Analyzing Player Ball Possession in Football Passing Networks Using Markov Chains and Steady-State Distribution

Andrew Shin

2024-08-26

Abstract:

In the dynamic game of football, players pass the ball to each other in order to maximize the chance of creating and capitalizing on scoring opportunities. In this project, I develop a probabilistic model using Markov chains to analyze and estimate the future player's ball possession in a passing network. By analyzing the steady-state distribution of the transition probability matrix, the model provides insights into how ball possession is distributed among individual players. In addition, I illustrate that the analysis of steady-state probability reveals how the likelihood of player ball possession in over the next 10 and 20 minutes differs from the probability distribution obtained through Monte-Carlo simulation.

Introduction

During a football match, both teams make an effort to move the ball around the pitch to find spaces where they can increase the probability of creating goal-scoring opportunities. Maintaining possession for a significant amount of time relies heavily on the passes exchanged between teammates. In the fast-paced environment of football, each player contributes differently to the team's overall pass distribution. This leads to a question of how the ball is likely to be shared or possessed among players over time. When a team initiates a passing sequence, each pass from one player to another represents a transition from one state to another, which can be modeled using Markov chains.

By applying the transition probability matrix and Markov Chains theory, I can gain insights into the system's dynamics, such as estimating the probability of a specific player receiving the ball after multiple passes, calculating the expected number of passes needed to reach a particular player, and predicting long-term behavior through the simulation of stochastic processes. Additionally, the steady-state distribution can estimate how ball possession is likely to be distributed among players over time. This approach can also highlight a central player in ball possession, indicating that the majority of passes are likely to be directed toward this player.

Methods

Data Collection

The data for this analysis was retrieved from StatsBomb Open Data, which provides up-to-date match data from numerous football leagues around the world. Each match data set typically includes 75 columns and around 4,000 rows on average, capturing event data in detail such as player movements and passes, almost on a per-second basis. For this study, I chose the one of Spain's matches in EURO 2024 and focused on passes exchanged between players in this match. I chose Spain's Round of 16 match because their passing accuracy (90.2%, 748/792) and ball possession (75%) were the highest among all the matches they played in this tournament.

Data Preprocessing

Data Cleaning: The raw data was cleaned to ensure simplicity and accuracy. I filtered out the columns that were not used for this analysis, retaining only the following columns: match time, player names, types of actions, and substitution information.

Assumption

For simplicity, this analysis focuses exclusively on the first half of the game, a period during which substitutions rarely occur. Given that Spain completed a total of 228 accurate passes in the first 23rd minutes, accounting for 30.5 percent of their final total accurate passes, this sample is sufficient to calculate the initial distribution.

To maintain the memoryless property of Markov chains, I assume that the decision to pass to a specific player is based solely on the current ball holder, without consideration of previous passes. The transition probability matrix calculated based on the passes up to the 23rd minute is assumed to remain constant for the rest of the first half, even though they are likely to vary throughout the entire match. This assumption allows me to compare the steady-state distribution to the one obtained through Monte-Carlo simulation.

The analysis tracks only the direct passing events between players, such as when player i passes to player j. Passes resulting from set plays, such as corner kicks, goal kicks, and throw-ins, are excluded. These types of passes are often influenced by tactical decisions that designate specific players for these roles, which could introduce bias into the data. Excluding them ensures that the analysis more accurately reflects the team's typical in-play passing network.

Construction Transition Probability Matrix (TPM)

I counted the number of successful passes from each player to every other player on the team up to 23th minute of the match. The TPM was constructed by normalizing the pass counts. Specifically, the probability of a pass from player i to player j was calculated as:

The transition probability $P(X_{t+1} = j \mid X_t = i)$ is given by:

$$P_{ij} = P(X_{t+1} = j \mid X_t = i) = \frac{\#p}{\#P}$$

where:

- $\#p$ is the number of passes player i made to player j,
- $\#P$ is the total number of passes player i made in the first 30 minutes of the match,
- t represents a finite time when one pass is delivered from the ball holder to the recipient,
- i and j represent the ball holder and the pass recipient, respectively, and thus correspond to the eleven players on the team. $i, j \in S$, where $S = \{player_1, \ldots, player_{11}\}$
- $\sum_j P_{ij} = 1$.

Analysis of Steady-State Distribution

1. Markov Chain Modeling: Each TPM was treated as the transition matrix of a Markov chain. The steady-state distribution vector, representing the long-term probabilities of the ball being with each player, was computed based on the first 23rd minute of the game.
2. Limiting Probabilities:

- For my project, I will rely on the theorem 4.1 in Introduction to Probability Models by Sheldon M. Ross, saying:

$$\text{For an irreducible ergodic Markov chain } lim_{n\to\infty} P_{ij}^n \text{ exists and is independent of } i.$$
$$\text{Furthermore, letting } \pi_j = lim_{n\to\infty} P_{ij}^n,\ j \geq 0$$
$$\text{then } \pi_j \text{ is the unique nonegative solution of } \pi_j = \sum_{i=0}^{\infty} \pi P_{ij},\ j \geq 0,\ \sum_{j=0}^{\infty} \pi_j = 1$$

- I can guarantee that a real square matrix with positive entries has a unique eigenvalue of largest magnitude and that eigenvalue is real by Perron–Frobenius theorem. Thus, consider probability distribution X over S such that $P(X = i) = \pi_1$, which implies the probability that a player i receives the ball is π_1 in a long-run.

3. Monte-Carlo Simulation: I will compare the steady-state distribution to the probability distribution obtained through Monte Carlo simulation. The initial probabilities for this simulation will also be based on the total number of passes up to the 23rd minute of the match.

Result

Transition Probability Matrix

I constructed the transition probability matrix by counting all passes exchanged between each player and the other 10 players, then dividing by the total number of passes that player made up to the 23rd minute. The transition probability matrix that I will use for this analysis is:

N-step Transition Probabilities

Let q_i be the probability that the chain is in state i at time 0, which indicates that the ball is with a player i at time 0. I can construct q_i by dividing the number of pass receptions made by a player i during the first 23 minutes by the total number of passes made by the team.

```
kable(P)
```

	Aymeric Laporte	Daniel Carvajal	Fabián Ruiz	Lamine Yamal	Marc Cucurella	Nico Williams	Pedri	Robin Le Normand	Unai Rodri	Simón	Álvaro Morata
Aymeric Laporte	0.0000	0.0682	0.0227	0.0000	0.3636	0.1591	0.0909	0.1591	0.1364	0.0000	0.0000
Daniel Carvajal	0.0400	0.0000	0.0000	0.2400	0.0000	0.0000	0.1200	0.3200	0.2800	0.0000	0.0000
Fabián Ruiz	0.1250	0.1250	0.0000	0.1250	0.0625	0.1250	0.0625	0.0625	0.3125	0.0000	0.0000
Lamine Yamal	0.0000	0.3333	0.0000	0.0000	0.0000	0.2000	0.3333	0.0667	0.0667	0.0000	0.0000

	Aymeric Laporte	Daniel Carvajal	Fabián Ruiz	Lamine Yamal	Marc Cucurella	Nico Williams	Pedri	Robin Le Normand	Unai Simón	Rodri	Álvaro Morata
Marc Cucurella	0.4722	0.0000	0.1667	0.0000	0.0000	0.1667	0.0278	0.0000	0.1389	0.0278	0.0000
Nico Williams	0.2308	0.0769	0.0769	0.0000	0.4615	0.0000	0.0769	0.0000	0.0769	0.0000	0.0000
Pedri	0.1333	0.0000	0.3333	0.2000	0.0667	0.0667	0.0000	0.0667	0.0667	0.0000	0.0667
Robin Le Normand	0.2500	0.3750	0.0000	0.1250	0.0000	0.0000	0.0000	0.0000	0.1667	0.0833	0.0000
Rodri	0.2121	0.0909	0.1818	0.1212	0.2424	0.0000	0.1212	0.0303	0.0000	0.0000	0.0000
Unai Simón	0.2500	0.2500	0.0000	0.0000	0.0000	0.0000	0.0000	0.2500	0.2500	0.0000	0.0000
Álvaro Morata	0.0000	0.0000	0.3333	0.0000	0.0000	0.0000	0.0000	0.0000	0.6667	0.0000	0.0000

We can then determine the probability that the system is in state i at time n using the following reasoning:

$$\begin{aligned}
&\text{Probability of the ball being in a player } j \text{ at time } n \\
&\sum_{i=1}^{11}(\text{probability that a ball is originally at player } i) \\
&\times(\text{probability of passing from player } i \text{ to player } j \text{ in n transitions}) \\
&= \sum_{i=1}^{11} q_i P_{ij}^n
\end{aligned}$$

```
kable(q, caption = "Initial Distribution")
```

Table 2: Initial Distribution

Player	p
Aymeric Laporte	0.17105
Daniel Carvajal	0.10526
Fabián Ruiz	0.08772
Lamine Yamal	0.07895
Marc Cucurella	0.14035
Nico Williams	0.08333
Pedri	0.08333
Robin Le Normand	0.08772
Rodri	0.14474
Unai Simón	0.01316
Álvaro Morata	0.00439

To illustrate the use of this notation, for example, we answer the following question: Suppose the initial probability distribution is given as above. Three passes from now, what fraction of all players will be passing to Fabián Ruiz? The desired probability is

$$q^T \cdot P^3_{Fabián\ Ruiz} = 0.00481$$

Hence, three passes from now, 0.48% of all players will be passing to Fabián Ruiz. Using this notion, we can calculate the probability of players' pass reception after multiple sequences of passes in a game.

Steady-State Distribution

To determine the long-run probabilities, I can compute the steady-state distribution by finding the eigenvector corresponding to the eigenvalue of 1.

```
kable(steady_states, caption = "Steady-state Distribution")
```

Table 3: Steady-state Distribution

Player	p
Aymeric Laporte	0.16839
Marc Cucurella	0.14535
Rodri	0.13876
Daniel Carvajal	0.10304
Fabián Ruiz	0.09094
Pedri	0.08767
Nico Williams	0.08442
Robin Le Normand	0.08365
Lamine Yamal	0.08091
Unai Simón	0.01101
Álvaro Morata	0.00585

The steady-state distribution represents the long-term behavior of the passing network, where each value corresponds to the probability that the ball is with a particular player after many passes.

Laporte has the highest stationary probability, indicating that in the long run, he is the most likely player to have possession of the ball. As a center-back, this might suggest that he is frequently involved in ball circulation, likely playing a key role in building play from the back. Rodri, as a central defensive midfielder, also has a high probability. His role often involves controlling the tempo of the game and distributing passes, which is consistent with a high stationary probability.

The distribution indicates a strong involvement of defenders and midfielders in maintaining possession. Players like Laporte, Rodri, and Cucurella are key figures in the passing network, likely reflecting their roles in controlling the game and redistributing the ball.

Interpretation of Steady-State Probabilities Recall that in the steady-state distribution, the probability that the system is in state j is π_j, which in this context suggests that the ball is in a player j. Using the Markov Chain theory, we have:

$$\pi_j(1 - p_{jj}) = \sum_{k \neq j} \pi_k p_{kj}$$

This states that the probability that a player k passes to a player j is equal to probability that a player k receives a pass from a particular player. In other words, this simply says that the "flow" of passing probability into each player must equal the flow of probability out of each player. This yields

```
kable(result)
```

x
0.1683899
0.1030454
0.0909410
0.0809051
0.1453532
0.0844178
0.0876728
0.0836486
0.1387608
0.0110088
0.0058476

Steady-State Distribution vs Empirical Distribution

The empirical distribution shows how the ball is distributed among players up to the 35th and 45th minutes. In Figure 5, I observed that the empirical distribution closely matches the steady-state distribution. It's surprising how well the empirical distribution aligns with the steady-state distribution.

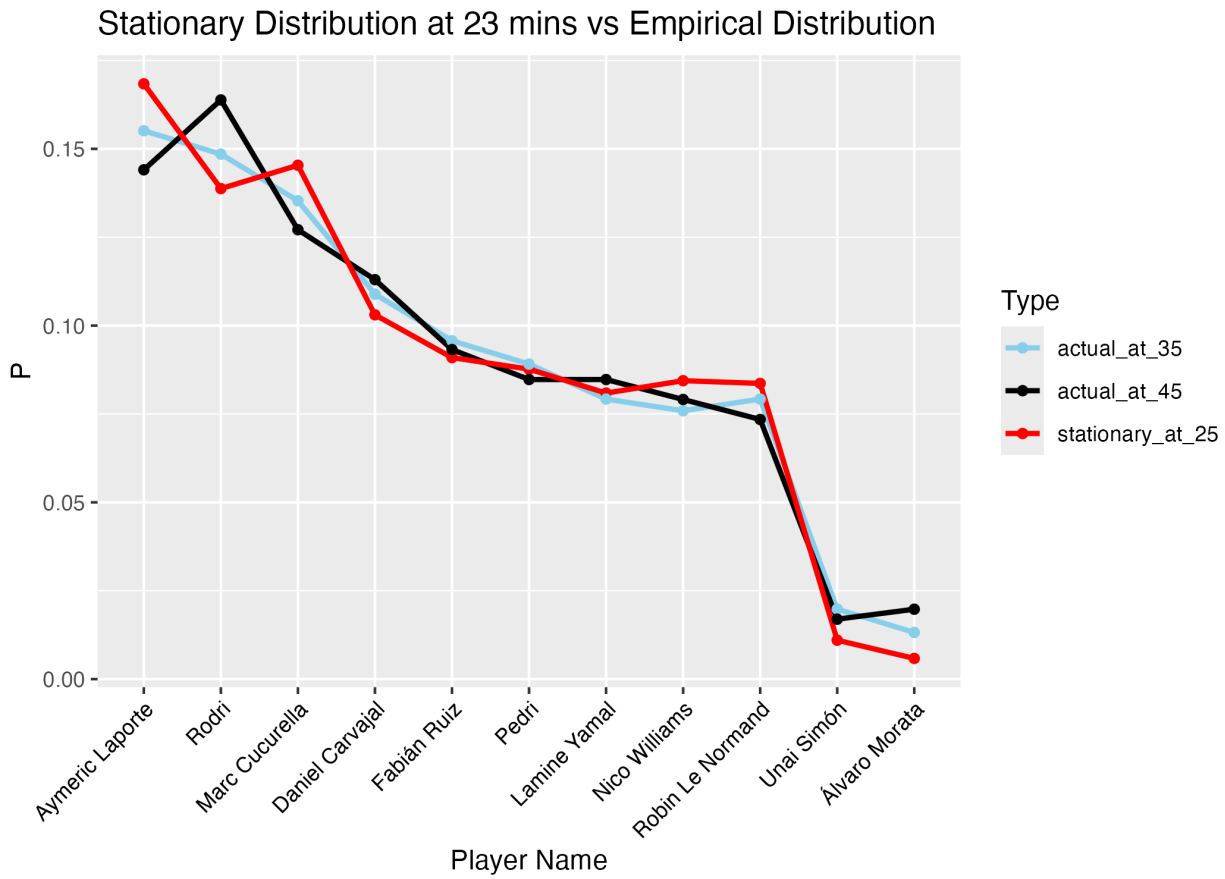

Figure 1: Plot of Steady-State Distribution

$$rmse_1 = \sqrt{mean((empirical_{35th\ min} - steady_{23rd\ min})^2)},\ rmse_2 = \sqrt{mean((empirical_{45th\ min} - steady_{23rd\ min})^2)}$$

$$mean(rmse_1, rmse_2) = 0.0107$$

The root mean square error (RMSE) is notably small, indicating that the steady-state distribution provides a strong estimate of ball distribution among players in the long run. Therefore, I can confidently conclude that the steady-state distribution is a reliable predictor of how the ball is likely to be distributed among players over extended periods.

Comparison with Monte Carlo Simulation

In this section, I will compare the error between the steady-state distribution and the empirical data to the error between the simulation distribution and the empirical data. The distribution I will use for the simulation is the initial distribution q, with sample sizes corresponding to the number of passes made between the 23rd and 35th minutes, and between the 23rd and 45th minutes.

By the 35th minute, Spain completed 303 accurate passes, indicating an additional 75 passes since the 23rd minute. Similarly, with a total of 354 passes completed by the 45th minute, reflecting 126 additional passes since the 23rd minute, Then I perform simulations.

Result of Monte Carlo Simulation

```
kable(df_mse, caption = "RMSE Comparison b/t steady-state and simulation")
```

Table 5: RMSE Comparison b/t steady-state and simulation

RMSE	value
steady_state vs empirical_35	0.0312975
steady_state vs empirical_45	0.0338983
sim vs empirical_35	0.0078529
sim vs empirical_45	0.0129833

```
kable(result_sim1, caption = "Result of Monte Carlo Simulation (35th min)")
```

Table 6: Result of Monte Carlo Simulation (35th min)

Iteration	Álvaro Morata	Aymeric Laporte	Daniel Carvajal	Fabián Ruiz	Lamine Yamal	Marc Cucurella	Nico Williams	Pedri	Robin Le Normand	Unai Rodri Simón
10	0.0000	0.1480	0.1093	0.0800	0.0707	0.1707	0.0973	0.0907	0.1040	0.12130.0080
10^2	0.0053	0.1715	0.1108	0.0855	0.0762	0.1441	0.0844	0.0810	0.0901	0.14610.0149
10^3	0.0043	0.1707	0.1054	0.0884	0.0790	0.1416	0.0825	0.0837	0.0879	0.14450.0131
10^4	0.0044	0.1718	0.1049	0.0876	0.0783	0.1405	0.0833	0.0836	0.0877	0.14480.0133

```
kable(result_sim2, caption = "Result of Monte Carlo Simulation (45th min)")
```

Table 7: Result of Monte Carlo Simulation (45th min)

Iteration	Álvaro Morata	Aymeric Laporte	Daniel Carvajal	Fabián Ruiz	Lamine Yamal	Marc Cucurella	Nico Williams	Pedri	Robin Le Normand	Unai Rodri Simón
10	0.0071	0.1659	0.1048	0.0921	0.0817	0.1484	0.0738	0.0770	0.0929	0.14680.0095
10^2	0.0043	0.1722	0.1064	0.0883	0.0809	0.1388	0.0787	0.0848	0.0907	0.15000.0149
10^3	0.0044	0.1722	0.1045	0.0890	0.0777	0.1400	0.0828	0.0835	0.0880	0.14540.0134
10^4	0.0044	0.1710	0.1054	0.0877	0.0789	0.1406	0.0836	0.0833	0.0872	0.14480.0131

Based on the RMSE, the steady-state distribution closely matches the empirical distribution at both 35th and 45th minute. However, both simulation distributions slightly aligns more with the empirical distributions.

Discussion

The steady-state distribution of the transition probability matrix not only illustrates how the passing network is currently distributed among the players, but it also provides an approximation of the pass distribution in the next several sequences. To validate this prediction, I calculated the empirical distribution at each time period and the root mean squared error (RMSE) between these two distributions. Interestingly, the error is sufficiently small, supporting the validity of the estimation. Furthermore, a comparison between the accuracy of the stationary distribution and the accuracy of a Monte Carlo simulation based on the initial distribution q reveals an intriguing result.

For the sake of simplicity, I focused only on the first half of the game, where substitutions rarely occur, and used the first 23 minutes to calculate the initial distribution and transition probability matrix. In the near future, I would split the game into more time segments, continuously update the transition matrix, and recalculate the steady-state distribution for the next 10-minute interval. I would then compare these predictions with the empirical ball possession and the Monte Carlo simulation results across the entire match.

In addition to n-step transition probabilities and identifying key players in steady-state distribution, transition matrix and steady-state vector contains numerous valuable probabilistic information, such as the likelihood of the ball reaching a specific player after several passing sequences and the average number of individual pass receptions. Exploring these probabilities further yields deeper insights into the dynamics of the passing network and its evolution throughout the match.