and 2b figures.

Considering the two constraints, we selected 16 as the most suitable number of topics. Figure 3 displays the topic footprints for this selection. The reason for this selection is that the topics in this row show spatial consistency (encompassing all topics or their similar ones that consistently appear in other runs) and are easily distinguishable (excluding topics that have redundant patterns that can be shown as one). In particular, the selection enclosed in a pink rectangle shows a consistent pattern that does not appear in the result with 15 topics but does appear in many other results, including the one with 16 topics. In addition, the brown selection shows that in the results with 17 or more topics, one topic is split into two topics that have high overlap. We note that this approach is intentionally designed to be domain-agnostic. It allows users to assess and choose the most suitable number of topics visually, without requiring prior domain knowledge.

A well-established method for analyzing ball possession and movement in a football pitch is to segment the pitch into discrete zones and then record actions or events according to their location. Following Camerino *et al.* [10], we segment the pitch along both horizontal (defensive-offensive) and vertical (left-right) axes to capture how play transitions from one area of the pitch to another. As shown in Figure 3, we draw these dividing lines to create two sets of zones:

- *Horizontal or Defensive-Offensive Bands:* These bands run from the top/most advanced offense to the bottom/deep defense of the pitch, subdividing the field into the following areas: Ultra Off (most advanced offense), Off (offense), Cent (central midfield), Def (defense), Ultra Def (deepest defense)
- *Vertical or Left-Right Channels:* These channels split the pitch laterally

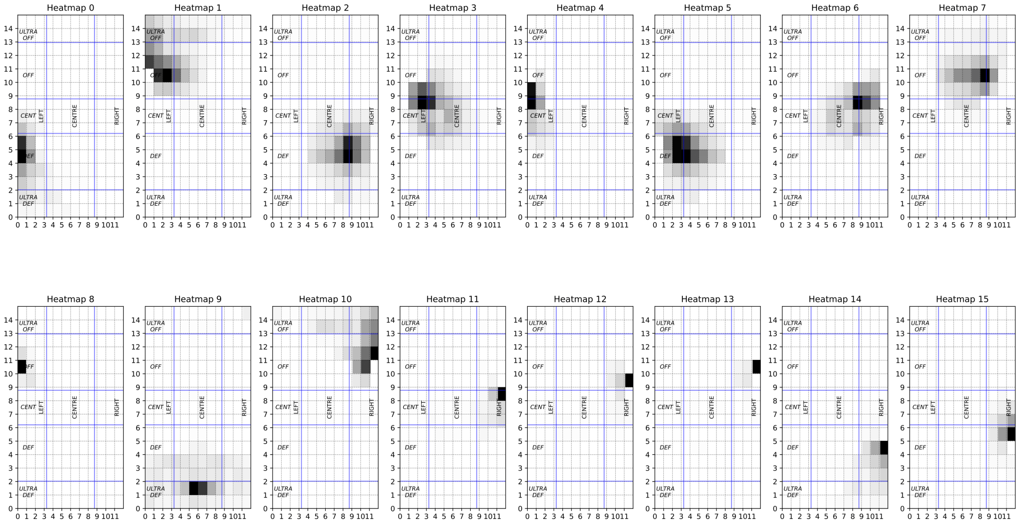


Figure 3: A detailed view of the topic footprints on a football pitch, with 16 topics.

into three sections: Left, Center, and Right.

In our application, each episode (i.e., document) is defined by a team's period of ball possession. As such, we do not track the specific sequence of events within an episode. Consequently, the topics we identify primarily reflect *where on the pitch the majority of actions and interactions occur*. For example, a "deep defense" zone merely indicates that a team holds possession in a deeper area of the pitch; it does not necessarily mean they are actively defending. They might be controlling the ball in their own half, retaining possession, or waiting for an opportune moment to advance. Similarly, a topic reflecting interactions in an offensive zone confirms that the team possesses the ball in advanced areas; it does not necessarily mean an ongoing attack. These scenarios could also arise from duels or other actions in the final third. While in *most cases* the zones align with typical defensive or offensive roles, they do not always correspond to a team's offensive/defensive mode.

6. A Suite of Visualization Techniques for Understanding Football Games Through Topics

In the following two sections (Sections 7 and 8), we discuss the third step of our workflow, i.e., visual analytics in the context of the football domain.

We present five distinct visualizations that utilize topics derived from topic modeling to analyze and summarize football games at various levels of granularity. The first visualization offers an in-depth exploration of an individual match, providing detailed insights into its dynamics (Section 7).

In contrast, the subsequent four visualizations are designed to deliver concise match summaries and facilitate comparisons between different matches and teams (Section 8). Together, these tools enable a comprehensive analysis of football games.

In the following two sections, we first introduce the details of our visual analytics techniques. We then use the visualizations to showcase some behaviors, either from one team in a match or a set of teams in all their matches. We leave more behavioral analysis to future work.

7. Exploring a Single Match

We present a visualization framework for analyzing the flow of topics within a single football match. Figure 4 illustrates an example from the second half of the Barcelona vs. Real Madrid match (May 2018), which ended in a 2–2 draw. The visualization is divided into two primary components:

- **Timeline-based charts (left side):** Displays each team's episodes in chronological order, showing which topics were active during each episode. It also highlights whether an episode was successful or had any key events.

- **Aggregated summaries (right side):** Provides a summary of how each topic contributed across All, Successful, and Unsuccessful episodes. Additionally, it includes a "difference plot" that highlights the differences in topic weights between the two teams, enabling direct comparison.

This layout allows analysts to see both how topics evolve over time (left side) and how topics' weights aggregate across various outcomes (right side). Further details about each side's analytical goals and design are provided in the following subsections.

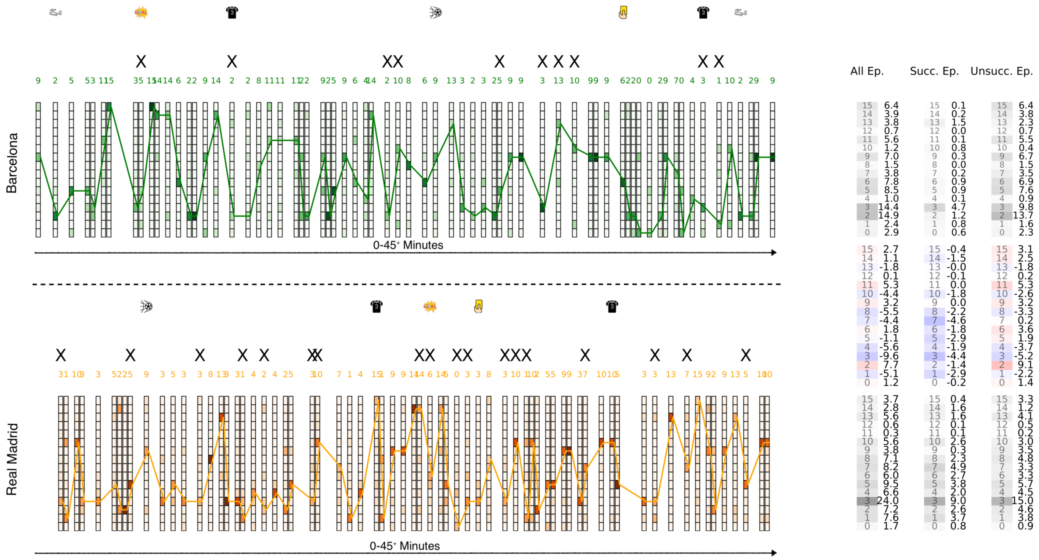


Figure 4: This figure shows data from the second half of the Barcelona vs. Real Madrid match (May 2018, 2–2 draws). **Left side:** segmented bar charts on a timeline depict active topics (IDs 0–15) for Barcelona (above the dashed line, green) and Real Madrid (below, orange). Each bar is labeled at the top with its dominant topic ID, along with markers for key match events and successful episodes. **Right side:** Summary columns compare aggregated topic weights (All, Successful, Unsuccessful) and include a difference plot to highlight contrasts between the two teams.

7.1. Timeline-based charts

7.1.1. Analytical Goal

We developed a visual tool that provides an in-depth view of both teams' playing behavior during a football match by tracking how topics evolve during each team's ball possession, identifying the dominant topics in each episode, and revealing the relationship between key events, successful episodes, and topics emergence.

This concise yet comprehensive overview is particularly helpful for football data analysts seeking to quickly assess a match's dynamics.

7.1.2. Design and Justification

Figure 4 illustrates the second half of the Barcelona vs. Real Madrid match (May 2018) extending a layout commonly seen in football reports. The x-axis represents the timeline, starting at 0 and extending beyond 45 minutes to cover the entire second half. The y-axis displays each team's episodes, with

Barcelona's information above the dashed divider and Real Madrid's below, forming two sub-figures.

Each vertical bar represents one ball-possession episode, color-coded by team (green for Barcelona, orange for Real Madrid) and arranged chronologically. We chose to show every episode as the same width to prevent episodes from covering each other. Still, with the same length, some short episodes would be covered by their next episode. In such cases, we slightly shift the second episode (and the following ones) to prevent masking. Each bar consists of 16 segments (IDs 0 at the bottom to 15 at the top), corresponding to topics defined in Figure 3, and is normalized to sum to 1. The darkness of each segment reflects the topic's weight, with the dominant topic (highest weight) labeled on top. The segments corresponding to the dominant topics are connected by a continuous line, making it easier to observe their changes over time.

Additionally, we marked the successful episodes with an "X" above the most dominant row. Successful episodes are characterized by significant progression into the opponent's part of the pitch. An episode is considered successful if its final event is a "shot" or "goal", or if the final event is one of the "pass", "duel", or "others on the ball" events and that event's location is within the last 20% of the pitch. We also defined match key events to be Received a Goal (⚽), Scored a Goal (GOAL), Own Goal (🥅), Dangerous Ball Loss (⚽), Red Card (🟥), Yellow Card (🟨), Second Yellow Card (🟨🟥), and Substitutions (3) and presented them with representative distinct icons. We use successful episodes and these markers as a tool to help the analyst understand game development and the relationship between topics and match events more easily.

7.1.3. Normalization

As mentioned in section 7.1, we normalize each episode so that its total weight is 1. This approach highlights which topics are most dominant within a given episode (one bar chart), independent of overall episode weight.

Let W be the document-topic matrix of a team in a given match. We normalize each row of W so that its values sum to 1, using

$$\hat{W}_{i,j} = \frac{W_{i,j}}{\sum_{k=1}^{K} W_{i,k}}, \quad \forall\, j \in \{1, 2, \ldots, K\}. \tag{1}$$

Here, $\hat{W}$ is the normalized version of W, where each row's elements sum to 1.

This facilitates the comparison of topic distributions across different episodes and helps analysts quickly identify the most dominant topics. Each row of $\hat{W}$ corresponds to one of the episodes (bar charts) in Figure 4.

7.2. Aggregated summaries

7.2.1. Analytical Goal

On the right side of the figure, segmented vertical bars summarize topic values across three categories: **All Episodes**, **Successful Episodes**, and **Unsuccessful Episodes** for both teams. This summary helps analysts quickly assess how each team's active topics change under different circumstances during the half.

To facilitate direct comparison, we included a "difference plot" between the teams' aggregated topics. It visualizes the differences in topic weights between the two teams.

This layout provides an intuitive overview of how each team relied on different topics overall and during specific phases of play (successful or unsuccessful), making it easy to compare their approaches in various scenarios.

7.2.2. Design and Justification

Each segmented bar on the right side, located in the first and last rows, represents an aggregated total for one category: **All Episodes**, **Successful Episodes**, or **Unsuccessful Episodes**. The *All* category sums the weights from all episodes, while *Successful* and *Unsuccessful* focus exclusively on episodes from their respective groups. The darkness of each segment reflects the topic's weight, with darker shades indicating higher values. These values are normalized using a gray-scale mapping (see next section) to ensure consistency and comparability across topics and teams. This design helps analysts quickly identify which topics were most prominent within each category (e.g., topics most used in successful plays).

The "difference" plot, positioned in the middle, between the two teams' summaries, visualizes the topics' weight differences by subtracting Team 2's topic weights from Team 1's topic weights. A red-white-blue color scale is applied: shades of red indicate positive differences (Team 1 has higher weights), shades of blue indicate negative differences (Team 2 has higher weights), and white represents no difference. This design enables analysts to quickly compare the teams' topic weights across various scenarios at a glance, without needing to toggle between the two summary plots.

Each segment is labeled with its topic ID (0-15), and its normalized numeric value is displayed to the right.

7.2.3. Normalization

All Episodes. For the *All* episodes normalization, we derive $\texttt{global\_min}_h$ and $\texttt{global\_max}_h$ by examining every team's aggregated topic weights in a **single half (h)** —either the first or second half— within a given league (e.g., Spain). This half-specific normalization ensures that the results are contextually accurate for the period being analyzed.

We then map each aggregated value v_i to a gray intensity $\hat{v}_i \in [0, 1]$ using the formula:

$$\hat{v}_i = \frac{v_i - \texttt{global\_min}_h}{\texttt{global\_max}_h - \texttt{global\_min}_h}. \tag{2}$$

Here, $\hat{v}_i = 0$ corresponds to the lightest shade (white), and $\hat{v}_i = 1$ corresponds to the darkest shade (black). Having the half index (h) shows normalization values are specific to a single half. This provides a more accurate representation of variations within the half rather than across the entire game.

Successful and Unsuccessful Episodes. The normalization for summary bars of *Successful* and *Unsuccessful* episodes follow the same procedure as *All Episodes*, but the raw sums are restricted to **only successful or unsuccessful episodes within a single half (h)**, respectively.

Color intensities in the "difference plot" are scaled so that the largest negative differences appear as the darkest blue, the largest positive differences as the deepest red, and intermediate differences as progressively lighter shades.

By applying this normalization, we ensure that color intensities are comparable in different columns and analysts can see at a glance which topics a team emphasizes most or least across the three categories (All, Successful, and Unsuccessful). The *All Episodes* column provides a broad overview of how each topic may relate to the final match outcome, while the *Successful* and *Unsuccessful* columns clarify which specific topics contributed to more effective or less effective episodes, respectively.

7.3. Observations and Interpretation

Looking at the Figure 4, we note that for successful events, usually, at least one of the topics 1 (attack from pitch's left side) or 10 (attack from pitch's right side) are activated; however, they are not necessarily the most dominant topic, since those two topics are usually the final active topic of a successful episode. Analyzing the other active topics that lead to successful episodes could give insights into how a team performs its attacks. In addition, the dynamics of topic weights along the timeline show that some topics are consistently appearing with high weights, while others occur infrequently. We can see pairs and even larger groups of topics that are prominently visible together within multiple episodes. For example, topics 9 and 14 are activated together on multiple occasions for both teams. Based on the topics' footprints (from Figure 3), this means that they made passes between their goal/penalty area and the right side of their defensive half, or these two topics became active because of other co-occurring events (e.g., duals in both areas).

Another way to utilize this figure is by focusing on key events and analyzing the topics that led to them. For instance, as shown in Figure 4, Barcelona scored a goal (GOAL) in their ninth episode. To understand how this happened, we analyzed the preceding episodes for both teams.

Examining Real Madrid's episodes reveals that they Received the goal (⚽) in their tenth episode. However, just before that, in their episode 9, they had a successful possession, indicating that they had advanced effectively into Barcelona's territory. The following episodes in the game shifted momentum as Barcelona took possession, maintained control for two consecutive episodes, and ultimately scored a goal.

When analyzing the most active topics during Barcelona's ninth episode (the scoring episode), we observed that topic IDs 5, 3, and 1 were heavily activated. These topics correspond to actions on the left side of the pitch (from Barcelona's perspective), spanning defense, midfield, and attack. This suggests that after Real Madrid's attempt, Barcelona executed a swift counterattack from the left side, leveraging these topics to score.

A review of the game footage confirmed this sequence: Real Madrid's promising attempt was a shot by Marco Asensio aimed at Barcelona's goal, which failed to find the net. Barcelona regained possession immediately and, within just two episodes, successfully scored against Real Madrid.

These observations illustrate how an analyst might link specific topic patterns to actual on-pitch strategies, fostering deeper insights into team performance.

8. Exploring Multiple Matches

While our previous visualization provided an in-depth look at a single match topics, we further developed the visual analytics technique and created a new set of visualizations that enable analysts to compare multiple matches across different teams or leagues. Specifically, these visualizations reveal:

- how a team's topic activations change when facing stronger or weaker opponents (8.1),
- how topic activations evolve over the course of a season (8.3),
- how each match's topics deviate from the team's norm (Sections 8.2 and 8.4), and
- how they correlate with final outcomes, particularly when playing at home or away.

Following the same structure as in the single-game visualization, the proposed visualizations represent aggregated match topics in various matrix orders, complemented by aggregates to provide deeper insights. In the following subsections, we show an at-a-glance, concise summary of each team's topic weights across every match of a season. We use a consistent grid/matrix layout and global normalization scheme, making it possible to spot patterns, outliers, and potentially informative differences in topic activation and match outcomes.

8.1. Ordered Teams Based on their Rankings

8.1.1. Analytical Goal

Our main analytical goal is to reveal how each team's topic weights change when facing stronger or weaker opponents, incorporating additional details such as match results and home/away status. By integrating this information, the tool enables analysts to identify potential patterns and correlations between match outcomes and topic activations.

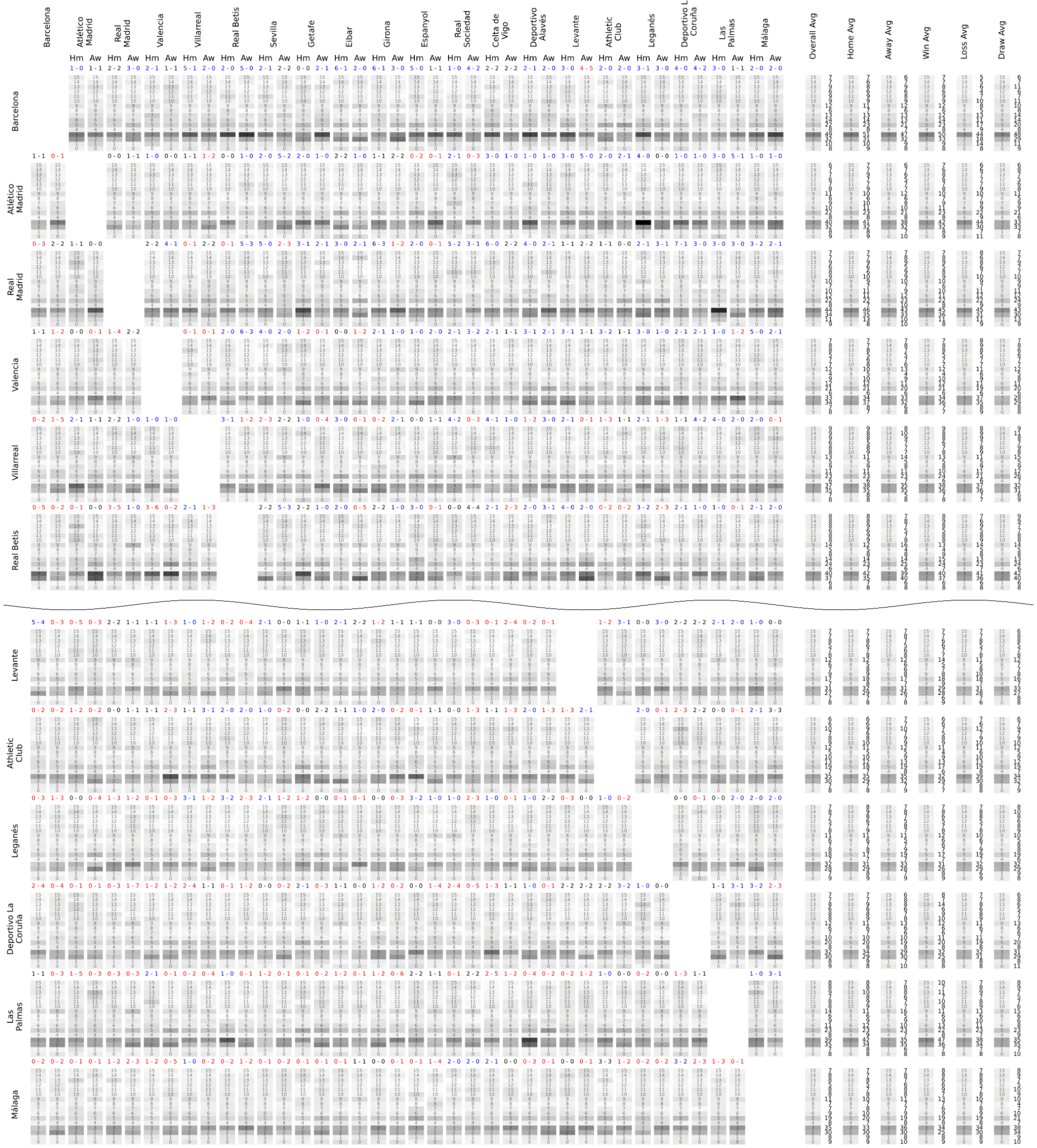


Figure 5: Topic weight visualizations for LaLiga 2017–2018 season, sorted by team ranking. The strongest teams are positioned at the top-left, and the weakest at the bottom-right. Only the top six teams (above the wavy line) and the bottom six teams (below it) are displayed because of space constraints. Each cell represents the row team's topic weights in a match against its opponent (column). On the right, summary columns present average topic distributions under different conditions (e.g., home and away). See Section 8.1 for more details.

8.1.2. Design and Justification

Matrix Layout. The visualization (Figure 5) employs a matrix sorted by league ranking, with the strongest team placed at the top-left and the weakest at the bottom-right. Rows and columns both follow this ranking, so moving along a single row from left to right shows how a team's topics shift against progressively weaker opponents. Each cell represents a single match and displays a bar chart for aggregation of the corresponding topic weight matrix W.

The match result at the top of each cell is color-coded (blue for a win, black for a draw, red for a loss) from the row team's perspective, and the labels "Hm" (home) or "Aw" (away) indicate the match venue. This information helps analysts see how venue or match outcomes might influence topics' patterns in matches. To accommodate Figure 5 on a single page, a wavy line is used to indicate the exclusion of middle-ranked teams, retaining only the top six and bottom six teams for display. The full version of this figure can be found in the provided link [1].

Summary Columns. Six additional columns (overall, home, away, win, loss, draw) appear on the right side of the matrix. They are computed by averaging each team's normalized topic weights under the corresponding condition. Each summary column is a bar chart with 16 segments (IDs 0–15), with the topic weight displayed on the right. These columns let analysts quickly gauge which topics a team relies on most or least overall.

This design A) shows whether top-ranked teams display distinctly different topic weights compared to lower-ranked teams; B) links topic activation with match results at a glance, using color-coded outcomes; and C) highlights team-level patterns (overall, home, away, win, loss, draw) that transcend individual opponent match-ups.

8.1.3. Normalization

We employ a global normalization scheme to ensure consistent color intensities across all matches and summaries in the matrix.

Assume that $\mathbf{W}_{\mathbf{M}\times K}$ is a comprehensive document-topic matrix, where $\mathbf{M}$ represents the total number of episodes (across all teams and matches

[1] High-resolution versions of all figures used in this paper are available at: https://lalehmoussavi.github.io/topic-modeling-for-behavioral-patterns/

of a league)[2] and K is the number of topics. The matrix $\mathbf{W}$ encompasses the document-topic matrices $W^1, W^2, \ldots, W^T$, where T is the total number of unique $(\text{team}, \text{match})$ combinations. Each sub-matrix W^t corresponds to one particular team in a specific match. Let M^t be the number of episodes in the t-th $(\text{team}, \text{match})$ combination, and let K be the number of topics. Formally:

$$\mathbf{W} \;=\; \begin{pmatrix} W^1 \\ W^2 \\ \vdots \\ W^T \end{pmatrix}, \tag{3}$$

where each element $W^t_{i,j}$ indicates the weight of topic j in episode i for the t-th $(\text{team}, \text{match})$ pair and must be a non-negative real number:

$$W^t_{M^t \times K} \overset{\text{def}}{=} W^t \in \mathbb{R}_{\geq 0}^{M^t \times K}, \quad t \in [1, \ldots, T], \tag{4}$$

Unlike Equation 2 in Section 7, which applies only to a single half, here we derive the global minimum and maximum from the full match. Specifically, to obtain each cell's aggregated row vector, we sum the pre-normalized topic weights across all episodes in a match to produce $\tilde{W}^t_{1 \times K}$. We then determine the global minimum $w_{\min}$ and maximum $w_{\max}$ across all $\tilde{W}^t$ vectors. We then normalize each $\tilde{W}^t$ vector as follows:

$$\hat{\tilde{W}}^t_{1,j} = \frac{\tilde{W}^t_{1,j} - w_{\min}}{w_{\max} - w_{\min}}. \tag{5}$$

Here, darker shades represent higher summarized topic weights (values closer to 1), while lighter shades indicate lower weights (values closer to 0).

This approach ensures standardized color intensities, allowing fair comparisons of topic relevance across different matches, teams, and outcomes.

8.1.4. Observations and Interpretation

One way to use this visualization is to analyze topic activation under different outcomes (win, draw, loss). By comparing the summary columns in

[2] Unlike in Section 4 where $\mathbf{W}$ contains topic weights of all matches in all leagues and tournaments in the dataset, in this section we use $\mathbf{W}$ to describe the topic weights of all matches of a single league.

Figure 5 and focusing on topic IDs 1 and 10 (attack) versus 9 (defense), we see a common pattern: many teams show higher activation of attack (1 and 10) and lower activation of defense (9) when losing. While this may appear counterintuitive, it aligns with the idea that once a team gets behind, it often prioritizes attacking to score goals.

For instance, Barcelona's summary columns indicate that topic 9 has weights of 12 when winning, 8 when losing, and 10 when drawing, whereas topic 1 (attack) has weights of 9 during wins, 14 during losses, and 11 during draws. Teams trailing in a match typically attack more aggressively to recover, while teams ahead adopt a more defensive stance. Although playing styles can vary, and not all teams follow the same pattern, this strategic balance is widely observed across many teams.

8.2. Ordered Teams Based on their Rankings - Deviation from the Average

We extended our previous visualization to display not only topic distributions per match but also deviations from the average, providing deeper insights into team behaviors and strategies across different game situations.

Figure 6 uses the same matrix structure as Figure 5, but each non-summary cell now shows the difference between each topic's value and its corresponding overall average. Deviations are visualized as stacks of colored bars: red for positive deviations (higher than average), blue for negative deviations (lower than average), and white for no deviation. Darker shades denote more extreme differences, with the darkest red and blue representing the maximum positive and negative deviations, respectively. The summary columns remain unchanged, providing consistent context for topic average distributions in different situations.

By visualizing these deviations, analysts can quickly identify over- or under-performance, spot patterns, and anomalies, and gain clear insights into how teams deviate from their average, simplifying complex data for better interpretability. Using this figure, the pattern described in Section 8.1.4—teams attacking more and defending less in losses—is further confirmed.

8.3. Ordered Matches Based on Time for Selected Teams in a League

To analyze how topics evolve over a season for selected teams, we designed a visualization similar to Figure 5, with columns arranged chronologically (from earliest to latest in the season). Users can select specific teams from the league to compare their topic evolutions across rows.

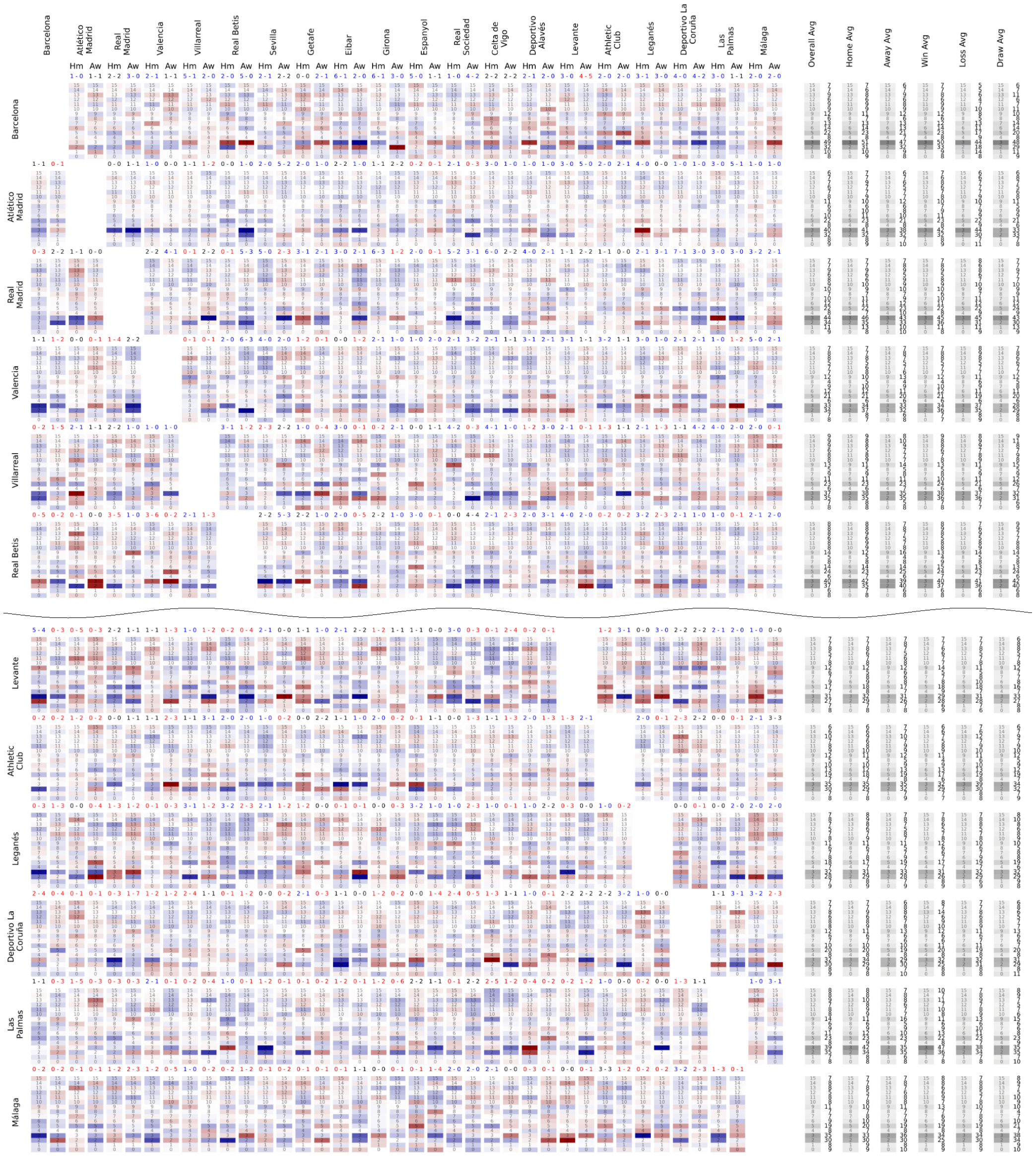
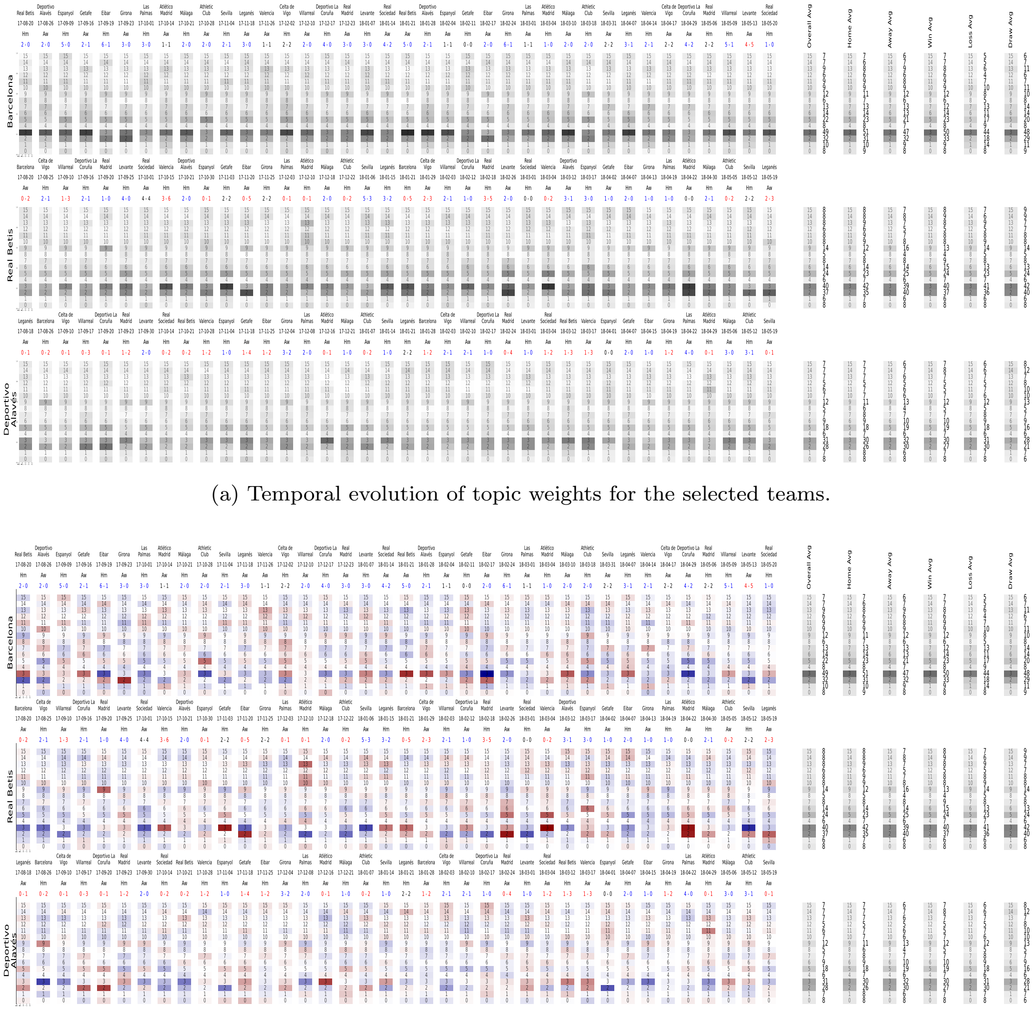


Figure 6: Deviations of teams' topic weights from their overall average, sorted by team ranking. This visualization is similar to Figure 5, but each cell displays the deviation of the row team's topic weights from its mean values. These deviations are visualized as stacks of colored bars: red for positive, blue for negative, and white for no deviation, with darker shades indicating larger differences. See Section 8.2 for more information.



(a) Temporal evolution of topic weights for the selected teams.

(b) Temporal evolution of topic deviations from the overall averages for the selected teams. Red segments indicate positive deviations, blue segments indicate negative deviations, and white denotes no deviation, with darker shades representing larger differences.

Figure 7: A matrix-based visualization of the temporal evolution of (a) topic activations, and (b) their deviations from the overall average for three teams: Barcelona (top), Real Betis (middle), and Deportivo Alavés (bottom). In both figures, rows represent the teams, and matches are ordered chronologically from left to right. Each cell includes labels above it, indicating the match result (blue for a win, red for a loss, black for a draw, from the row team's perspective), whether the match was at home or away, the match date, and the opponent's name.

Figure 7a showcases three teams–Barcelona (top row), Real Betis (middle row), and Deportivo Alavés (bottom row)–representing a range of league rankings, though any teams can be selected.

The left columns display matches chronologically, while the right columns summarize team topics. Bar charts are normalized using league-wide topic weight ranges, consistent with earlier visualizations.

8.4. Ordered Matches Based on Time for Selected Teams in a League—Deviation from the Average

We extend the temporal view in Figure 7a by creating a deviation-from-average visualization. Instead of raw topic weights, analysts can observe how each match's topic values differ from a team's average. Matches remain arranged chronologically in the left columns, and the right columns remain unchanged and continue to display summary metrics. As in previous deviation visualizations (Section 8.2), we color-code the difference between the raw value and the overall average value.

By focusing on *deviations* rather than raw values, this visualization emphasizes how each match compares to a team's own baseline. If, for instance, a certain topic (e.g., attack from left) consistently appears in dark red for a team in a certain period of the season, it suggests they have relied on that tactic far more than usual in those matches. In contrast, topics that remain in shades of blue indicate under-utilization relative to the average. Taken together, the color-coded differences, the match context (opponent, date, home/away), and the outcome (win/loss/draw) reveal patterns in a team's adaptability, strengths, and weaknesses over time.

Listing matches in chronological order further aids in spotting trends such as mid-season strategic shifts or responses to injuries, or coaching changes. Consequently, analysts gain a deeper understanding of how each team's topic usage evolves over the course of a season, along with how it deviates from its average in individual matches.

One practical example of using such visualization is to identify shifts in topic variation during the season and link them to external factors. In Deportivo Alavés's case, a noticeable increase in certain attacking topics (specifically topic IDs 1 and 10, shown in red) appears in their matches after December 4, 2017. This trend neatly aligns with the appointment of Abelardo Fernández on December 1 as the team's head coach, which sparked a mid-season tactical shift. Under Abelardo's guidance, Alavés moved away from their conservative, low-block style and adopted a more direct, forward-focused approach.

9. Analytical Tasks in the Football Dataset

Task	General Approach	Application in the Case Study
T1 Identifying common behavioral patterns	Apply topic modeling to predefined episodes to extract recurring patterns that serve as representative units of behavior.	Episodes are defined as sequences of ball-possession events by a team (Sec. 3). Topic modeling is applied to these episodes (Sec. 4), and the resulting patterns are visualized via spatial heatmaps, showing frequently activated pitch zones (Sec. 5.1).
T2 Tracking the distribution of patterns across individuals or groups	Analyze how the distribution of patterns varies across individuals or groups within a unified view.	We propose a matrix layout to track topic distributions for both individual teams and leagues throughout the season. An example of this design is shown in Sec. 8, Fig. 5.
T3 Observing the progression of patterns at different temporal scales	Analyze how pattern distributions evolve over varying time scales, from fine-grained episode timelines to aggregated larger trends.	Topic progression is shown at both detailed and broader scales—ranging from episode-level timelines during a match (Sec. 7.1, Fig. 4) to match-level averages during a season (Sec. 8.3, Fig. 7a).
T4 Comparing how contextual conditions influence the occurrence of patterns	Contrast topic weights across different scenarios.	An example from our study is Fig. 4 in Sec. 7. On the left side, we incorporate key match events to examine their influence on the topics' weights, while on the right side, we compare the aggregated topic weights when episodes have been successful or unsuccessful.
T5 Detecting and quantifying deviations from typical behavior	Measure divergence between topic weights in a specific situation and its baseline values.	Deviation-from-average visualizations in Sec. 8.2, Fig 6, and Sec 8.4, Fig 7b highlight when teams deviate from their average patterns during different matches.

Table 2: Mapping of analysis tasks (T1–T5) to general approaches and their applications in our case study.

To structure the analytical goals of our workflow more clearly, we define five generic analysis tasks, ranging from discovering common patterns to identifying deviations from norm behaviors. Table 2 outlines how the computational and visual components of our workflow support these tasks and illustrates their application in the context of football data. The first column specifies each task (T1–T5), the second describes the general approach used to address it, and the third shows how it is implemented in our football case study.

10. Conclusions and Future Work

In this study, we introduced an end-to-end workflow that progresses from raw spatio-temporal event data to meaningful insights, structured around three key components (data transformations, computational methods, and visual analytics). We demonstrated this workflow by exploring the dynamics of football games as an example of important yet under-explored spatio-temporal event data.

For the first step (data transformations), we transformed football game event logs into episodes of ball possession.

For the second step (computational methods), we used topic modeling on these episodes to identify recurring patterns. Topic modeling is a relatively new and evolving research area that helps abstract and conceptualize spatio-temporal data. One of the key challenges in applying topic modeling is selecting the optimal number of topics such that they are spatially consistent and easily distinguishable. To address this, we proposed a visual analytics approach based on dimensionality reduction, applied to topics derived from multiple modeling runs. Our findings demonstrated the effectiveness of this approach in identifying interpretable topics for further analysis.

Finally, for the third step (visual analytics), we used the results of topic modeling in a suite of visual analytics techniques to support tasks such as representing common patterns (T1), tracking their distribution across teams or leagues (T2), examining how these patterns evolve over time (T3), comparing behavior under different match conditions (T4), and detecting deviations from typical play (T5). More specifically, in the football case, the derived topics were visualized through spatial heatmaps and incorporated into various visualizations to analyze football games at multiple levels of granularity. These tools provided detailed insights into both teams within a match, highlighting key events and successful episodes. They also revealed teams' tactical behaviors under varying conditions, such as competing against stronger or weaker opponents or playing at home versus away. Additionally, they enabled us to observe mid-season trends, tactical adjustments, and deviations from typical patterns in games played under different scenarios.

One promising future work is expanding the vocabulary to include terms that capture the order of events in the pitch and the players involved —such as transitions from zone z_i to zone z_j or from player p_1 to player p_2. This enriched representation could provide deeper insights into player interactions and ball movement patterns, uncovering critical connections between players and activity in specific areas of the pitch. This approach would allow for the discovery of nuanced positional and tactical behaviors that are not captured in the current representation.

One avenue for future research is automating the analysis of the relationship between topics and game outcomes using machine learning models. These models could learn patterns in topic activations and predict their influence on match results (e.g., wins, draws, or losses). Automating this process would enable coaches to dynamically adapt their strategies based on the opponent's tactics or specific in-game situations. For instance, the model could recommend adjustments to player positioning or team tactics by identifying topic activations that have historically led to favorable outcomes. Such advancements could further enhance the practical utility of this approach, supporting data-driven decision-making in football.

Another direction for future work is conducting a structured user study with domain experts (e.g., coaches, analysts) to formally evaluate the usefulness and usability of our workflow and visualizations.

Additionaly, a potential future research direction is to develop automated non-visual methods for selecting the optimal number of topics. This would allow us to compare non-visual results with those selected by users based on our visual analytics approach, helping to assess the robustness.

Finally, we plan to examine our workflow and techniques used in this paper in other spatio-temporal domains beyond football. This will allow us to further assess the generalizability of our approach.

Acknowledgments

This work was supported by Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B).

References

[1] Andrienko, G., Andrienko, N., Anzer, G., Bauer, P., Budziak, G., Fuchs, G., Hecker, D., Weber, H., Wrobel, S., 2019. Constructing spaces and times for tactical analysis in football. IEEE Transactions on Visualization and Computer Graphics 27, 2280–2297. doi:10.1109/TVCG.2019.2952129.

[2] Andrienko, G., Andrienko, N., Hecker, D., 2023a. Extracting movement-based topics for analysis of space use, in: EuroVA: International Workshop on Visual Analytics, The Eurographics Association. doi:10.2312/eurova.20231091.

[3] Andrienko, G., Andrienko, N., Heurich, M., 2011. An event-based conceptual model for context-aware movement analysis. International Journal of Geographical Information Science 25, 1347–1370. doi:10.1080/13658816.2011.556120.

[4] Andrienko, N., Andrienko, G., 2006. Exploratory analysis of spatial and temporal data: a systematic approach. Springer Science & Business Media. doi:10.1007/3-540-31190-4.

[5] Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., Wrobel, S., 2020. Visual analytics for data scientists. Springer. doi:10.1007/978-3-030-56146-8.

[6] Andrienko, N., Andrienko, G., Gatalsky, P., 2003. Exploratory spatio-temporal visualization: an analytical review. Journal of Visual Languages & Computing 14, 503–541. doi:10.1016/S1045-926X(03)00046-6.

[7] Andrienko, N., Andrienko, G., Shirato, G., 2023b. Episodes and topics in multivariate temporal data. Computer Graphics Forum 42, e14926. doi:10.1111/cgf.14926.

[8] Atluri, G., Karpatne, A., Kumar, V., 2018. Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys (CSUR) 51, 1–41. doi:10.1145/3161602.

[9] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

[10] Camerino, O.F., Chaverri, J., Anguera, M.T., Jonsson, G.K., 2012. Dynamics of the game in soccer: Detection of t-patterns. European journal of sport science 12, 216–224. doi:10.1080/17461391.2011.566362.

[11] Chen, S., Andrienko, N., Andrienko, G., Adilova, L., Barlet, J., Kindermann, J., Nguyen, P.H., Thonnard, O., Turkay, C., 2020. Lda ensembles for interactive exploration and categorization of behaviors. IEEE Transactions on Visualization and Computer Graphics 26, 2775–2792. doi:10.1109/TVCG.2019.2904069.

[12] El-Assady, M., Sperrle, F., Deussen, O., Keim, D., Collins, C., 2018. Visual analytics for topic model optimization based on user-steerable speculative execution. IEEE transactions on visualization and computer graphics 25, 374–384. doi:10.1109/TVCG.2018.2864769.

[13] Holmes, I., Harris, K., Quince, C., 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS one 7, e30126. doi:10.1371/journal.pone.0030126.

[14] Krüger, R., Thom, D., Wörner, M., Bosch, H., Ertl, T., 2013. Trajectorylenses–a set-based filtering and exploration technique for long-term trajectory data, in: Computer Graphics Forum, Wiley Online Library. pp. 451–460. doi:10.1111/cgf.12132.

[15] Kruskal, J.B., 1964. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29, 1–27. doi:10.1007/BF02289565.

[16] Laxman, S., Sastry, P., Unnikrishnan, K., 2005. Discovering frequent episodes and learning hidden markov models: A formal connection. IEEE Transactions on Knowledge and Data Engineering 17, 1505–1517. doi:10.1109/TKDE.2005.181.

[17] Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A., Zheng, Q., 2017. Probabilistic non-negative matrix factorization and its robust extensions for topic modeling, in: Proceedings of the AAAI Conference on Artificial Intelligence. doi:10.1609/aaai.v31i1.10832.

[18] Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

[19] McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 doi:10.48550/arXiv.1802.03426.

[20] Moussavi, L., Andrienko, G., Andrienko, N., Slingsby, A., 2024. Interplay of visual analytics and topic modeling in gameplay analysis. Computer Graphics & Visual Computing (CGVC) 2024 doi:10.2312/cgvc.20241213.

[21] Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., Giannotti, F., 2019a. Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology (TIST) 10, 1–27. doi:10.1145/3343172.

[22] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F., 2019b. A public data set of spatio-temporal match events in soccer competitions. Scientific data 6, 236. doi:10.6084/m9.figshare.9711164.

[23] Pappalardo, L., Massucco, E., 2019. Soccer match event dataset. figshare. collection.

[24] Penn, M.J., Donnelly, C.A., Bhatt, S., 2023. Continuous football player tracking from discrete broadcast data. arXiv preprint arXiv:2311.14642 .

[25] Shekhar, S., Jiang, Z., Ali, R.Y., Eftelioglu, E., Tang, X., Gunturi, V.M., Zhou, X., 2015. Spatiotemporal data mining: A computational perspective. ISPRS International Journal of Geo-Information 4, 2306–2338. doi:10.3390/ijgi4042306.

[26] Vayansky, I., Kumar, S.A., 2020. A review of topic modeling methods. Information Systems 94, 101582. doi:10.1016/j.is.2020.101582.

[27] Wang, Q., Zhu, H., Hu, W., Shen, Z., Yao, Y., 2015. Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2197–2206. doi:10.1145/2783258.2788577.

[28] Wang, X., Grimson, E., 2007. Spatial latent dirichlet allocation. Advances in neural information processing systems 20.

[29] Whiting, M., Cook, K., Grinstein, G., Fallon, J., Liggett, K., Staheli, D., Crouser, J., 2015. Vast challenge 2015: Mayhem at dinofun world, in: 2015 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE. pp. 113–118. doi:10.1109/VAST.2015.7347638.

[30] Wood, J., Dykes, J., Slingsby, A., Clarke, K., 2007. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. IEEE transactions on visualization and computer graphics 13, 1176–1183. doi:10.1109/TVCG.2007.70570.