

Movement Dynamics in Elite Female Soccer Athletes: The Quantile Cube Approach

Kendall L. Thomas¹ and Jan Hannig¹

¹Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, US

Abstract

This paper presents the *quantile cube*, a novel three-dimensional summary representation designed to analyze external load using GPS-derived movement data. While broadly applicable, we demonstrate its utility through an application to data from elite female soccer athletes across 23 matches. The quantile cube segments athlete movements into discrete quantiles of velocity, acceleration, and movement angle across match halves, providing a structured and interpretable framework to capture complex movement dynamics. Statistical analysis revealed significant differences in movement distributions between the first and second halves for individual athletes across all matches. Principal Component Analysis identified matches with unique movement dynamics, particularly at the start and end of the season. Dirichlet-multinomial regression further explored how factors such as athlete position, playing time, and match characteristics influenced movement profiles. Our analysis reveals external load variations over time and provides insights into performance optimization. The integration of these statistical techniques demonstrates the potential of data-driven strategies to enhance athlete monitoring and workload management in women's soccer.

Keywords: Dirichlet-multinomial regression; GPS tracking; Hellinger distance; multivariate analysis; Principal Component Analysis;

1 Introduction

Wearable technology has fundamentally transformed how athletic performance is monitored and analyzed, especially among elite athletes. These devices generate vast quantities of data, providing insights into the physical demands placed on athletes and enabling more precise adjustments to training regimens (Cummins et al., 2013). Building on this foundation, integration with advanced data analytics has become essential for extracting actionable insights, especially when evaluating training volume and intensity (Bourdon et al., 2017).

A key component of this process is external load monitoring via Global Positioning System (GPS) technology. These systems quantify movement metrics such as velocity, acceleration, and distance covered during matches and training. Such metrics capture critical aspects of movement intensity and dynamics, allowing for the assessment of workload distribution, fatigue development, and positional movement demands. While total distance and average speed are often reported, more nuanced metrics such as acceleration and deceleration efforts and individualized speed thresholds may provide more sensitive indicators of player workload and fatigue (Snyder et al., 2024).

Despite the potential of these metrics, challenges remain. Many wearable device companies provide proprietary “training load” metrics that integrate multiple performance variables without disclosing their exact formulas or component definitions. For example, speed threshold definitions, which categorize movement into zones such as high-speed running or sprinting, can vary across devices, teams, or sports. Some use absolute velocities (e.g., $> 5 \text{ m/s}$), while others reference an athlete’s maximum speed (e.g., $> 70\%$ or $> 90\%$). Consequently, comparisons across devices can be difficult, leading to inconsistencies in training and performance interpretation, and thereby raising questions about metric reliability.

While most of the research on workload patterns has largely focused on male athletes, studies investigating women’s soccer are steadily increasing. De Lucia et al. (2024) reported gender differences in GPS-derived workload metrics such as sprint distance,

accelerations, and player load per minute. Extending this work, Kuhlman et al. (2025) highlighted sport- and position-specific workload variations across women’s collegiate soccer, lacrosse, and field hockey. Within women’s soccer, traditional GPS metrics do not always correlate strongly with match outcomes (Gailor et al., 2024). Time-segmented analyses reveal early-onset fatigue, which affects high-speed running, acceleration, and deceleration during match play (Snyder et al., 2024). This highlights short-term performance declines that composite metrics over a full session, such as total load, can obscure. Collectively, these findings emphasize the need for refined, interpretable, and position-specific monitoring frameworks tailored to female athletes.

Temporal and positional variations in workload demands have been well documented in male professional players. For example, Barrera et al. (2021) observed reductions in high-speed running and other external load metrics during the second half of professional matches, reflecting fatigue and tactical adjustments. Similarly, Wehbe et al. (2014) reported that midfielders cover greater total and high-intensity running distances than defenders, highlighting position-specific load profiles. While these studies focus on male athletes, they provide a valuable comparative framework for investigating similar dynamics in elite female soccer athletes using advanced modeling techniques.

Modern statistical approaches have emerged to address the computational challenges of analyzing large, longitudinal GPS datasets. Traditional analyses often assume independent and identically distributed data, which rarely holds in real-world contexts (Luo and Song, 2020). Recent methods, including linear state-space mixed models (Luo and Song, 2023) and incremental inference via dynamic updates (Luo et al., 2023), leverage the summation of summary statistics over data batches to dynamically update point estimates and standard errors. However, reliance on summary statistics can obscure important extremes of the data distribution, which are critical for capturing nuanced patterns and generating actionable insights. To overcome these limitations, researchers are increasingly adopting sophisticated approaches that integrate multiple data sources for a more comprehensive understanding of longitudinal workloads.

Complementing these statistical advances, the integration of GPS-derived movement

metrics with multi-dimensional and machine learning frameworks has shown considerable promise for improving workload monitoring and injury prediction in male athletes (Valance et al., 2020; Rossi et al., 2018). These approaches leverage both external load data (e.g., velocity, acceleration, distance covered) and internal load indicators (e.g., subjective well-being, heart rate) to create richer, more predictive models of athlete performance and injury risk. However, the complexity of these models can limit their practical application for coaching and training staff, as interpretability and real-time usability are often constrained. Ferraz et al. (2023) emphasize the urgent need for integrative frameworks that combine external and internal load data in a manner that is both statistically robust and practically actionable. Moreover, the implementation and validation of such methods in women’s sports remains limited, leaving a critical gap in evidence-based monitoring strategies for female athletes. Developing interpretable, multi-dimensional models tailored to the unique physiological, tactical, and positional demands of female soccer athletes represents a key step toward bridging this gap and translating advanced analytics into meaningful coaching and training interventions.

To address these gaps, we propose a novel method to integrate GPS-based external load metrics with athlete and match characteristics in elite female soccer. Our primary objective is to develop interpretable statistical models that quantify movement patterns—specifically velocity, acceleration, and movement angle—and examine their relationships with athlete performance and match outcomes. Unlike traditional zone-based thresholds, which rely on arbitrary cutoffs and vary across devices, our approach leverages the full empirical distribution of movement features to produce player-specific and statistically principled profiles. By combining probabilistic modeling with transparent statistical methods, this approach bridges the gap between data collection and practical application, providing a data-driven foundation for optimizing training protocols.

The paper is organized as follows: Section 2 details the data and proposed summaries for downstream analysis, Section 3 presents the methods and results, Section 4 discusses findings and practical implications, and Section 5 concludes with a summary of strengths, limitations, and directions for future research.

2 Data

The data was collected by the Applied Physiology Lab in the Exercise Science Department at UNC Chapel Hill and shared under Institutional Review Board (IRB) 23-2673. A summary of the notation used throughout this section is provided in the Notation Table (see Appendix A).

GPS tracking data were obtained from all match sessions over one season for 33 elite female soccer athletes. Only match sessions in which an athlete played at least 25 minutes in both the first and second halves were retained. Overtime periods were removed to ensure uniform match durations and comparability. Athletes who met this full-match play criterion in more than five match sessions were then selected. This filtering process resulted in a subset of nine athletes and 23 matches, yielding 198 valid athlete-match sessions. Note that not every selected athlete participated in every included match.

Each raw GPS dataset corresponded to a single athlete in a single match (i.e., one athlete-match session), and contained one data point per second, consisting of a timestamp along with longitude and latitude coordinates for the athlete's location. For example, if an athlete played 80 minutes in a match session, the raw dataset comprising one athlete-match session would contain 4800 rows of timestamped positional data. Each athlete-match session contributed two halves to the analysis, resulting in a total of $n = 198 \times 2 = 396$ athlete-match-halves. Figure 1 (left) shows a 50-second example of this raw data overlaid on a satellite map (Google Maps API, 2025).

To calculate velocity, acceleration, and angle of movement from the raw GPS coordinates for one athlete-match, the longitude and latitude values were converted to (x, y) coordinates in meters using standard spatial transformations (Pebesma, 2018). A third-degree interpolating spline was fitted to the data at ten points per second to model the athlete's movements (Figure 1 , right). Velocity (in m/s) and acceleration (in m/s^2) were derived from the first and second derivatives of the spline, respectively. The angle of movement was calculated as the angular difference between the velocity vector (direction of movement) and the acceleration vector (direction of change in velocity), capturing

the degree of turning or directional change. The angle was computed modulo 360 and subsequently shifted to the range of -180 to 180 degrees for directional interpretability. To remove low-magnitude noise, velocity values below 0.01 m/s and acceleration values below 0.001 m/s² were thresholded to zero. Due to the right-skewed nature of the raw distributions, $\log_{10}(1 + \text{velocity})$ and $\log_{10}(1 + \text{acceleration})$ transformations were applied for interpretability. From this point forward, the transformed values will be referred to simply as velocity and acceleration, except where specified in Table 1.

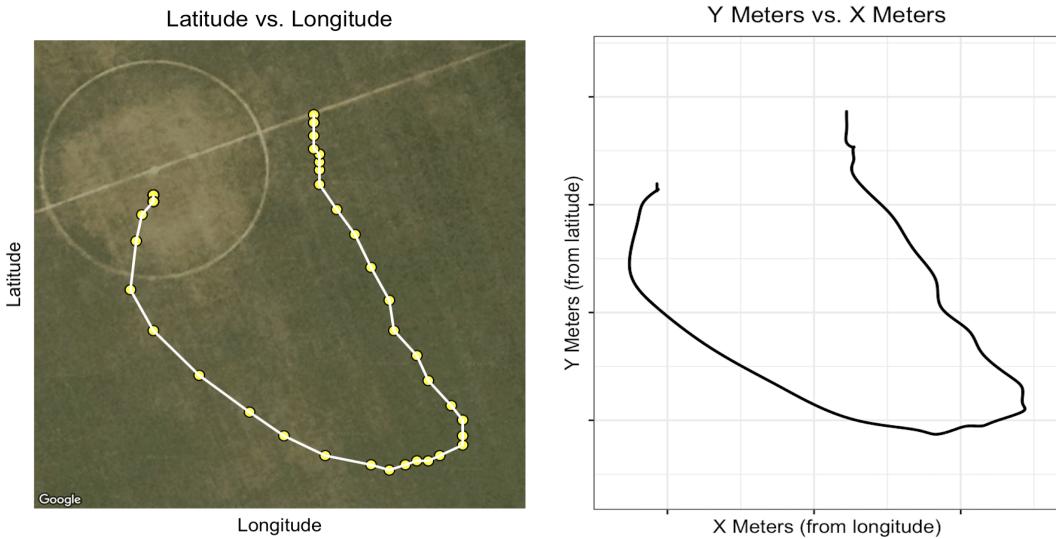


Figure 1: Left: Raw GPS data for 50 seconds of movement for one athlete overlaid on a satellite map (source: Google Maps API (2025)). Right: Interpolating spline (in meters) fit to the same 50 seconds of movement for one athlete seen on the left. For confidentiality purposes, longitude, latitude, and transformed coordinates are not displayed.

2.1 The Quantile Cube

Raw wearable GPS data provide detailed, high-frequency measurements of athlete movement, including velocity, acceleration, and direction. However, direct analysis of these data is challenging due to noise, complexity, and variability across athletes and matches. To address this challenge, we introduce the quantile cube, a novel three-dimensional summary representation that discretizes key movement features into quantiles, capturing their joint distribution over time. Specifically, the quantile cube partitions velocity, acceleration, and movement angle into a structured grid of quantile bins along each dimension, forming a cube-shaped summary that represents how movement intensities and

directions vary throughout a match. This approach enables a clear characterization of the time athletes spend in different types of movements and facilitates robust comparisons and trend detection within and across players and match contexts. To our knowledge, this is the first application of a quantile-based three-dimensional summary framework in sports movement analysis, providing a flexible and interpretable foundation for downstream statistical modeling and inference.

To form the quantile cube, the spline-derived data, containing velocity, acceleration, and angle of movement at ten points per second, was aggregated across all 396 athlete-match-halves. For velocity and acceleration, five bins (0-20th, 20-40th, 40-60th, 60-80th, and 80-100th percentiles) were selected to provide a detailed characterization of movement intensity. The number of bins was selected to provide a compromise between a continuous representation of movement effort over time and having a large enough number of observations within each bin. The corresponding quantiles are shown in Table 1.

Quantile (%)	0%	20%	40%	60%	80%
Velocity (m/s)	0.0100	0.3289	0.9006	1.5026	2.5983
Acceleration (m/s²)	0.0000	0.4220	0.8159	1.2930	2.0367

Table 1: Five quantiles for velocity and acceleration (values shown are raw, prior to the log-transformation).

For the angle of movement, four quantiles (0th, 25th, 50th, and 75th) were computed, starting from a shifted baseline of -30 degrees. This segmentation, illustrated in Figure 2, aligns with the four cardinal movement directions: forward, right, backward, and left. The numeric quantile cut-points for angle are reported in Table 2.

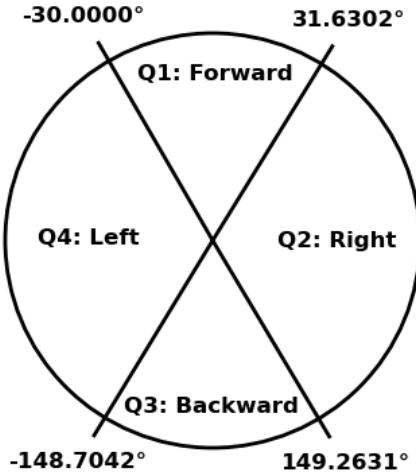


Figure 2: Segmentation of movement angles into four quantiles, starting from a shifted baseline of -30 to align with the four cardinal directions: forward, right, backward, and left.

Quantile (%)	0%	25%	50%	75%
Angle (°)	-30.0000	31.6302	149.2631	-148.7042

Table 2: Four quantiles for movement angle, starting from a shifted baseline of -30°.

Conceptually, the quantile cube acts like a three-dimensional histogram that records how much time an athlete spends at combinations of velocity, acceleration, and movement angle. This provides an intuitive summary of movement style that can then be compared across halves, matches, or players.

Using the defined quantile boundaries derived from the full set of 396 athlete-match-halves, a quantile cube for each half of every athlete’s match was constructed. Each athlete-match-half’s spline-derived data were discretized using these fixed global boundaries, ensuring consistent binning across all sessions. Each dimension of the cube represents one of the key metrics: velocity, acceleration, and angle of movement. Color intensity within each cell indicates the proportion of time the athlete spent in that specific combination of velocity, acceleration, and angle quantiles.

The quantile cube can be visualized as shown in Figure 3, where the inset zooms into a single velocity-acceleration bin to illustrate how the four angle quantiles are further

subdivided within that bin. For example, this figure shows that the largest proportion of time (0.04629) was spent in the highest velocity and acceleration quantiles while turning left. In contrast, the smallest proportion (0.0008) occurred in the lowest velocity quantile and highest acceleration quantile while moving backward. A detailed toy example in Appendix B provides a step-by-step explanation of the quantile cube construction process.

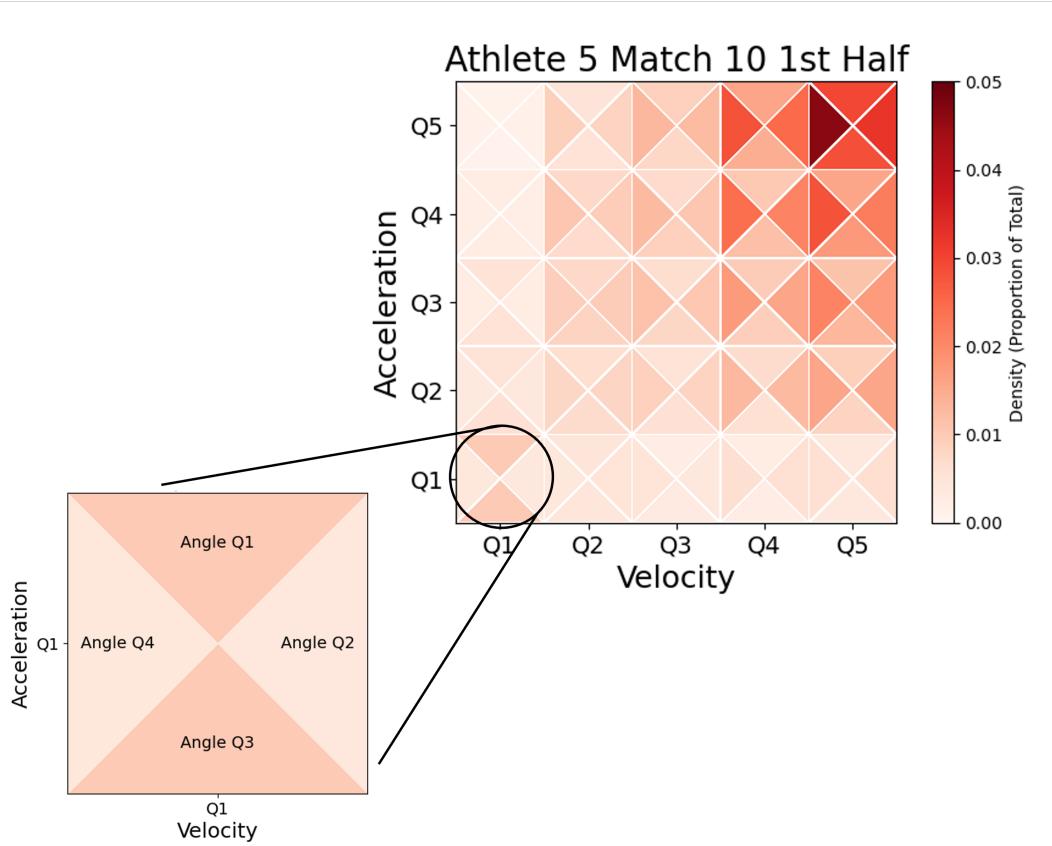


Figure 3: Visual representation of the quantile cube for the density of movements in the first half of Match 10 for Athlete 5. The main plot shows the distribution of movements across velocity (x-axis), acceleration (y-axis), and angle of movement quantiles, with color intensity indicating the proportion of time spent in each bin. The inset zooms into the first quantile for velocity and acceleration, illustrating the subdivision of the four angle quantiles.

The quantile cube can be represented either in deciseconds of time or as proportions of total time spent in each segment. The decisecond representation reflects the absolute time spent in each movement category, whereas the proportional representation captures the athlete’s movement distribution across the velocity-acceleration-angle space. The results are organized into an $n \times d$ matrix \mathbf{Y} , where each of the $n = 396$ rows corresponds to a vectorized quantile cube from an individual athlete-match-half. Each row

contains $d = 100$ features, with each entry indicating the time spent (in deciseconds) within the corresponding movement quantile. The dimensionality ($d = 100$) is defined by the Cartesian product of quantile bins across features: 5 velocity quantiles \times 5 acceleration quantiles \times 4 angle quantiles = 100 total combinations. Each feature therefore corresponds to a unique combination of these bins, capturing the joint distribution of movement intensity and direction. The data preprocessing steps from raw GPS data to the quantile cube representation are summarized in the flowchart provided in Appendix C (Figure 14).

2.2 Covariates

In addition to the GPS data, covariates associated with each match and athlete were obtained, forming an $n \times r$ matrix \mathbf{X} with $n = 396$ rows corresponding to athlete-match-half observations and $r = 13$ covariates. The ten match-level covariates included match ID, the location (home, away, neutral), half (1st or 2nd), result (win, loss, or tie), goals scored at halftime and full time, goals conceded at halftime and full time, and score differential at halftime and full time. The three athlete-level covariates included the athlete ID, position (defender, midfielder, forward), and playing time by half.

3 Methods and Results

Our analysis of the constructed quantile cubes followed a structured three-step pipeline designed to systematically characterize and model athlete movement patterns. In Section 3.1, we quantified differences in movement distributions between the first and second halves of each match for every athlete using the Hellinger distance metric. This step captured temporal changes in external load profiles across match halves. In Section 3.2, Principal Component Analysis (PCA) was applied to the 100-dimensional quantile cube to reduce dimensionality while preserving key variation and to identify dominant movement patterns. Finally, in Section 3.3, Dirichlet-multinomial regression (DMR) was employed to model the probabilistic relationships between movement distributions and relevant covariates, including player position, playing time, and match factors. The following

subsections provide detailed descriptions of the methods and results for each step, and a summary of the notation used is provided in the Notation Table (see Appendix A).

3.1 Quantifying Differences in Movement Distributions Between Match Halves

This section quantifies changes in athletes' movement patterns between the first and second halves of matches by comparing their underlying movement distributions. Due to the high-dimensional and complex nature of the data, classical tests for distributional differences are inappropriate. As illustrated in Figure 4, an athlete may spend a higher proportion of time in the higher velocity and acceleration quantiles during the first half, whereas in the second half, the athlete spends more time in the lower velocity and acceleration quantiles. This example highlights shifts across the entire distribution and motivates the use of a distributional metric to capture these changes.

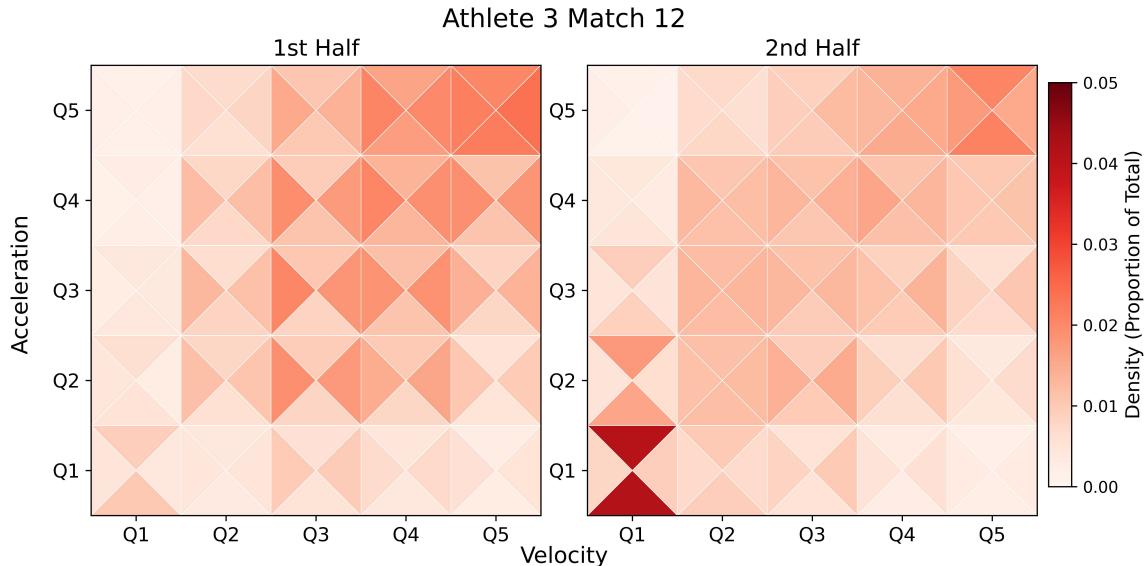


Figure 4: Quantile cubes for Athlete 3 in the first (left) and second (right) halves of Match 12, illustrating shifts in the distribution of velocity and acceleration across the match.

Several existing parametric methodologies are available to measure differences in distributions or means between samples, such as the t -test, ANOVA, and Hotelling's T^2 . However, these tests assume certain conditions, such as common variance, independence, and multivariate normality, which are not satisfied by our data (Casella and Berger, 2002). Moreover, focusing solely on changes in the mean overlooks the full range of

fluctuations contributing, including extreme values that may drive critical changes in external load. High-dimensional statistics addresses scenarios where the number of variables r exceeds the sample size n . For example, Bai and Saranadasa (1996) introduce a high-dimensional two-sample test that adjusts for the breakdown of classical methods under such conditions, and Chen and Qin (2010) developed a test specifically designed for high-dimensional applications such as gene-set testing, where r can be arbitrarily large. Although effective in extreme high-dimensional settings, these approaches still emphasize aggregate changes and may overlook important extreme fluctuations. Therefore, in our moderate-dimensional setting, alternative methodologies are needed to capture the full distribution of movement data, including these extremes.

Given these considerations and the multinomial-nature of the quantile cube data, we propose using the Hellinger distance metric to compare distributions (van der Vaart, 1998, pp.211-212). The Hellinger distance is a true metric for measuring the difference between two probability distributions. If $P = (p_1, \dots, p_d)$ and $Q = (q_1, \dots, q_d)$ are discrete probability distributions defined on the same finite set $\{1, \dots, d\}$, then the Hellinger distance is given by

$$H(P, Q) = \sqrt{\frac{1}{2} \sum_{i=1}^d (\sqrt{p_i} - \sqrt{q_i})^2}. \quad (1)$$

Intuitively, the Hellinger distance provides a single number summarizing how different two distributions are. Values close to zero indicate similar halves, while larger values indicate greater differences.

The Hellinger distance metric offers several advantages over alternative metrics, such as the Kullback-Leibler divergence. Its symmetry and boundedness ($0 \leq H(P, Q) \leq 1$) facilitate interpretable and robust comparisons, and its formulation using square roots makes it particularly suitable for multinomial settings (van der Vaart, 1998, pp.211-212). The square root transformation naturally moderates the influence of variance across categories, downweighting differences arising from high-variance or low-count bins. This variance-adapting property ensures the stability and meaningfulness of our summaries

even when multinomial counts differ, making the Hellinger distance an optimal choice for assessing distributional differences between match halves.

For the analysis, for each athlete $a \in \{1, \dots, 9\}$ and match $m \in \{1, \dots, 23\}$, movement distributions for the first and second halves, denoted by $\hat{p}_{a,m}^{(1)}$ and $\hat{p}_{a,m}^{(2)}$, were estimated. Each $\hat{p}_{a,m}^{(i)}$ is a row of \mathbf{Y} , i.e., a d -dimensional vector of non-negative values summing to one, representing the proportions of time spent in each cell of the quantile cube. Let $P_{a,m}^{(i)}$ denote the underlying probability distribution of $\hat{p}_{a,m}^{(i)}$, corresponding to the quantile cube for athlete a in match m during half $i \in \{1, 2\}$. A formal hypothesis test was applied to determine whether the movement distributions in the first and second halves were statistically equivalent:

- **Null hypothesis (H_0):** The distributions are the same, i.e., $P_{a,m}^{(1)} = P_{a,m}^{(2)}$.
- **Alternative hypothesis (H_1):** The distributions differ, i.e., $P_{a,m}^{(1)} \neq P_{a,m}^{(2)}$.

The observed test statistic for each athlete-match pair was

$$\lambda_{a,m} = H(\hat{p}_{a,m}^{(1)}, \hat{p}_{a,m}^{(2)}).$$

Hypothesis testing was conducted using a resampling procedure based on the Hellinger distance. For each pair (a, m) , let t_1 and t_2 denote the athlete's playing time in deciseconds in the first and second halves, respectively. New count vectors of sizes t_1 and t_2 were simulated by sampling without replacement from the overall movement distribution estimated from all athlete-match-halves, generating simulated first- and second-half samples under the null hypothesis of no distributional difference. For each simulation, the corresponding d -dimensional proportion vectors $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ were calculated from the simulated data, and the Hellinger distance was computed. Repeating this process 10,000 times produced an empirical null distribution of Hellinger distances for each athlete-match pair.

To control the family-wise error rate (FWER) due to multiple comparisons, the Bonferroni correction was applied (Kaltenbach, 2012, p.72) with a significance threshold of $\alpha_a = 0.05/g_a$, where g_a is the number of valid matches played by athlete a . The critical

value $c_{a,m}$ for each athlete-match pair was defined as the $(1-\alpha_a)$ quantile of the null distribution. If $\lambda_{a,m} > c_{a,m}$, the null hypothesis was rejected for that athlete-match pair.

The hypothesis tests showed significant distributional differences for every athlete-match pair, even after applying the conservative Bonferroni correction. All matches exhibited significant differences between first- and second-half movement distributions, with no cases where halves were statistically indistinguishable after multiple comparison adjustment. Figure 5 illustrates these results for Athlete 1, and Figure 6 summarizes the findings across all athletes using the Bonferroni correction.

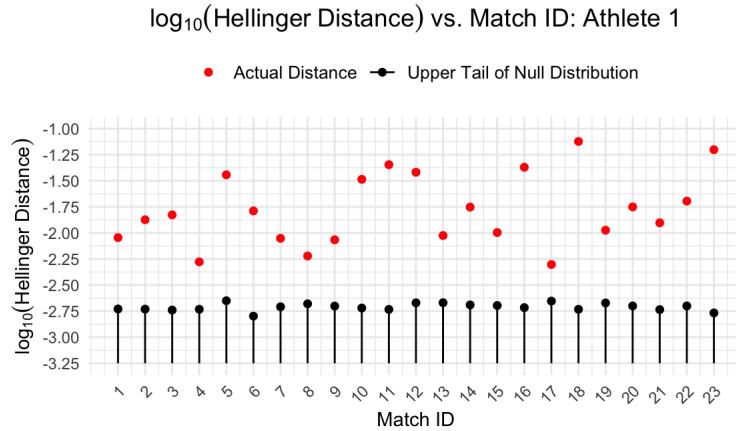


Figure 5: Hellinger distance by match ID between the first and second halves for the actual match data (red circles) and the upper bound of a $((1 - (0.05/23)) \cdot 100)\%$ confidence interval from the null distribution (black circles) for Athlete 1. For all 23 matches, the observed distances exceed the upper bound, indicating that first- and second-half movement distributions differ.

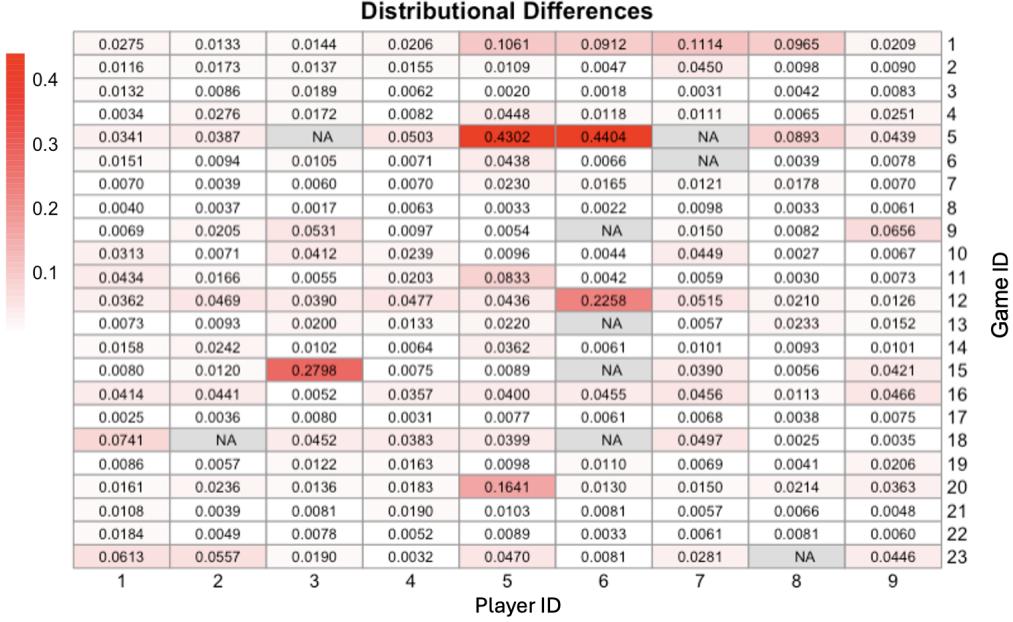


Figure 6: Difference between the Hellinger distance for actual game data and the null distribution at the $(1 - \alpha_a)^{th}$ quantile for all athletes across all matches. NA indicates that the athlete’s playing time did not meet the selection criteria for the match ID. The cells are colored according to the difference value, with larger values (darker red) indicating greater differences between the actual match and null distribution. Exact difference values are also provided.

3.2 Dimensionality Reduction of Movement Patterns in Match Contexts

To identify the dominant patterns underlying athlete movement behaviors and to reduce the computational complexity of the 100-dimensional quantile cube data, PCA was applied to the observed count matrix \mathbf{Y} (defined in Section 2.1). This dimensionality reduction approach allows extraction of the key modes of variation, summarizing movement distributions while providing interpretable insights into match-specific and athlete-specific dynamics (Jolliffe, 2002). The resulting principal components (PCs) offer a structured framework for detecting anomalous movement patterns, distinctive match characteristics, and systematic variations in external load profiles across different competitive contexts.

PCA decomposition of \mathbf{Y} yielded 100 PCs, each linked to a 396-dimensional score vector representing the projection of individual athlete-match-halves onto the component space. To determine the optimal number of components for downstream analysis, a cutoff of 90% variance explained was applied, resulting in the retention of the first seven PCs (Figure 7). This criterion ensures that the reduced representation captures the majority

of systematic variation while minimizing noise and redundant information from lower-variance components.

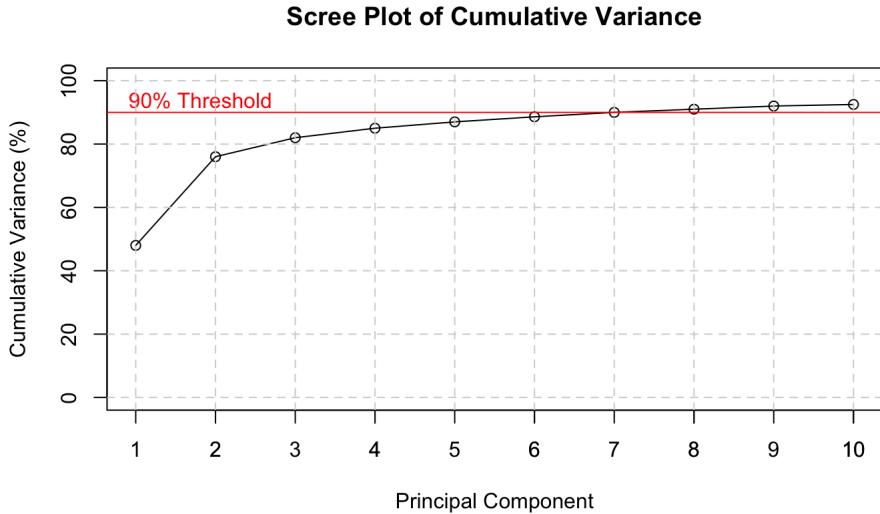


Figure 7: Cumulative variance explained by the top 10 PCs from the quantile cube analysis. The decline in explained variance after the 7th component supports the selection of a lower-dimensional representation for downstream analysis.

However, variance explained alone does not guarantee the practical interpretability or actionable relevance of the PCs for understanding athlete performance. To assess their meaningfulness, component scores were plotted against match- and athlete-level characteristics from the design matrix \mathbf{X} (defined in Section 2.2). While second principal component (PC2) and third principal component (PC3) explained substantial variance, their loadings were diffusely distributed across velocity-acceleration-angle bins, showing no coherent patterns linked to specific covariates. Scatterplots confirmed no discernible clustering or separation, indicating these components primarily capture subtle, distributed variations rather than systematic behavioral differences. In contrast, first principal component (PC1) and fourth principal component (PC4) demonstrated clear interpretability, exhibiting clustering patterns directly associated with specific matches and movement dynamics. Therefore, only PC1 and PC4 are presented here as they provided the most actionable insights into external load variations.

PC1, accounting for the largest proportion of variance, captured distinctive movement characteristics observed during the second half of Match 1, with loadings revealing a sys-

tematic reduction in time spent in the middle quantiles of velocity and acceleration. This indicates a shift toward more polarized movement patterns, characterized by either low-intensity positioning or high-intensity bursts, with less time spent in moderate-intensity activities. Table 3 presents the ten highest-magnitude loadings for PC1, dominated by features combining the third velocity quantile (moderate running speeds) with the fourth acceleration quantile (high acceleration) across forward and backward movement directions (first and third angle quantiles). This negative loading pattern reflects reduced time spent in movement categories requiring moderate velocity paired with high acceleration. The uniqueness of this pattern is further confirmed by comparing PC1 scores across all athlete-match-halves: Figure 8 shows that observations from Match 1’s second half were systematically more negative relative to the overall distribution, highlighting a significant deviation from typical movement patterns observed throughout the season.

PC4 captured the distinctive movement characteristics of Match 23, the final match of the season. Table 4 presents the component’s highest-magnitude loadings, characterized by combinations of low velocity (first and second quantiles) with maximal acceleration (fifth quantile), particularly in the forward and backward directions. These positive loadings indicate increased time spent in low-velocity, high-acceleration movements throughout the whole match. Figure 9 confirms that PC4 scores for Match 23 were systematically elevated relative to the season-long distribution, highlighting the uniqueness of the final match’s movement patterns.

	Variable	PC1 Loading		Variable	PC4 Loading
1	Q3_vel_Q4.acc.Q3.angle	-0.1306	1	Q1.vel.Q5.acc.Q1.angle	0.3297
2	Q3.vel.Q4.acc.Q1.angle	-0.1300	2	Q1.vel.Q5.acc.Q3.angle	0.3179
3	Q4.vel.Q3.acc.Q3.angle	-0.1293	3	Q2.vel.Q5.acc.Q1.angle	0.2215
4	Q4.vel.Q3.acc.Q1.angle	-0.1290	4	Q2.vel.Q5.acc.Q3.angle	0.2111
5	Q4.vel.Q4.acc.Q1.angle	-0.1273	5	Q1.vel.Q5.acc.Q4.angle	0.2053
6	Q4.vel.Q4.acc.Q3.angle	-0.1271	6	Q1.vel.Q5.acc.Q2.angle	0.2022
7	Q4.vel.Q2.acc.Q4.angle	-0.1253	7	Q4.vel.Q5.acc.Q3.angle	0.1734
8	Q4.vel.Q2.acc.Q1.angle	-0.1253	8	Q3.vel.Q5.acc.Q3.angle	0.1709
9	Q2.vel.Q5.acc.Q2.angle	-0.1253	9	Q3.vel.Q5.acc.Q1.angle	0.1692
10	Q3.vel.Q3.acc.Q3.angle	-0.1253	10	Q4.vel.Q5.acc.Q1.angle	0.1642

Table 3: Top 10 absolute loadings for PC1

Table 4: Top 10 absolute loadings for PC4

PCA successfully reduced the dimensionality from 100 to 7 components, retaining

90% of the variance in the quantile cube data. Among these, PC1 and PC4 showed the strongest associations with specific matches and interpretable movement patterns, with PC1 reflecting the distinctive characteristics of Match 1 and PC4 reflecting those of Match 23.

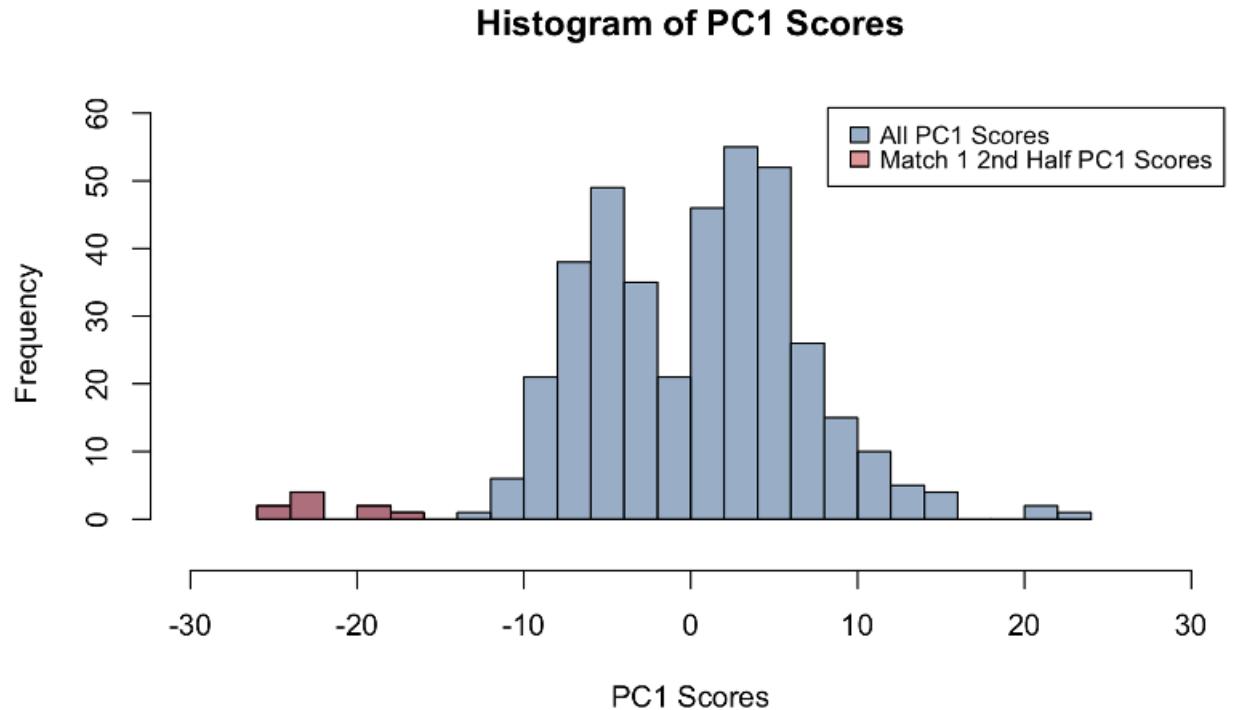


Figure 8: Histogram comparing the distribution of PC1 scores for all athlete-match-halves (blue) versus the second half of Match 1 (red). Scores from the second half of Match 1 are shifted toward more negative values relative to the overall distribution, indicating reduced time spent in the quantiles most strongly associated with this component during that period.

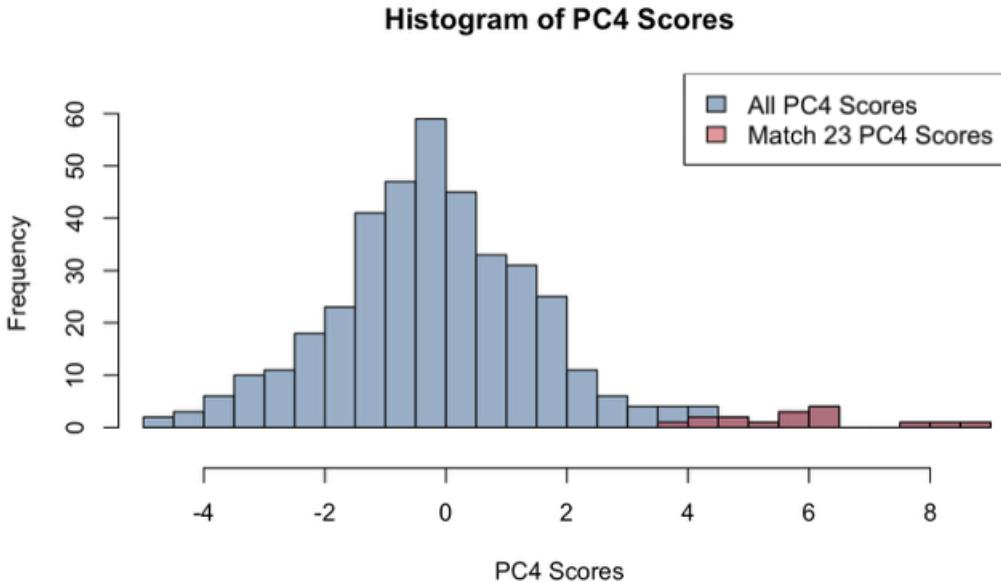


Figure 9: Histogram comparing the distribution of PC4 scores for all athlete-match-halves (blue) versus Match 23 (red). The Match 23 PC4 scores are shifted toward higher values relative to the overall distribution, suggesting this match involved increased time spent in low-velocity, high-acceleration quantiles associated with this component.

3.3 Modeling Movement Distributions as a Function of Player and Match Characteristics

Having established systematic distributional differences between match halves and identified key movement patterns through dimensionality reduction, the next step was to quantify how these movement distributions varied as a function of player and match characteristics using formal statistical modeling. To characterize the relationships between athlete movement distributions and contextual factors, DMR, a flexible modeling framework specifically designed for compositional count data with overdispersion, was employed. The quantile cube data exhibit two key features that necessitate this specialized approach: (1) compositional dependence, where time allocated to one movement category directly constrains time available for others; and (2) overdispersion, where the observed variance in movement category counts exceeds the variance predicted by standard multinomial models due to individual athlete differences, match-specific contexts, and temporal clustering effects.

The Dirichlet-multinomial (DM) distribution addresses both features by modeling

the underlying category probabilities as random draws from a Dirichlet distribution, naturally accommodating the compositional constraints while allowing greater variance than multinomial models (Mosimann, 1962; Chen and Li, 2013). The DMR framework extends this model to a regression setting, enabling systematic incorporation of external covariates such as match half, player position, and playing time. This approach provides a principled method to quantify how movement patterns vary with contextual factors while respecting the inherent structure of compositional movement data.

3.3.1 Model Specification

For each observation (row) i in the count matrix \mathbf{Y} , the movement distribution $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{id})$ was modeled using the DM distribution:

$$f_{DM}(\mathbf{y}_i|\boldsymbol{\eta}) = \frac{\Gamma(N_i + 1)\Gamma(\sum_{j=1}^d \eta_j)}{\Gamma(N_i + \sum_{j=1}^d \eta_j) \prod_{j=1}^d \frac{\Gamma(y_{ij} + \eta_j)}{\Gamma(\eta_j)\Gamma(y_{ij} + 1)}}. \quad (2)$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_d)$ are positive concentration parameters for each movement category and $N_i = \sum_{j=1}^d y_{ij}$ is the total movement time for athlete-match-half i (Mosimann, 1962; Chen and Li, 2013). These η_j are unknown parameters of the model to be estimated from the observed count data. Intuitively, they control the expected proportions and variability of each movement category.

To incorporate covariate effects, each concentration parameter η_j for movement category $j \in \{1, \dots, d\}$ was modeled using a log-linear formulation:

$$\eta_j := \exp \left(\beta_{j0} + \sum_{k=1}^r \beta_{jk} x_{ik} \right), \quad (3)$$

where β_{j0} represents the baseline for movement category j , β_{jk} quantifies the effect of covariate k on category j , and x_{ik} denotes the value of covariate k for observation i (Mosimann, 1962; Chen and Li, 2013). This formulation makes the role of η_j explicit, modeling them as functions of covariates rather than fixed values. This ensures positivity while

allowing flexible, interpretable relationships between covariates and movement patterns across all $d = 100$ quantile cube categories. Put simply, this regression framework allows us to see how different factors, such player position, match result, or playing time, affect the overall distribution of an athlete’s movements, while accounting for the fact that time spent in one type of movement limits time available for others and that real data are more variable than a simple model would assume.

3.3.2 Covariate Encoding and Selection

All covariates in the design matrix \mathbf{X} were included as potential predictors in the regression model. Categorical variables were encoded using standard dummy-variable approaches: binary factors (e.g., match half) as 0/1 indicators, and multi-level factors (e.g., athlete position, match location) with one reference level omitted. Playing time was mean-centered and log-transformed to improve interpretability and stabilize estimation.

All candidate models were estimated using the **MGLMfit** function from the MGLM package in R (Zhang et al., 2017; Zhang and Zhou, 2022), which applies maximum likelihood estimation to obtain the concentration parameters η_j through the log-linear regression coefficients β_{jk} in Equation 3. Candidate models were systematically compared, testing different combinations of covariates. The optimal model, selected for both statistical performance and interpretability, included three key predictors: match half ($1^{st}, 2^{nd}$), player position (defender, midfielder, forward), and mean-centered log(playing time). This specification allowed the detection of systematic changes in movement patterns across these primary contextual factors while maintaining model simplicity.

3.3.3 Model Results

Parameter estimation identified significant covariate effects across multiple movement categories. Coefficients from Equation 3 were considered statistically significant when $|\beta_{jk}/SE_{\beta_{jk}}| > 3$, where $SE_{\beta_{jk}}$ is the standard error of β_{jk} , computed from the observed Fisher information matrix provided by **MGLMfit** (Zhang et al., 2017; Zhang and Zhou, 2022). This threshold was determined based on the empirical distribution of standardized

coefficients in the dataset, offering a conservative approach to highlight meaningful effects.

Match half and playing time effects (Figure 10) revealed systematic changes in movement patterns. During the second half, more negative coefficients dominated the higher velocity quantiles across acceleration and angle categories, indicating decreased time spent in high-intensity movement patterns as matches progressed, consistent across all movement directions. Additionally, athletes with above-average playing time exhibited distinct movement signatures compared to those with shorter durations, spending significantly less time in the lowest velocity and acceleration quantiles, particularly in forward and backward directions, suggesting that longer-playing athletes maintain higher baseline activity levels throughout the match.

Positional differences exhibited clear and interpretable patterns (Figure 11). Compared to defenders (reference category), midfielders spent more time in lower velocity and acceleration quantiles, combined with elevated time in the highest velocity quantile across all acceleration levels. This bimodal pattern suggests midfielders alternate between periods of lower-intensity positioning and high-intensity running. Forwards demonstrated a contrasting pattern, spending less time in middle and lower velocity quantiles across most acceleration categories, with one notable exception in the first velocity quantile at maximum acceleration. Additionally, forwards exhibited significantly reduced left-right movement compared to forward-backward movement relative to defenders.

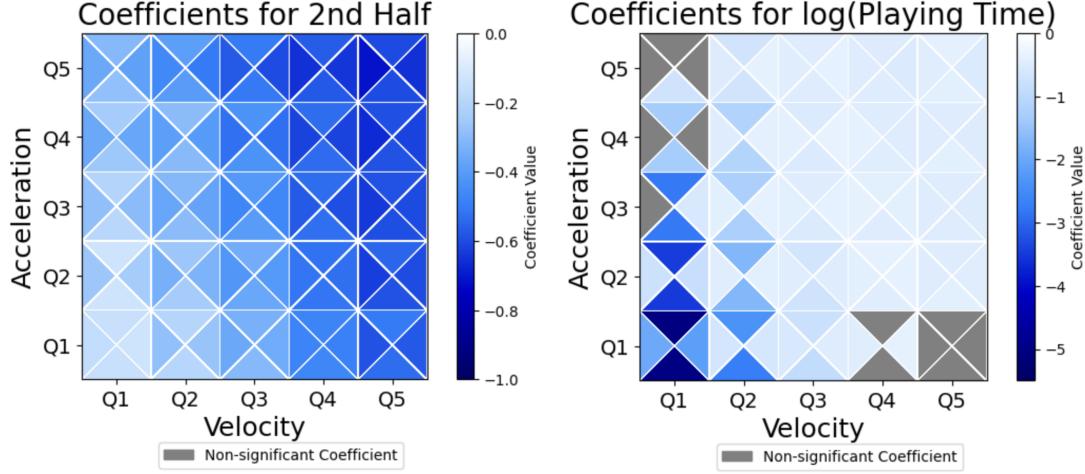


Figure 10: Quantile cube illustrating the DMR coefficients for the second half (left) and $\log(\text{Playing Time})$ (right). Non-significant coefficients are shown in gray, while significant coefficients are color-coded based on effect size, with intensity indicating magnitude.

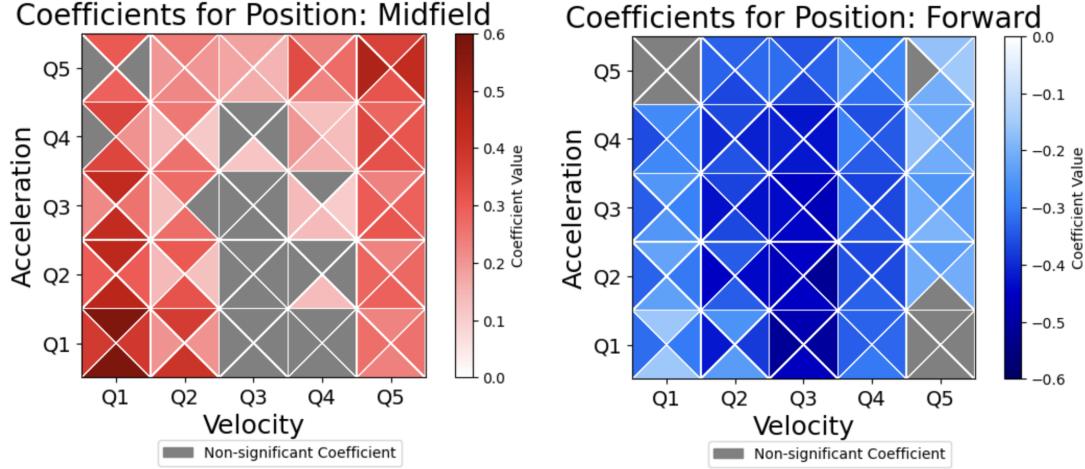


Figure 11: Quantile cube depicting the DMR coefficients for player position, where the defender is treated as the base class. Coefficients for comparisons with midfielders are shown on the left, and those for forwards are shown on the right. Non-significant coefficients are shaded in gray, while significant coefficients are color-coded based on effect size and direction: blue for negative effects and red for positive effects, with intensity reflecting magnitude.

4 Discussion

This paper presents a novel adaptation of established methodologies for assessing external load in elite female soccer athletes, with a focus on improving the interpretability of movement patterns during match play. By leveraging wearable GPS data, we examined

the relationships between velocity, acceleration, and angle of movement with athlete and match characteristics. Our findings demonstrate how a probabilistic, distribution-based analytic framework can uncover patterns that conventional metrics overlook, providing new insights into performance and fatigue in elite women's soccer.

Our analysis revealed significant differences in athlete movement patterns between the first and second halves of matches. Specifically, within our sample of elite women's soccer athletes, quantile cube distributions differed significantly between halves, reinforcing the influence of match duration on movement dynamics. This finding aligns with previous research, such as Barrera et al. (2021), which reported reductions in external load metrics like high-speed running during the second half of professional male soccer matches, and uses a distributional lens to extend Barrera et al.'s conclusions to women's soccer. These results suggest that, in this dataset, second-half differences were not confined to one or two performance metrics but reflected a broader reshaping of movement intensity profiles that may be driven by neuromuscular fatigue or tactical decisions.

Additionally, the use of PCA to reduce the dimensionality of the quantile cubes revealed context-specific deviations across the season. Most notably, the second half of Match 1 (season opener) and Match 23 (postseason tournament match) displayed movement profiles distinct from typical seasonal patterns. Match 1 did not involve a major rival and may reflect early-season conditioning rather than tactical pressure, whereas Match 23, a decisive postseason loss, probably reflects accumulated fatigue combined with heightened match intensity. These deviations point to contextual factors such as early-season conditioning, late-season fatigue, or tactical adaptations. This highlights the importance of considering both temporal (within-match) and contextual (seasonal, environmental, or competitive) factors in workload analysis. For instance, environmental conditions like moderate altitude reduce high-intensity running and overall distance in collegiate female soccer matches (Bohner et al., 2015), while workload demands vary between in-conference and out-of-conference play (Bozzini et al., 2020). Such findings underscore the value of incorporating context directly within analytic frameworks rather than relying on fixed reference values.

The DMR model confirmed substantial differences between first and second halves, with athletes spending less time in the higher quantiles of both velocity and acceleration as the match progressed, providing statistical evidence of second-half intensity decline. This pattern likely reflects acute fatigue or strategic pacing (Snyder et al., 2024; Andersson et al., 2008; Barrera et al., 2021). Importantly, athletes with greater playing time per half spent less time in the lower velocity and acceleration bins, suggesting these players maintain a higher baseline workload despite extended minutes, an encouraging indicator for endurance and load management strategies.

The coefficients of the DMR also revealed positional differences for midfielders and forwards in relation to defenders. Midfielders exhibited a bimodal load profile, alternating between low-intensity positioning and high-intensity bursts. This dynamic role is consistent with prior work in women's and men's soccer showing that midfielders cover greater total distance and wider intensity ranges than defenders (Vescovi and Favero, 2014; Wehbe et al., 2014; Panduro et al., 2022). Forwards, on the other hand, spent significantly less time in middle velocity quantiles across accelerations, particularly with lateral movements, supporting their role in alternating between recovery and forward-directed sprints. Prior research in men's soccer has shown that defenders perform fewer high-speed runs but sustain notable acceleration demands (Wehbe et al., 2014), highlighting the importance of velocity–acceleration metrics and the limitations of relying solely on distance- or sprint-based measures.

When interpreting our findings, it is essential to consider contextual and competition level factors demonstrated in related research. Seasonal and postseason workload variations indicate changes in running intensity and volume that can impact athlete performance and fatigue (Wells et al., 2015). Competition at international levels presents higher demands in high-speed running and sprints, particularly affecting midfielders and defenders, further supporting the necessity of individualized training and monitoring strategies (Mara et al., 2017; Griffin et al., 2021; Datson et al., 2017). These considerations, alongside environmental and individual recovery factors, emphasize the complexity of athlete monitoring while highlighting opportunities for integrated, multivariate analytics.

4.1 Practical Implications

Taken together, these results have important practical implications for athlete monitoring and management. Our work introduces an accessible statistical method that transforms raw GPS measurements into probabilistic insights about athlete movement patterns. The DMR model allows integration of positional and match factors, as well as environmental or physiological covariates, enabling predictions of movement responses under varying conditions. This facilitates evidence-based decisions about training load, recovery, and substitution.

Importantly, we believe these insights can translate directly into future practical applications for coaching and athlete management:

- **Enhanced substitution strategies:** Coaches can use probabilistic movement profiles generated by the model to identify players whose movement patterns deviate significantly from their baseline in real time, signaling acute fatigue or injury risk before overt signs appear.
- **Tailored training prescriptions:** Training drills can be informed by quantified positional movement demands. For example, midfielders exhibiting wide velocity and acceleration distributions may benefit from conditioning emphasizing both endurance and explosive speed, while defenders might focus on drills stressing acceleration bursts and recovery.
- **Context-aware workload management:** Understanding how environmental and match-specific factors affect workload allows practitioners to adjust training intensity and recovery strategies accordingly. For instance, if altitude reduces high-intensity running capacity as suggested by prior research, training loads can be moderated before competitions at such venues (Bohner et al., 2015).
- **Integrated internal and external load monitoring:** By combining movement distributions with athlete wellness data (e.g., soreness, sleep quality), sports science teams can implement individualized recovery protocols and readiness assessments, ultimately promoting injury prevention and sustainable performance.

Building on these applied scenarios, implementing this methodology in real time involves integrating continuous GPS data streams from training and matches, establishing player-specific baselines, and detecting statistically significant deviations from typical movement profiles. This approach elevates GPS monitoring from simple tracking to a rigorous statistical tool, where decisions about player management are guided by formal inference rather than arbitrary thresholds. In practice, it enables sports science teams to deploy dashboards or automated alert systems that notify coaching staff of atypical movement patterns warranting intervention, supporting proactive strategies such as substitution, load adjustment, or medical evaluation to better safeguard athlete health and performance.

5 Conclusion

This study provides a novel, probabilistic framework that advances the analysis of external load in women’s soccer. By introducing the quantile cube approach combined with Hellinger distance, PCA, and DMR, we demonstrate how complex GPS trajectories can be summarized into interpretable, distribution-based measures of movement. These methods capture nuanced variation and enable individualized inference, offering richer insights than traditional aggregate or threshold-based metrics.

Importantly, our female-specific dataset offers a unique perspective that addresses a critical gap in existing research, which predominantly focuses on male athletes. This study responds to longstanding calls for gender-specific workload analytics and contextual interpretations aimed at enhancing training strategies and reducing injury risk (Mujika et al., 2009). Disparities in physical fitness and performance capacities between genders and competitive levels are well documented, further underscoring the necessity of female-specific data and analytic frameworks like the ones developed here to support effective and tailored workload management.

Despite these strengths, several limitations should be noted. First, our sample size was modest (nine athletes from a single team over one season), limiting the generalizability of findings. The exclusion of wide-vs-central positional subdivisions and goalkeepers

further restricts the applicability of findings. Taking into account these positions as separate categories may reveal additional differences in movement patterns, as studies have found that in male soccer athletes, wide position players typically produce higher acceleration efforts than central position players (Ingebrigtsen et al., 2015). Additionally, match session selection was based on a threshold of at least 25 minutes per half; while this ensures data quality, it may introduce selection bias by excluding shorter substitution stints and atypical playing patterns.

Our analysis did not include training sessions, recovery protocols, or internal load factors such as heart rate, perceived exertion, or biochemical markers—key components of comprehensive athlete monitoring. Further, important contextual influences such as weather conditions, opposition strength, fixture congestion, match importance (postseason vs. regular season), and pitch quality were not included in our models, despite evidence that these factors influence external workload. Future research should seek to incorporate altitude and match-type effects (Bohner et al., 2015; Bozzini et al., 2020), as these have been shown to affect high-intensity work and overall match demands.

Technological limitations must also be considered: while GPS devices provide accurate tracking of most movement patterns, they may underperform when capturing abrupt or highly multidirectional changes. The addition of inertial measurement units (IMUs)—which integrate accelerometers and gyroscopes—could enable high-frequency, real-time data collection for more precise analysis of rapid, multidimensional movements that traditional GPS devices may overlook (Mudeng et al., 2022). Advances in monitoring technology now allow measurement of physiological variables (such as heart rate, heart rate variability, and neuromuscular fatigue markers), which would further enrich future workload analyses.

Finally, the selection and binning of quantiles in the quantile cube framework, while aiming for interpretability and robustness, are somewhat arbitrary and may need tuning for other teams or application scenarios. Future research should employ larger, multisite cohorts, include training and recovery data, integrate contextual and physiological covariates, and validate these models longitudinally—including prospective injury and recovery

outcomes—to maximize translational relevance for athlete health and performance.

In conclusion, this study demonstrates that a multidimensional, quantile-based approach enhances interpretability, statistical rigor, and practical utility of GPS-derived external load in women's soccer. By bridging wearable technology, probabilistic modeling, and applied sports science, we establish a methodological foundation for individualized, data-driven athlete management aimed at optimizing performance and reducing injury risk. Continued research should further validate and refine these approaches across larger and more diverse cohorts, linking external load patterns with internal physiology, recovery, contextual variables, and long-term outcomes, to maximize their translational impact on athlete health and performance.

Acknowledgements

We extend our gratitude to Elena Cantu, Sam Moore, and Dr. Abbie Smith-Ryan from the Applied Physiology Lab in the University of North Carolina at Chapel Hill's Department of Exercise and Sport Science for their generous contributions in gathering and providing access to the data.

Funding

Jan Hannig's research was supported in part by the National Science Foundation under Grant No. DMS-1916115, 2113404, and 2210337.

References

- Andersson, H., Raastad, T., Nilsson, J., Paulsen, G., Garthe, I. and Kadi, F. (2008), 'Neuromuscular fatigue and recovery in elite female soccer: effects of active recovery', *Medicine & Science in Sports & Exercise* **40**(2), 372–380.
- Bai, Z. and Saranadasa, H. (1996), 'Effect of high dimension: By an example of a two-sample problem', *Statistica Sinica* **6**, 311–329.

Barrera, J., Sarmento, H., Clemente, F. M., Field, A. and Figueiredo, A. J. (2021), ‘The effect of contextual variables on match performance across different playing positions in professional portuguese soccer players’, *International Journal of Environmental Research and Public Health* **18**, 5175.

Bohner, J. D., Hoffman, J. R., McCormack, W. P., Scanlon, T. C., Townsend, J. R., Stout, J. R., Fragala, M. S. and Fukuda, D. H. (2015), ‘Moderate altitude affects high intensity running performance in a collegiate women’s soccer game’, *Journal of Human Kinetics* **47**, 147–154.

Bourdon, P. C., Cardinale, M., Murray, A., Gastin, P., Kellmann, M., Varley, M. C., Gabbett, T. J., Coutts, A. J., Burgess, D. J., Gregson, W. and Cable, N. T. (2017), ‘Monitoring athlete training loads: Consensus statement’, *International Journal of Sports Physiology and Performance* **12**(Suppl 2), S2161–S2170.

Bozzini, B. N., McFadden, B. A., Walker, A. J. and Arent, S. M. (2020), ‘Varying demands and quality of play between in-conference and out-of-conference games in division i collegiate women’s soccer’, *Journal of Strength and Conditioning Research* **34**(12), 3364–3368.

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Thomson Learning Inc. Citing Chapters 8 and 11.

Chen, J. and Li, H. (2013), ‘Variable selection for sparse dirichlet-multinomial regression with an application to microbiome data analysis’, *The Annals of Applied Statistics* **7**(1).

Chen, S. X. and Qin, Y.-L. (2010), ‘A two-sample test for high-dimensional data with applications to gene-set testing’, *The Annals of Statistics* **38**(2), 808–835.

Cummins, C., Orr, R., O’Connor, H. and West, C. (2013), ‘Global positioning systems (gps) and microtechnology sensors in team sports: A systematic review’, *Sports Medicine* **43**(10), 1025–1042.

- Datson, N., Drust, B., Weston, M., Jarman, I. H., Lisboa, P. J. and Gregson, W. (2017), ‘Match physical performance of elite female soccer players during international competition’, *Journal of Strength and Conditioning Research* **31**(9), 2379–2387.
- De Lucia, B. J., Feit, M. K., Fillbach, A. and Fields, J. B. (2024), ‘Comparison of game external loads across a men’s and women’s soccer season’, *Medicine & Science in Sports & Exercise* **56**(10S), 198.
- Ferraz, A., Duarte-Mendes, P., Sarmento, H., Valente-Dos-Santos, J. and Travassos, B. (2023), ‘Tracking devices and physical performance analysis in team sports: a comprehensive framework for research—trends and future directions’, *Frontiers in Sports and Active Living* **5**.
- Gailor, M., Rimer, E., King, K. and Stamatis, A. (2024), ‘Performance metrics and game outcomes in division i women’s soccer: A sport analytics case study’, *Medicine & Science in Sports & Exercise* **56**(10S), 198–199.
- Google Maps API (2025), ‘Google maps platform documentation’, <https://developers.google.com/maps/documentation>. [Accessed: January 14, 2025].
- Griffin, J., Newans, T., Horan, S., Keogh, J., Andreatta, M. and Minahan, C. (2021), ‘Acceleration and high-speed running profiles of women’s international and domestic football matches’, *Frontiers in Sports and Active Living* **3**.
- URL:** <https://www.frontiersin.org/articles/10.3389/fspor.2021.604605>
- Ingebrigtsen, J., Dalen, T., Hjelde, G. H., Drust, B. and Wisløff, U. (2015), ‘Acceleration and sprint profiles of a professional elite football team in match play’, *European Journal of Sport Science* **15**(2), 101–110.
- Jolliffe, I. T. (2002), *Principal Component Analysis*, Springer.
- Kaltenbach, H.-M. (2012), *A Concise Guide to Statistics*, SpringerBriefs in Statistics, Springer, Heidelberg, Dordrecht, London, New York.

- Kuhlman, N. M., Jagim, A. R., Jones, M. T., Feit, M. K. and Fields, J. B. (2025), ‘A comparison of match external load demands across women’s collegiate field sports’, *Journal of Strength and Conditioning Research* **39**(2), 234–241. Epub 2024 Oct 24.
- Luo, L. and Song, P. X.-K. (2020), ‘Renewable estimation and incremental inference in generalized linear models with streaming data sets’, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**(1), 69–97.
- Luo, L. and Song, P. X.-K. (2023), ‘Multivariate online regression analysis with heterogeneous streaming data’, *Canadian Journal of Statistics* **51**(1), 111–133.
- Luo, L., Wang, J. and Hector, E. C. (2023), ‘Statistical inference for streamed longitudinal data’, *Biometrika* **110**(4), 841–858.
- Mara, J. K., Thompson, K. G., Pumpa, K. L. and Morgan, S. (2017), ‘Quantifying the high-speed running and sprinting profiles of elite female soccer players during competitive matches using an optical player tracking system’, *Journal of Strength and Conditioning Research* **31**(6), 1500–1508.
- Mosimann, J. E. (1962), ‘On the compound multinomial distribution, the multivariate β distribution, and correlations among proportions’, *Biometrika* **49**(1/2), 65–82. Publisher: [Oxford University Press, Biometrika Trust].
- Mudeng, V., Hakim, I. M., Suprapto, S. S. and Choe, S. W. (2022), ‘An alternative athlete monitoring system using cost-effective inertial sensing instrumentation’, *Journal of Electrical Engineering & Technology* **17**(6), 3581–3592. Epub 2022 Sep 26.
- Mujika, I., Santisteban, J., Impellizzeri, F. M. and Castagna, C. (2009), ‘Fitness determinants of success in men’s and women’s football’, *Journal of Sports Sciences* **27**(2), 107–114.
- Panduro, J., Ermidis, G., Røddik, L. et al. (2022), ‘Physical performance and loading for six playing positions in elite female football: full-game, end-game, and peak periods’, *Scandinavian Journal of Medicine & Science in Sports* **32**(Suppl. 1), 115–126.

Pebesma, E. (2018), ‘Simple features for r: Standardized support for spatial vector data’, *The R Journal* **10**(1), 439–446.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernàndez, J. and Medina, D. (2018), ‘Effective injury forecasting in soccer with GPS training data and machine learning’, *PLOS ONE* **13**(7).

Snyder, B. J., Maung-Maung, C. and Whitacre, C. (2024), ‘Indicators of fatigue during a soccer match simulation using gps-derived workload values: Which metrics are most useful?’, *Sports* **12**(1), 9.

Vallance, E., Sutton-Charani, N., Imoussaten, A., Montmain, J. and Perrey, S. (2020), ‘Combining Internal- and External-Training-Loads to Predict Non-Contact Injuries in Soccer’, *Applied Sciences* **10**(15).

van der Vaart, A. W. (1998), *Asymptotic Statistics*, 1 edn, Cambridge University Press.
URL: <https://www.cambridge.org/core/product/identifier/9780511802256/type/book>

Vescovi, J. D. and Favero, T. G. (2014), ‘Motion characteristics of women’s college soccer matches: Female athletes in motion (faim) study’, *International Journal of Sports Physiology and Performance* **9**(3), 405–414.

Wehbe, G. M., Hartwig, T. B. and Duncan, C. S. (2014), ‘Movement analysis of australian national league soccer players using global positioning system technology’, *Journal of Strength and Conditioning Research* **28**(3), 834–842.

Wells, A. J., Hoffman, J. R., Beyer, K. S., Hoffman, M. W., Jajtner, A. R., Fukuda, D. H. and Stout, J. R. (2015), ‘Regular- and postseason comparisons of playing time and measures of running performance in ncaa division i women soccer players’, *Applied Physiology, Nutrition, and Metabolism* **40**(9), 907–917. Epub 2015 May 6.

Zhang, Y., Zhou, H., Zhou, J. and Sun, W. (2017), ‘Regression models for multivariate count data’, *Journal of Computational and Graphical Statistics* **26**(1), 1–13.

URL: <http://dx.doi.org/10.1080/10618600.2016.1154063>

Zhang, Y. and Zhou, Y. (2022), *MGLM: Multivariate Response Generalized Linear Models*. R package version 0.2.1.

URL: <https://CRAN.R-project.org/package=MGLM>

A Notation Table

Symbol	Description
n	Number of observations (athlete-match-halves), $n = 396$
d	Dimension of the quantile cube vector, $d = 5 \times 5 \times 4 = 100$
r	Number of covariates in the regression model, $r = 13$
i	Index for observation $i = 1, \dots, n$
a, m, h	Athlete ID ($a \in \{1, \dots, 9\}$), Match ID ($m \in \{1, \dots, 23\}$), and Half ($h \in \{1, 2\}$)
t_i	Playing time in deciseconds for observation i
$\mathbf{X} = (x_{ik})_{n \times p}$	Design matrix of covariates
x_{ik}	Value of covariate k for observation i
k	Index for covariate, $k = 1, \dots, 13$
$\mathbf{Y} = (y_{ij})_{n \times d}$	Observed count matrix of quantile cube vectors
y_{ij}	Time (in deciseconds) spent in quantile bin j by observation i
j	Index for quantile cube bins, $j = 1, \dots, 100$
$H(P, Q)$	Hellinger distance between two discrete distributions P and Q
$\hat{p}_{a,m}^{(1)}, \hat{p}_{a,m}^{(2)}$	Empirical distributions from the quantile cube for first and second halves
$\lambda_{a,m}$	Observed Hellinger distance between halves for athlete a in match m
$c_{a,m}$	Bonferroni-corrected critical value (threshold) from the empirical null distribution for athlete a in match m used in the Hellinger distance test
π_j	Probability of occupying quantile bin j in DMR model
$\boldsymbol{\pi} = (\pi_1, \dots, \pi_d)$	Vector of movement probabilities over quantile bins
η_j	Dirichlet concentration parameter for quantile bin j
$\boldsymbol{\eta} = (\eta_1, \dots, \eta_d)$	Vector of Dirichlet concentration parameters
β_{j0}	Intercept for bin j in the DMR model
β_{jk}	Coefficient for covariate k on bin j

Table 5: Summary of notation used throughout the paper.

B Toy Example of a Quantile Cube

We present a toy example of a quantile cube to demonstrate the formation in a clear and manageable way. In this example, we use two quantiles for velocity, two quantiles for acceleration, and four quantiles for angle. This simplified version is intentionally smaller than the full quantile cube described in the main text of the paper. By reducing the number of quantiles and points, we provide a concrete, easy-to-follow example that illustrates how raw data are mapped into the quantile cube, how counts are aggregated within each bin, and how proportion vectors are derived. This approach allows readers to gain intuition about the process without being overwhelmed by the complexity of a full-scale dataset, while still demonstrating all the key steps of the quantile cube methodology.

We illustrate how five example points, listed in the first three columns of Table 6, are mapped into this simplified quantile cube. Each point includes three features: velocity (v), acceleration (a), and angle (θ). The quantile ranges, summarized in the table in Figure 12, are based on a larger theoretical dataset, and the figure also provides a visual legend for the four angle quantiles. The last three columns of Table 6 provide the corresponding quantile assignment for each feature of every point.

Feature	Quantile	Definition
Velocity (v)	Q1	$\leq 3 \text{ m/s}$
	Q2	$> 3 \text{ m/s}$
Acceleration (a)	Q1	$\leq 1 \text{ m/s}^2$
	Q2	$> 1 \text{ m/s}^2$
Angle (θ)	Q1	$-45^\circ \text{ to } 45^\circ$
	Q2	$45^\circ \text{ to } 135^\circ$
	Q3	$135^\circ \text{ to } -135^\circ$
	Q4	$-135^\circ \text{ to } -45^\circ$

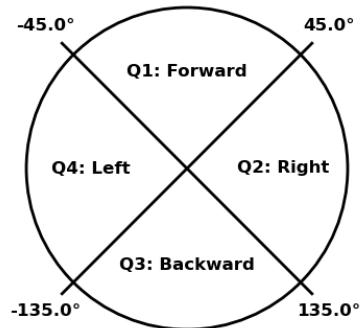


Figure 12: Quantile definitions for velocity, acceleration, and angle. The table shows numeric cutoffs, while the figure illustrates the angular quantile regions.

Point	Velocity (m/s)	Acceleration (m/s ²)	Angle (°)	v Quantile	a Quantile	θ Quantile
1	2.0	0.5	0	Q1	Q1	Q1
2	4.0	1.2	50	Q2	Q2	Q2
3	3.0	0.8	180	Q1	Q1	Q3
4	3.5	1.5	-160	Q2	Q2	Q3
5	2.5	0.7	-30	Q1	Q1	Q1

Table 6: Toy data points used for the quantile cube example.

Since we now have two quantiles for velocity, two for acceleration, and four for angle, the quantile cube has $2 \times 2 \times 4 = 16$ bins. Each bin is identified by a triplet (v_q, a_q, θ_q) representing the quantile assignments for velocity, acceleration, and angle, respectively. We count the number of points in each bin and compute proportions as the count divided by the total number of data points (5):

Bin (v_q, a_q, θ_q)	Count	Proportion
(Q1,Q1,Q1)	2	0.4
(Q1,Q1,Q2)	0	0.0
(Q1,Q1,Q3)	1	0.2
(Q1,Q1,Q4)	0	0.0
(Q1,Q2,Q1)	0	0.0
(Q1,Q2,Q2)	0	0.0
(Q1,Q2,Q3)	0	0.0
(Q1,Q2,Q4)	0	0.0
(Q2,Q1,Q1)	0	0.0
(Q2,Q1,Q2)	0	0.0
(Q2,Q1,Q3)	0	0.0
(Q2,Q1,Q4)	0	0.0
(Q2,Q2,Q1)	0	0.0
(Q2,Q2,Q2)	1	0.2
(Q2,Q2,Q3)	1	0.2
(Q2,Q2,Q4)	0	0.0
Total	5	1.0

Table 7: Counts and proportions of points in each quantile cube bin.

The final 16-dimensional representation can then be expressed as a raw count vector or as a proportion vector. The vector elements are ordered systematically: for each velocity quantile (Q1, then Q2), we cycle through acceleration quantiles (Q1, then Q2), and for each velocity-acceleration combination, we cycle through all angle quantiles (Q1, Q2, Q3, Q4). This gives us the following representations:

Raw count vector:

$$[2, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]$$

Proportion vector:

$$[0.4, 0.0, 0.2, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.2, 0.2, 0.0]$$

From the proportional representation, we can then visualize the quantile cube. Figure 13 provides a walk-through of the creation, starting with the schematic, then into the point assignment, and concluding with the final toy quantile cube visualization.

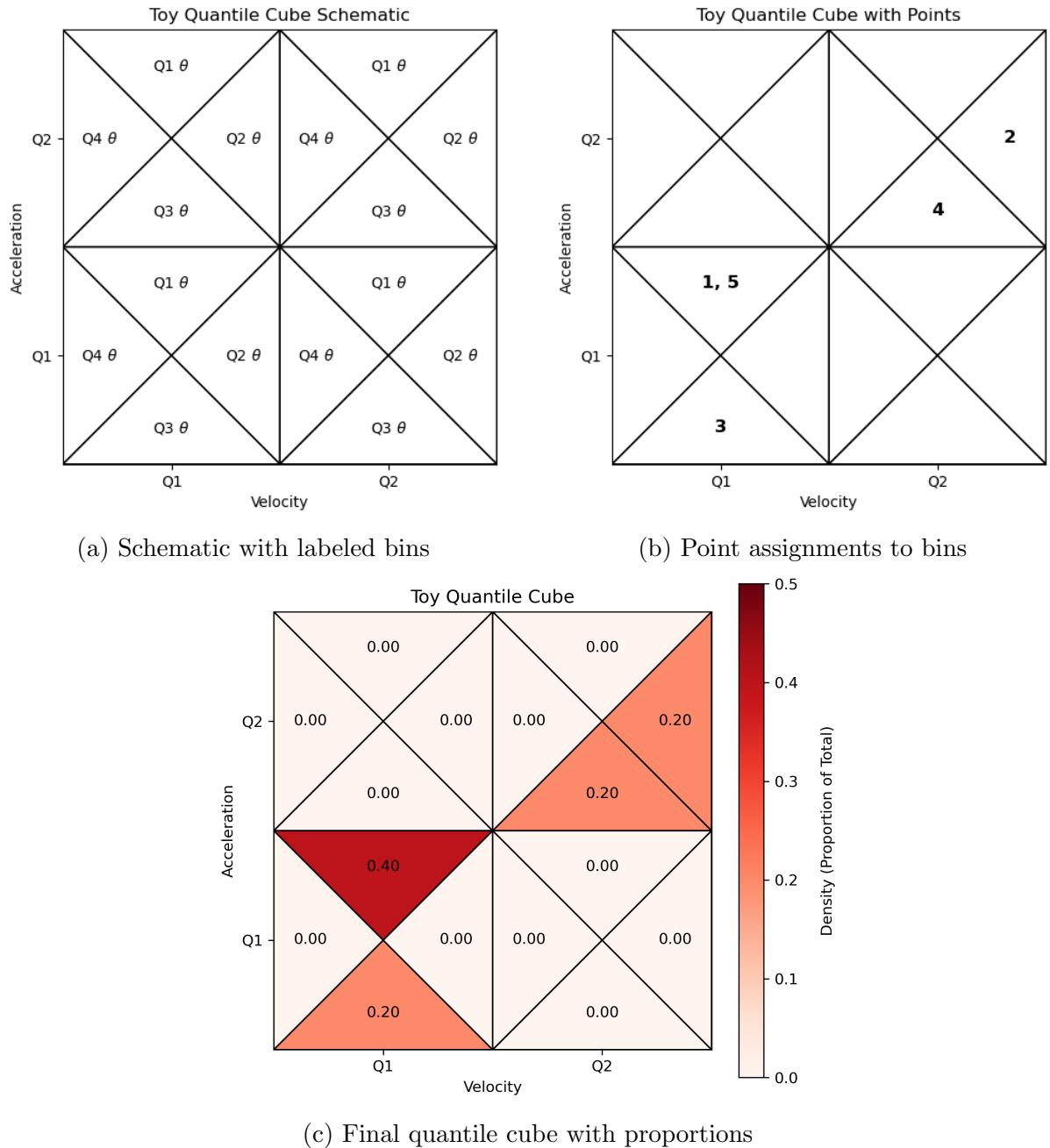


Figure 13: Illustration of the quantile cube formation. Panel (a) shows the schematic with labeled bins, panel (b) shows the data point numbers that fall into each bin, and panel (c) shows the proportions in each bin as a colored visualization of the quantile cube.

C Data Preprocessing Flowchart

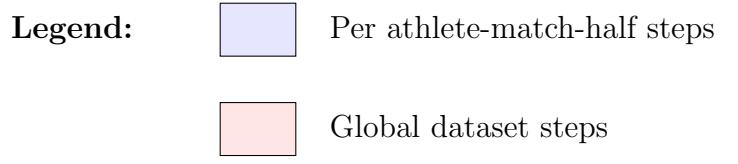
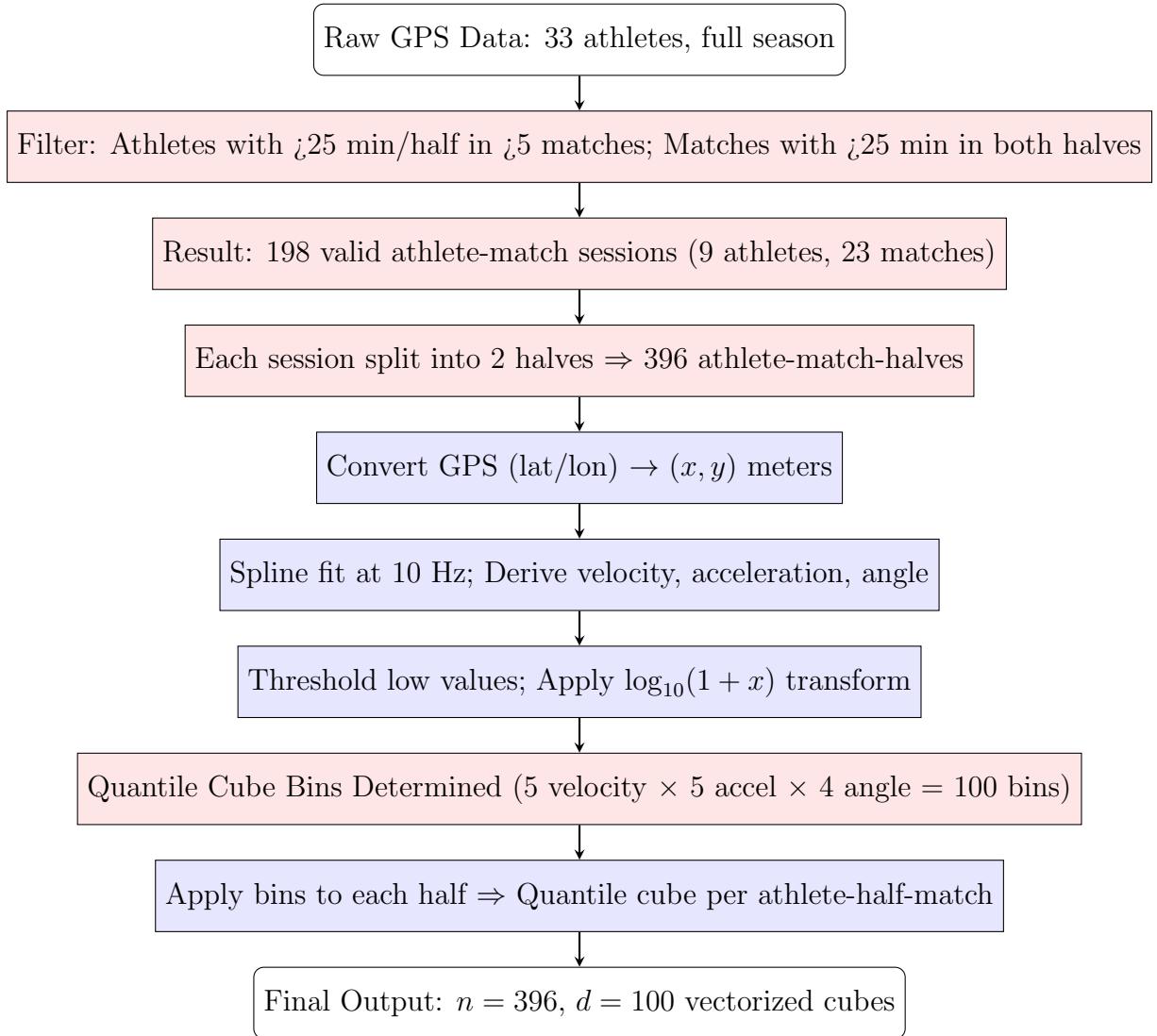


Figure 14: Flowchart of data preprocessing steps from raw GPS data to quantile cube representation. Blue boxes indicate processing per athlete-match-half; light red boxes indicate global processing steps across the dataset.