Inferencia Bayesiana de un Modelo Markoviano de Fútbol con aplicación en Scouting

Pablo Galaz [*]
Sebastián Mena [*]
Denis Sauré [*]

Resumen

En este trabajo proponemos un enfoque analítico que utiliza datos granulares de fútbol profesional para modelar encuentros de fútbol considerando interacciones a nivel de jugador para predecir el desarrollo de un encuentro. Este enfoque de modelamiento representa un avance significativo respecto a la literatura, enfocada hasta ahora principalmente en predecir de manera agregada el resultado de un encuentro, puesto que permite analizar la influencia de jugadores individuales en el rendimiento colectivo de un equipo, por lo que cuenta con múltiples aplicaciones en la administración de un equipo de fútbol, como por ejemplo el proceso de scouting. El modelo propuesto visualiza el desarrollo de un partido como una cadena de Markov en tiempo discreto, en el cual las probabilidades de transiciones dependen de una forma no trivial en parámetros que hablan de las características cognitivo-perceptuales y técnicas de los jugadores. El enfoque propuesto utiliza inferencia bayesiana para estimar la distribución a posterior de los parámetros que definen las características de los jugadores. Ilustramos la factibilidad y el uso potencial de este enfoque utilizando datos de la temporada 2017-2018 de la Premier League de Inglaterra. Una vez calibrado, utilizamos el modelo propuesto para realizar múltiples análisis de sensibilidad, propios de las labores asociadas al proceso de scouting. Nuestros resultados hablan del gran potencial para la práctica del uso este tipo de modelos en particular, y de *sports analytics* en general.

Palabras Clave: Inferencia bayesiana, Modelo markoviano, Simulación, Scouting, Fútbol.

[*]Departamento de Ingeniería Industrial, Universidad de Chile

1. Motivación

La disciplina de *sports analytics* se centra en el uso de datos y estadísticas para apoyar la toma de decisiones en el deporte mediante modelos analíticos. Si bien su uso es un lugar común desde ya hace más de una década en las ligas profesionales más importantes de Estados Unidos, su uso en el fútbol se vio un tanto demorado precisamente por la falta de datos detallados acerca del transcurso de un evento deportivo. Sin embargo, recientemente han comenzado a surgir distintas técnicas para registrar los *eventos*, jugada-a-jugada, que ocurren durante un partido de fútbol, lo que permite analizar el rendimiento individual de los futbolistas. Actualmente, el nivel más detallado de datos, conocido como *eventing*, consiste en un registro de cada movimiento del balón durante un partido, incluyendo los jugadores involucrados, el minuto de juego, el lugar del campo, en conjunto con un número de métricas de desempeño, como por ejemplo, el *expected goal*, asociado a los tiros a portería [8]. El uso de técnicas es cada vez maś masivo, al punto que actualmente es posible acceder a datos con este nivel de detalle para la mayoría de las ligas profesionales de fútbol del mundo. Entre los actuales proveedores de este tipo de data a nivel mundial se encuentran OPTA, Instat, WyScout, por mencionar algunos.

La abundancia de los datos, junto al gran grado de detalle de estos, ha propiciado un desafío de *Big Data* en el tratamiento de esta información, el que es posible enfrentar utilizando herramientas analíticas. A modo de ejemplo, a nivel táctico, unas de las decisiones más importantes en fútbol son aquellas acerca de la disposición espacial de los jugadores durante diferentes etapas del juego [19]. En este contexto, utilizando herramientas analíticas es ahora posible determinar y analizar las zonas de influencias de los futbolistas en fase defensiva, considerando su dirección y velocidad de desplazamiento, lo que permite identificar zonas de pases con diferentes niveles de riesgos y, así, poder explotar zonas de bajo control del rival. De forma similar, a nivel estratégico, es posible utilizar la información disponible para categorizar a los futbolistas de acuerdos a sus perfiles y a su desempeño, lo que posibilita identificar perfiles de futbolistas atractivos para un club, un input crucial para realizar el proceso de búsqueda de nuevos jugadores, conocido como *scouting* [20].

Tradicionalmente, las unidades técnicas de análisis en los clubes de fútbol dependen fuertemente del análisis de vídeo para realizar labores de scouting, en un enfoque más bien cualitativo de evaluación y búsqueda de futbolistas. Sin embargo, los entrenadores y analistas típicamente son capaces de recordar

menos de la mitad de los incidentes claves que surgen durante un partido [11]. Este recuerdo se ve afectado principalmente por varios factores: en general se tiende a seguir el balón y se pierde información lejana a él; puntos de vistas y prejuicios establecidos, pues algunos entrenadores solo ven lo que quieren o esperan ver; efectos de las emociones, como el estrés y la ira, que afectan a la concentración y pueden distorsionar la impresión del entrenador sobre el partido; además de las limitaciones propias de la memoria humana. El análisis cuantitativo de los datos surge entonces como una forma de evidenciar o evitar sesgos cualitativos mencionados. Para ello es necesario postular modelos analíticos que traduzcan datos en información útil. La irrupción de este tipo de modelos causo la revolución de sport analytics en el béisbol, documentado por Michael Lewis en su libro Moneyball [10], donde cuenta como General Manager de los Oakland Athletics adoptan una política de contratación basada en datos obteniendo rendimientos por sobre lo esperado. En el básquetbol tenemos el ejemplo del proyecto *Carmelo* desarrollado por *FiveThirtyEight*, el cual busca predecir el rendimiento futuro de los deportistas en base a modelos probabilísticos con datos históricos de la NBA [5].

1.1. Objetivo y supuestos.

En este trabajo realizamos una primera aproximación al desarrollo de una herramienta, basada en un modelo analítico, que permita apoyar el proceso de *scouting* de los clubes de fútbol. Para eso, mediremos el impacto que genera la inclusión de un futbolista en un club, en reemplazo de otro, en el rendimiento deportivo del equipo, en base a métricas agregadas del rendimiento del equipo. Una de las métricas utilizadas, de gran interés para la dirigencia del club, es la probabilidad de ser campeón de un torneo. Un punto de partida en nuestro modelo es reconocer la heterogeneidad de los jugadores en términos tanto de sus perfiles técnicos como de sus procesos de toma de decisión dentro del campo de juego. Esta última característica, de aspecto cognitivo-perceptual, permite capturar la diferencia en desempeño entre jugadores de similares capacidades técnicas. Esto es, modelamos la toma de decisiones de los futbolistas condicionado a distintas circunstancias propias del juego y, en el ámbito técnico, modelamos la capacidad de ejecución de la acción decidida por el jugador.

En nuestro modelo, utilizamos las características de los jugadores para trazar, de forma estilizada/simplificada, el transcurso de un partido de fútbol. Para esto, modelamos un encuentro como un proceso Markoviano, donde las características cognitivas-perceptuales y técnicas de cada futbolista determinan, de manera probabilista, las transiciones entre distintos estados de fase del

juego. En particular, en términos de toma de decisiones durante un partido, los jugadores constantemente deben escoger (de manera probabilista) entre tres alternativas: intentar un pase, intentar un regate o rematar al arco. En términos de la ejecución técnica, la capacidad técnica del jugador y del rival que se enfrenta en el momento (ya sea arquero, defensa rival, etc.) determina (de forma probabilista) el éxito o no de la acción.

Nuestro modelo base de un partido de fútbol depende de múltiples parámetros, que gobiernan las acciones de los jugadores y sus resultados. Estos parámetros deben ser inferidos desde la data, para poder utilizarse en el análisis. Para esto utilizaremos el paradigma de inferencia Bayesiana: asumiremos que los parámetros desconocidos son aleatorios, y los equiparemos con distribuciones a priori; luego calcularemos la distribución a posterior de los parámetros, condicionales en la data, y utilizaremos dicha distribución para estimar los parámetros, por ejemplo, tomando sus valores esperados. Para modelar la toma de decisión de los jugadores y el éxito de dichas acciones utilizamos modelos de decisión logística multinomial. Esto resulta en modelos donde la inferencia no se puede realizar en forma analítica, por lo que utilizamos la técnica numérica conocida como Monte Carlo Markov Chain (MCMC) para aproximar las distribuciones a posterior de los parámetros en base a simulación: ver, por ejemplo [13].

Ilustramos la aplicación de nuestro modelo utilizando datos de nivel *eventing* correspondientes a todos los partidos de la temporada 2017-2018 de la *Premier League* (primera división de Inglaterra) [16]. Recordamos que éstos datos mantienen un registro de cada interacción ocurrida durante el partido que involucre al balón, incluido en que coordenadas del campo de juego ocurre el evento, el tiempo de juego, el jugador asociado, si la acción fue exitosa o no y, en caso de aplicar, jugadores rivales asociados.

1.2. Contribución y resultados.

La principal contribución hecha por nuestro trabajo consiste en el desarrollo de una herramienta que permite medir el impacto absoluto y/o relativo de incluir un jugador en un equipo determinado, en el rendimiento global de un equipo en una competencia. Esto tiene un potencial significativo para mejorar inmensamente los procesos de scouting en el fútbol profesional en países como Chile, y en toda Sudamérica. Es importante notar que, por lo menos en la literatura académica (más detalles en la siguiente sección), el tipo de modelos de transcurso de partidos desarrollados en el pasado pone el foco en la predicción del resultado global del encuentro; dichos modelos no son de manera directa capaces de responder preguntas fundamentales relacionadas a scouting

y a caracterizar la habilidad de los jugadores en términos multidimensionales.

Una segunda contribución, quizás tanto o más importante, aparece como subproducto del proceso de calibración de nuestro modelo: la herramienta permite caracterizar individualmente los perfiles cognitivos-perceptuales y técnicos de cada jugador, posibilitando un análisis detallado y multidimensional de los jugadores. Esto permite, por un lado, ampliar el espectro de decisiones que pueden ser apoyadas por análisis de este estilo, y por otro, el conocer mejor a los jugadores de un plantel, realizar reforzamientos de algunas habilidades técnicas, o focalizar el estilo de juego individual en aspectos donde los jugadores cuentan con ventajas comparativas.

En relación con lo anterior, y a modo de ejemplo, nuestro modelo permite identificar el efecto de la inclusión de un futbolista en un equipo en el cambio de la probabilidad de ser campeón: incluir a Mohamed Salah en el Stoke City, aumenta la probabilidad de ser campeón de ese equipo en un 6.34 %; la inclusión de Kevin de Bruyne en el Arsenal, aumenta su probabilidad de ser campeón en un 2.46 %. El modelamiento innovador y detallado del transcurso de un partido en función de las habilidades individuales de los jugadores identifica, por ejemplo, qué jugador es más beneficioso para cada uno de los clubes en base a sus propias necesidades y estilos de juego, lo que se traduce en un scouting basado en criterios objetivos y cuantificables. También, abre la posibilidad de medir el valor real de un futbolistas en términos monetarios para un club, en base a los objetivos y potenciales logros que pueda obtener, no solo salir campeón, sino que clasificar a torneos internacionales o evitar el descenso.

1.3. Estructura del manuscrito.

En la Sección 1.3 repasaremos la literatura relacionada y en la Sección 1.3 presentamos el modelo propuesto, junto al esquema de inferencia Bayesiana. En la Sección 3.3 mostraremos los resultados de calibración, e ilustramos el análisis posible de realizar utilizando nuestra propuesta. Finalmente, en la sección 4.2, presentamos una discusión de nuestros supuestos y resultados, junto a nuestras conclusiones.

2. Revisión de la Literatura

La estadística ha ido de la mano del fútbol desde su comienzo, principalmente con el desarrollo de las ligas profesionales. Sin embargo, es tan solo en los 80's que comenzaron a surgir los primeros esfuerzos para registrar, mas allá de es-

tadísticas agregadas, los sucesos que iban ocurriendo en los partidos de fútbol, con el objetivo de analizar a un potencial equipo rival o analizar el rendimiento individual de un futbolista [4, 9, 17]. La tecnología asociada a estos esfuerzos ha evolucionado desde registro manual, pasando por grabaciones de sonido y el uso de teclados especializados [14], hasta el uso de software especializado, que permiten incluso el acceso a la data en tiempo real [18]. Actualmente, el registro de un partido de fútbol típicamente involucra alrededor de ocho millones de datos. Debido al gran volumen de datos que genera la actividad, la información debe ser accedida y manipulada utilizando herramientas de *Big Data*. Esto impone una barrera tecnológica mínima necesaria para la utilización de la información generada para la toma de decisiones y análisis de datos en los clubes de fútbol.

Este gran volumen de datos entrega nuevas posibilidades en el mundo de la predicción de resultados deportivos. Un modelo base para esta materia es el que busca representar la conversión de goles de los equipos a través de procesos de Poisson, calibrando el modelo con variables como goles del equipo local, goles del equipo visitante y el efecto de la localía [3]. Sin embargo, este modelo y sus variantes utilizan datos agregados de los equipos. Esto tiene la limitante de que no se puede incorporar información individual de los jugadores para la predicción de resultados deportivos y, por otra parte, no permite individualizar el rendimiento de los jugadores para realizar *scouting*.

En este contexto, [2] estudia el proceso de detección de talentos que siguen 125 buscadores de talento (*scouts*) de distintas categorías en Países Bajos para poder determinar las características más importantes que, según ellos, son predictores del talento: a través de encuestas de respuestas abiertas, lo autores identifican los aspectos técnicos de los jugadores (37 %) la categoría de atributos más importante, seguida de los aspectos tácticos y percepción-cognitiva (22 %). Una vez definidas estas variables, se procede al análisis para determinar qué jugadores del mercado podrían ser atractivos para maximizar el rendimiento deportivo. En este ámbito, [6] propone medir el impacto que un jugador genera en un club en particular a través del modelamiento de un partido a través de un proceso estocástico Markoviano para medir dicho impacto. Nuestro trabajo puede ser visto como una continuación de la propuesta en [6], pero incluyendo inferencia Bayesiana para la calibración de los modelos propuestos, entre otros aspectos. Utilizando herramientas analíticas y datos de posesión de balón, [7] analiza líneas de pases que se generan cuando un jugador tiene el balón, clasificándolas como i) pases penetrantes, ii) pases de apoyo y iii) pases de seguridad, evaluando, en función del tiempo disponible las línea de pase disponibles . Otra área de aplicación, que no exploramos en esta revisión, es posible utilizar técnicas de analítica avanzadas para intentar predecir

lesiones en el fútbol. Las lesiones tienen un gran impacto en el deporte, por los costos de rehabilitación y la pérdida de ese elemento para la competencia durante un tiempo determinado. A modo de ejemplo, [15] propone clasificar a los jugadores en base a sus cargas de entrenamiento, utilizando redes neuronales convoluciones que utilizan datos en forma de series de tiempo multivariadas, con el objetivo de detectar potenciales ventanas de tiempo donde un jugador podría lesionarse y prevenir esta afección.

3. Modelo Matemático

En este sección presentamos un modelo estocástico de desarrollo de un partido de fútbol. Como mencionamos en la Sección **??**, el nivel de detalle del modelo debe permitir medir el efecto de un jugador en la alineación del equipo, de forma de poder estimar su impacto en la probabilidad de salir campeón de un torneo. Esto, de forma de facilitar, por ejemplo, la valoración de un precio justo para alguna transacción de jugadores entre dos clubes, o intentar predecir como sería el rendimiento de un futbolista que está dento de los planes de contratación de un club. A modo de resumen, el modelo que proponemos visualiza el desarrollo de un partido como una cadena de Markov en tiempo discreto, donde un estado incluye, por ejemplo, donde se ubica el balón y que jugador lo tiene, además de incluir variables que detallan el estado global del partido. Las transiciones entre un estado y otro de la cadena están gatilladas por las acciones de los jugadores. En este trabajo consideraremos tan solo tres acciones: pases, regates y tiros. En particular, dicha transición ocurre producto de una combinación de eventos independientes: en primer lugar, un evento asociado al proceso de decisión del jugador (ámbito cognitivo-perceptual), quien selecciona que acción realizar (pase, regate o tiro), en función de variables contextuales del partido; y luego un evento relacionado con la ejecución de tal acción, es decir, si la realiza de éxito o no. Planteamos modelos de decisión logísticos para las probabilidades asociadas a estos eventos, las que dependen de parámetros asociados, entre otras cosas, a los jugadores de forma individual. Para estimar estos modelos adoptamos un enfoque de inferencia Bayesiano, el cual detallamos más adelante en esta sección, y que ilustramos en la Sección 3.3 utilizando los datos de la Premier League de Inglaterra, temporada 2017-2018.

3.1. Modelo Markoviano de un Partido de Fútbol

3.1.1. Información Preliminar.

La construcción del modelo depende fuertemente de los datos disponibles para su calibración. Considerando esto, comenzamos esta sección con una descripción de los datos a nivel de eventos.

Descripción de los datos. A grandes rasgos, los datos contienen información general asociada a las plantillas de los equipos, eventos de cada uno de los partidos, e información del rendimiento deportivo de los jugadores. Para cada uno de los partidos del torneo se tiene el detalle de cada evento deportivo que ocurre durante el encuentro, esto es: pases, tarjetas, tiros, faltas, regates, detenciones, entre otros. En promedio ocurren casi 1.500 eventos relacionados al balón por partido, considerando las cinco ligas más grandes de Europa. Un evento se compone de múltiples características: la zona del campo de juego donde ocurre (punto (x_i, y_i) del evento i geo-referenciado en la cancha), el tiempo relativo al inicio del partido cuando ocurre el evento, el jugador asociado a este, si realiza la acción de forma exitosa o no, y si existe un segundo jugador asociado. Con este nivel de datos desagregados puede reconstruir el desarrollo del partido a nivel de la estructura y progreso del juego y por ende, el del campeonato completo.

Resumen del modelo, estados y transiciones Representamos el desarrollo del partido como una cadena de Markov en tiempo discreto $\{X_n : n \in \mathbb{N}\}$, donde X_n representa un estado de desarrollo del partido tras n transiciones o períodos. Si bien no modelamos directamente la duración de un período, si controlamos el número de períodos que contiene un partido mediante una discretización de la duración del mismo (detalles más abajo). Para que un proceso estocástico como $\{X_n : n \in \mathbb{N}\}$ sea una cadena de Markov es necesario que X_n contenga toda la información necesaria para caracterizar (probabilísticamente) la evolución futura del proceso. En nuestro modelamiento, las transiciones entre estados de la cadena de Markov son gatilladas por las decisiones tomadas por el jugador que tiene el balón al comienzo de un periodo, y por el subsecuente éxito o fracaso de dicha acción. Modelamos la decisión del jugador, y el resultado de la ejecución como eventos independientes, condicional en el estado del sistema.

3.1.2. Estados de la cadena de Markov.

En este trabajo, utilizaremos la siguiente representación, la cual implícitamente asumiremos otorga la condición de Markov al desarrollo del partido:

$$X_n = (X_n^1, X_n^2, X_n^3, X_n^4), \quad n \in \mathbb{N},$$

donde

- X_n^1 representa el jugador que tiene el balón al comienzo del período: las transiciones entre estados, se gatillan por las acciones tomadas por este jugador, y el éxito en su ejecución. Ver más detalles, abajo.
- X_n^2 representa la zona de la cancha donde se encuentra el jugador: para evitar trabajar con una representación continua (en dos dimensiones) de la ubicación del balón, lo que implica un grado de sofisticación mayor para determinar las transiciones espaciales de la cadena, dividimos el campo de juego en 12 zonas, producto de dividir la cancha horizontal/verticalmente en 4/3 zonas. Para más detalles respecto a esta discretización, ver [12].
- X_n^3 representa el tiempo transcurrido desde el inicio del partido: tal como en el caso anterior, evitamos trabajar en tiempo continuo utilizando una discretización del tiempo de juego. En particular, supusimos que cada acción toma un tiempo determinista Δ, por lo que el tiempo de juego periodo a periodo avanza en esa cantidad. El número total de períodos en un partido se ajustó considerando la historia de encuentros entre los equipos. Para más detalles, ver [12][1].
- X_n^4 contiene información adicional asociada al contexto general del partido: en particular, incluimos en esta información el resultado parcial del partido, es decir, si el equipo que tiene el balón está ganando, empatando o perdiendo.

Información adicional que no cambia durante el partido, como por ejemplo el estado de localía o visita, también es considerada, pero no como parte del estado.

3.1.3. Transiciones de estados

La transición de un estado a otro involucra, en orden cronológico, los siguientes pasos.

1. Primero, el jugador decide que acción ejecutar.
2. Segundo, dependiendo de la acción, un jugador secundario es seleccionado para participar de la acción: que en el caso de pases, se selecciona un jugador del mismo equipo que recibirá el pase, junto a una zona de

[1]Notamos que este enfoque es equivalente remover X_n^3 de la descripción de estado, e incluir una dependencia en el periodo n en las probabilidades de transición. Preferimos el enfoque actual, que facilita la descripción del modelo.

recepción del pase - la selección se realiza considerando la frecuencia con la que pases con esa zona de origen se realizan a esos jugadores/zonas; para el caso de regates, se selecciona un jugador de equipo rival, considerando la frecuencia con la cual dichos jugadores se ven involucrados en regates en esa zona del campo; en el caso de tiros, el jugador secundario es siempre el arquero del equipo rival[2].

3. Tercero, se determina si la acción es exitosa o no. Finalmente, si la acción no es exitosa, se determina qué jugador del equipo rival recupera el balón; exceptuando los pases, este jugador es el jugador secundario asociado a la jugada; en el caso de un pase, el jugador se elige (de manera frecuentista) utilizando la frecuencia con la cual los jugadores interceptan pases, dependiendo de la zona de la cancha involucrada.

De la descripción anterior, vemos que hay dos eventos cuyas probabilidades asociadas falta especificar: la selección de la acción a realizar, y el éxito de dicha acción. A continuación, modelamos estas probabilidades adoptando un enfoque paramétrico, que luego incorporaremos dentro de un enfoque de inferencia bayesiano.

- **La decisión del jugador.** En términos de las acciones a realizar, para la construcción de este modelo se consideran los pases y regates, que representan más del 75 % de los eventos que ocurren en un partido de fútbol y los tiros que dan pie a representar los goles, el evento con mayor importancia de este deporte [12]. El resto de los eventos corresponden a interrupciones y reposiciones de balón, los cuales no son abordadas en este modelo. Para modelar la decisión del jugador utilizaremos modelos de decisión discreta (ver, e.g., [22]), capaces de reflejar el hecho que los jugadores no siempre toman la misma decisión cuando se enfrentan a situaciones similares, y si bien tienden a privilegiar ciertas acciones sobre otras, en general existe un grado de aleatoriedad en la decisión, necesaria para evitar que los rivales puedan predecir sus acciones. En particular, en este trabajo asumiremos que los jugadores utilizan un modelo de decisión Logit multinomial [22].

[2]Notamos que las probabilidades de selección de jugadores secundarios se calibran usando una lógica frecuentista, por lo que no se incluyen dentro del marco de inferencia bayesiana a detallar mas adelante en esta sección. Este supuesto se realiza para reducir la complejidad de la tarea de inferencia.

Dado un estado X_n al comienzo del periodo n, consideramos la variable aleatoria $Y_n \in \{\text{p (pase), t (tiro), r (regate)}\}$. Supondremos que

$$\begin{aligned}
\mathbb{P}\{Y_n = p|X_n\} &= \frac{\exp(Z_p(X_n))}{1+\exp(Z_p(X_n))+\exp(Z_t(X_n))},\\
\mathbb{P}\{Y_n = t|X_n\} &= \frac{\exp(Z_t(X_n))}{1+\exp(Z_p(X_n))+\exp(Z_t(X_n))},\\
\mathbb{P}\{Y_n = r|X_n\} &= \frac{1}{1+\exp(Z_p(X_n))+\exp(Z_t(X_n))},
\end{aligned}$$

donde

$$Z_i(x_n) := \beta_{i,0} + \beta_{i,X_n^1} + \beta_{i,X_n^2} + \beta_{i,X_n^3} + \beta_{i,X_n^4}, \quad i \in \{p,t\} \tag{1}$$

Esto es, la decisión de que acción realizar depende de una *propensión* de cada jugador a realizar cada acción, la que es afectada transversalmente por la zona de la cancha donde se encuentra el balón, del tiempo transcurrido del partido, y el resultado parcial. Notamos que el factor de localía se incluye en el termino inicial $\beta_{i,0}$, que considera dos valores posibles, dependiendo el jugador se encuentra en un equipo que juega en calidad de local o visita (no se registran partidos en terreno neutral en nuestros datos).

- **El resultado de la acción.** Modelamos el resultado de una acción mediante un modelo logístico, donde juegan un rol distintos factores dependiendo de las características de la acción. En general, el modelo utilizado es:

$$\mathbb{P}(\text{Exito}|A) = \frac{\exp(W(A))}{1+\exp(W(A))},$$

donde A contiene las características de la acción (jugador, tiempo, zona, etc.). A continuación describimos el modelo asociado a cada una de las acciones consideradas.

Pases. En el caso de pases tenemos que $A = (X_n, J, Z)$ donde J representa el jugador al cual el pase esta dirigido, y Z la zona del campo donde se recibirá el pase. En este caso tenemos que

$$W(A) := \alpha_{p,0} + \alpha^o_{p,X_n^1} + \alpha^d_{p,J} + \alpha_{p,X_N^2,Z} + \alpha_{p,X_n^3} + \alpha_{p,X_n^4}. \tag{2}$$

Vemos que el éxito del pase depende tanto del ejecutante como del receptor del pase, los sectores del campo involucrados, y las condiciones del encuentro.

Tiros. En el caso de tiros tenemos que $A = (X_n, J)$ donde J representa el arquero del cuadro rival. En este caso tenemos que

$$W(A) := \alpha_{t,0} + \alpha^o_{t,X^1_n} + \alpha^d_{t,J} + \alpha_{t,X^3_n} + \alpha_{t,X^4_n}. \tag{3}$$

Entonces, vemos que la probabilidad de conversión del tiro depende tanto del ejecutante como del arquero rival, y las condiciones del encuentro.

Regates. En el caso de regates tenemos que $A = (X_n, J)$ donde J representa el jugador al cual se encara durante el regate. Tal como en el caso anterior, tenemos que

$$W(A) := \alpha_{r,0} + \alpha^o_{r,X^1_n} + \alpha^d_{r,J} + \alpha_{r,X^3_n} + \alpha_{r,X^4_n}. \tag{4}$$

Entonces, vemos que la probabilidad de sobrepasar al jugador rival depende tanto del ejecutante como del rival, y las condiciones del encuentro.

3.2. Inferencia Bayesiana del Modelo

De la sección anterior, vemos que la selección probabilista de la acción a realizar queda determinado por el vector de paramétros $\beta := (\beta_t, \beta_p)$ aludidos en la ecuación (1), donde $\beta_i := (\beta_{i,j} : j \in J)$ donde el conjunto J incluye todos los jugadores, zonas del campo, periodos, resultados parciales, y situaciones de localía posibles, $i \in \{t, p\}$. Notamos que, así como no es necesario especificar parámetros asociados a la acción de regate (esto, pues la magnitud de parámetros asociados a tiro y pase se entienden como expresados relativos a aquellos - no especificados - asociados a regate), para la correcta identificación del modelo tampoco es necesario especificar parámetros para todas las opciones en J, y es posible expresar estos parámetros relativos a un jugador, un tiempo, una zona, y un resultado en particular. De la misma forma, la probabilidad de éxito de la acción seleccionada queda determinada por el vector de parámetros $\alpha := (\alpha_t, \alpha_p, \alpha_r)$ aludidos en las ecuaciones (2)(3) y (4), donde $\alpha_i = (\alpha_{i,j} : j \in J_i)$ donde el conjunto J_i incluye todas las situaciones en las que se puede realizar la acción $i \in \{t, p, r\}$. Al igual que en el caso anterior, no todos los coeficientes necesitan ser especificados para la correcta identificación del modelo.

Incluso con las consideraciones mencionadas arriba, el número de parámetros necesario identificar para especificar el modelo es grande. Más importante aun, muchas de las acciones que son directamente observables en los datos dependen de la interacción entre múltiples parámetros, por lo que la inferencia de tipo frecuentista es compleja. Con esto en mente, proponemos un esquema de inferencia bayesiana. Esto es, supondremos que los parámetros (α, β)

son a priori variables aleatorias: cada realización de esos parámetros aleatorios determina un modelo distinto. Sin embargo, supondremos que la realización toma lugar antes del comienzo del torneo (i.e. antes de que se recojan los datos), y se mantiene constante a través del torneo, por lo que los datos son originados condicionales en una realización en particular de los parámetros; el trabajo de inferencia consiste en estimar cual es esta realización.

Dado el modelo probabilista y de generación de datos, la estimación de los parámetros consiste en el cálculo de la distribución a posterior de estos, condicional en los datos observados. Esto es, si (α, β) distribuyen *a priori* de acuerdo a una densidad $f(\cdot)$, y $L(\text{data}|(\alpha, \beta))$ representa la verosimilitud asociada a los datos observados, condicional en los parámetros (α, β), entonces la densidad *posterior* de los parámetros condicional en los datos, $f(\alpha, \beta|\text{data})$, es tal que

$$f(\alpha, \beta|\text{data}) \propto L(\text{data}|\alpha, \beta) \cdot f(\alpha, \beta) \tag{5}$$

Este esquema de inferencia mostrado en (5) se apoya fundamentalmente en el teorema de Bayes, el que en el caso que el modelo es identificable a partir de los datos, garantiza la consistencia de los estimadores. En el caso de nuestro modelo, esta distribución a posterior no es calculable de forma analítica, y debe ser estimada numéricamente. Para esto utilizamos el enfoque de Monte Carlo Markov Chain [13], el que aproxima la distribución posterior utilizando técnicas de simulación. El método queda definido, entre otras cosas, por la distribución a priori dada a los parámetros. Para esto, en nuestros experimentos numéricos, presentados en la Sección 3.3, utilizamos un prior normal, independiente para cada parámetro. Esto es, suponemos que

$$f(\alpha, \beta) = \phi(\alpha, \beta, \mu, \Sigma),$$

donde μ y σ son un vector y una matriz diagonal, positiva, de las dimensiones apropiadas, ϕ representa la densidad de un vector normal multivariado, también de las dimensiones correctas. Para mas detalles respecto a la selección de las distribuciones a priori de los parámetros (esto es, los valores de μ y Σ utilizados), ver [12].

3.3. Limitaciones del modelo

El modelo propuesto cuenta con variadas limitaciones, algunas originadas por los datos utilizados, y otras por los supuestos simplificadores que nos permiten representar un partido de fútbol como una cadena de Markov. Desde el punto de vista de los datos, es importante considerar que estos solo se refieren a la información del jugador o los jugadores que participan directamente con

el balón, pero no se cuenta con información del resto de los jugadores que no participan activamente con el balón (que se puede argumentar es tan importante como el resto), por lo que no se captura toda la información de que está ocurriendo en el partido. Por otra parte, este modelo es una representación simplificada de un partido de fútbol y hay ciertos elementos que no son considerados, como por ejemplo, la formación de cada equipo (lo que puede cambiar la propensión de un jugador a tomar una u otra acción), o de forma similar, la posición de cada jugador dentro del terreno de juego (considere el comportamiento de un mismo jugador cuando juega como central versus volante). Estas limitaciones serán discutidas más en detalle en la Sección 4.2

4. Resultados

En este sección ilustramos la aplicación del enfoque propuesto, utilizando los datos de la temporada 2017-2018 de la Premier League de Inglaterra, de la cual se disponen datos a nivel eventing [16]. Utilizando la dinámica del modelo propuesto, programamos rutinas de simulación capaces de muestrear el desarrollo de un partido de acuerdo a las reglas probabilistas de nuestro modelo, posibilitando la simulación de torneos completos y la realización de ejercicios de sensibilidad como por ejemplo, estudiar el efecto de intercambiar un par de jugadores entre equipos rivales. Todas las rutinas fueron programas en el lenguaje Python [23]; las rutinas de inferencia bayesiana fueron programadas desde Python usando el lenguaje de modelamiento Stan [21], a través de la librería PyStan.

4.1. Comparación con otro modelo

Comenzamos el análisis realizando una simulación del primer partido del torneo, entre los equipos Arsenal y Leicester City, cuyo resultado fue 4-3 a favor del local. La Tabla 1 muestra estadísticas agregadas de 10.000 simulaciones de dicho encuentro.

Equipo	**Promedio eventos**	**Promedio goles**	**Desv. estándar goles**
Arsenal	794,78	3,14	1,23
Leicester City	657,02	2,66	1,10

Tabla 1: Promedio de eventos y goles: Arsenal vs Leicester City.

Nuestra simulación concluye que el Arsenal anota (en promedio) cerca de

1/2 gol más que el Leicester cuando estos dos equipos se encuentran. En ese sentido, si bien el resultado real del encuentro fue 4-3 a favor del local, nuestro modelo indica que la victoria del Arsenal es una realización de múltiples posibilidades, y que no existe un dominio total sobre el rival, como lo ilustra la desviación estándar del número de goles anotados.

Si bien nuestro modelo esta diseñado para representar de buena forma los sucesos ocurridos durante un partido, y no pronosticar el resultado final, el que puede muchas veces estar desconectado del desarrollo de un encuentro (el fútbol es uno de los deportes masivos mas difícil de predecir debido a esta desconexión, y al relativo bajo número de anotaciones por encuentro [1]) es interesante realizar la comparación contra modelos alternativos que si tienen ese propósito (pero que no sirven para apoyar labores de scouting, por lo mismo). Con esto en mente, comparamos nuestro modelo contra una variación del modelo de Poisson propuesto por [3]. Dicho modelo se alimenta solamente de los resultados finales de los partidos del torneo, más las condiciones de localía. Con esto, el modelo calibra **tasas de ataque y defensa** para cada equipo, más un factor de localía; para simular un partido, estas tasas se utilizan para calcular tasas de procesos de Poisson (uno para cada equipo), los que representan los procesos de anotación de goles de cada equipo; la simulación del resultado del partido se genera a través del muestro de dos variables aleatorias distribuidas Poisson.

Para comparar modelos consideramos la métricas de **acierto de resultado** (local, empate, o visita) y **acierto diferencia de goles** (goles del local menos goles de visitante) que indican el porcentaje del partidos (calculado sobre base a 10.000 simulaciones del torneo) en los que un modelo genera un resultado/diferencia de goles que coincide con el resultado/diferencia real del partido. La Tabla 2 muestra los resultados de la comparación de los modelos.

Parámetros	**C. Markov**		**Poisson**	
	1-19	**20-38**	**1-19**	**20-38**
% acierto al resultado	39,09	39,12	44,28	44,28
% acierto diferencia de goles	18,38	19,26	17,64	18,54

Tabla 2: Rendimiento de los modelos de cadena de Markov y Poisson para 10.000 iteraciones según diferentes parámetros.

En la Tabla 2 los modelos son primero calibrados utilizando los partidos de la primera rueda (fechas 1-19), y simulando las partidos de la segunda rueda (fechas 20-38) para el calculo de las métricas, y luego invirtiendo los roles de la primera y segunda ruedas. En ambos casos notamos que nuestro modelo tiene un rendimiento aproximado del 39 % en contraste con el 44 % del

modelo de Poisson. Es decir, este último es capaz de predecir en un 44 % el resultado (local, empate o visita) de los partidos, calibrando el modelo sólo con la mitad de los datos (fecha 20 a 38). Por otro lado, vemos que nuestro modelo predice la diferencia de goles en un 18-19 % de los casos, superando levemente el resultado del modelo de Poisson. Este resultado es alentador, considerando que el modelo de Poisson esta calibrado explícitamente para replicar diferencias de goles. En ese sentido, nuestro modelo se comporta en forma comparable al modelo estándar para predecir resultados agregados, pero tiene la capacidad adicional de permitir hacer análisis a nivel de jugador, que es lo procedemos a ilustrar a continuación.

4.2. Análisis de Sensibilidad: Intercambio de Jugadores

Para ilustrar las posibilidades de análisis que permite el modelo, en esta sección consideramos el ejercicio donde intercambiamos dos jugadores entre equipos rivales, y estudiamos el efecto de dicho intercambio en la probabilidad de dichos equipos de ganar el campeonato[3]. Los pares de jugadores fueron escogidos por jugar en una misma posición y en base a un análisis cualitativo realizado por los autores. En total consideramos 6 dichos intercambios. Para comenzar, simulamos el torneo 100, 1.000, y 10.000 veces (sin intercambio alguno) para aproximar la probabilidad base de cada equipo de salir campeón. Estas estimaciones se muestran en la Tabla 3, junto al resultado real (en puntos) del torneo.

Una observación destacable desde la tabla 3 son las probabilidades de salir campeón de los equipos de Manchester. Nuestro modelo indica que el M. United tiene un "mejor plantel" relativo al M. City, por lo que la probabilidad de salir campeón es 10 % mayor. Por otro lado, los datos indican que durante esa temporada el M. United hizo cerca de 6 goles menos de los que debiese haber hecho (considerando *expected goals*), y por otro lado el M. City hizo 17 goles más de lo que debiese haber hecho. En forma más coloquial, los jugadores del M. United tuvieron un rendimiento bajo el promedio esperado y los delanteros del M. City tuvieron un rendimiento por sobre lo esperado, en términos de conversión de goles basado a la calidad de las oportunidades generadas (*expected goals*). Esto es parte de lo estocástico que puede ser el fútbol.

[3]Vale la pena notar que nuestro modelo ignora ciertas consideraciones estratégicas, y supone que los equipos siempre ponen el mismo esfuerzo en cada encuentro. En ese sentido, el modelo no considera el objetivo de los equipos, y el análisis presente no incorpora de forma especial el objetivo de ganar el campeonato, por lo que otras medidas, como por ejemplo la probabilidad de descender de categoría, también pueden ser analizadas.

Tabla de posiciones real		Probabilidad de campeonar		
Equipo	**Ptos.**	**100 its.**	**1.000 its.**	**10.000 its.**
		%	%	%
Manchester City	**100**	25,00	20,34	20,50
Manchester United	**81**	**27,00**	**31,11**	**31,20**
Tottenham Hotspur	**77**	11,00	18,58	18,52
Liverpool	**75**	11,00	9,32	9,29
Chelsea	**70**	8,00	4,81	5,35
Arsenal	**63**	2,00	1,23	1,24
Burnley	**54**	6,00	4,16	4,12
Everton	**49**	0,00	0,93	0,81
Leicester City	**47**	10,00	7,92	7,45
Newcastle United	**44**	0,00	0,21	0,19
Crystal Palace	**44**	0,00	0,00	0,01
Bournemouth	**44**	0,00	0,17	0,11
West Ham United	**42**	0,00	0,97	0,67
Watford	**41**	0,00	0,00	0,00
Brighton & H. A.	**40**	0,00	0,00	0,04
Huddersfield Town	**37**	0,00	0,00	0,00
Southampton	**36**	0,00	0,00	0,02
Swensea City	**33**	0,00	0,00	0,00
Stoke City	**33**	0,00	0,25	0,26
West Bromwich A.	**31**	0,00	0,00	0,04

Tabla 3: Probabilidad de salir campeón de los equipos, luego de 100, 1.000 y 10.000 iteraciones del torneo.

1. **Harry Maguire por Rob Holding (defensas).** El primer intercambio de jugadores que se realiza es el de Harry Maguire, jugador perteneciente al Leicester City por Rob Holding, perteneciente al equipo Arsenal. Ambos son jugadores que juegan en la posición de defensas centrales y participaron en 3.420 minutos y 820 minutos respectivamente. La Figura 1 muestra el desempeño de ambos considerando los parámetros: porcentaje de duelos defensivos ganados, porcentaje de duelos aéreos ganados, porcentaje de pases cortos correctos, porcentaje de pases largos correctos y porcentaje de minutos jugados en relación a todo el torneo. Si bien, ambos son jugadores relativamente parejos en relación al desempeño que tuvieron durante el campeonato, el jugador Harry Maguire

le saca ventajas a Holding en el porcentaje de duelos aéreos ganados, porcentaje de duelos defensivos ganados y estuvo en cancha más de cuatro veces los minutos en los que participó Rob Holding. Por otro lado, Holding, es levemente superior en los parámetros de pases cortos y pases largos, con un acierto cercano al 80 % y 60 % respectivamente.

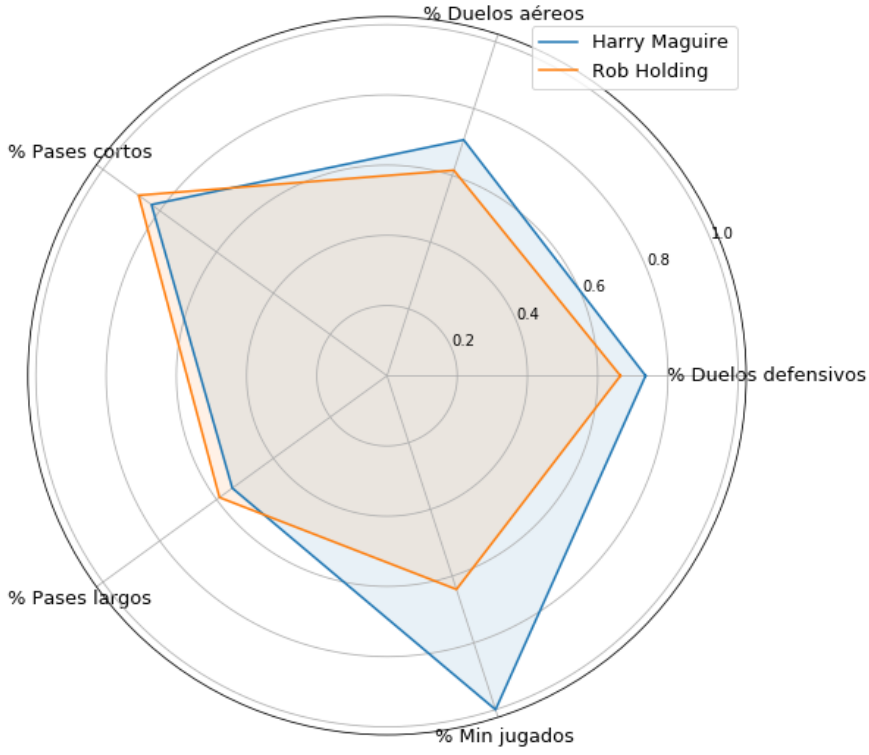

Figura 1: Comparación de rendimiento defensivo del jugador Harry Maguire (Leicester City) y Rob Holding (Arsenal).

La Tabla 4 muestra la variación en las probabilidades de ganar el campeonato para el Arsenal y el Leicester producto de este intercambio; para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Se observa que con la salida de Harry Maguire y la inclusión de Holding en el Leicester City, las probabilidades de campeonar de este equipo bajan en 0,75 puntos porcentuales, de forma inversa la inclusión de Maguire en el Arsenal hace que este equipo se refuerce de mejor manera y sus probabilidades aumentan en 1,16 puntos porcentuales.

Tabla 4: Probabilidades de campeonar intercambiando a Maguire por Holding.

Equipo	Probs. sin cambio	Probs. con cambio	Delta
Arsenal	1,24	2,40	+ 1,16
Leicester City	7,45	6,70	- 0,75

2. **David Silva por Dele Alli.** Los resultados del intercambio de David Silva (2.438 min. jugados) jugador del equipo campeón Manchester City por el jugador del Tottenham Hotspur Dele Alli (2.971 min. jugados), se muestran a continuación. En primer lugar, se analiza el desempeño de ambos jugadores en el torneo, tal y como se muestra en la Figura 2.

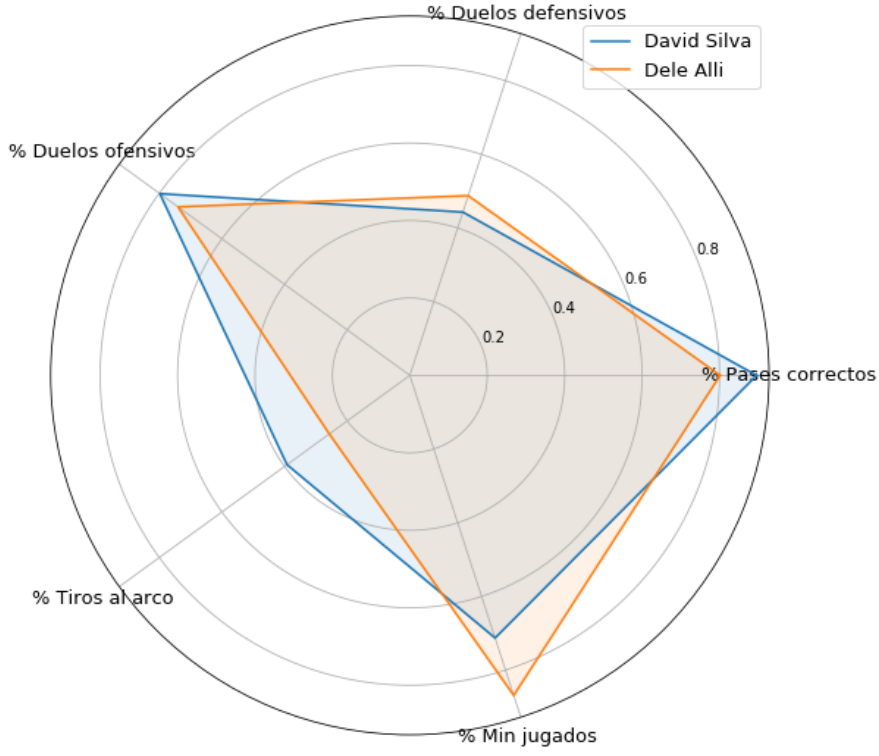

Figura 2: Comparación de rendimiento en el mediocampo del jugador David Silva (Manchester City) y Dele Alli (Tottenham Hotspur).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 5 el Manchester C. disminuye en 9,90 puntos porcentuales su probabilidad de campeonar, mientras que el Tottenham con la inclusión de D. Silva aumenta sus probabilidades en 20,58 puntos por-

centuales. Este es un buen ejemplo, para determinar que Silva podría ser un gran refuerzo en un equipo como el Tottenham dado que con su inclusión las probabilidades de campeonar de este último equipo aumentan considerablemente: para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Tabla 5: Probabilidades de campeonar intercambiando a Silva por Alli.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	10,60	- 9,90
Tottenham Hotspur	18,52	39,10	+ 20,58

3. **Kevin de Bruyne por Mark Noble (mediocampistas).** Los resultados del intercambio entre Kevin De Bruyne (3.084 min. jugados) jugador del equipo Manchester City y Mark Noble (2.404 min. jugados), histórico medicampista del West Ham United desde el año 2004, se detallan a continuación. En la Figura 3 se realiza una comparación del desempeño durante el torneo de estos dos jugadores.

   Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 6 el Manchester C. disminuye en 5,50 puntos porcentuales su probabilidad de campeonar, mientras que el West Ham U. con la incorporación de K. De Bruyne aumenta sus probabilidades en 1,83 puntos porcentuales. Aquí se observa lo importante del proceso de *scouting*, ya que la inclusión de un jugador puede incluso cuadriplicar las probabilidades de campeonar, tal como lo sería K. De Bruyne en el West Ham United: para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Tabla 6: Probabilidades de campeonar intercambiando a De Bruyne por Noble.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	13,70	- 5,50
West Ham United	0,67	2,50	+ 1,83

4. **Romelu Lukaku por Álvaro Morata (delanteros).** Los resultados del intercambio entre el traspaso más caro del torneo, Romelu Lukaku (2.869 min. jugados), jugador del equipo Manchester United y Álvaro Morata (2.068 min. jugados), jugador del Chelsea y de la selección española, se detallan a continuación. En la Figura 4 se muestra una comparación del desempeño durante el torneo entre estos dos jugadores. Lukaku es superior en todos los parámetros a analizar con excepción de

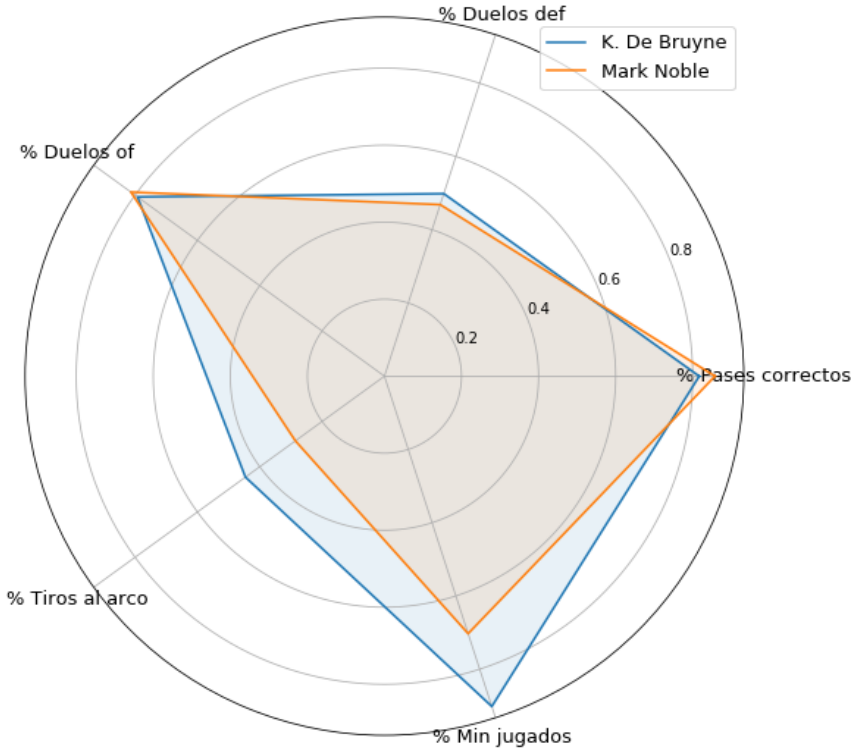

Figura 3: Comparación de rendimiento en el mediocampo del jugador Kevin De Bruyne (Manchester City) y Mark Noble (West Ham United).

los pases correctos, donde Morata tiene una mayor efectividad.

Tal como lo muestra la Figura 4 la salida de Lukaku y la integración de Morata en el Manchester United beneficia al Chelsea, ya que Lukaku es un delantero con mejor rendimiento. En la Tabla 7 se ve que el intercambio entres estos dos ocasiona que el Manchester United se vea perjudicado, disminuyendo en 13,10 puntos porcentuales su probabilidad de campeonar, en contraposición al Chelsea que se vería beneficiado aumentándola en 5,87 puntos porcentuales: para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Tabla 7: Probabilidades de campeonar intercambiando a Lukaku por Morata.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester United	31,20	18,10	- 13,10
Chelsea	5,53	11,40	+ 5,87

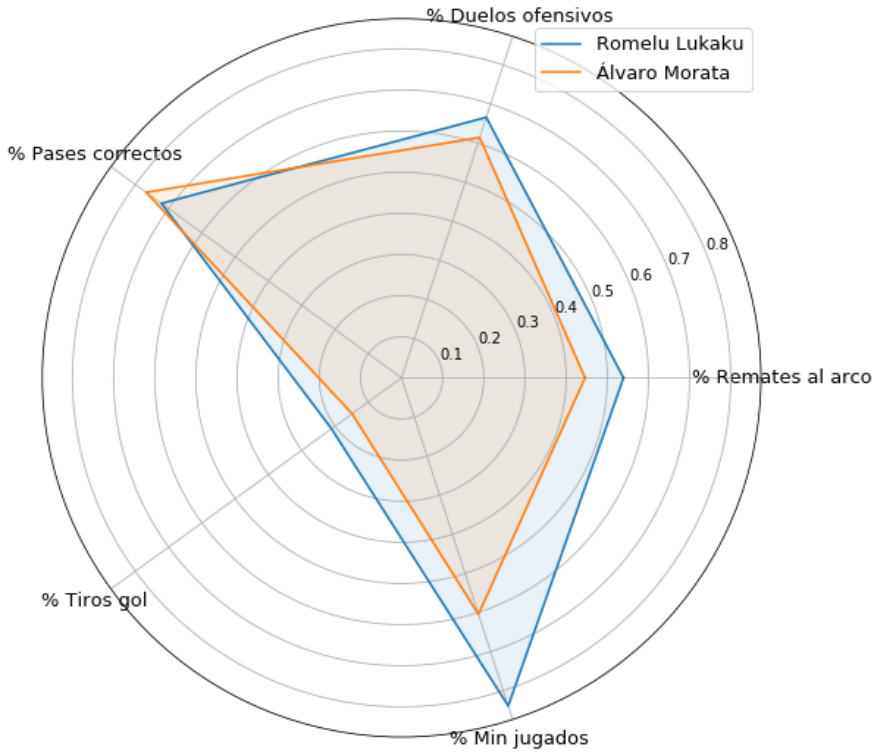

Figura 4: Comparación de rendimiento ofensivo del jugador Romelu Lukaku (Manchester United) y Álvaro Morata (Chelsea).

5. **Sergio Agüero por Alexandre Lacazette (delanteros).** Los resultados del intercambio de Sergio Agüero (1.968 min. jugados), goleador histórico del equipo campeón Manchester City (21 goles en el torneo) y el jugador del Arsenal Alexandre Lacazette (2.211 min. jugados), se muestran a continuación. En primer lugar, se revisa el desempeño de ambos jugadores en el torneo, en el gráfico de la Figura 5.

   Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 8 el Manchester City disminuye en 6,80 puntos porcentuales su probabilidad de campeonar, mientras que el Arsenal con la incorporación de Sergio Agüero aumenta sus probabilidades en 2,46 puntos porcentuales. Si bien Lacazette tiene un mayor porcentaje de tiros al arco, Agüero hizo siete goles más en siete partidos menos jugados, lo que lo convierte en un atacante muy efectivo con 0,84 goles por partido, justo detrás del goleador del torneo Mohamed Salah, mientras que Alexandre Lacazette tiene 0,44 goles por partido, casi la mitad: para ver

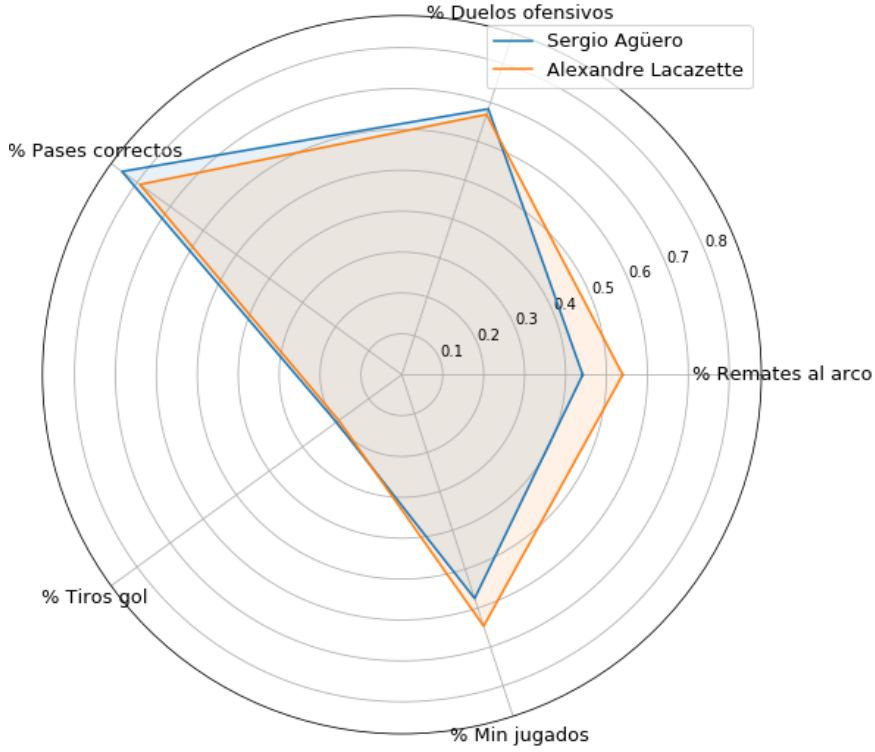

Figura 5: Comparación de rendimiento ofensivo del jugador Sergio Agüero (Manchester City) y Alexandre Lacazette (Arsenal).

el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Tabla 8: Probabilidades de campeonar intercambiando a Agüero por Lacazette.

Equipo	**Probs. sin cambio**	**Probs. con cambio**	**Delta**
Manchester City	20,50	13,70	- 6,80
Arsenal	1,24	3,70	+ 2,46

6. **Mohamed Salah por Mame Diouf (delanteros).** Los resultados del intercambio entre el goleador del torneo Mohamed Salah (2.922 min. jugados), jugador del Liverpool y Mame Diouf (2.601 min. jugados), histórico atacante del Stoke City desde el año 2014 al 2020, se detallan a continuación. En la Figura 6 se realiza una comparación del desempeño durante el torneo de estos dos jugadores. En donde se observa claramente que Salah es un jugador mucho más completo en relación a estos

parámetros, ya que supera en todo nivel a M. Diouf.

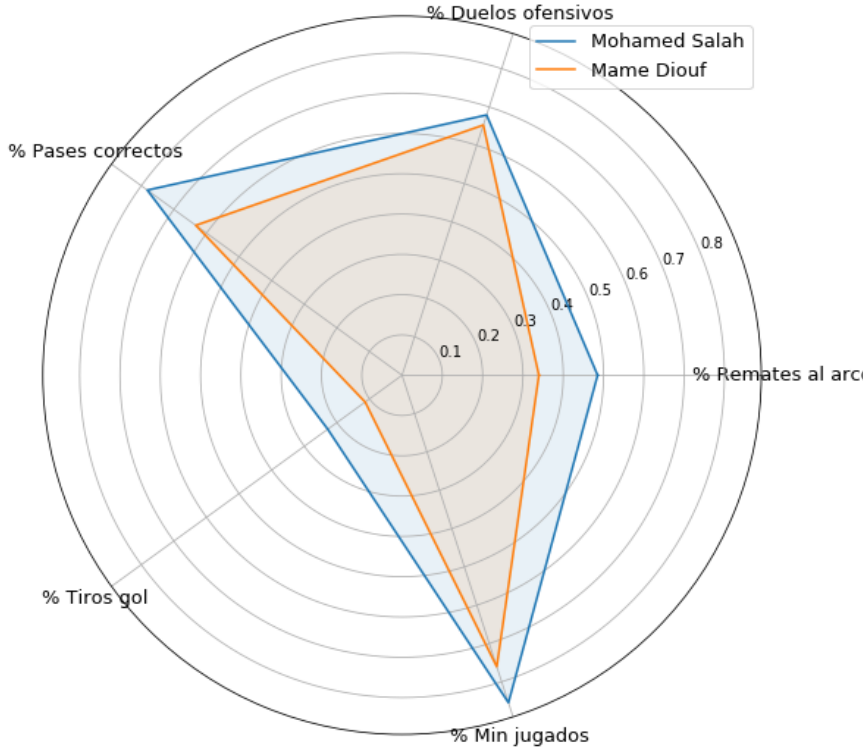

Figura 6: Comparación de rendimiento ofensivo del jugador Mohamed Salah (Liverpool) y Mame Diouf (Stoke City).

Con respecto a la variación en la probabilidad de campeonar, como se muestra en la Tabla 9 el Liverpool disminuye en 8,89 puntos porcentuales su probabilidad de campeonar, mientras que el Stoke City con la incorporación de Mohamed Salah aumenta sus probabilidades en 6,34 puntos porcentuales. La incorporación del goleador (32 goles) del campeonato en un equipo que terminó en el puesto 19° con tan solo 35 goles convertidos, casi los mismos goles convertidos por M. Salah en 36 partidos hace que las probabilidades de campeonar aumenten casi 25 veces: para ver el detalle completo de las nuevas probabilidades de campeonar de todos los equipos, ver [12].

Tabla 9: Probabilidades de campeonar intercambiando a Salah por Diouf.

Equipo	Probs. sin cambio	Probs. con cambio	Delta
Liverpool	9,29	0,40	- 8,89
Stoke City	0,26	6,60	+ 6,34

5. Discusión y conclusiones

La recolección exhaustiva de datos a nivel eventing en la mayoría de las ligas profesionales del planeta plantea un desafió de Big Data en este deporte: la creación de herramientas analíticas para el apoyo a la toma de decisiones. Por otro lado, la literatura actual en fútbol analytics plantea modelos que no son adecuados para apoyar labores de scouting, puesto que se basan en el comportamiento agregado de los equipos, sin utilizar estadísticas individuales de los jugadores.

Con esto en mente, en este trabajo planteamos un enfoque de modelamiento que permite medir el rendimiento de los futbolistas a nivel individual y cuantificar el impacto que la inclusión de un futbolista genera en un determinado equipo, para luego ser utilizado en el proceso de *scouting*. Nuestro enfoque plantea la calibración del modelo utilizando inferencia bayesiana, y estima la capacidad cognitiva-perceptual (toma de decisiones) y capacidad técnica (ejecución de una acción) de cada jugador, permitiendo también ver el impacto de la inclusión de un futbolista en un equipo en particular medido como la probabilidad de ser campeón de dicho equipo. Este tipo de herramientas permite apoyar la toma de decisiones sobre cuáles son los futbolistas que más aportan al equipo, en base al estilo de juego de este y cómo este jugador impacta en a nivel agregado. Se observa que la *salida* de jugadores de buen rendimiento en equipos grandes impacta en forma significativa y negativa al equipo, mientras que su inclusión en equipos de menor categoría impacta en forma positiva pero no en la misma proporción que el equipo afectado. Esto beneficia más a los equipos que disputan el torneo en forma directa al equipo que se le sustrajo un buen elemento (ejemplo: reemplazo de (3) De Bruyne por Noble o (4) Lukaku por Morata). Según la experiencia en clubes de los autores, al momento de contratar un jugador se revisan 3 componentes: 1) reporte cuantitativo del jugador, 2) reporte cualitativo en base a videos y 3) situación contractual. Es posible ligar los resultados de este modelo al reporte cuantitativo que el tomador de decisiones requiere y complementar con los otros dos elementos. Además, esta herramienta sirve para acotar la búsqueda

de jugadores y utilizar menos horas pero con mayor foco en la revisión de videos de futbolistas.

El modelo propuesto presenta variadas limitaciones. Por un lado, el modelamiento solo describe el movimiento del balón sobre el campo de juego, ignorando el juego que se realiza sin el balón; esto se debe al tipo de dato con el que se trabaja (el que representa lo que ocurre con los jugadores que tienen el balón pero no con aquellos jugadores que no tienen el balón en un instante en particular); sin embargo, tecnologías para el seguimiento individual de cada jugador ya se encuentran disponibles, estas son utilizadas principalmente para entrenamientos, y no se encuentran disponibles a nivel comercial. Prevemos que una segunda revolución de datos se originara una vez ese tipo de dato se haga disponible.

Adicionalmente nuestro modelamiento ignora el sistema de juego (formación) utilizada por los equipos o el entrenador de turno (con las implicaciones táctico-estratégicas que esto conlleva), y tampoco considera la demarcación funcional de cada jugador (posición dentro de la formación), ni tampoco el nivel de "presión" bajo el cual transcurre el encuentro (no es lo mismo jugar contra un equipo débil al comienzo del campeonato versus jugar contra el lider, peleando el campeonato, durante las ultimas fechas, en un equipo tradicionalmente fuerte). Sin duda, estos aspectos pueden condicionar las acciones que un jugador realiza durante el partido. Si bien estamos trabajando para incluir estos aspectos en un modelo, esto requiere un trabajo previo de etiquetación el rol de los jugadores en cada partido/formación, e identificación de una métrica razonable para medir la presión que enfrentan los jugadores. Aspectos adicionales que contribuyen potencialmente a mejorar nuestro modelo son, por ejemplo: i) hacer un manejo más detallado del transcurso del tiempo, incorporando heterogeneidad en la duración de las distintas acciones de los jugadores; ii) considerar como la disposición técnica de los entrenadores (sin necesariamente cambiar formación) afecta la toma de decisiones de los jugadores. Prevemos que futuros esfuerzos de investigación se centraran en estos aspectos, y en la calibración del modelo utilizando otras ligas/temporadas.

Otros trabajos que se pueden realizar a partir de este es la inclusión de más ligas de fútbol profesional para ampliar el espectro de jugadores observables (mercado de futbolistas), utilizar otras métricas agregadas de impacto en los clubes tales como probabilidad de clasificar a copas internacionales y evitar el descenso, y por último, evaluar otras métricas del poder predictivo del modelo para mejorarlo y ser más precisos en este ámbito.

Agradecimientos: Esta investigación ha sido financiada en parte por el Instituto Sistemas Complejos de Ingeniería ISCI (ICM-FIC: P05-004-F, CO-

NICYT: FB0816).

Referencias

[1] C. Anderson and D. Sally. *The Numbers Game: Why Everything you know about soccer is wrong*. Penguin Books, 2013.

[2] Tom L. G. Bergkamp, Wouter G. P. Frencken, A. Susan M. Niessen, Rob R. Meijer, and Ruud. J. R. den Hartigh. How soccer scouts identify talented players. *European Journal of Sport Science*, 0(0):1–11, 2021. PMID: 33858300.

[3] Dixon & Coles. Modelling association football scores and inefficiencies in the football betting market. *J Stor*, 0(0), 1997.

[4] W. Dufour. Los métodos de objetivación del comportamiento motor en la recogida de datos en fútbol. *Apunts*, 1982.

[5] FiveThirtyEight. Carmelo nba player predictions. accessed: 10/01/2022, 2018.

[6] P. Galaz. Datazul: Un primer caso de analytics aplicado al fútbol profesional en chile. accessed: 10/01/2022, 2020.

[7] L. Gómez-Jordana, J. Milho, A. Ric, R. Silva, and P. Passos. Landscapes of passing opportunities in football – where they are and for how long are available? *Barça Sports Analytics Summit*, 2019.

[8] S. Gregory. `https://www.statsperform.com/resource/expected-goals-in-context/`. accessed: 10/01/2022, 2020.

[9] M. Hughes. Computerized notation analysis in field games. *Taylor and Frances*, 1988.

[10] M. Lewis. *Moneyball: the art of winning an unfair game*. 2004.

[11] G. Martín. Accumulated confirmed cases of coronavirus nationwide. accessed: 10/01/2022, 2020.

[12] S. Mena. Impacto de los futbolistas de la premier league 2017-2018 en la probabilidad de salir campeon a traves de un metodo de simulacion con inferencia bayesiana. accessed: 01/05/2022, 2022.

[13] R. Neal. Probabilistic inference using markov chain monte carlo methods. Technical report, University of Toronto, 1993.

[14] E. Olsen and O. Larsen. *Science and Football III: Use of match analysis by coaches*. 1997.

[15] L. Pappalardo, L. Guerrini, A. Rossi, and P. Cintia. Explainable injury forecasting in soccer via multivariate time series and convolutional neural networks. *Barça Sports Analytics Summit*, 2019.

[16] L. Pappalardo and E. Massucco. Soccer match event dataset. *Figshare*, 2019.

[17] D. Partridge and IM. Franks. A detailed analysis of crossing opportunities from the 1986 world cup (part i). *Soccer Journal*, 1989.

[18] Á. Ric, R. Peláez, and Barça Innvotation HUB. *Football Analytics: Now and Beyond : a Deep Dive Into the Current State of Advanced Data Analytics*. 2019.

[19] L. Shaw and M. Glickman. Dynamic analysis of team strategy in professional football. *Barca Sports Analytics Summit*, 2019.

[20] B. Spencer, M. Hawkey, and M. Robertson. Using contextual player movement and spatial control to analyse player passing trends in football. *Barça Sports Analytics Summit*, 2019.

[21] Stan Development Team. Stan modeling language users guide and reference manual. Version 2.29, 2022.

[22] Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition, 2009.

[23] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.