Bayesian Pass Prediction Zone and Logistic Regression in Football: Spain in UEFA Euro 2024 Final

Author: Maged Saeed Abdo Mostafa Kharshom & Precious Prince

Department of Economics, Management and Quantitative Methods, University of Milan

May 2025

UNIVERSITÀ
DEGLI STUDI
DI MILANO

"Technique is not being able to juggle a ball 1000 times. Anyone can do that by practising. Then you can work in the circus. Technique is passing the ball with one touch, with the right speed, to the right foot of your teammate."
— Johan Cruyff

Introduction

Modern football is a game of fine margins, where understanding subtle tactical shifts can determine the outcome of a major final. Predictive modeling and statistical inference have become invaluable tools for analysts in this context. This project investigates the passing behaviour of the Spanish national team during the UEFA Euro 2024 Final using a two-part Bayesian approach that blends spatial analysis with probabilistic modeling.

The project is approached from the perspective of an England team analyst, seeking to assess whether Spain's in-game tactical behaviour, particularly their passing tendencies, remains consistent or changes under the pressure of a final. This has critical implications: if passing patterns remain stable, early-match data can be leveraged to counter Spain's strategy. However, if Spain adapts mid-game, reactive tactics are needed.

Part 1 focuses on real-time tactical evaluation. Here, passing data from Spain's previous matches is treated as prior knowledge, which is then updated using first-half data from the final to form a Bayesian posterior distribution. This process models the likelihood of a successful pass under various contextual conditions (e.g., field zones, pass type, pressure). The goal is to evaluate, mid-match, whether Spain's tactics are consistent enough for England to exploit or whether a strategic rethink is necessary at halftime.

Part 2 performs a more comprehensive post-match tactical comparison. It employs Bayesian logistic regression on event-level pass data to learn posterior distributions over pass success probability for Spain in two datasets: all pre-final matches, and the final itself. By comparing these posterior distributions across key features (e.g., zone of origin, pressure, body part), the analysis reveals how Spain's tactical profile shifted or remained stable in the final relative to their tournament average.

Both parts rely heavily on spatial segmentation of the pitch into "pass prediction zones," enabling localised interpretation of pass dynamics. The Bayesian framework offers not just point estimates but full distributions, which are crucial for dealing with uncertainty in sports contexts.

By combining spatial intelligence with Bayesian reasoning, this project provides a robust, real-time and retrospective method to evaluate team behaviour and tactical evolution, helping coaching staff make informed decisions under uncertainty.

Motivation

Football at the elite level is as much a game of strategic intelligence as it is of physical execution. Coaches and analysts are constantly seeking ways to outmanoeuvre opponents—not just with superior players, but through a deeper understanding of tactical dynamics, decision-making, and adaptability under pressure.

This project is driven by the need to translate real-time football events into actionable strategic insights, especially in high-stakes scenarios like the final of a major international tournament.

One of the core assumptions in football analysis is that players, under similar circumstances, often behave in predictable patterns, many of which are subconscious. For instance, a midfielder under pressure might consistently attempt a short pass to a nearby teammate in a specific zone. Recognising these tendencies early in a match allows opposing teams to prepare and counter them effectively. However, relying solely on first-half data to infer second-half behaviour can be misleading, particularly against intelligent opponents like Spain, who are known for mid-match tactical adaptations.

This is where Bayesian methods shine. Unlike static models, Bayesian inference allows analysts to combine prior knowledge with new evidence. In real-time, this means using past matches as "priors" and updating them as the game unfolds, enabling analysts to evaluate in-game whether a team is adhering to their usual tactics or deviating significantly. This insight is crucial for in-match coaching decisions, such as altering pressing intensity, changing formation, or targeting specific zones on the pitch.

In real-world use, a model like this could be deployed by a national team's analysis department during a tournament:

- In Part 1, an England analyst could monitor Spain's passing behaviour in the first half of the final and, by comparing it to a posterior derived from historical data, determine whether Spain's tactics are consistent or have shifted. If Spain shows a deviation, such as bypassing midfield zones, they typically build through. England's coaching staff could be advised to abandon their original press plan and shift to a more reactive approach.
- In Part 2, logistic Bayesian regression provides post-tournament tactical auditing. By comparing posterior distributions of passing success from earlier matches to those from the final, teams can assess whether tactical risks taken in the final were effective or if pressure led to breakdowns in execution. For Spain's analysts, this could inform whether their strategic identity held firm under tournament pressure, or if further tactical flexibility is needed.

Beyond international tournaments, this methodology has practical applications in club football, where data access is even more granular, and decision-making must happen weekly. Clubs could:

- Use prior-season or first-half data to predict likely second-half zones of attack from opponents.
- Evaluate the consistency of their own team's passing structure across varying contexts (e.g., home vs. away, under pressure, after substitutions).
- Identify individual player trends in positional passing choices and success under pressure for recruitment or tactical planning.

Ultimately, the motivation for this project lies in bridging the gap between raw event data and meaningful tactical intelligence. In an era where milliseconds and meters matter, using Bayesian models to infer, predict, and adapt could be the key to turning analytical insight into competitive advantage on the pitch.

Theoretical Background: Models, Distributions, and Algorithms Used

This section outlines the theoretical foundation of the models, statistical distributions, and algorithms employed in both parts of the project. The analysis involves two interconnected objectives: predicting second-half pass distributions using a Bayesian framework (Part 1) and evaluating contextual features affecting pass success using Bayesian logistic regression (Part 2).

Part 1: Bayesian Updating for Box-to-Box Pass Prediction

Objective

The first part of the project aims to model and visualize the spatial distribution of passes made by Spain during the UEFA Euro 2024 Final. The approach uses probabilistic transition modeling across pitch zones to compare pre-match expectations, first-half behavior, Bayesian-updated predictions, and actual second-half passing data.

Model Description

The football pitch is divided into 24 equal zones (referred to as boxes). The probability of a pass transitioning from one zone to another is modeled using a probabilistic transition matrix. For each zone (box), the model estimates the probability of a pass ending in each of the other 23 zones.

Let:

- $P_{prior}(i \rightarrow j)$: Probability of a pass from box i to box j based on historical data.
- $P_{1st}(i \rightarrow j)$: Observed probability from the first half of the match.
- $P_{post}(i \rightarrow j)$: Posterior probability combining prior and 1st half data.
- $P_{2nd}(i \rightarrow j)$: Observed probability from the second half.

The posterior is computed using Bayesian updating, where the posterior probability is proportional to the product of the prior and the likelihood:

$$P_{post}(i \rightarrow j) \propto P_{prior}(i \rightarrow j) \times P_{1st}(i \rightarrow j)$$

This is an application of Bayes' Theorem in a discrete spatial context. The resulting posterior helps analysts infer how Spain is likely to distribute passes in the second half, using first-half evidence to update prior beliefs.

Distributions

The transitions between boxes can be conceptually modeled using the Dirichlet-multinomial distribution, although explicit inference is not performed here. The Dirichlet prior is suitable for modelling categorical distributions (like transitions between boxes), while the multinomial likelihood represents the count of passes observed between zones.

- Dirichlet Distribution: A multivariate generalisation of the beta distribution, used to model a probability distribution over multiple categories (in this case, 24 zones). It is commonly used as a prior for categorical outcomes.
- Multinomial Distribution: Represents the likelihood of observing counts of events across multiple categories, given a categorical probability distribution.

These distributions are implicitly assumed when computing and updating transition probabilities.

Algorithms and Implementation

The Streamlit application uses the following algorithmic steps:

1. Data Preprocessing:

   - Load and preprocess CSV data containing transition probabilities.
   - Ensure all destination boxes are included for each source box (fill missing entries with zeros).

2. Visualization:

- For each selected starting box, a probability map is drawn using mplsoccer, showing the size and intensity of predicted passes to destination zones.
- A heatmap is generated for each probability matrix (prior, 1st half, posterior, 2nd half) using seaborn's heatmap.

3. Bayesian Update:

- Posterior probabilities are precomputed or calculated externally and loaded for visualization.
- The model assumes prior and likelihood are weighted appropriately (e.g., via empirical Bayesian averaging or count-based updates).

Part 2: Bayesian Logistic Regression for Pass Outcome Classification

Objective

The second part of the project models the probability of a pass being successful using contextual features such as spatial location, pressure status, pass type, and more. The binary outcome (successful or unsuccessful pass) is modeled using Bayesian logistic regression, enabling full posterior inference over model parameters.

Model Description

Bayesian logistic regression is used to model the probability that a pass will be successful. It is a generalised linear model where the log-odds of the binary outcome are expressed as a linear combination of input features:

$\log(P(y=1)/(1-P(y=1))) = \beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$

Or, equivalently, in terms of probability:

$P(y=1) = 1/(1+\exp(-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)))$

Where:

1. y is the binary outcome (1 = successful pass, 0 = unsuccessful).
2. β_0 is the intercept term.
3. β_i are the regression coefficients.
4. x_i are the input features.

Distributions Used

1. Normal Distribution (Priors):

- Coefficients (βi) and intercept (β0) are given Gaussian priors: $\beta_i \sim N(0,\sigma^2)$
- This reflects a belief that coefficients are likely to be centred near zero, with uncertainty represented by a wide standard deviation (σ=10).

2. Bernoulli Distribution (Likelihood):

- The binary outcome is modeled as:
- $y \sim \text{Bernoulli}(p)$,where $p=1/(1+e^{-z})$, $z=\beta_0+x^T\beta$

Algorithm and Sampling Method

The model is implemented using PyMC, a probabilistic programming framework in Python that allows specification and sampling of Bayesian models.

Markov Chain Monte Carlo (MCMC):

- PyMC uses advanced MCMC methods, specifically the No-U-Turn Sampler (NUTS), a Hamiltonian Monte Carlo (HMC) variant. This method efficiently explores high-dimensional posterior distributions.
- A total of 2000 samples are drawn (1000 tuning, 1000 posterior), which helps ensure convergence and accurate uncertainty estimation.

Posterior Inference and Evaluation

- Trace Sampling: The posterior distribution over each coefficient is stored in a trace object.
- Posterior Summaries: Mean, standard deviation, and 95% Highest Density Intervals (HDI) are computed using arviz.summary.
- Posterior Plots: Visualised using arviz.plot_posterior, allowing inspection of the range and credibility of each model parameter.

This approach provides not just predictions, but also confidence intervals and uncertainty quantification, which are critical in real-world decision-making under uncertainty.

Comparative Use of Models Across the Two Parts

Aspect	Part 1: Spatial Transition Modeling	Part 2: Logistic Regression
Task	Predict destination zones of passes	Predict success/failure of a pass
Model Type	Discrete probabilistic model (Bayesian updating)	Bayesian Generalized Linear Model (GLM)
Distribution	Dirichlet prior + Multinomial likelihood (implicit)	Normal prior + Bernoulli likelihood
Algorithm	Grid-based updating and visualization	MCMC sampling using NUTS (HMC)
Output	Posterior probability distributions over zones	Posterior distributions over coefficients

Here's a complete and structured Data section for your report, rewritten clearly to reflect your entire data pipeline, both for the transition probability model (Part 1) and the logistic regression model (Part 2):

Data

This project uses event-level football data sourced from the StatsBomb Open Data API, focusing specifically on Spain's matches in the UEFA Euro 2024.

Data Collection and Filtering

- Only matches involving Spain were selected using match IDs retrieved from the StatsBomb API.

- For each selected match, the event data was extracted, and only pass events (both successful and unsuccessful) made by Spain were retained.
- The dataset was filtered to include only essential columns:
- ['location', 'match_id', 'period', 'minute', 'pass_end_location', 'pass_outcome', 'pass_type', 'possession_team', 'team', 'type']

Coordinate Extraction and Spatial Grid

To enable spatial modeling, the following transformations were performed:

- The location and pass_end_location fields were converted from lists to tuples for consistency.
- From these, four spatial features were extracted:
  - x, y: start coordinates of the pass
  - endx, endy: end coordinates of the pass

These coordinates were used to assign each pass to pitch zones using a custom 6×4 grid layout (24 total zones). Each pass was labeled with:

- box_start: the zone the pass originated from
- box_end: the zone the pass ended in

Data Splitting for Temporal Comparison

To assess tactical evolution and real-time adaptation, the data was split into:

- Pre-final matches (training)
- First half of the final
- Second half of the final

This allowed comparisons between historical tendencies and in-game strategic shifts during the final.

Part 1: Pass Transition Model

Using the spatial zone labels, a transition probability matrix was constructed:

- Passes were grouped by their origin (box_start) and destination (box_end) zones.
- Transition counts were aggregated and converted into probabilities:
- transition_counts = prior_match_pass.groupby(['box_start', 'box_end']).size().reset_index(name='count')
- transition_counts['probability'] = transition_counts['count'] / transition_counts.groupby('box_start')['count'].transform('sum')
- A 24x24 matrix was generated to visualize the likelihood of transitions between all zones.

Separate matrices were computed for:

- All passes prior to the final
- Passes from the first half of the final
- Passes from the second half of the final

Finally, all matrices were merged to compute a posterior transition matrix by summing counts and recalculating probabilities. This comparison highlights how Spain’s spatial passing patterns evolved during the match.

Part 2: Logistic Regression Model

For pass outcome prediction, additional features were engineered:

- Normal Pass: 1 if the pass type was “Normal Pass”; else 0
- Leg Pass: 1 if the pass was made with either foot; else 0
- Under Pressure: 1 if the player was under pressure; else 0
- Pass Outcome: Target variable, 1 if the pass was completed, 0 otherwise

Final Dataset for Regression

The features and target used in the Bayesian logistic regression model were:

- Features:
  ['box_start', 'box_end', 'Normal Pass', 'Leg Pass', 'Under Pressure']
- Target:
  'Pass Outcome'

X = train_xp[features]
y = train_xp[target]

The model uses passes from pre-final matches for training, and applies the learned posterior distribution to predict and analyze passes in both halves of the final.

Results

Interpretation of Results (Part 1)

The results in Part 1 are highly situational and nuanced, making them less suitable for definitive statistical interpretation. Their meaning depends heavily on tactical context, player roles, and in-game scenarios. Therefore, a detailed interpretation is best left to an experienced football coach who can assess the data in relation to specific match strategies and qualitative insights.

Interpretation of Results (Part 2)

The Bayesian logistic regression models provide a detailed understanding of how different features influence the probability of a successful pass in football. The analysis compares Spain’s performance across two phases — before the final and during the final match against England — to evaluate how key passing factors evolved.

Before the Final

The baseline probability of pass success was high, with the intercept indicating strong average performance when all predictors were at their mean. Passes starting further upfield ('box_start') positively influenced success, suggesting that initiating passes from advanced positions helped maintain possession. In contrast, passes ending deeper in the attacking third ('box_end') showed a strong negative effect, implying that final-third passes carried higher risk.

Among pass types, 'Leg Pass' had a moderate positive effect on success, indicating controlled and deliberate plays. 'Normal Pass' had uncertain impact, with their interval overlapping zero, while being under pressure significantly reduced pass success — a finding consistent with in-game expectations.

Final vs England

In the final, the intercept increased slightly, suggesting that overall pass success was marginally higher. However, some feature dynamics shifted. The impact of 'box_start' slightly decreased, while 'box_end' became even more negatively associated with success, possibly reflecting England’s more compact defensive structure.

Interestingly, 'Normal Pass' became more effective, likely reflecting Spain's strategic emphasis on structured passing under high-stakes conditions. Conversely, 'Leg Pass' lost their earlier advantage and showed no clear impact on pass success in the final, potentially due to changes in defensive pressure or tactical setup. Being under pressure still negatively affected performance, though its impact lessened slightly, suggesting improved handling of pressure by Spanish players.

Pre-Final (Training) Model

Coefficient	Mean	94% HDI	Interpretation
Intercept	2.8	[2.7, 3.0]	High baseline success (~94% success rate at feature means).
box_start	1.0	[0.85, 1.2]	Positive effect: Forward-origin passes more successful.
box_end	-2.0	[-2.2, -1.8]	Strong negative effect: Long passes (high box_end) riskier.
Normal Pass	~0.03	[-0.08, 0.14]	Uncertain effect: No clear impact on success.
Leg Pass	0.25	[0.15, 0.36]	Moderate positive: Foot passes outperform others (e.g., headers).
Under Pressure	-0.35	[-0.45, -0.25]	Significant negative: Pressure reduces success probability by ~8%.

Final Match (vs. England) Model

Coefficient	Mean	94% HDI	Change vs. Prior
Intercept	3.0	[2.6, 3.5]	Slightly higher baseline success.
box_start	0.85	[0.4, 1.4]	Weaker effect → Spain used more varied starting positions.
box_end	-2.3	[-2.9, -1.7]	More negative → England's press made long passes riskier.
Normal Pass	0.32	[0.05, 0.58]	Now significant → Simpler passes worked better.
Leg Pass	-0.034	[-0.29, 0.26]	Lost impact → No advantage over non-leg passes.
Under Pressure	-0.25	[-0.51, ~0.00]	Less severe → Spain adapted better under pressure.

Summary

These findings illustrate how passing success in football is highly context dependent. Spain adapted to the conditions of the final, with some strategies becoming more effective and others less so. The model reveals not only which features matter, but also how their importance can shift based on the opponent and match context. Such insights are critical for coaching decisions and performance analysis in competitive football.