GRADO EN ECONOMÍA

CURSO 2023-2024

TRABAJO FIN DE GRADO

Simulador de partidos de fútbol con cadenas de Markov

Markov Chains based football match simulator

Raúl Collazo Berasategui

José Luis Gallego Gómez

Julio 2024

“Para mi familia y amigos, que siempre estuvieron presentes
y empujaron por mí cuando fue necesario.

Espero que os sintáis tan orgullosos de mí
como yo me siento de que forméis parte de mi día a día.”

Índice

RESUMEN

En un contexto futbolístico en el que la disposición de grandes bases de datos se ha vuelto cada vez más accesible, las métodologías de predicción han visto su demanda aumentada dada la utilidad de estas herramientas para clubes y miembros técnicos para tomar decisiones óptimas. En este trabajo, se ha desarrollado un simulador capaz de evaluar la capacidad de toma de decisiones de movimientos y zonas de los futbolistas en el campo. Fundamentado a partir de dos cadenas de Markov independientes entre sí, el simulador es alimentado con datos de LaLiga durante la temporada 2022-2023 para construir una herramienta útil para la optimización de decisiones tácticas de equipos. Para medir la sensibilidad del simulador, se ilustra un partido FC Barcelona - Real Madrid donde se implementan alteraciones en las alineaciones. Los resultados señalan cambios significativos en las matrices de transición al modificar el planteamiento táctico de los equipos.

In a football context in which the availability of large databases has become increasingly accessible, match results predictive methodologies have seen an increase in their demand due to their utility for clubs and technical staff for making optimal decisions. In this work, a simulator is developed, capable of assessing the decision-making abilities of football players' movements and pre-defined zones on the field. Based on two independent Markov chains, the simulator is trained with data from LaLiga during the 2022-2023 season in order to build a tool useful for optimizing tactical decisions of teams. To measure the simulator's sensitivity, a match between FC Barcelona and Real Madrid is presented where alterations in the line-ups are implemented. The results indicate significant changes in the transition matrices when the team line ups are altered.

1 INTRODUCCIÓN

El deporte rey, el fútbol, ha evolucionado considerablemente a lo largo de las décadas, en especial en los últimos treinta años, no sólo en términos tácticos en los que se disputan los partidos sino también en cómo se gestiona y analiza. En un contexto en el que la tecnología y el poder computacional avanza de forma exponencial, el mundo del fútbol ha logrado también su adaptación a los nuevos tiempos redefiniendo sus distintas áreas a través de la implementación del *Big Data*

La ciencia de datos aplicada al deporte o *sport data science* se centra en el análisis de grandes volúmenes de datos y estadísticas para fundamentar la toma de decisiones deportivas apoyados en modelos matemáticos. El empleo de datos no es una ciencia nueva en el deporte pues su uso es común desde hace décadas en los deportes más populares en Estados Unidos (baseball, baloncesto o fútbol americano). Sin embargo, desde la década de comienzo de los 2000, ha surgido esta nueva vertiente del deporte a raíz de los avances tecnológicos en recopilación de información.

La irrupción de los departamentos de análisis de datos en clubes de fútbol han supuesto el desarrollo y potenciación de diversas áreas que se trabajan en un club profesional. En primer lugar, los clubes de fútbol han incorporado a nivel directivo la toma de decisiones comerciales basadas en el análisis exhaustivo de los datos, no sólo para la optimización de operaciones básicas como la venta de entradas, camisetas etc. sino también el desarrollo de estrategias para la promoción en ligas y crecimiento económico relacionadas con el segundo aspecto: el *scouting*.

La figura del ojeador o *scout* en el fútbol siempre ha sido relevante por sus funciones de captación de talento jóven para grandes clubes. Inicialmente, la profesión del scout estaba basada en la experiencia e intuición. Bergkamp et al. (2022) evalúan el trabajo de 125 ojeadores y concluyen que alcanzan decisiones subóptimas. Con la implementación de técnicas de análisis de datos, el ojeador ya no depende únicamente de sus habilidades de observación. Ahora, ya no es necesario asistir presencialmente o revisar cintas de vídeos, el proceso de superivisión de futbolistas se ha optimizado a tal punto que es posible gestionar la evaluación de miles de jugadores con métricas avanzadas sin barreras geográficas, abriendo nuevas oportunidades para la detección de talentos en mercados

poco tradicionales.

Como consecuencia de estas nuevas metodologías, los clubes han desarrollado nuevas estrategias basadas en el análisis de datos, no siendo esta práctica exclusiva del mundo del fútbol. Es bien conocida la historia de los Oakland Athletics de baseball que recoge el libro Moneyball de Lewis (2004) y su posterior adaptación al cine. También, son poulares en la NBA los procesos de selección del *draft* de jugadores universitarios Sailofsky (2018). En el fútbol son múltiples los ejemplos de jugadores desconocidos que fueron descubiertos gracias a departamentos de datos de clubes y que rápidamente se adaptaron a la élite del deporte. Por mencionar alguno, el centrocampista francés N´Golo Kanté fue jugador durante varios años de un equipo de la segunda división francesa donde no destacó por su baja estatura (1.68m). Sin embargo, sus valores extraordinarios en acciones defensivas como la anticipación llamaron la atención del Leicester City, equipo que se alzó campeón Premier League en la temporada 2015-2016, siendo Kanté el jugador revelación de la temporada Kuper and Szymanski (2018). En España también los clubes que refuerzan sus funciones de ojeo con la incorporación de los datos. Es el caso del Athletic Club de Bilbao, dada su filosofía de fichaje de jugadores de origen vasco Gutiérrez Chico (2018), la utilización de metodología analítica les permite filtrar aquellos jugadores y optimizar la configuración de la plantilla cada temporada. A nivel europeo, Firildak and Akin (2020) y Larsen et al. (2020) evalúan las estrategias de reorientación de las academias de Ajax y Borussia Dortmund hacia un modelo de negocio basado en la inversión en cantera para la búsqueda de mayor rentabilidad a través de futuras ventas.

Por último, la revolución originada por el mundo de los datos ha afectado también significativamente la forma en que los clubes analizan el desempeño y rendimiento físico de sus futbolistas permitiendo a los analistas examinar cada aspecto del juego o la predicción de lesiones a través de la modelización del historial lesivo, carga física o cansancio acumulado Pappalardo et al. (2019).

El *Big Data* ha revolucionado las diversas áreas de la escena futbolística. Sin embargo, existe un campo donde el efecto de la irrupción de las ciencias de datos es más pronunciado: las simulaciones. Desde videojuegos de fútbol, casas de apuestas hasta el *staff* técnico en un banquillo de un partido recurren a estas metodologías para la predicción de resultados ya que las simulaciones permiten experimentar virtualmente con diferentes escenarios. A partir de modelización estadística, los simuladores son alimentados de enormes cantidades de datos que incluyen movimientos de futbolistas, patrones de juego o tácticas agregadas.

Este trabajo toma de los artículos de Formento (2022), Zou et al. (2017) y Galaz et al. (2021) la metodología de simulación, basada en cadenas de Markov. El interés en la simulación de estos trabajos propuestos reside en la predicción de resultados de partidos y el análisis de su sensibilidad durante las dinámicas de un partido a través de alteraciones en las configuraciones iniciales.

De los tres trabajos mencionados, el realizado por Galaz et al. (2021) es el mayor exponente de lo que se pretende implementar en este trabajo. Se replica la metodología de simulación markoviana para la distinción entre decisiones del futbolista, a la que además se añade una cuarta decisión que modeliza la pérdida involuntaria del balón por parte del futbolista. También, se incorpora una distinción de las zonas del terreno de juego donde se desarrolla el partido. Asimismo, se incluye una sección de análisis de la sensibilidad del simulador. En lugar de un intercambio de futbolistas entre equipos como se propone en Galaz et al. (2021), el ejercicio planteado en este trabajo para medir la sensibilidad del simulador consiste en modificaciones de la alineación inicial de un equipo para enfrentarse a otro.

Con este trabajo comprobamos cómo una de las ventajas más notables que ofrece la simulación es su capacidad para probar y ajustar estrategias en el campo antes de la disputa de un partido. Por ejemplo, en este trabajo se compara la alineación propuesta del 4-3-3 del FC Barcelona contra el Real Madrid frente a una alternativa ficticia propuesta de un 3-4-3 más ofensivo. A través de esta simulación, se analiza después cómo afectan los ajustes del entrenador en la alineación al comportamiento del equipo en rendimiento de goles, juego por zonas del campo o efectividad en distintos

aspectos del juego.

La estructura de este trabajo es la siguiente: Primero, se realiza una revisión de investigaciones previas relacionadas con el análisis de datos en el deporte y simulación en el deporte, centrada especialmente en el fútbol. Seguidamente, se introduce un apartado de datos en el que se añade contexto sobre el origen y formato de los mismos. Además, se incluye una subsección descriptiva con el objetivo de ilustrar el amplio abanico de posibilidades que abre este formato de datos. En la tercera sección del trabajo se detalla la metodología implementada, la descripción de los supuestos que configuran el simulador y las regresiones logit que también se incorporan al mismo. Después, se presentan los resultados de simulación del partido seleccionado y se añade un ejercicio de medición de la sensibilidad del mismo para demostrar la validez de este. Por último, el trabajo finaliza con un apartado de conclusiones y recomendaciones donde se sugieren líneas futuras para implementar al simulador.

2 MARCO METODOLÓGICO

En esta sección del trabajo se proporciona un marco teórico y contextual sobre la implementación del análisis de datos al fútbol y sus aplicaciones en términos de simulación. Para ello, se revisan en este apartado las contribuciones académicas así como aportaciones recientes del sector profesional de carácter relevante.

2.1 Métricas populares como variables proxy a implementar en el simulador: Expected Goal y Expected Threat

En el contexto del análisis moderno del fútbol son dos las métricas que han ganado protagonismo debido a su capacidad para cuantificar aspectos del juego que antes se consideraban intagibles o díficilmente medibles: el **Expected Goal** (xG) o Goles Esperados y **Expected Thread** (xT) Secuencia Esperada.

El término *Expected Goal* se refiere a la probabilidad de que un disparo a portería termine en gol. Fue popularizada en 2012 por el trabajador de Opta (empresa proveedora de datos) Sam Green. El impacto de la creación de esta variable resultó revolucionario pues había creado una métrica que cuantifica la calidad de cada tiro. El xG se construye a partir de modelos de probabilidad, generalmente logit, que consideran variables propias de cada disparo como la distancia, ángulo, posición del portero rival etc. y Tippett (2019) y Sumpter (2017). Esta variable permite cuantificar la calidad de las oportunidades de los jugadores al disparar y también ofrece una visión agregada del rendimiento ofensivo del equipo al finalizar el partido. Al igual que con los datos, existen distintos proveedores de esta métrica y consecuentemente, una modelización propia.

En el ámbito académico son varias las publicaciones que proponen ejemplos de modelización de esta métrica como en Mead et al. (2023), que además implementa características no consideradas generalmente como la habilidad o cuestiones psicológicas del jugador en el momento de realizar el disparo.

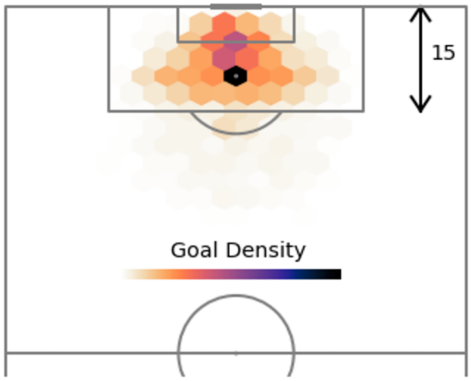


Figura 1: Mapa de calor de disparos que terminaron en gol. Fuente: Mead et al. (2023)

También son frecuentes artículos donde se realizan ejercicios de validación de variables que influyen en la construcción del xG como la distancia entre otros Kullowatz (2017) o Fernandez et al. (2020)

Por otro lado, el *Expected Threat* (xT) mide la probabilidad de que una posesión iniciada en una determinada región del terreno de juego finalice en gol. Esta métrica se extienda más allá de los disparos a portería como el xG ya que ahora se considera la contribución aportada por cada jugada anterior a la construcción de la ocasión de gol. Al considerarlo, el xT proporciona una versión detallada de cómo las acciones individuales de regate, pases filtrados o movimientos en carrera contribuyen al avance del equipo a la consecución de un gol.

Se trata de una métrica muy reciente en el sector, propuesta por el *Data Scientist* del Arsenal FC en su web personal Singh (2018) rápidamente ganó popularidad y se ha convertido en una de las herramientas más prácticas dentro del gremio de analistas. Generalizando a un campo de fútbol dividido en 192 subzonas de mismo tamaño, la fórmula del *Expected Threat* es:

$$xT_{x,y} = (s_{x,y} \times g_{x,y}) + \left(m_{x,y} \times \sum_{z=1}^{16} \sum_{w=1}^{12} T_{(x,y)\to(z,w)} xT_{z,w} \right) \tag{1}$$

donde $s_{x,y}$ es la probabilidad de que el jugador dispare desde la zona (x, y), $g_{x,y}$ es la probabilidad de marcar un gol si se dispara desde la zona (x, y) y $m_{x,y}$ es la probabilidad de que el jugador transicione con la pelota desde la zona (x, y) hacia cada zona (z, w).

Resulta evidente comprobar que la configuración de esta variable está estrechamente vinculada con los procesos markovianos en los que más adelante se basará la creación del simulador. Dada su utilidad en el análisis de las secuencias de los equipos, son varios los artículos que evalúan la capacidad de futbolistas con estas métricas, como Wisdom and Javed (2023), que ponen de ejemplo el xT como recurso útil entre otras métricas. Además del *Expected Goal* y *Expected Threat* han surgido también otras variables que sirven para cuantificar otros aspectos del juego como la precisión en pases o regates de los jugadores. Finnoff et al. (2002), Anzer and Bauer (2022) o Mackay (2017) son alguno de los ejemplos que inspiran más adelante la configuración del simulador y modelos propuestos durante este trabajo.

2.2 Técnicas de simulación

Este subapartado se centra en la exploración de configuraciones y aplicaciones de simuladores en el fútbol. Tal y cómo se introdujo al inicio de este trabajo, las simulaciones tienen diversos usos como las predicciones que se emplean para analizar el impacto de decisiones tácticas o cambios específicos. A continuación, se exponen diversos trabajos que tratan este área:

En Goddard (2005), se diseña a partir de modelos de regresión que consideran variables similares a las que incorpora los modelos posteriores de xG un simulador capaz de predecir los resultados de los partidos. Del mismo modo, Hucaljuk and Rakipović (2011) o Prasetio and Harlili (2016) recurriendo también a regresiones econométricas construyen herramientass capaces de predecir los resultados de partidos con información de la temporada 2015-2016 de la Premier League inglesa o datos de jugadores de la base del videojuego FIFA, respectivamente.

Tabla 1: Variables empleadas para la predicción de un resultado. Fuente: Goddard (2005)

Variable	**Descripción**
$F_{i,d,s}$	Goles totales anotados de local en la última temporada, en la misma división.
$A_{i,d,s}$	Goles totales encajados de local en la última temporada, en la misma división
S_i	Goles anotados en el partido más reciente
C_i	Goles concedidos en el partido más reciente.
R_i	Resultado del partido más reciente.
$SIGH_{i,j}$	Variable binaria. 1 si el partido tiene relevancia para i y j.
$SIGH_i$	Variable binaria. 1 si el partido tiene relevancia para i y j.
$DIST_{i,j}$	Distancia entre los estadios
Att_i	Asistencia de espectadores al estadio.

Desde una perspectiva vinculada a la metodología que se emplea en este trabajo, ya anticipamos que la métrica del *Expected Threat* estaba estrechamente relacionada con la simulación mediante cadenas de Markov, en Fernandez et al. (2020) se ilustra como con matrices de probabilidades de movimientos de los futbolistas en el campo es posible simular una secuencia de jugadas que puede terminar en gol.

En cuanto a simuladores de partidos constuidos exclusivamente a partir de metodología markoviana destacamos los trabajos de Formento (2022) y Zou et al. (2017) en el que se desarrollan simuladores de partidos en las técnicas mencionadas y muestran las dinámicas del mismo. Pero, sin lugar a duda, el artículo en el que se ha centrado este trabajo como fuente de inspiración es el planteado por Galaz et al. (2021). En él, el simulador se configura a partir de las siguientes ideas:

La decisión de un futbolista se reduce a tres movimientos: pasar, regatear y disparar. Esto se determina debido a la frecuencia de este tipo de eventos durante el desarrollo del partido. Cuando el jugador tiene el balón en posesión, se enfrenta a un modelo logit multinomial en el que decide que movimiento realizar. Además, una vez ha ejecutado su movimiento se determina la probabilidad de éxito de la ejecución a través de un modelo logit para avanzar en la cadena.

El objetivo último de Galaz et al. (2021) consiste en no sólo simular un campeonato de la Premier League en la temporada 2017-2018 sino medir la sensibilidad del simulador a través de intercambios de futbolistas entre plantillas (Maguire-Holding, Morata-Lukaku, David Silva-Dele Alli) e inferir la probabilidad de campeonar en liga y medir el impacto de dichos intercambios de jugadores.

3 METODOLOGÍA

3.1 Datos

En esta sección se decribe la naturaleza de los datos que son utilizados durante el trabajo así como la fuente de donde son extraídos. De igual modo, se incluye un análisis exploratorio inicial y una contextualización de los mismos.

3.1.1 Fuente

Los datos empleados en la realización de este trabajo fueron suministrados por Opta, con la que se dispone de un acuerdo de colaboración para la realización de este trabajo. Opta es una empresa británica especializada en la recopilación, análisis y distribución de datos deportivos que ofrece una amplia variedad de productos y servicios fundamentados en sus datos para distintas audiencias: clubes deportivos, casas de apuestas, medios de comunicación o aficionados. Opta suministra principalmente datos de fútbol aunque también abarca otros deportes como el cricket, rugby o baloncesto entre otros. En el ámbito futbolístico, Opta se enmarca como el distribuidor oficial de competiciones domésticas de élite como la Premier League (Inglaterra), LaLiga (España) o Ligue 1 (Francia) y sus respectivas divisiones inferiores.

En su apartado técnico, Opta se caracteriza por un proceso de recopilación de datos altamente detallado que combina metodología automatizada con supervisión humana. Entre algunos de los recursos que se dispone para capturar la información, se cuenta con sistemas de cámaras especializadas y algoritmos de seguimiento de movimiento que verifican que el dato producido es preciso.

3.1.2 Contextualización y estructura de los datos

Para la realización de este trabajo se utilizan datos de la liga española (LaLiga) para la temporada 2022-2023. En esta competición se enfrentan veinte equipos entre sí a través de las 38 jornadas que componen el calendario de una temporada. La participación de los equipos en este torneo viene determinado por los resultados de la liga de la temporada anterior: Los 17 equipos con mayor puntuación de la temporada pasada y 3 equipos ascendidos de la segunda división española (LaLiga 2). Durante el torneo, cada club se enfrentará en dos partidos contra los otros diecinueve participantes, disputando un partido en su estadio y otro en el estadio rival. La clasificación al final de la temporada queda determinada por la puntuación obtenida del resultado del total de partidos: Victoria (+3 puntos), Empate (+1 punto) y Derrota (+0 puntos). Para este trabajo, la muestra de datos se compone de los datos generados durante las 38 jornadas, estando cada jornada del campeonato compuesta por 10 partidos. En otras palabras, se dispone de la información de todas las acciones generadas durante un encuentro para un total de 380 partidos.

El formato de datos con el que se trabajará se conoce como *eventing*. Este formato de datos en el fútbol hace referencia al amplio catálogo de acciones que se desarrollan durante la disputa de un partido de fútbol. Esto incluye goles, pases, duelos, tarjetas recibidas, entre otros. Los avances conseguidos en captura de la información permiten la extracción de variables de información valiosa sobre cada evento. Tomando como ejemplo de evento un pase del jugador A al jugador B, es posible capturar la posición en coordenadas del campo donde el jugador A ejecuta el pase, el instante en el que se produce, clasificar el tipo de pase, la distancia recorrida, el resultado que tuvo esa acción o la localización final del balón en el campo cuando la acción finaliza entre otros.

Para la lectura del archivo que suministra Opta y acceder a la información desagregada primero debemos comprender la naturaleza del archivo en la que los datos son presentados en bruto: Opta genera un archivo XML para cada partido. En él se encuentra la información completa del partido pero estructurada de forma ramificada, partiendo de lo más general a lo más específico. Para lograr la lectura, se dispone de un bucle de programación de lenguaje de R que recorre los archivos XML de los 380 partidos y transforma toda la información en una matriz de datos. Para facilitar la lectura y configuración de la base de datos, este trabajo empleó el paquete de R aplicado a fútbol soccergraphR desarrollado por Lagos (2024).

Como se menciona anteriormente, la información de cada partido es entregada agregada en ramificaciones que parten desde lo general a lo específico. A continuación, se incluye un conjunto de tablas que ayudan a comprender cada uno de los tres estratos en los que se subdivide la información relativa a cada partido. La Tabla 1 hace referencia a los metadatos del partido.

Tabla 2: Metadatos

Variable	**Descripción**
ID	Identificador único de cada partido.
Additional Info	Añadido para partidos a puerta cerrada o con aforo limitado.
Home Score	Marcador del equipo local.
Away Score	Marcador del equipo visitante.
Home Team ID	Identificador único del equipo local.
Away Team ID	Identificador único del equipo visitante.
Home Team Name	Nombre del equipo local.
Away Team Name	Nombre del equipo visitante.
Game Date	Fecha de disputa del partido.
Matchday	Jornada del calendario de liga.
Season ID	Identificador único de la temporada del partido.

En el siguiente estrato se encuentran los eventos que suceden durante el partido. En la Tabla 2 se recoge la información relativa a las variables más significativas que definen cada evento.

Tabla 3: Variables relativas a un evento

Atributo	**Descripción**
ID	Identificador único de cada evento.
Type ID	Clasificación según tipo del evento.
Period	Periodo en el que se produce el evento.
Min	Minuto del partido en el que se produce el evento.
Sec	Segundo del partido en el que se produce el evento.
Team ID	Identificador único del equipo del jugador que produce el evento.
Player ID	Identificador único del jugador que produce el evento.
Player Name	Nombre del futbolista que produce el evento.
Outcome	Operador booleano. Toma 1 si el evento se completa con éxito.
X	Coordenada X de la longitud del campo donde se produce el evento.
Y	Coordenada Y de la amplitud del campo donde se produce el evento.

Por último, Opta incluye en el registro de eventos los calificadores. Son más de 300 variables binarias que ayudan a caracterizar cada evento. En total, pueden producirse en un partido hasta 84 tipos distintos de eventos que van desde goles, pases, tiros, hasta intervenciones arbitrales. A modo de ejemplo, si filtramos dentro de los eventos referidos a pases encontramos calificadores que caracterizan el tipo de pase (balón largo, centro, tiro libre directo/indirecto, saque de esquina), parte del cuerpo con la que se ejecuta el golpeo del balón, ángulo, ubicación de inicio y final o distancia recorrida. Si un pase precedió a un disparo, será tomado como asistencia y podremos también conocer cuál fue el resultado de ese disparo. Es con la información de estos datos de eventos y sus respectivos calificadores los que más adelante se emplearán como instrumento para la creación de variables para los modelos econométricos.

3.1.3 Análisis exploratorio

Si bien es cierto que la estructura presentada en esquema de árbol compila de forma eficiente la información, al trabajar con una base de datos de tal magnitud, se presentan los datos en formato de tabla. Tal y como se introdujo en la Tabla 2, cada evento generado dispone de su propio identificador o tipo que permite filtrar de manera eficiente y facilita la visualización.

Inicialmente la base de datos cuenta aproximadamente con 660.000 eventos. Filtrando descendentemente los eventos más relevantes de juego (omitiendo por ejemplo eventos relativos a reanudaciones del juego cuando el balón sale del campo) se obtiene la siguiente Tabla 3. Consulte Tipos de Evento en el Anexo para conocer en detalle la definición de cada evento. Como en Galaz et al. (2021),

Tabla 4: Eventos

Evento	**Cantidad**	**Porcentaje sobre el total**
Pase	357.793	54.13 %
Recuperación de balón	38.864	5.88 %
Toque de balón	25.417	3.84 %
Duelo aéreo	20.520	3.10 %
Falta	19.680	2.97 %
Regate	19.196	2.91 %
Despeje	14.096	2.13 %
Balón dividido	13.646	2.06 %
Entrada	12.152	1.83 %
Saque de Esquina	7.274	1.10 %
Ocasión salvada	4.548	0.68 %
Fallo	3.693	0.55 %
Goles	955	0.14 %
Poste	196	0.03 %

se define un partido de fútbol como una síntesis de tres eventos principales: **Pases**, **Duelos** y **Disparos**. De esta manera, cuando un futbolista recibe un balón de otro jugador se le presentan tres alternativas: pasar de nuevo, disparar a portería o regatear (duelo). Los pases representan más de la mitad (54 %) de los eventos que se produjeron en la competición durante la temporada 2022-2023 y son la principal acción con la que se desencadena el resto de jugadas. Los disparos son el recurso futbolístico con el que se anotan los goles y se determina el resultado del partido. Por último, los regates o duelos son ocasiones en las que se enfrenta un jugador de cada equipo con el objetivo de ganar/defender un espacio del campo que es sensible a generar una ocasión con ventaja para un equipo.

En la siguiente página se incluye una tabla que contiene los principales estadísticos de las tres acciones de los futbolistas consideradas para cada club que participó en la temporada.

Tabla 5: Principales estadísticos por partido de equipos según los eventos considerados

Equipo	Media Pases	Desviación Estándar	Media Regates	Desviación Estándar	Media Disparos	Desviación Estándar
Athletic Club	487.74	88.63	20.58	5.68	14.50	5.01
Mallorca	367.34	80.20	15.03	4.97	8.47	3.21
Rayo Vallecano	341.12	216.10	14.48	8.80	11.65	6.14
Barcelona	537.50	258.84	19.49	9.90	13.32	6.77
Elche	236.42	208.67	13.68	9.56	8.10	5.84
Real Betis	478.29	98.56	21.32	6.25	11.08	4.55
Espanyol	224.65	184.77	11.25	8.36	8.07	5.73
Celta de Vigo	484.68	103.18	19.37	6.23	12.47	4.24
Real Sociedad	500.79	88.75	19.13	6.43	12.37	3.29
Cádiz	354.16	74.17	15.13	4.57	9.29	3.96
Sevilla	410.04	225.21	14.71	7.33	11.61	6.42
Osasuna	441.45	90.97	17.68	7.75	11.79	3.91
Almería	251.10	198.22	14.40	5.57	9.76	4.74
Real Madrid	655.87	113.48	27.61	8.48	17.03	6.40
Atlético de Madrid	534.68	98.25	17.95	5.83	13.89	4.63
Getafe	369.66	76.29	12.97	4.18	9.87	3.91
Villarreal	383.74	216.66	19.02	10.93	11.83	6.22
Real Valladolid	225.96	217.28	13.02	9.36	7.81	6.25
Valencia CF	458.21	110.80	20.61	7.07	12.89	4.63
Girona	357.79	216.12	15.70	8.28	9.85	5.51

El resto de los eventos producidos en la competición que no pueden ser agrupados dentro de esta triple clasificación no son considerados para la simulación pues no se pretende modelizar en la simulación interrupciones del partido como las faltas. Sin embargo, todas las reanudaciones posteriores a una interrupción del partido de acuerdo a las Reglas de Juego de la International Football Association Board (2024) como un tiro libre directo o indirecto (reanudación de una falta), así como un saque de meta o de esquina serán tratados y clasificados como pases o disparos.

El abanico de análisis que permite este formato de datos es amplio. A continuación, se presenta una serie de gráficos con los que se ilustra el potencial económico-deportivo de este tipo de bases de datos. Para ello, se adopta la perspectiva de club del campeón de la competición en la edición que abarca la base de datos: el F.C. Barcelona.

En cada temporada se otorga un premio individual al máximo goleador de la competición. Este galardón se conoce popularmente como ”Trofeo Pichichi”. En la temporada que se corresponde a nuestra muestra el máximo anotador de la competición fue el delantero polaco Robert Lewandowski, jugador del FC Barcelona. A continuación, se muestra en la Figura 1 la distribución de los 135 disparos que realizó Robert Lewandowski a lo largo de la temporada con los que logró anotar 23 goles. Según el resultado final de cada disparo, han sido representados en Goles-Atajados-Errados.

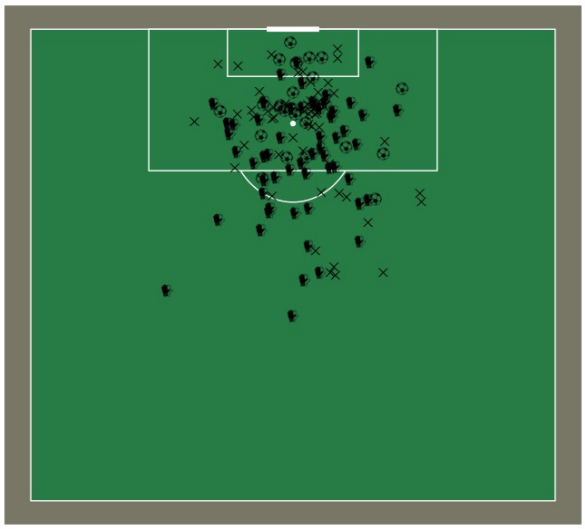


Figura 2: Mapa de Disparos Robert Lewandowski

Añadiendo un grado mayor de profundidad a la visualización de los datos, podemos realizar un enfoque táctico de un partido. Por ejemplo, tomaremos como objeto de estudio el partido FC Barcelona - Real Madrid disputado el 19 de marzo de 2023. A partir de los eventos de pases de los jugadores del once inicial del FC Barcelona en y promediando la posición de cada futbolista en el campo, al representar se obtiene:

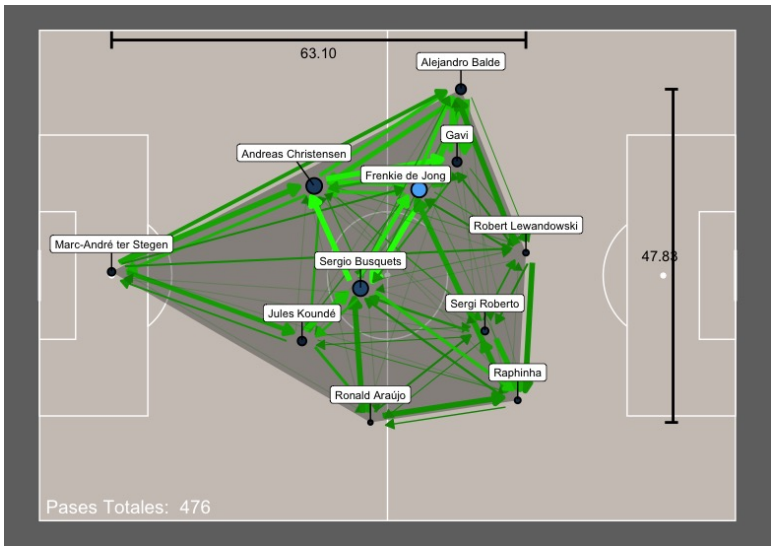


Figura 3: Mapa de pases

Este gráfico permite el análisis y entendimiento táctico del esquema planteado por el entrenador y el comportamiento del equipo en cuanto a relación con el balón cuando se dispone de la posesión del balón. Inicialmente, el entrenador del Barcelona, Xavi Hernández, alineó a los jugadores formando una estructura de 4-3-3 (defensas-mediocampistas-delanteros). Sin embargo, cuando el equipo mantenía el balón durante el partido, la estructura del bloque formado por los jugadores se alteró. Destacamos entre varios análisis del juego, que el lateral izquierdo en ese partido, Alejandro Balde, añadió profundidad y amplitud al ataque mientras que el extremo, Gavi, con el que Balde compartía banda, jugó más céntrico.

También, a partir de la información que extraemos de cada evento, podemos configurar métricas que nos permite cuantificar la profundidad o amplitud del bloque formado por el equipo y entender el estilo de juego del equipo. Incluso, es posible establecer relaciones de asociación entre parejas de futbolistas a partir del conteo de sus interacciones mediante los pases y representarlo en el campo. A la vista de estas interacciones, señaladas en el gráfico mediante flechas donde, a mayor relación de pase entre dos jugadores mayor intensidad del color de la flecha, concluimos desde una perspectiva deportiva, que el estilo de juego del equipo, como cabría esperar, estuvo caracterizado por el protagonismo de los centrocampistas.

3.2 Diseño de un simulador

En esta sección del trabajo, se desarrolla y aplica una metodología de simulación con el objetivo de analizar las dinámicas de un partido de fútbol a partir de los eventos que se producen. El principal elemento diferenciador del simulador consistirá en sus propiedades para simular las secuencias y observar el comportamiento de los equipos en términos de toma de decisiones futbolísticas así como las regiones por las que se desarrolla cada secuencia de eventos. Adicionalmente, en cada eslabón de la secuencia, se evalúa la probabilidad de acierto de la decisión tomada por el futbolista (pasar, regatear en un duelo o disparar a portería) a través de un modelo logit. Tanto el simulador como los modelos de probabilidad incluidos en el mismo son validados utilizando los datos de los 380 partidos disputados durante la temporada.

3.2.1 Supuestos precios para la construcción del simulador.

Para el desarrollo del simulador, se establecen una serie de supuestos fundamentales que permiten simplificar la complejidad de la modelización de todos los elementos que componen un partido

de fútbol, todo ello sin comprometer su eficacia y relevancia práctica. Los supuestos han sido seleccionados cuidadosamente para lograr un equilibrio entre precisión y practicidad, permitiendo que el simulador reproduzca el desarrollo de un partido, considerando condiciones reales que proporcionan una base robusta para un posterior análisis y extracción de conclusiones. Los supuestos son los siguientes:

Duración del partido: Conforme a las Reglas de Juego de International Football Association Board (2024), un partido tiene una duración de 90 minutos, a los que se añade un tiempo adicional en cada mitad del partido con el propósito de compensar las interrupciones que ocurren durante el desarrollo del partido como sustituciones, lesiones u otros lances del juego. En nuestro simulador, la duración del partido se configura en base a la duración promedio observada en los 380 partidos que componen la competición. Las interrupciones están implícitamente consideradas, ya que el conjunto de datos ya incluye la información relativa a estas. En el Anexo se muestra la distribución de secuencias por partidos durante todo el campeonato.

Zonas que componen el campo de juego: Para simplificar el desarrollo de un partido, el campo de fútbol ha sido dividido en dos mitades: defensiva y ofensiva. La mitad ofensiva, con el objetivo de analizar distintas vertientes estratégicas de ataque, se subdivide a su vez en tres zonas: izquierda, centro y derecha.

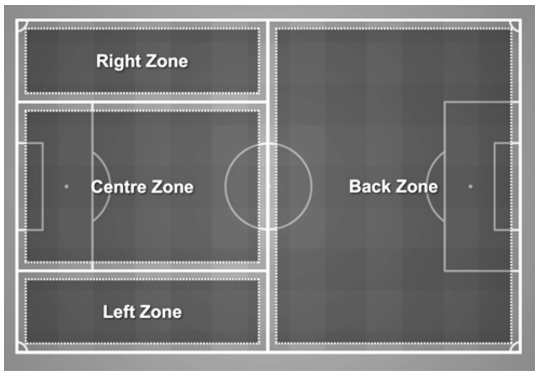


Figura 4: Zonas del campo. Fuente: Opta

Composición de los equipos: Cada equipo disputa el encuentro únicamente con 11 jugadores. No se incorporan las sustituciones al modelo. Además, se presupone que todos los jugadores que constituyen las plantillas de cada club se encuentran en plenas condiciones físicas y están disponibles para ser alineables, eliminando así la necesidad de modelar lesiones o baja condición física.

Efectos por localía: Es bien conocido en el mundo del deporte que jugar en casa puede tener un positivo sobre el equipo que actúa como local debido a diversos factores como la familiaridad con el entorno, por ejemplo, las condiciones climáticas, o el apoyo de la afición. En el contexto del simulador desarrollado en este trabajo, se ha decidido no hacer una diferenciación entre equipo local y visitante. Por ello, se asume que cada partido se juega en un estadio neutro donde no existen ventajas para ninguno de los dos equipos.

Sin embargo, mientras que el simulador opera bajo este supuesto de neutralidad en términos de localía, dada la configuración de la base de datos, los modelos logit de probabilidad empleados para predecir el acierto de la decisión del jugador sí que contemplan este factor entre otros.

Decisión del jugador: Como en Galaz et al. (2021), el simulador reduce la toma de decisiones del futbolista cuando está en posesión del balón a tres alternativas: pasar el balón a un compañero, disparar a portería para anotar un gol o regatear a un oponente. Además, se incorpora la posibilidad de

que el jugador pierda el balón debido a un robo por parte del equipo rival o evento homólogo. Este evento se modela como un cuarto estado en la cadena de Markov, denominado como ”Pérdida”. La incorporación de este último estado permite que el simulador considere tanto las decisiones positivas de avance en el juego como las circustancias negativas.

3.2.2 Cadenas de Markov

La simulación de un partido en este trabajo está caracterizada por la implementación de Cadenas de Markov que permiten la modelización y predicción de las dinámicas de las secuencias de jugadas de un equipo de fútbol. Simultáneamente, se emplean dos cadenas de Markov independientes entre sí para capturar las decisiones de los futbolistas (pase, regate, disparo, pérdida) y la zona del campo donde se desarrolla la jugada (mitad defensiva, mitad ofensiva - izquierda, mitad ofensiva - derecha y mitad ofensiva - centro).

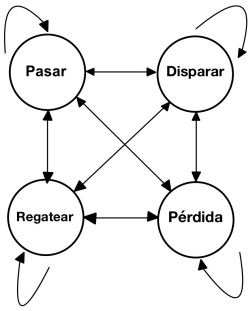


Figura 5: Esquema Cadena de Markov

Una cadena de Markov es un proceso estocástico discreto con una propiedad esencial que la define: ”la falta de memoria”. En otras palabras, la probabilidad de transicionar de un estado a otro depende exclusivamente del estado presente y no de los estados anteriores. A esto se le conoce como la propiedad de Markov.

Según esta propiedad, la probabilidad de transición de un estado a otro dentro de la cadena depende únicamente del estado actual y no de la secuencia que de eventos que preceden. Matemáticamente, si X_n representa el estado en el tiempo n, la propiedad de Markov queda expresada como:

$$P(X_{n+1} = x \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x \mid X_n = x_n) \tag{2}$$

Donde P denota la probabilidad condicional y $x, x_n, x_{n-1}, \ldots, x_0$ son posibles estados de la cadena.

En términos de secuencias de jugadas de fútbol, es razonable pensar que las dinámicas de un partido son compatibles con esta modelización. Para conocer en mayor detalle las características de este tipo de procesos estocásticos, se ha recurrido a Neal (1993), Feller (1968) y Kalbfleisch and Lawless (1984).

Las dinámicas que siguen las cadenas de Markov se representan a través de la matriz de transición, donde cada elemento p_{ij} contenido en ella representa la probabilidad de transicionar del estado x_i al estado x_j. Dado un conjunto de estados genéricos $X = \{x_1, x_2, \ldots, x_n\}$, la matriz de transición para una cadena de Markov queda definida por los elementos no negativos:

$$\mathbf{M} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

donde cada elemento de la matriz es calculado a través de:

$$p_{ij} = P(X_{t+1} = j \mid X_t = i) \tag{3}$$

La implementación de cadenas de markov en el análisis de datos de fútbol es recurrente. No sólo en términos de simulación como en Galaz et al. (2021),también en la creación de métricas avanzadas para evaluar el rendimiento ofensivo de equipos en departamentos de analistas como Yam (2019) o en Sumpter (2017).

Dadas las dos cadenas de Markov consideradas para nuestra simulación, se describe a continuación el diseño de las matrices de transición 4x4 y los resultados (para el equipo medio representativo de toda la competición) de las probabilidades de transición para cada una de ellas:

La primera matriz, recoge las probabilidades de transición entre estados en los que el futbolista, una vez en posesión del balón, realiza el próximo movimiento parte de la secuencia. En la matriz, las probabilidades vienen expresadas en el siguiente orden: Disparo, Regate, Pase y Pérdida.

$$\mathbf{M_{acción}} = \begin{bmatrix} 0.045 & 0.016 & 0.916 & 0.021 \\ 0.039 & 0.034 & 0.857 & 0.018 \\ 0.018 & 0.029 & 0.937 & 0.014 \\ 0.007 & 0.042 & 0.923 & 0.027 \end{bmatrix}$$

De la misma forma, la segunda matriz de transición, que recopila las probabilidades de transición entre las distintas zonas en las que se ha dividido el campo de fútbol en el orden Back-Center-Left-Right es:

$$\mathbf{M_{zonas}} = \begin{bmatrix} 0.613 & 0.157 & 0.117 & 0.111 \\ 0.321 & 0.374 & 0.155 & 0.148 \\ 0.345 & 0.322 & 0.256 & 0.075 \\ 0.351 & 0.328 & 0.070 & 0.248 \end{bmatrix}$$

Se destaca que los resultados obtenidos para las probabilidades mostradas en las matrices de transición consideran los veinte clubes que componen el campeonato de liga. Si se simula un partido determinado, las probabilidades de cada equipo serán distintas entre sí, reflejando las diferencias existentes en calidad entre las plantillas de los clubes o el estilo de juego. Además, existirán también diferencias dentro de cada equipo según el once que se configure para la simulación.

En resumen, la simulación de un partido queda definida por:

1. El partido inicia con una secuencia que comienza con un Pase en la zona ”Back”.
2. En cada eslabón que compone la cadena de markov, el jugador al recibir el balón decide la región y movimiento hacia donde irá.
3. Se determina el resultado de la acción tomada por el jugador. Si es existosa, se avanza hacia el siguiente estado. Por definición, el estado Pérdida siempre finaliza la secuencia.
4. Cada vez que un futbolista realiza un disparo y anota gol se termina la secuencia. Se reanuda el partido de idéntica manera a como se inicial el partido.

5. Cuando un futbolista toma una decisión y ésta no es existosa, el equipo contrario recupera el balón en la misma región donde el jugador del otro equipo perdió la posesión.

3.2.3 Estimación de un modelo logit para determinar el resultado de la acción

Dadas las tres posibles acciones que puede tomar un futbolista cuando tiene el balón en posesión que no incurren en una pérdida de la pelota, modelamos el resultado de la decisión tomada a través de un modelo de probabilidad no lineal logístico. Para el modelo logit, la función de probabilidad se define como:

$$P(Y = \acute{E}xito \mid \mathbf{X}) = \frac{1}{1 + e^{-\mathbf{X}\beta}} \tag{4}$$

donde Y es la variable dependiente binaria que determina el éxito de la acción tomada por el fútbolista, β es el vector de coeficientes y $\mathbf{X}$ es el vector de variables explicativas que caracterizan cada acción posible por el futbolista. Para conocer en detalle el proceso de estimación del modelo logit, se recurrió a Wooldridge (2010) y Greene (2003).

3.2.4 Pases

Los pases son sin lugar a duda el evento más frecuente durante un partido de fútbol. En esta base de datos, correspondiente a la temporada anterior de LaLiga española, supusieron un 54 % del total de eventos observados durante los 380 partidos que componen el campeonato. Dada su relevancia y protagonismo en la creación de cada estilo táctico, son varios los autores que han tratado de modelizar distintos aspectos fundamentales del pase. Mencionando algunos ejemplos, Anzer and Bauer (2022), Spearman et al. (2017) y Arbués-Sangüesa et al. (2020) plantean diferentes modelos con el propósito de modelizar la probabilidad de acierto de un pase.

Para el modelo logit específico para los pases, se consideran las siguientes variables para la estimación: **CCEE** es una variable dicotómica que reúne a aquellos equipos que clasificaron a competiciones europeas al finalizar el campeonato, **min** refiere al minuto del partido en el que se realiza el pase. Las variables **X** e **Y** son las coordenadas ajustadas (debido a que en cada mitad del tiempo reglamentario se disputa en un lado distinto) del pase. **Back**, **Center**, **Left** y **Right** como se describió en la subsección de supuestos de la simulación, son las regiones en las que se ha dividido el campo. Son cuatro variables binarias que recogen si el pase ha sido efectuado en esa zona en concreto.

Para medir los efectos del formato de pase que el jugador da, tenemos las siguientes variables: **Centro** es una variable binaria que informa sobre el tipo de pase, tomando valor uno si el pase fue un centro y cero en caso contrario. La distancia que recorre el balón (en metros) es capturada a través de la variable **Longitud**. Además, disponemos de **Ángulo**, que mide (en radianes) el ángulo del pase con respecto a la dirección que sigue la jugada. La variable **local** es otra variable binaria que informa a cerca del equipo del futbolista que da el pase, si actúa como local, toma valor uno. Para también analizar el efecto del resultado en el jugador que da el pase, disponemos de la variable **Resultado**, que informa de si el equipo en cada instante va ganando, perdiendo o empatando el partido.

Por último, se han creado las siguientes cuatro variables binarias: **GK**, **DF**, **CM** y **DC**. Estas informan si el pasador es un portero, defensor, centrocampista o delantero, respectivamente.

Cabe mencionar que la inclusión de algunas variables consideradas en los modelos logit para determinar el acierto de cada movimiento de los futbolistas no han sido implementadas en la simulación con el motivo de simplificar la misma o bien, debido a la incompatibilidad al discretizar las transiciones entre eventos dados los supuestos del partido.

3.2.5 Disparo

La segunda decisión que puede tomar el futbolista cuando está en posesión del balón es realizar un disparo a la portería rival. El disparo es el evento más importante en el fútbol, pues desencadenan los goles que determinan el resultado de cada partido.

En el fútbol moderno, la probabilidad de anotar un gol en determinadas circustancias se ha convertido en el elemento clave para analistas. La herramienta más reconocida en este ámbito es el concepto de **Expected Goal** (xG), que ofrece una cuantificación de la calidad de las oportunidades de gol basándose en la probabilidad de que el disparo termine en gol.

Los Expected Goals (xG) son calculados a través de modelos estadísticos logísticos como los planteados por Pollard and Reep (1997), que analizan diversos factores asociados al disparo que se pretende evalúar. Para el cálculo, se consideran variables como la distancia, ángulo de disparo, jugador que ejecuta, presencia de defensor u ocasión que precede al disparo entre otros. Los avances e implementación de nuevas tecnologías en captura de datos han permitido el refinamiento de estos modelos.

Dada la estructura de la base de datos, debido a restricciones presupuestarias, se dispone de limitaciones en el ”tracking”de datos sobre los disparos. A la hora de replicar una aproximación fiel al modelo genérico de xG, se han considerado las siguientes variables para su construcción:

De nuevo, para reflejar las diferencias en calidad de plantillas, empleamos la variable **CCEE**, que considera los equipos que clasificaron a final de temporada para competiciones europeas. Esta vez, en vez de considerar todas las regiones en las que se subdivide el campo en el simulador, disponemos de dos variables diferentes para medir las distancias del balón. La primera, **distport** mide la distancia euclídea entre el punto del golpeo del balón a la línea de meta. La otra variable considerada es **Box**, una binaria que informa si el disparo fue realizado dentro del área rival. La justificación del uso de estas variables puede ser explicado a través de la siguiente imagen:

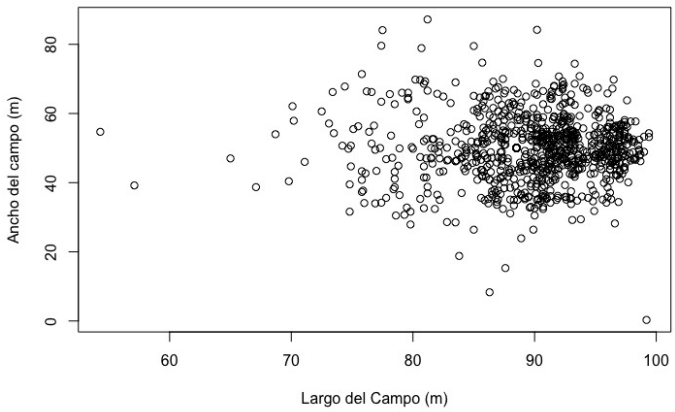


Figura 6: Mapa de disparos que terminaron en gol.

Tal y como se observa en la figura, han sido representados todos los disparos que terminaron en gol durante toda la competición de LaLiga 2022-2023. En total, fueron 955 goles. Existe un denominador común al observar los disparos que entraron: Todos ellos fueron desde la mitad ofensiva del campo, y en su mayoría desde el área rival, por lo que la inclusión de estas dos variables en lugar de la diferenciación de zonas como en el modelo logit de pases queda justificado.

Se ha considerado al igual que para los pases la inclusión del efecto temporal en la precisión. La

variable **min** recoge el momento en el que se ejecuta el tiro. También, se incorpora al modelo los efectos sobre la probabilidad de gol del **resultado** del marcador al disparar o el actuar como **local** durante el partido. Por último, entendemos que los jugadores especializados en anotar goles son los **delanteros**, por lo que mediremos el efecto de que el disparo lo haya efectuado un jugador que ocupa esta posición a través de una variable binaria.

3.2.6 Regate

La última decisión que modelizaremos en este trabajo que el futbtolista puede tomar cuando tiene el balón son los regates. En el fútbol, aunque cada vez escasea más la presencia del futbolista driblador, el regate es una habilidad técnica fundamental que permite a los jugadores guardar la posesión de la pelota, superar a defensores y generar oportunidades de ataque ya que, al completar un regate a un oponente, un futbolista puede crear espacios en defensas compactas, cambiar el ritmo de juego y desencadenar un avance hacia la portería y anotar un gol.

En esta última subsección, se aborda el desarrollo del modelo con el se predice el éxito de un intento de regate. Si bien es complicado realizar una implementación o diseño de modelos para predecir el éxito de un regate, han sido varias las aproximaciones o técnicas que consideran el regate como elemento clave para la creación de métricas avanzadas. Es el caso del *Expected Threat* impulsado por Singh (2018).

Para este trabajo, dado el tipo de los datos disponibles, que no incorporan el tracking de jugadores, no es posible configurar un modelo sofisticado para predecir el acierto del regate a través de la inclusión de variables que tengan en cuenta factores esenciales como la presencia de muchos defensores sobre el jugador que intenta el regate. Es por ello, que con las herramientas disponibles, se plantea el siguiente conjunto de variables para la estimación:

Primero, representando todos los regates completados durante la competición podemos extraer unas ideas preliminares. En la Figura 6 observamos mayor intensidad para la realización de regates en las bandas del campo. También, resulta llamativo, aunque es coherente, la ausencia de regates en las áreas por su dificultad si hablamos de transiciones ofensivas o por el riesgo asumido en acciones defensivas. Para recoger el efecto de estas zonas en el éxito del regate, recuperamos la división original del campo en cuatro zonas, recurriendo a las variables **right** y **left**. De nuevo,

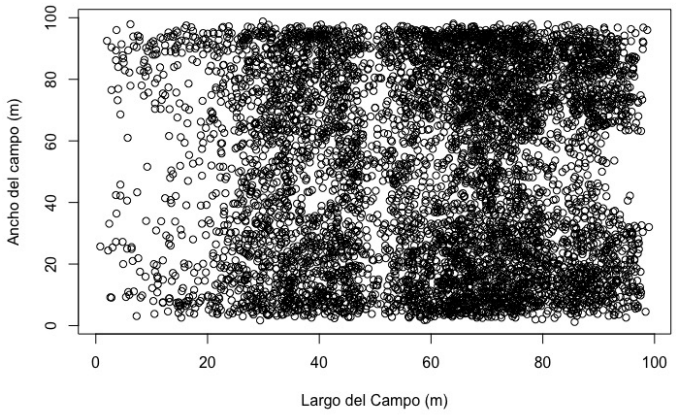


Figura 7: Mapa de regates completados.

consideraremos el efecto del tiempo y la actuación como local con las variables **min** y **local**. También, se incluye la variable **resultado** ya empleada anteriormente. Por útlimo, aunque sí resulta interesante diferenciar individualmente los jugadores, aunque fuese reduciendo para cada club, al considerar

toda la competición la diferenciación del jugador se limita al uso también de las variables **delantero** y **centrocampista** al ser los principales regateadores de los equipos y, una vez más, la variable **CCEE** para reflejar la diferencia de calidad de plantilla entre clubes.

4 RESULTADOS

En esta sección se presenta los resultados de los modelos implementados y ilustra la aplicación del simulador en un partido determinado. Por su relevancia mediática, dentro de todos los encuentros disputados en el campeonato, se ha escogido el partido entre el FC Barcelona y el Real Madrid, conocido popularmente como El Clásico.

4.1 Resultados de la estimación

En la Tabla 6 se adjuntan los resultados de la estimación del modelo para los pases:

Variable	**Coeficiente**	**Error Estándar**	**z-Valor**	**Pr**$(>\|z\|)$
CCEE	3.469e-01	1.757e-02	19.749	$<$ 2e-16 ***
min	-3.998e-05	2.944e-04	-0.136	0.891983
X-coord	3.653e-03	6.384e-04	5.722	1.05e-08 ***
Y-coord	-2.086e-04	3.400e-04	-0.614	0.539545
Back	3.840e+00	1.463e-01	26.244	$<$ 2e-16 ***
Center	2.823e+00	1.504e-01	18.770	$<$ 2e-16 ***
Left	2.960e+00	1.516e-01	19.526	$<$ 2e-16 ***
Right	2.946e+00	1.510e-01	19.509	$<$ 2e-16 ***
Pase_Centro	-3.016e+00	3.285e-02	-91.820	$<$ 2e-16 ***
Longitud	-2.406e-02	6.698e-04	-35.922	$<$ 2e-16 ***
Ángulo	5.589e-03	4.813e-03	1.161	0.245527
Local	6.168e-02	1.607e-02	3.838	0.000124 ***
Resultado	1.417e-02	1.151e-02	1.231	0.218264
GK	-2.465e-01	1.454e-01	1.696	0.089931 .
DF	-2.224e-01	1.407e-01	-1.581	0.113785
CM	-1.723e-01	1.405e-01	-1.226	0.220311
DC	-5.726e-01	1.414e-01	-4.049	5.15e-05 ***

Tabla 6: Estimación del modelo logit de pases

A la vista de los resultados de la estimación, se pueden extraer las siguientes conclusiones: En primer lugar, la variable **CCEE** es estadísticamente significativa a niveles muy bajos. Presenta un efecto marginal positivo. Su interpretación refleja la existencia de un efecto positivo sobre la probabilidad de acierto en el pase de aquellos futbolistas que pertenezcan a clubes que clasificaron a competiciones europeas, siendo esto, una representación de las diferencias de calidad de plantillas entre clubes. La variable **min** y **resultado** no son significativas, mientras que actúar como **local** sí lo es. A efectos de su traducción al lenguaje deportivo, podemos concluir que no existe un efecto que altere la probabilidad dependiendo del minuto en que se realice el pase así como tampoco existe un efecto según el resultado.

Centrando ahora en analizar las variables relacionadas con la localización donde se efectúa el pase, se concluye que todas las variables son muy significativas y con efecto marginal positivo con excepción de la variable que refiere a la coordenada que hace referencia al ancho del campo, **Y-coord**, que no es significativa. Dadas las estimaciones, el signo del efecto positivo de la variable **X-coord** resulta contraintuitivo, ya que entendemos que existe mayor dificultad en acertar un pase a mayor proximidad a la portería rival.

En cuanto a las variables que caracterizan cada pase, observamos que la **longitud** que recorre el balón durante el pase resulta significativa al niveles nulos. Presenta un signo negativo, por lo que como cabría esperar, a mayor distancia que se pretende trasladar el balón habrá menor precisión, y por tanto, menor probabilidad de acertar el pase. Por el contrario, el **ángulo** con el que se dirige el pase respecto a la orientación de la jugada no es significativa.

Por último, las variables relativas a la posición del futbolista presentan un signo negativo en su efecto marginal, siendo únicamente significativas las binarias de delantero y portero, a niveles muy bajos y al 10 % respectivamente.

Los resultados modelo logit para medir la probabilidad de gol de los disparos:

Variable	**Coeficiente**	**Error Estándar**	**z-Valor**	**Pr**$(>\|z\|)$
CCEE	-0.026324	0.270509	-0.097	0.92248
min	-0.027053	0.004680	-5.781	7.44e-09 ***
distport	-0.056886	0.005155	-11.036	$<$ 2e-16 ***
Box	4.072460	0.236064	17.251	$<$ 2e-16 ***
Local	-0.587756	0.238088	-2.469	0.01356 *
DC	0.879283	0.291267	3.019	0.00254 **
Resultado	0.222652	0.189693	1.174	0.24049

Tabla 7: Estimación del modelo logit de disparo

Una vez obtenidas las estimaciones del modelo, podemos concluir lo siguiente a cerca de las variables implementadas: primero, a diferencia del logit de pases, no hay significatividad en la variable **CCEE**, por lo que parece no existir diferencias de calidad entre equipos en términos de anotación de goles. El hecho de jugar como **local** en el estadio es significativo únicamente al 5 %. Esta vez, la variable **min** sí que es muy significativa y presenta un coeficiente negativo. A medida que se aproxima el final del partido, resulta más complicado anotar un gol.

Respecto a las variables de localización del disparo, ambas son significativas también a niveles bajos y el signo de los coeficientes es coherente. A mayor **distancia a porteria** menor probabilidad de gol. Por otra parte, el hecho de que el tiro se ejecute dentro del área aumenta la probabilidad de gol. Finalmente, la variable **delantero** es significativa al 1 % con coeficiente positivo, es decir, si el futbolista que dispara es delantero, habrá mayor probabilidad de que el disparo acabe entrando a portería.

Finalmente, las estimaciones del modelo logit de regate son:

Variable	**Coeficiente**	**Error Estándar**	**z-Valor**	**Pr**$(>\|z\|)$
CCEE	0.0826620	0.0355552	2.325	0.0201 *
min	0.0001509	0.0006178	0.244	0.8070
Back	-0.0160862	0.0465711	-0.345	0.7298
Center	-0.1410632	0.0503656	-2.801	0.0051 **
Left	-0.4738514	0.0539946	-8.776	$<$ 2e-16 ***
Right	-0.4850699	0.0530998	-9.135	$<$ 2e-16 ***
Local	0.0769966	0.0335560	2.295	0.0218 *
DC	0.2015735	0.0364291	5.533	3.14e-08 ***
Resultado	0.0158978	0.0240803	0.660	0.5091

Tabla 8: Estimación del modelo logit de regate

Se concluye del modelo que, primero, el minuto en el que se realiza el regate no es significativo para

determinar la probabilidad de acierto, como tampoco lo es el resultado del partido en el instante en el que se realiza. Pero sí que es significativa al 5 % y tiene un efecto positivo el actuar como local. Segundo, respecto a la zona donde se realiza el regate, las bandas tienen un efecto negativo y muy significativo sobre la probabilidad de éxito, al igual que en el tercio central de la mitad ofensiva del campo en el que subdividimos el terreno de juego inicialmente. En tercer y último lugar, las variables de distinción del futbolista son también significativas: El ser delantero afecta positivamente a la probabilidad de regatear. Del mismo modo, pertenecer a un equipo que jugaría competiciones a nivel continental también tendría un efecto significativo al 5 % y positivo sobre la probabilidad.

Por último, como método de evaluación de los modelos propuestos, se ha construido para cada logit sus respectivas curvas ROC. Se trata de una herramienta gráfica especialmente útil para modelos binarios que representa la relación entre la tasa de verdaderos positivos (sensibilidad) y los falsos positivos (especificidad).

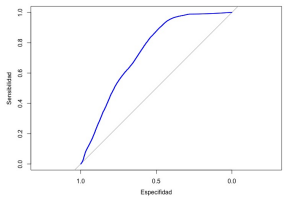


Figura 8: Curva ROC logit pases

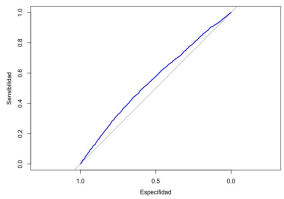


Figura 9: Curva ROC logit regates

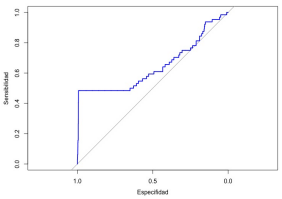


Figura 10: Curva ROC logit disparos

Para el modelo logit de pases, la curva se extiende significativamente hacia arriba y hacia la izquierda antes de acercarse al borde superior del gráfico. Esto indica un buen rendimiento, cuanto más cercana esté la curva al borde superior izquierdo (esquina), mejor es la capacidad del modelo. Para el modelo de los regates, la curva ROC está muy próxima a la diagonal. Podríamos interpretar que la acción de regatear con éxito a un contrario de acuerdo al modelo propuesto, tiene una probabilidad similar al lanzamiento de una moneda. Por otro lado, la curva ROC del logit de disparos describe un comportamiento irregular: La elevación vertical temprana en la curva sugiere que el modelo tiene un umbral efectivo donde logra clasificar correctamente un alto número de instancias positivas sin aumentar significativamente el número de falsos positivos. Sin embargo, la tendencia escalonada que sigue al salto inicial muestran incrementos más pequeños en la sensibilidad. Este patrón puede indicar que mientras el modelo sigue identificando correctamente algunos casos adicionales de disparos como goles, también empieza a enfrentar limitaciones en su capacidad para distinguir entre eventos positivos y negativos a medida que el umbral se ajusta. Esto podría ser corregido con la implementación de variables al modelo únicamente accesibles a través de datos de tracking.

4.2 Ejemplo de simulación

Tal y como está configurada la competición de una liga, un equipo se enfrenta al resto de clubes dos veces, una en su estadio actúando como local y otra asistiendo como equipo visitante. El primer FC Barcelona - Real Madrid que se disputó en la temporada 2022-2023 fue el 16 de octubre de 2022, en el Santiago Bernabéu. El partido finalizó con un marcador de 3-1 en favor del Real Madrid. Por otro lado, el partido disputado en el Camp Nou, el 19 de marzo de 2023, terminó con el club catalán llevándose la victoria en los minutos finales con un resultado de 2-1. Este último partido será el que inspirará el ejercicio de réplica de simulación:

A continuación se adjuntan las alineaciones que presentó cada equipo al partido:

De acuerdo a los supuestos establecidos que construyen el modelo, cada partido se disputa únicamente con **11 jugadores** y que las sustituciones no están implementadas. Dadas estas alineaciones,

Figura 11: Alineación FC Barcelona. Creada con la herramienta BuildLineup

Figura 12: Alineación Real Madrid. Creada con la herramienta BuildLineup

simulamos **10.000 partidos** con el método de Montecarlo, obteniendo así distintos ejemplos de enfrentamientos entre ambos equipos.

En el Anexo se encuentran las matrices de transición entre zonas del campo y decisiones de futbolistas del FC Barcelona y Real Madrid. Si en primer lugar se comparan las probabilidades de transición entre zonas para ambos equipos, destaca que el Real Madrid recurre en mayor proporción que el Barcelona a orientar su juego hacia el tercio ofensivo izquierdo cuando tiene el balón en posesión. Por el contrario, cuando el Barcelona tiene ubicado el balón en los tercios ofensivos de las bandas recurre a transicionar hacia el centro y atrás, de acuerdo a su estilo de juego característico.

Respecto a la toma de decisiones de los futbolistas, a partir de la matriz específica de cada equipo presentan patrones similares en cuanto a los movimientos con el balón dada la calidad semejante entre ambas plantillas. Sin embargo, destacamos que el Real Madrid conserva mejor el balón y no asume tantos riesgos con él. A modo de ejemplo, cuando la jugada precede de un pase, el Real Madrid es tiene la mitad de probabilidad de tener una pérdida que el FC Barcelona.

En cuanto a los goles anotados por cada equipo durante las 10.000 simulaciones, a partir de los siguientes histogramas observamos anotaciones prácticamente idénticas dada la semejanza en la calidad de las plantillas:

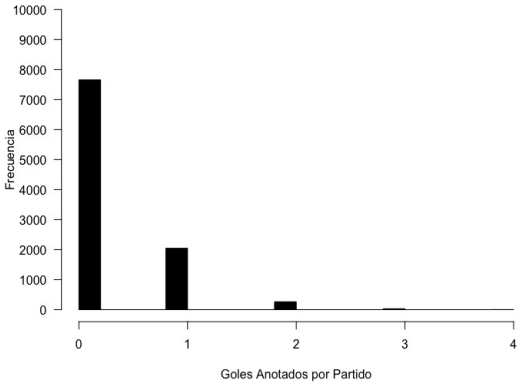


Figura 13: Histograma Goles por partido FC Barcelona

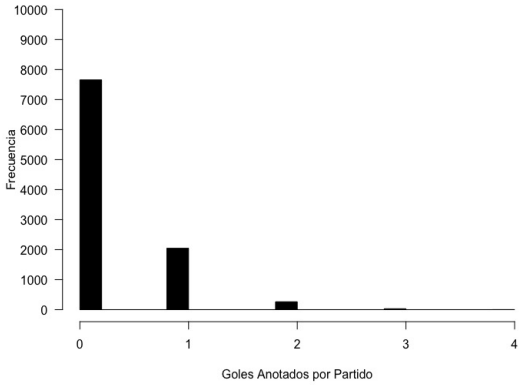


Figura 14: Histograma Goles por partido Real Madrid

Donde observamos como predomina como resultado principal el empate a cero goles, en un total de 6.128 partidos, mientras que el segundo resultado más frecuente fue el 1-0 (1.373 partidos ganó el Barcelona y 1.636 el Real Madrid). Si se consulta el resultado exacto del partido que utilizamos como ejemplo en el que el Barcelona ganó por 2 goles a 1 al Real Madrid, en nuestro simulador, este resultado se alcanza en 35 ocasiones.

Por último, del resultado del total de las simulaciones, dadas las alineaciones iniciales con las que se plantea el primer partido, en la siguiente tabla se presenta el promedio de acierto en el modelo logit en cada decisión según el equipo:

Equipo	**Pase**	**Regate**	**Gol**
FC Barcelona	87.1 %	45.7 %	16.0 %
Real Madrid	88.2 %	40.0 %	15.7 %

Tabla 9: Probabilidad media de acierto de acción según modelo logit

A la vista de la tabla, concluimos que no existen diferencias significativas entre el acierto promedio entre los equipos. Únicamente hay una diferencia ligera en el regate donde el FC Barcelona tiene 5 puntos porcentuales superiores respecto al Real Madrid. Este resultado de semejanzas refuerza el hecho de que el empate sea el resultado más frecuente en las simulaciones.

Finalizando esta primera subsección de resultados, del total de simulaciones se obtiene la siguiente tabla-resumen:

Equipo	**Media Goles por partido**	**Error Estándar**	**% de Victoria**
FC Barcelona	0.22	0.47	15.70 %
Real Madrid	0.26	0.51	19.30 %

Tabla 10: Estadísticas de partidos tras 10.000 simulaciones de Montecarlo

4.3 Análisis de sensibilidad del simulador

Con el objetivo de ilustrar el amplio abanico de posibilidades de análisis que permite el modelo, en esta última subsección planteamos un ejercicio para medir la sensibilidad del simulador a modificaciones de las alineaciones. Para ello, planteamos lo siguiente:

Uno de los supuestos que componen la definición de este simulador es que únicamente los partidos se disputan con 11 jugadores. Sin embargo, ya que no modelizamos la posibilidad de interrupciones durante el partido para efectúar cambios, se había considerado como otro supuesto la plena disponibilidad de la plantilla, permitiendo al usuario alinear a cualquier jugador sin considerar los efectos en el rendimiento del equipo por la ausencia de futbolistas debido a lesiones, cansancio o sanciones.

A modo de ejemplo, alteraremos la alineación que propuso Xavi Hernández para el partido del 19 de marzo de 2023. Planteamos para comprobar los efectos sobre el simulador una alineación totalmente ofensiva. La alineación es la siguiente: Una vez introducida la nueva alineación en el simulador, se

Figura 15: Alineación personalizada del FC Barcelona. Creada con la herramienta BuildLineup

repiten los 10.000 partidos, manteniendo el resto de variables constantes de forma que únicamente se alteren las probabilidades de transición del estilo de juego de este nueva alineación del FC Barcelona. Las nuevas matrices actualizadas a la alineación propuesta es:

$$\mathbf{M_{acción}} = \begin{bmatrix} 0.048593350 & 0.02046036 & 0.5191816 & 0.41176471 \\ 0.044554455 & 0.11716172 & 0.7508251 & 0.08745875 \\ 0.021586894 & 0.03257308 & 0.9088339 & 0.03700610 \\ 0.009761388 & 0.02169197 & 0.8253796 & 0.14316703 \end{bmatrix}$$

Al compararla con la matriz de transición entre acciones originales resaltamos una disminución de la probabildad de pérdida después de cada acción. Esta nuevo reparto en las probabilidades en las tomas de decisión sugiere que las pérdidas se sustituyen en parte por un aumento del peso en la probabilidad de disparo pero sobre todo, en un pase. En cuanto a las zonas del campo en las que el

balón transiciona, la matriz actualizada es la siguiente:

$$\mathbf{Mzonas\_A} = \begin{bmatrix} 0.5633291 & 0.2263650 & 0.09272448 & 0.1175813 \\ 0.3108834 & 0.4230769 & 0.11740007 & 0.1486396 \\ 0.2920021 & 0.4288782 & 0.15780998 & 0.1213097 \\ 0.2884927 & 0.4185575 & 0.08387358 & 0.2090762 \end{bmatrix}$$

Se observa que con la actualización de la plantilla hay una rotación del juego desde la zona de atrás hacia el tercio ofensivo central que se considera en la división inicial del campo. Este resultado concuerda con lo esperado ya que reducimos el número de defensores para introducir un mayor número de atacantes.

En cuanto a los resultados de las simulaciones destaca: Primero, al plantear una alineación más ofensiva, hay una ligera disminución de los partidos que terminan en empate a cero. Los empates sin goles se reducen aproximadamente un 4 %. Sin embargo, sigue siendo el resultado más frecuente pese a que es razonable pensar que, manteniendo un planteamiento ofensivo hay más ocasiones de gol tanto a favor como en contra, al asumir mayores riesgos en ataque.

En segundo lugar, no apreciamos en el histograma diferencias significativas con los goles anotados por partido en el nuevo histograma: Por último, presentamos dos tablas que nos ayudan a observar

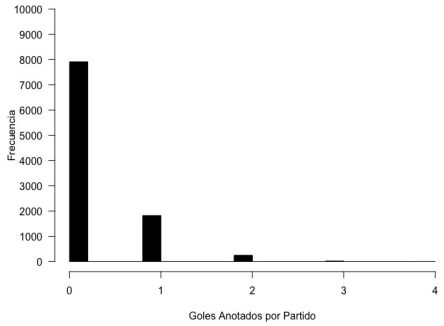


Figura 16: Nuevo Histograma de goles por partido FC Barcelona

la variación en los resultados de los partidos y por tanto, la sensibilidad del simulador a alteraciones en las alineaciones.

Equipo	Media Goles por partido	Error Estándar	% de Victoria
FC Barcelona	0.238	0.493	16.20 %
Real Madrid	0.283	0.52	20.01 %

Tabla 11: Nuevas estadísticas de partidos tras 10.000 simulaciones

En resumen, tras aplicar los cambios de la plantilla, observamos un pequeño aumento de aproximadamente del 10 % en los goles anotados por partido y de 1.5 puntos porcentuales en la probabilidad de victoria de ambos equipos, siendo ésta poco relevante por la insignificancia de los cambios. Por otro lado, en cuanto a las probabilidades de acierto medio de las acciones del FC Barcelona: A la vista de estas tablas, sorprende la disminución del promedio de acierto de los disparos que terminan en gol dada el carácter ofensivo de esta nueva alineación. Sin embargo, sí que parece que el promedio de acierto en el regate se ve afectado (en un aumento de tres puntos porcentuales) al implementar

Equipo	Pase	Regate	Gol
FC Barcelona	87.1 %	45.7 %	16.0 %
FC Barcelona (Ofensivo)	86.0 %	48.1 %	15.3 %

Tabla 12: Probabilidad media de acierto de acción según modelo logit

jugadores ofensivos, pero sigue alejado de unas expectativas iniciales donde el cambio se esperaba más significativo.

Finalmente, aunque no ha sido objeto de estudio en este trabajo, pero sí es implementado en Galaz et al. (2021), cabe mencionar que existe también en este simulador la posibilidad de incorporación de futbolistas de otros clubes a nuevas plantillas, permitiendo así la medición de su impacto de un potencial fichaje.

5 CONCLUSIONES

La recopilación de grandes cantidades de datos de *eventing* en las ligas de fútbol supone una revolución de valor incalculable debido al poder e inmensas posibilidades para el desarrollo de herramientas útiles para los distintos departamentos de equipos profesionales.

Se ha planteado el desarrollo de un simulador efectivo que permite puntuar tanto individualmente como de forma colectiva la decisión del entrenador a la hora de elegir una alineación para enfocar un partido determinado. Este simulador ha implementado a través de dos cadenas de Markov independientes entre sí la capacidad del futbolista en términos de toma de decisión y la orientación táctica que sigue el juego en las zonas predefinidas, creando una herramienta sensible a alteraciones en las alineaciones y consecuentemente útil para el apoyo de toma de decisiones. De acuerdo al ejercicio propuesto para ilustrar la sensibilidad, aunque no es una variación significante en el porcentaje de victoria dada la igualdad entre las plantillas enfrentadas, sí que se observan los efectos de alteraciones en la composición del esquema táctico y los jugadores alineados en las matrices de transición.

El modelo presenta varias limitaciones dados los supuestos simples que configuran el simulador. En primer lugar, se redujo a únicamente cuatro posibles alternativas los movimientos de los fútbolistas: Pasar, Regatear, Disparar o Perder el balón. Aunque son incorporados como pases, el simulador no distingue entre formatos de pases por lo que se ignoran los saques de esquina o tiros libres. Del mismo modo, un penalti es tratado como un disparo desde 11 metros de distancia.

Segundo, las zonas del campos con el ánimo de simplificar el código del simulador se redujeron a cuatro: La mitad defensiva compone una única región mientras que la mitad ofensiva es dividida en tres regiones, distinguiendo entre las bandas y el tercio central.

También, el simulador no consideró la inclusión de interrupciones para las sustituciones, cada vez más importantes en el fútbol dada la profundidad de las plantillas. Aunque es posible subdividir la simulación en etapas del partido e introducir manualmente cambios de jugadores, por motivos de accesibilidad al código las secuencias que componen un partido son simuladas en una única vez y no en dos tiempos como los partidos.

Otros trabajos que surgen de este es la inclusión de otras ligas europeas para ampliar el espectro de equipos elegibles y simular competiciones como torneos eliminatorios como la UEFA Champions League o Copa del Rey. Además, resulta también atractivo la configuración de un simulador en el que las probabilidades de transición sean dinámicas según el resultado e instante del partido en vez extraídas a partir desde una visión frecuentisita.

Agradecimientos: Este trabajo ha sido posible gracias a la colaboración de Jesús Lagos Milla y ”Chechu” Fernández Conde.

Referencias

Anzer, G. and Bauer, P. (2022). Expected passes. *Data Mining and Knowledge Discovery*, 36(1):295–317.

Arbués-Sangüesa, A., Martín, A., Fernández, J., Ballester, C., and Haro, G. (2020). Using Player's Body-Orientation to Model Pass Feasibility in Soccer. *arXiv e-prints*, page arXiv:2004.07209.

Bergkamp, T., Frencken, W., Niessen, A., Meijer, R., and den Hartigh, R. (2022). How soccer scouts identify talented players. *European Journal of Sport Science*, 22(7):994–1004. Epub 2021 Apr 29.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. Wiley, 3rd edition. Chapter XV.

Fernandez, J., Gopaladesikan, S., Spearman, W., Shaw, L., Peralta, F., Thomas, A., and Bauer, P. (2020). Soccermatics: Mathematical adventures in the beautiful game. https://soccermatics.readthedocs.io/en/latest/index.html. Recorded during the Covid-19 lockdown by a group of club analysts and leading academics, including lead data scientists: Javier Fernandez (Barcelona), Sudarshan Gopaladesikan (formerly Benfica, now Atalanta), William Spearman (Liverpool), Laurie Shaw (Manchester City), Fran Peralta (formerly Hammarby, now Atletic Bilbao), Alex Thomas (The English FA), Pascal Bauer (German DFB). Accessed: 19 June 2024.

Finnoff, J., Newcomer, K., and Laskowski, E. (2002). A valid and reliable method for measuring the kicking accuracy of soccer players. *Journal of Science and Medicine in Sport*, 5(4):348–353.

Firildak, A. C. and Akin, H. (2020). *Footballpreneurship: The role of scouting and youth academies in football entrepreneurship and value creation from young talents: A case study on AFC Ajax and Borussia Dortmund*. Phd thesis, Linneaus University.

Formento, E. (2022). *Markov Chain Model for Football Analytics*. Master's thesis, Vrije Universiteit Amsterdam.

Galaz, P., Mena, S., and Saure, D. (2021). Inferencia bayesiana de un modelo markoviano de fútbol con aplicación en scouting. *Revista Ingeniería de Sistemas*, 35.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Gutiérrez Chico, F. (2018). Entre asensios y diarras: el athletic bilbao y la construcción de identidades. *Revista Latina de Sociología*, 8(3):160–171.

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 1623–1627, Opatija, Croatia.

International Football Association Board (2024). Laws of the game 2024/25.

Kalbfleisch, J. and Lawless, J. (1984). Least-squares estimation of transition probabilities from aggregate data. *Canadian Journal of Statistics*, 12(2):169–182.

Kullowatz, M. (2017). Validating the asa xgoals model. https://www.americansocceranalysis.com/home/2017/3/6/validating-the-asa-xgoals-model. American Soccer Analysis, Accessed: 26 June 2024.

Kuper, S. and Szymanski, S. (2018). *Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the US, Japan, Australia, Turkey–and Even Iraq–Are Destined to Become the Kings of the World's Most Popular Sport*. Hachette UK.

Lagos, J. (2024). *soccergraphR: Analysis and visualization from Opta XML data*. R package version 0.1.2.

Larsen, C. H., Storm, L. K., Sæther, S. A., Pyrdol, N., and Henriksen, K. (2020). A world class academy in professional football: The case of ajax amsterdam. *Scandinavian Journal of Sport and Exercise Psychology*, page 33.

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. WW Norton & Company.

Mackay, N. (2017). Predicting goal probabilities for possessions in football. Master's thesis, Vrije Universiteit Amsterdam.

Mead, J., O'Hare, A., and McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4):e0282295.

Neal, R. M. (1993). Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Pappalardo, L., Guerrini, L., Rossi, A., and Cintia, P. (2019). Explainable injury forecasting via multivariate time series and convolutional neural networks. In *Proceedings of the Barça Sports Analytics Summit 2019*.

Pollard, R. and Reep, C. (1997). Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46(4):541–550.

Prasetio, D. and Harlili, D. (2016). Predicting football match results with logistic regression. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5, Penang, Malaysia.

Sailofsky, D. (2018). Drafting errors and decision making theory in the nba draft. Master's thesis, Brock University, St. Catharines, Ontario.

Singh, K. (2018). Introducing expected threat (xt): Modelling team behaviour in possession to gain a deeper understanding of buildup play. Consultado el 7 de mayo de 2024.

Spearman, W., Basye, A., Dick, G., Hotovy, R., and Pop, P. (2017). Physics-based modeling of pass probabilities in soccer. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 123–130.

Sumpter, D. (2017). Using markov chains to evaluate football players' contributions. https://soccermatics.medium.com/using-markov-chains-to-evaluate-football-players-contributions-57a107cc09e6. Accessed: 2024-05-21.

Tippett, J. (2019). *The Expected Goals Philosophy: A Game-Changing Way of Analysing Football*. Independently published.

Wisdom, C. and Javed, A. (2023). Machine learning for data analytics in football: Quantifying performance and enhancing strategic decision-making. https://ssrn.com/abstract=4558733 or http://dx.doi.org/10.2139/ssrn.4558733. Available at SSRN.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edition.

Yam, D. (2019). Attacking contributions: Markov models for football. https://statsbomb.com/articles/soccer/attacking-contributions-markov-models-for-football/. Accessed: 2024-05-06.

Zou, Q., Li, Q., Guo, H., and Shi, J. (2017). A discrete-time and finite-state markov chain model for association football matches. *Communications in Statistics - Simulation and Computation*, 47(8):2476–2485.

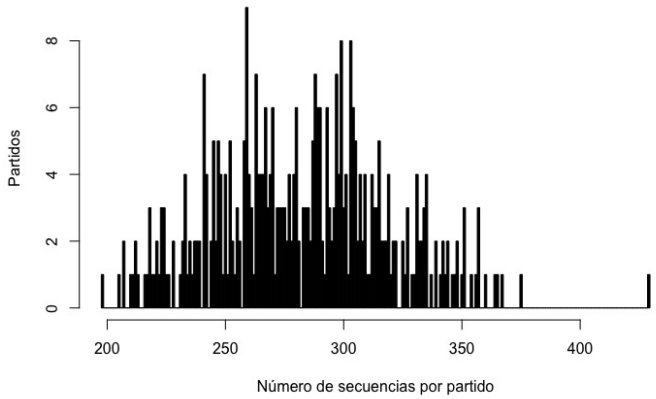
ANEXO

Tipos de Evento

Tabla 13: Tipos de evento

Evento	Definición
Pase	Intento de entrega de la pelota de un jugador a otro del mismo equipo.
Recuperación de balón	Un jugador recoge un balón y recupera la posesión para su equipo.
Toque de balón	Un jugador realiza un mal toque de balón y pierde la posesión.
Duelo aéreo	Dos jugadores de equipos opuestos disputan un balón aéreo.
Falta	Indica que se ha cometido una falta.
Regate	Intento de regatear a un oponente.
Despeje	Acción defensiva en la que un jugador aleja el balón de una zona peligrosa.
Balón dividido	Enfrentamiento disputa equilibrada entre dos jugadores por la posesión del balón.
Entrada	Acción defensiva en la que un jugador intenta despojar al oponente del balón con un contacto físico legal.
Saque de Esquina	Indica que el balón abandona el campo y se concede un saque de esquina.
Ocasión Salvada	Disparo atajado.
Fallo	Cualquier disparo a portería que se desvía ancho o por encima.
Gol	Se atribuye un gol al jugador que lo anota.
Poste	El balón impacta en el poste/larguero de la portería.

Distribución del número de secuencias por partido.



Matrices de transición

FC Barcelona

La matriz con las probabilidades de transición entre zonas **(Back-Center-Left-Right)** del FC Barcelona es:

$$\mathbf{M_{zonas}} = \begin{bmatrix} 0.5915687 & 0.2010258 & 0.09019264 & 0.1172129 \\ 0.3052291 & 0.4116846 & 0.12433998 & 0.1587464 \\ 0.3041082 & 0.3952906 & 0.17434870 & 0.1262525 \\ 0.3093159 & 0.3904658 & 0.07168850 & 0.2285298 \end{bmatrix}$$

Mientras que la matriz con las probabilidades correspondientes a transicionar entre decisiones **Disparo-Regate-Pase-Pérdida** es:

$$\mathbf{M_{acción}} = \begin{bmatrix} 0.071428571 & 0.02285714 & 0.4457143 & 0.46000000 \\ 0.030120482 & 0.10040161 & 0.6987952 & 0.17068273 \\ 0.018733493 & 0.02555126 & 0.8922671 & 0.06344819 \\ 0.003293808 & 0.01581028 & 0.8241107 & 0.15678524 \end{bmatrix}$$

Real Madrid

Para el Real Madrid, la matriz con las probabilidades de transición entre zonas **(Back-Center-Left-Right)** es:

$$\mathbf{M_{zonas}} = \begin{bmatrix} 0.5732894 & 0.1958569 & 0.1453233 & 0.08553045 \\ 0.2821126 & 0.4389032 & 0.1670961 & 0.11188811 \\ 0.2585059 & 0.4079142 & 0.2544379 & 0.07914201 \\ 0.2797897 & 0.4106308 & 0.1057243 & 0.20385514 \end{bmatrix}$$

Por último, la matriz con las probabilidades correspondientes a transicionar entre decisiones **Disparo-Regate-Pase-Pérdida** es:

$$\mathbf{M_{acción}} = \begin{bmatrix} 0.03269755 & 0.002724796 & 0.5994550 & 0.36512262 \\ 0.04581359 & 0.093206951 & 0.7630332 & 0.09794629 \\ 0.02224816 & 0.037492275 & 0.9101833 & 0.03007622 \\ 0.00286944 & 0.038737446 & 0.8665710 & 0.09182209 \end{bmatrix}$$