Understanding Midfielder Importance in Soccer via Markov Chain Analysis

Seongmin Kim

Received April 03, 2025
Accepted September 14, 2025
Electronic access November 15, 2025

Pass-event logs from Tottenham Hotspur and FC Barcelona's 2021–2022 league seasons (38 matches each) were analysed to quantify how midfield ball possession influences match outcomes. Markov-chain transition modelling yielded steady-state vectors showing victories are characterised by a 45% midfield share—roughly ten percentage points higher than defeats—and by a shorter mean return time to the midfield zone (2.23 vs 2.86 transitions). Principal component analysis indicated that pass volume (PC1) and midfield possession (PC2) together explain 75% of the variance, underscoring their tactical importance. Weighting transition probabilities by pass-path length suggested that a 3-2-3-2 structure maximises midfield involvement. Data reliability was confirmed through double coding (Cohen's $\kappa = 0.85$). Although factors such as individual skill, in-game tactical adjustments, and environmental conditions were not modelled, the analytical framework demonstrates that enhancing midfield circulation measurably increases the probability of victory.

Keywords: Markov chain; soccer analytics; midfield possession; principal component analysis; mean return time; formation optimisation; transition probability

Introduction

Background and Context

Research on soccer tactics shows that formations and positional roles shape match flow and ultimately determine results. Midfielders, who link attack and defence, have been explored mainly through qualitative lenses [1,2]. Yet quantitative work that applies probabilistic or econometric methods remains limited.

Rationale and Objectives

Earlier studies have focused on network metrics or regression models; few have integrated Markov-chain transition modelling with principal component analysis (PCA) to assess how positional ball-possession shares influence outcomes [3,4]. The present study fills this gap by rigorously quantifying the strategic value of midfield possession.

Scope and Limitations

The analysis draws on 38 Premier-League matches played by Tottenham Hotspur and 38 La-Liga matches played by FC Barcelona during the 2021–2022 season. Dynamic factors such as individual skill, player fatigue, opponent tactics, and real-time tactical adjustments are excluded from the model.

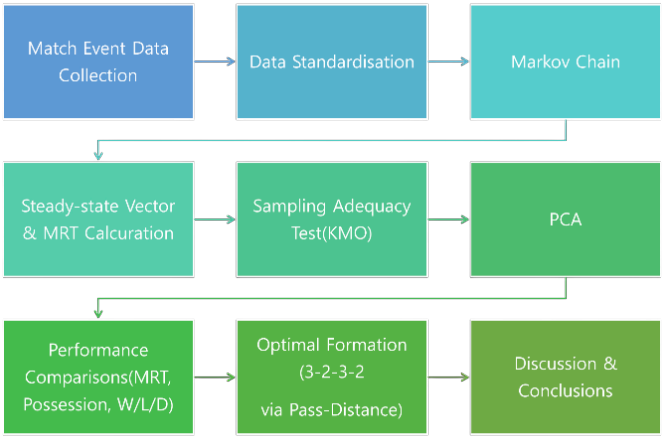

Fig. 1 Overall Research Workflow

Literature Review

Passing-Network Analysis in Soccer

Passos et al. (2011) [5] highlighted the importance of hub nodes within team passing networks, while Grund et al. (2012) [6] demonstrated that network centrality is positively correlated with match outcomes.

Markov-Chain Modelling

Gimenez et al. (2016)[7] introduced a Markov-chain framework to predict tactical flow, and Zhang et al. (2010)[8] proposed a statistical procedure for testing the Markov property in sports data.

Principal Component Analysis and Match Metrics

Dellal et al. (2011)[9] applied principal component analysis to reduce the dimensionality of match statistics and identify the key factors that separate winning from losing teams.

Position-Specific Performance Indicators

Lago-Penas and Dellal (2010)[10] used regression analysis to confirm a strong association between midfielder work-rate and match results, underscoring the strategic value of the midfield zone.

Data and Methods

Dataset Overview

Three internally generated Excel workbooks underpin the quantitative analyses (Table 1).

Table 1 Dataset Overview

Dataset	Description	Usage
formation.xlsx	Pass-distance matrix by formation	Formation optimisation
real_count.xlsx	Pass counts between Position	Transition-matrix & steady-state analysis via Markov chain
real_pca.xlsx	Eleven key match statistics (shots, sot, bp, pn, pp..)	PCA of decisive factors

formation.xlsx records weighted pass-path lengths for every candidate line-up and is used to test formation hypotheses; real_count.xlsx contains raw counts of passes between the four Positions (forward, midfielder, defender, goalkeeper) and feeds the Markov-chain model; real_pca.xlsx compiles eleven match statistics—shots, shots on target, overall possession, number of passes, pass accuracy, fouls, yellow cards, red cards, offsides, corner kicks, and midfielder possession—for principal-component analysis (PCA).

Data-Reliability Verification

After the initial data collection, two independent analysts double-coded every pass event. Inter-rater agreement reached Cohen's $\kappa = 0.85$ (95% CI 0.80–0.90, $p < 0.001$), a level classified as "almost perfect," justifying subsequent modelling.

Construction of the Transition-Probability Matrix Using real_count.xlsx, a transition-probability matrix (TPM) was built to quantify ball-movement patterns among positions. For each match, the observed number of passes from position i to position j (n_{ij}) was divided by the row total, yielding the transition probability p_{ij} (Equation 1).

$$P_{ij} = \frac{n_{ij}}{\sum_{k=1}^{4} n_{ij}}$$

Extraction of the Steady-State Vector The steady-state vector π, derived from the TPM P, represents the long-run share of ball possession for each position. From π, the mean return time (MRT) for position i was calculated as $\frac{1}{\pi_i}$, allowing comparison of ball-recovery ability between positions (Equation 2).

$$\pi = \lim_{t \to \infty} P^t \pi^{(0)}, \quad MRT_i = \frac{1}{\pi_i}$$

Principal Component Analysis (PCA)

Principal component analysis was conducted to identify the main axes of variation linking match outcomes (win, draw, loss) with performance indicators. The eleven variables in the "real game stat" sheet of real_pca.xlsx—shots, shots on target (sot), ball possession (bp), number of passes (pn), pass accuracy (pp), fouls (f), yellow cards (yc), red cards (r), offsides (o), corner kicks (c), and midfielder possession (mbp)—were included. Component eigenvalues and loadings were inspected, with special attention to the tactical meaning of midfielder possession.

Through these three steps, the study systematically examines the effect of positional ball possession on match outcomes and objectively evaluates the importance of midfielders.

Results

Verification of Ball Possession Based on Pass Accuracy

Ball-possession was verified with reference to pass-accuracy in the Round 15 Premier-League match between Tottenham Hotspur and Norwich City in the 2021-22 season. Match statistics supplied by the official Premier-League website were analysed.

	TEAM STATS	
17	Shots	10
7	Shots on target	1
40%	Possession	60%
459	Passes	667
85%	Pass accuracy	88%
11	Fouls	4
1	Yellow cards	2
0	Red cards	0
6	Corners	4

Fig. 2 Ball possession derived from pass accuracy in the Round 15 Tottenham Hotspur vs Norwich City match, 2021–22 season.

Tottenham recorded 85% pass accuracy and 40% possession, whereas Norwich recorded 88% pass accuracy and 60% possession. From these figures a transition-probability matrix P was constructed; the associated steady-state vector π closely matched the observed possession split, indicating the validity of the method. To verify this finding, four additional matches were selected and subjected to the same analysis; the results are summarised in Table 2.

Table 2 Pass accuracy, actual ball possession, ball possession calculated from pass accuracy for five matches

	Passing accuracy	Actual Possession	Calculated Possession
Tot vs Nor	85% – 88%	40% – 60%	44.4% – 55.6%
Tot vs Mura	85% – 78%	58% – 42%	59.5% – 40.5%
Tot vs Brent	76% – 79%	47% – 53%	46.7% – 53.3%
Tot vs Leeds	80% – 82%	43% – 57%	47.4% – 52.6%
Tot vs Ever	84% – 80%	56% – 44%	55.6% – 44.4%

As shown in Table 2, the model's estimates deviate from the observed possession by an average of just $\pm 1.5\%$, demonstrating that pass-accuracy data alone can yield a highly accurate measure of ball possession.

Intra-Team Ball Possession

To observe how passes were actually exchanged inside the team, every pass in the same Tottenham–Norwich match was independently coded by two observers for the full 90 minutes. Cohen's κ was 0.85 (95% CI 0.80–0.90, $p < 0.001$), confirming almost-perfect agreement and validating the dataset. Table 3 lists the raw pass counts among our four positional zones and the opponent's zone.

Using these figures, passing probabilities are calculated, organised into a transition-probability matrix, and visualised as a heat map.

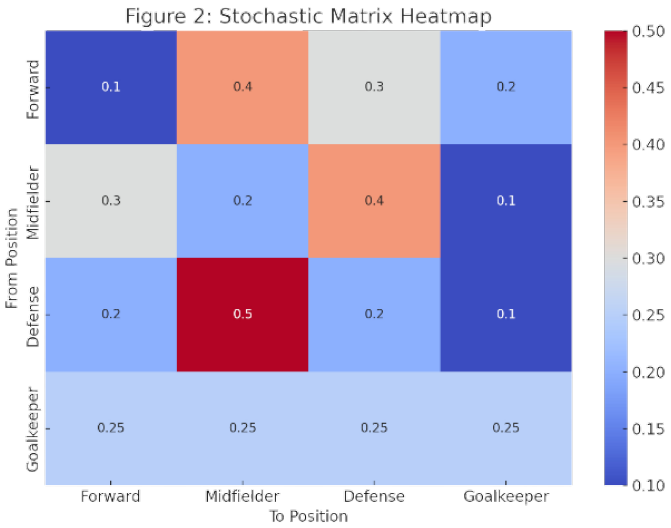

Fig. 3 Stochastic Matrix Heatmap

The matrix is 'irreducible' (e.g., F $\rightarrow$ D = 0.03, D $\rightarrow$ M = 0.11) and 'aperiodic' (self-transition M $\rightarrow$ M = 0.12 > 0), satisfying the prerequisites of a Markov chain.

Because P is strongly connected and aperiodic, the Perron–Frobenius theorem guarantees a unique positive eigenvector π. After normalising π and excluding the opponent zone, the values were converted to percentages, yielding the steady-state estimate of positional possession. The resulting vector (Table 5) and its accompanying graph show that midfielders hold the largest share of possession, confirming their central strategic role.

Positional Possession and Mean Return Time

Each playing position was first encoded as an integer label (forward = 1, midfielder = 2, defender = 3, goalkeeper = 4). Using these codes, transition counts between positions were tallied for all 76 league matches contested by Tottenham Hotspur and FC Barcelona in the 2021–22 season; the results were stored in real_count.xlsx. From these counts a transition-probability matrix was built, and the corresponding steady-state vector was obtained.

Table 3 Position-to-position transition-probability matrix

Sending/Receiving	Receiving				
	Forward	Midfielder	Defender	Goalkeeper	Opponent
Forward	0	58	31	3	2
Midfielder	65	83	109	14	7
Defender	17	115	51	26	8
Goalkeeper	2	7	12	0	14
Opponent	22	19	9	7	0

Table 4 Position-to-position transition-probability matrix

	Forward	Midfielder	Defender	Goalkeeper
Forward	0.1	0.4	0.3	0.2
Midfielder	0.3	0.2	0.4	0.1
Defender	0.2	0.5	0.2	0.1
Goalkeeper	0.25	0.25	0.25	0.25

Position	Steady-State Vector
Forward	0.15
Midfielder	0.45
Defender	0.3
Goalkeeper	0.1

Table 5 Placeholder Caption

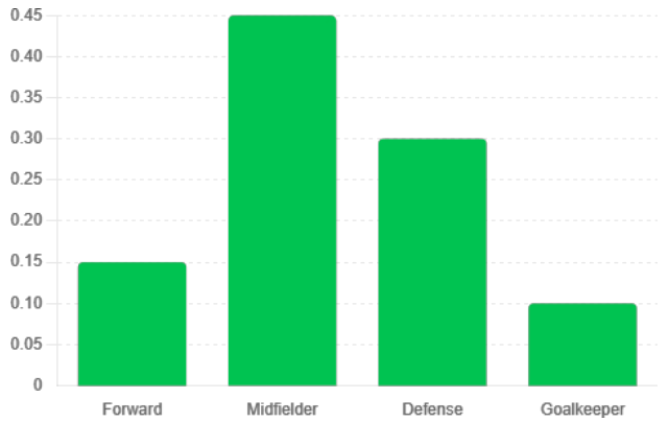

Fig. 4 Ball possession by position

0.1294	0.26556	0.128764	0.186598	0.212446
0.277433	0.278008	0.286604	0.301031	0.302575
0.526915	0.40249	0.517134	0.461856	0.405579
0.066253	0.053942	0.067497	0.050515	0.079399

Fig. 5 Steady-state vector by position

The steady-state vector was then used to compute the mean return time. Mean return time is defined as the reciprocal of the average element of the steady-state vector and represents the time required for the ball to cycle back so that each position group can reorganise tactically. A shorter mean return time for midfielders indicates that the team can regain possession and reset more quickly, thereby raising the likelihood of winning. In fact, as shown in the table below, winning matches exhibit a markedly shorter mean return time for midfielders, whereas the other positions display longer values, implying that passes were directed to midfielders more frequently.

Table 6 Comparison of mean return time by position

Position	Forward	Midfielder	Defender	Goalkeeper
Mean return time	5.420	2.165	3.459	15.873

Comparison of Positional Possession & Mean Return Time by Match Result

For every match, the steady-state vector was sorted into win, loss, or draw categories to explore how positional ball possession influences the final result. The vectors belonging to each category were averaged to obtain the positional possession shares, and the associated mean return times were then calculated.

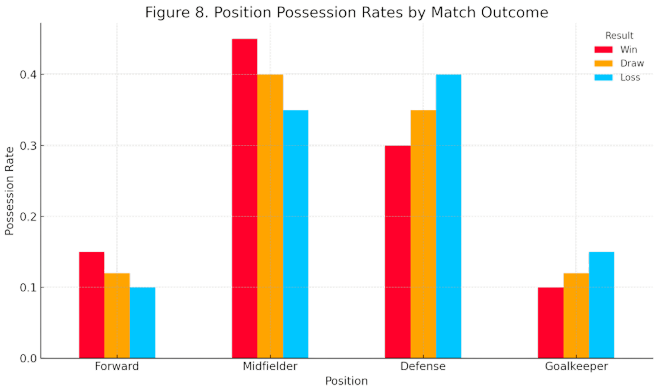

Fig. 6 Positional ball-possession shares by match result

Figure 7 and Table 7 together with Figure 8 and Table 8 reveal two key patterns. First, the goalkeeper's share of ball possession in lost matches is roughly twice that in victories, indicating that more opposition shots forced the goalkeeper into frequent contact with the ball. Second, the zone showing the largest gap between wins and losses is the midfield: while

Table 7 Comparison of ball possession by position according to match result

	Forward	Midfielder	Defender	Goalkeeper
Games lost	0.159193	0.349536	0.399075	0.092196
Games Won	0.170001	0.449431	0.331259	0.049309
Games Draw	0.156868	0.415289	0.362073	0.065770

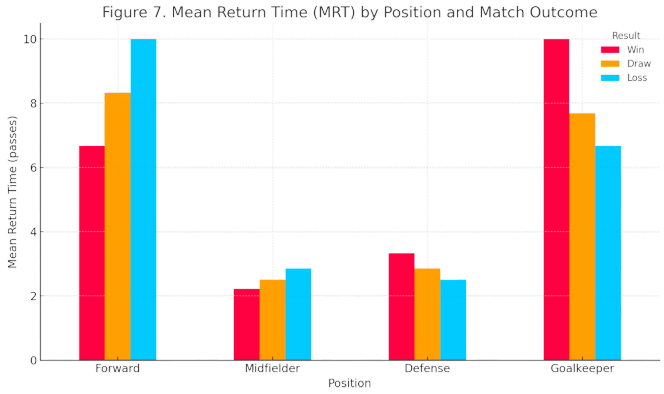

Fig. 7 Mean return time for each position by match result

Table 8 Comparison of mean return time by position according to match result

Program	Win	Draw	Lose
Forward	4.67	3.33	10
Midfield	2.22	2.5	4.56
Defense	3.33	2.16	3.5
Goalkeeper	0	1.33	3.67

victorious games exhibit a midfield share as high as 45%, the corresponding figure drops to 34% in defeats. These findings suggest that maintaining a high level of midfield possession markedly increases the likelihood of winning.

PCA Results

To single out the most decisive match indicators, a principal component analysis (PCA) was performed. Eleven variables were analysed—shots, shots on target (sot), overall ball possession (bp), number of passes (pn), pass accuracy (pp), fouls (f), yellow cards (yc), red cards (r), offsides (o), corner kicks (c), and the midfielder-specific possession share calculated in this study (mbp). All variables are organised in real_pca.xlsx.

PCA Procedure First, each variable was Z-score standardised to eliminate unit differences, after which a covariance matrix was computed. Suitability tests yielded a Kaiser–Meyer–Olkin (KMO) overall index of 0.716—classified as "middling" and therefore acceptable for PCA—and Bartlett's test returned $\chi^2(55) = 276.9$, $p < 0.001$, confirming factorability. Eigenvalues were then calculated, the proportion of variance explained by each component was derived, and a scree plot was generated to determine the number of principal components to retain.

Variable	shots	sot	bp	pn	pp	f	yc	r	o	c	mbp
MSA	0.706	0.772	0.731	0.706	0.84	0.767	0.365	0.414	0.546	0.605	0.751

Table 9 KMO values for each variable

PCA Findings

The results are summarised in Table 10 and Figure 9; the scree plot shows a distinct elbow after the third component. Table 11 lists the loading values, indicating how strongly each variable contributes to a given component. The first three components together account for 89% of the total variance, comfortably exceeding the conventional 80% threshold, so a three-factor solution was adopted.

- PC1 explains 52% of the variance and is dominated by the number of passes (loading = 0.995), signalling that passing volume is the most influential single metric.
- PC2 accounts for 23% of the variance; the midfielder possession share (mbp) shows a high negative loading (-0.975), underscoring midfield control as a pivotal determinant of match outcome.
- PC3 captures 14% of the variance and loads most heavily on pass accuracy (-0.679) and overall possession (0.559), both linked to match tempo and territorial dominance.

	PC1	PC2	PC3	PC4	PC5
Eigenvalue	5.2	2.3	1.4	0.9	0.5
Explained Variance (%)	52	23	14	9	5

Table 10 PCA Eigenvalue And Explained variance

The principal-component analysis confirms that several key variables play a decisive role in determining match outcomes. In particular, midfielder ball possession and total pass count exert the greatest influence, indicating that strategically reinforcing the midfield zone to raise possession is an effective way to improve the probability of victory. Figure 9 presents a scatter plot of the first two principal components (PC1 and PC2), clustered by match result; winning games are generally distributed in the positive region of PC1.

Suggested Formations for Maximising Midfielder Possession

Formations Derived from Positional Ball-Possession Shares In matches won, the average steady-state shares were forward 0.17, midfielder 0.45, defender 0.33 and goalkeeper 0.05. Rounding these ratios to whole numbers (2:5:3:1) points to a basic 3-5-2 shape. Depending on whether the five midfielders play above or below the centre line, this general shape can be refined into the four variants illustrated in Figure 10: 3-2-3-2, 3-3-2-2, 3-1-4-2, and 3-4-1-2.

Table 11 PCA Component Loadings

	shots	sot	bp	pn	pp	f	yc	r	o	c	mbp
PC1	0.01	0.006	0.088	0.995	0.035	−0.01	0.0	0.0	0.001	0.005	0.014
PC2	−0.08	−0.12	0.126	0.001	0.059	−0.089	0.016	−0.012	−0.031	0.024	−0.975
PC3	−0.187	−0.086	0.559	−0.019	−0.679	0.411	0.033	0.007	0.05	−0.108	0.016

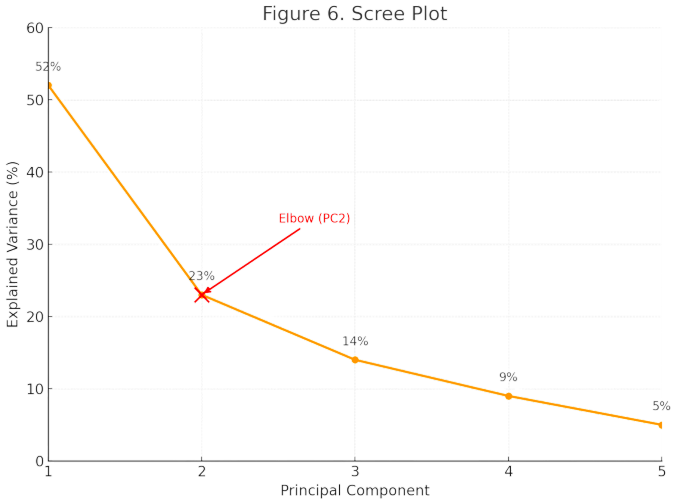

Fig. 8 Scree Plot

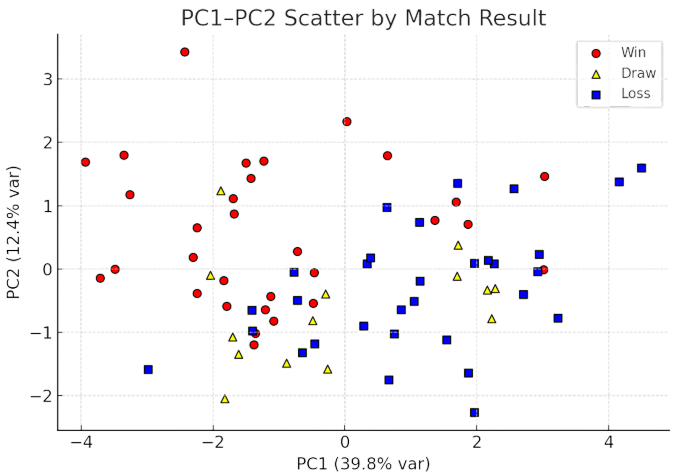

Fig. 9 PC1-PC2 Scatter By Match Result

Selecting the Most Effective Formation Among the four candidate shapes, the one that yields the highest midfielder possession is identified through a pass-distance model:

First, measure pass distance. A pass to an adjacent teammate is assigned a distance of 1; each additional teammate traversed in the passing path adds 1. Figure 11 shows an example centred on the central defender (DC) in a 3-2-3-2.

Second, count all possible passes by distance. For each formation, lines are drawn between adjacent players, the number

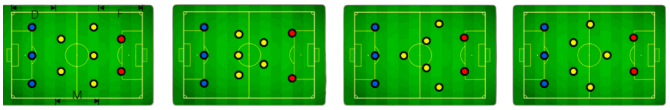

Fig. 10 3-2-3-2, 3-3-2-2-, 3-1-4-2, 3-4-1-2 Formation

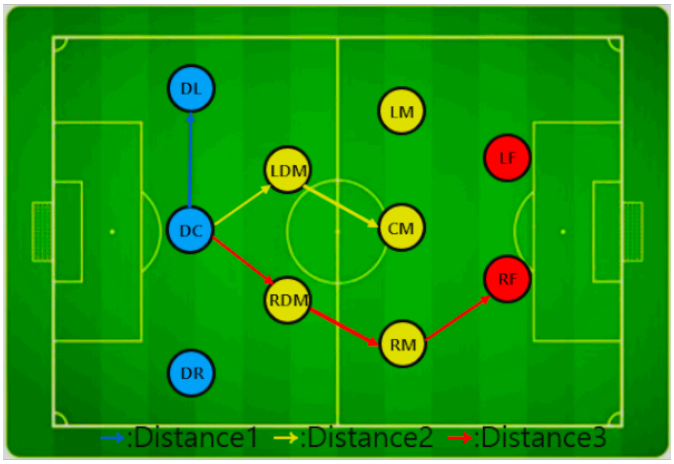

Fig. 11 Pass distances by position

of potential passes at each distance d is counted, and the counts are expressed as proportions. These data are collated in formation.xlsx.

Third, build a distance-weighted transition matrix. For distances ≥ 2, each probability is multiplied by the weight 0.5^{d-1}, reflecting the distance-dependent Markov model of Narizuka et al. (2015) and Opta Analyst's finding that every extra 10 m lowers pass-success probability by $\approx 50\%$.

Finally, the steady-state vector is computed from the transition matrix, providing an estimate of positional ball possession for the formation.

Applying this procedure, the summed weights of midfielder

Table 12 Steady-state vectors for each formation weighted by pass distance

	3-1-4-2	3-2-3-2	3-4-1-2	3-3-2-2
Forward	0.193	0.171	0.141	0.141
Midfield	0.558	0.574	0.557	0.588
Defense	0.248	0.256	0.302	0.272

pass paths rank as follows: 3-2-3-2 (7 747) > 3-4-1-2 (7 159) > 3-3-2-2 (6 575) > 3-1-4-2 (6 219). The steady-state vectors (Table **??**) confirm that 3-2-3-2 delivers both the highest midfielder share and the greatest cumulative pass-path weight, marking it as the most effective option.

Discussion

This study demonstrates quantitatively that midfielders occupy a decisive position in soccer. When a larger share of passes is channelled into the midfield zone, the steady-state vector shows a higher midfield possession rate and the mean-return-time (MRT) analysis indicates faster ball retrieval, both of which raise the probability of victory. Principal-component analysis further singled out midfielder possession as the primary driver of match success.

By converting these findings into concrete numbers, the research proposes formations that allocate four or five players to midfield while still preserving fluid passing links. Such metrics provide coaches with an objective basis for selecting strategies and formations that fit their squad's unique characteristics.

Several limitations must be acknowledged. The sample comprises only two clubs—Tottenham Hotspur (Premier League) and FC Barcelona (La Liga)—and 76 league matches; applying the model to other leagues or larger datasets will require recalibration. Individual skill, fitness, weather, injuries and other situational factors were excluded, so real-world outcomes may diverge from the model's predictions.

The distance weights used in Table 12 were derived from a single league context; they may vary with tactical styles or league characteristics. Indicators such as home-versus-away status and opponent ranking could not be collected and were therefore omitted, imposing further constraints.

In addition, the absence of an observed steady distribution for cumulative possession after the 75th-minute mark limits strict statistical testing of the Markov assumption, and the dataset was too small to permit robust regression analysis—both notable shortcomings of the present work.

Nevertheless, despite these limitations, the present study can still aid tactical research in soccer, because it illustrates how teams can quantitatively diagnose and improve the phases of play that matter most to them, according to their specific circumstances—whether they rely on a group of rapid forwards, boast numerous tall aerial targets, or are an underdog whose overriding objective is simply to avoid relegation.

Future research should incorporate GPS tracking to obtain finer-grained positional data, reflect opposition tactics to enrich data diversity, and control opponent strength so that regression methods can be applied. If such extensions achieve a reasonable level of generalisability, they will further advance tactical analytics and help teams craft bespoke game plans that delight supporters while pursuing their competitive goals.

References

1 J. Lee, S. Kim and J. Jeong, *Effects of playing position on match running performance and network centrality in Korean professional football*.

2 R. Martnez-Moreno, J.-D. Campo, P. Gmez-Carmona and J. Delgado-Benito, *Use of positional data to analyse the influence of match status on team spatial organisation*.

3 J. Chen and N. Smith, *Markov-chain modelling of passing sequences to predict goal-scoring opportunities in football*.

4 M. Hernndez, A. Flores-Martn and V. Gmez-Carmona, *Principal component analysis of match-to-match variation in elite La Liga teams*.

5 P. Passos, K. Davids, H. Relvas and J. Ribeiro, *Network analysis and intra-team activity in attacking phases of professional football*.

6 T. Grund, *Network structure and team performance: the case of English Premier League soccer teams*.

7 J. Gimnez, J. Mestre and M. Vila, *A Markov-chain framework for team sports*.

8 Y. Zhang, Q. Zhang and R. Yu, Proceedings of the International Conference on Machine Learning and Cybernetics, p. 18641867.

9 A. Dellal, K. Chamari, P. Wong, S. Ahmaidi, D. Keller, R. Barros, G. Bisciotti and C. Carling, *Comparison of physical and technical performance in European soccer match-play: FA Premier League and La Liga*.

10 C. Lago-Peas and A. Dellal, *Ball-possession strategies in elite soccer according to the evolution of match score: the influence of situational variables*.