FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Análisis estadístico de la posesión del balón en los Mundiales de fútbol Catar 2022 y Australia/Nueva Zelanda 2023

Trabajo final de grado presentado como requisito para la obtención del título Licenciado en Estadística

Martín Grau Pérez Santomauro

Facultad de Ciencias Económicas y de Administración
Universidad de la República

Tutores

Ignacio Álvarez-Castro
Andrés Sosa

Tribunal

Ignacio Álvarez-Castro
Luciana Cantera
Juan Kalemkerian

Montevideo, 15 de agosto de 2025

Agradecimientos

Carlos y Rossana, Manu, Santi y Sofi, Mariu y Lucho y Lu; gracias por siempre aguantar la cabeza y, en definitiva, por siempre estar ahí. No hace falta explayarse demasiado, el agradecimiento es diario y va más allá de este trabajo.

Por último, pero no menos importante, gracias Nacho y Andrés por darme para adelante y confiar en mí, incluso más de lo que yo suelo confiar en mí mismo.

Lula, vas a estar siempre con nosotros.

La conservación es, casi siempre, más una cuestión
de estilos que de pensamientos.
— **Cristina Peri Rossi, *La nave de los locos* (1984)**

Unos dicen jugar bien o jugar mal
Jugar es jugar
Que historia ni cuento
—**El alemán, *Maestro* (2016)**

Resumen ejecutivo

Este trabajo se enfoca en el análisis de las posesiones de balón de las selecciones participantes en los mundiales de fútbol Catar 2022 (masculino) y Australia/Nueva Zelanda 2023 (femenino), utilizando datos de *eventing* provistos por la empresa *StatsBomb*. El objetivo principal es identificar y caracterizar patrones de juego a través de técnicas de aprendizaje estadístico no supervisado, con énfasis en análisis de clúster mediante el algoritmo *k-means*. Se construye una base de 9370 posesiones filtradas para garantizar trayectorias continuas del equipo en posesión, calculando una amplia variedad de características para cada secuencia que describen distintos aspectos de éstas. Tras varios análisis preliminares, se seleccionan 6 clústeres de modo de capturar una variedad de estilos. Cada clúster se caracteriza en función de las variables creadas y su caracterización se puede resumir en: (1) juego basado mayoritariamente en traslados y por los laterales, (2) secuencias con gran cantidad de pases en campo propio, (3) jugadas elaboradas que alcanzan el área (y terminan en tiros al arco), (4) secuencias con gran cantidad de pases en campo rival, (5) acciones mayoritariamente de pelota quieta ofensiva y (6) envíos largos desde campo propio. De modo de complementar el análisis se utilizan herramientas de visualización tales como el gráfico de jugadas y el *parallel plot*. Como aportes adicionales, se explora el éxito de las posesiones según los clústeres así como de posesiones consecutivas ya sea si se alternan los equipos dueños de la posesión o si corresponden a la misma selección.

Palabras clave: Aprendizaje Estadístico No Supervisado, Análisis de Clúster, k-means, Sports Analytics, Football Analytics, Análisis de posesiones, Visualización estadística, eventing

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La analítica de deporte (*Sports Analytics*) es un área de conocimiento que integra diferentes enfoques y técnicas aplicadas con el fin de evaluar el rendimiento en las distintas competencias deportivas. Esto permite extraer información del desarrollo de los partidos en tiempo real y de aspectos sutiles a simple vista. Esta área ha crecido notoriamente estos últimos años motivada principalmente por los avances y desarrollos tecnológicos. Precisamente, es posible medir y registrar con mayor precisión lo que ocurre en cada momento de las competiciones de los distintos deportes. Esto ha derivado en una gran cantidad de trabajos de investigación con aplicación de distintas técnicas matemáticas y estadísticas tanto para medir rendimientos individuales como colectivos. De hecho, las distintas fases del juego en fútbol, básquetbol, fútbol americano o rugby han sido objeto de estudio de analistas e investigadores académicos (Fujii, 2021). Es posible afirmar que el máximo exponente es la *National Basketball Association* (NBA) ya que desde hace años reporta una gran cantidad de estadísticas en tiempo real permitiendo complementar lo que observan los cuerpos técnicos de cara a la toma de decisiones instantáneas. De hecho, en esa liga existen artículos académicos de gran relevancia para la disciplina que datan de hace unos 40 años, los cuales analizan en gran medida los tiros de los equipos durante los partidos y de las secuencias de pases derivan en esos tiros (Gilovich, Vallone, y Tversky, 1985). También resulta relevante el análisis de las distintas secuencias de juego de los equipos a tal punto que su agrupación permite definir estilos de juego y la capacidad, por ejemplo, para anotar puntos (Yu, Wu, Mengersen, y Hobbs, 2022). Además, al igual que sucede con el fútbol, la posibilidad de registrar no solo la trayectoria de la pelota sino que también la de cada jugador en cada instante del partido (*tracking*) permite así modelar dicho movimiento (Mortensen y Bornn, 2019). Estos análisis se encuentran también en otros deportes, tal es el caso del rugby, estos métodos también han adquirido notoria relevancia de modo de obtener numéricamente los aspectos más relevantes del juego (Parmar, James, Hearne, y Jones, 2018).

Puntualmente nos centramos en fútbol, deporte en el cual los avances ante-

riormente descritos han crecido notoriamente en los últimos años. El desarrollo de nuevas tecnologías permite medir con mayor precisión los distintos aspectos del juego incluso en tiempo real. Por lo tanto, medir y cuantificar los datos de lo que sucede en cada uno de los partidos adquiere una relevancia notable con el fin de analizar aspectos del juego difícilmente observables por el ojo humano. Esta área se presenta como una alternativa eficiente para evitar la observación de gran cantidad de partidos para analizar virtudes y falencias tanto del equipo propio como de los rivales. En este contexto, tanto los datos de *eventing* como los de *tracking* desempeñan un papel fundamental en la estructura deportiva de los equipos de alto nivel, y su uso se extiende progresivamente a las categorías formativas. Para ser más específicos, los datos de *eventing* ofrecen un registro detallado de cada acción realizada con el balón, mientras que los datos de *tracking* aportan información de la ubicación exacta de cada jugador en todo momento del partido, incluso de aquellos que no participan directamente en la jugada. Esta información permite extraer variables físicas de los deportistas y analizar en detalle sus interacciones y desplazamientos durante el juego (Otero-Saborido, Aguado-Méndez, Torreblanca-Martínez, y González-Jurado, 2021). De esta manera, al combinar ambas fuentes de información, se pueden extraer patrones de juego, individuales y colectivos, a través del estudio de las distintas fases del juego, así como las distintas formaciones tanto en fase ofensiva como defensiva, entre otras. De modo de ilustrar un poco más las diferencias entre ambas fuentes de información, se puede tomar como referencia cómo varían las visualizaciones de los mapas de calor según los datos que se utilicen (Garrido, Burriel, Resta, del Campo, y Buldú, 2022).

La relevancia de estas nuevas tecnologías y el análisis de datos en el fútbol han derivado en el crecimiento de empresas que se encargan no solo de relevar la información durante las competiciones, sino que también de su procesamiento. De hecho, tanto los datos de *eventing* como los de *tracking* en el mundo del fútbol son recabados por empresas consultoras especializadas en datos deportivos tales como pueden ser *StatsBomb (Hudl)*, *Opta*, *Driblab*, entre otras. Si bien entre ellas pueden existir ciertas diferencias en cuanto a las consideraciones de algunas situaciones puntuales del juego, la esencia de sus datos resulta ser la misma.

En el mundo del fútbol, existen varios ejemplos, ya sea de análisis matemáticos y estadísticos provenientes tanto de empresas privadas, incluso llevados a cabo por los propios clubes, como desde la academia. Estos aspectos sustentan, al menos en parte, la elección de enfocar el análisis en secuencias de posesión y sus respectivas características. Además de la producción académica, algunos clubes han desarrollado iniciativas propias orientadas a la investigación aplicada. El *Barça Innovation Hub*, promovido por el FC Barcelona, es un ejemplo de cómo una institución deportiva puede integrar conocimiento científico a su estructura organizativa, por ejemplo, para analizar qué jugadores propios explotan mejor las zonas liberadas por los jugadores rivales, y cuáles de estos son los que permiten más estas situaciones (Llana, Madrero, y Fernández, 2020).

En el caso del Brentford FC, el club ha creado el *Football Research Centre*, una unidad especializada en análisis de datos que ha sido clave en la identificación (*scouting*) y desarrollo de jugadores a través de métricas objetivas. Desde el centro encontraron que en la liga inglesa se realizan mayor cantidad de acciones de alta intensidad, teniendo en cuenta la posición de los futbolistas en comparación con la liga francesa (Morgans y cols., 2025) lo que sirve para personalizar el entrenamiento y puesta a punto de los jugadores del equipo, particularmente para aquellos que vienen desde Francia.

Desde la academia y la comunidad científica, se han multiplicado los artículos y publicaciones relacionados con el análisis de rendimiento en el fútbol. En ese sentido, los distintos enfoques, así como las técnicas empleadas, son de gran variedad. Sin embargo, aquí se propone realizar un breve repaso en lo que refiere al estudio de los distintos estilos de juego. Precisamente, el comportamiento de los equipos y de su estructura de pases resulta relevante a la hora de evaluar el estilo y, en consecuencia, la manera de obtener los resultados. De hecho, las secuencias de pases se presentan como uno de los aspectos más relevantes a la hora de intentar explicar posibles rendimientos y resultados: de éstas se desprenden los patrones de juego y la manera de alcanzar el arco rival. Más precisamente, la secuencia de pases puede afectar directamente no solo la calidad de las situaciones sino que la probabilidad del desenlace de los tiros al arco. De hecho, mover rápidamente la pelota en las zonas de alto riesgo puede repercutir en un aumento significativo de la probabilidad de éxito en los remates (Cao, 2024). De modo de profundizar en el análisis de los estilos de juego de modo de calificar a los equipos, resulta separar 2 grandes grupos de equipos: aquellos que dominan la posesión y el control de la pelota y los que no, y el estudio de las acciones preponderantes a cada estilo de juego de modo de estudiar la efectividad de estos equipos (Soroka y cols., 2023). Asimismo, se pueden estudiar estilos de juego de modo de agrupar en clústeres ya sea ligas de diferentes países (Plakias y cols., 2023) o jugadores en particular ya sea en base a sus rendimientos (Akhanli y Hennig, 2023) o en base al estudio de los *flow motifs* ya mencionados (Bekkers y Dabadghao, 2019). De manera más precisa, diversos estudios han analizado el crecimiento de las publicaciones en este campo y destacan la precisión de algunas de las aplicaciones de *machine learning* enfocadas al análisis del rendimiento. Estas se subdividen en áreas como el análisis técnico-táctico a partir de secuencias de pase, la identificación de roles posicionales y la estimación de goles esperados (xG) (Rico-González, Pino-Ortega, Méndez, Clemente, y Baca, 2023).

Vale la pena destacar que en los años 50 nos encontramos con el que quizás fue el primer caso conocido de la (quizás incorrecta) utilización de datos en el mundo del fútbol de modo de aplicar a la forma de juego. Fue por esos años que Charles Reep, un poco desencantado por la manera de jugar de los equipos ingleses, se propuso asistir a todos los partidos del conjunto Swindon Town FC de la temporada de modo de registrar lo que sucedía en cada una de las jugadas del partido (acaso el primer *play-by-play* de la historia). Fue a partir de esas anotaciones que Reep llegó a la conclusión de que cuantos menos pases

hiciera el equipo, mayor probabilidad tendrían de marcar gol y por lo tanto, de ganar el partido (Pollard, 2002). Concretamente, observó que la mayoría de los goles derivaban de jugadas con 3 pases previos o menos: *"No más de 3 pases"* pregonaba. Por lo tanto, él argumentaba que la posesión era intrascendente. De este modo, Charles intentó implantar el juego directo hacia los delanteros de modo de llegar lo antes posible (en menos de 3 pases) al arco rival. Si bien sus aportes generaban controversia, su filosofía alcanzó las esferas de la selección inglesa imponiéndose el juego directo como estilo de juego de esa época (Martínez Arastey, 2019). Sin embargo, en la actualidad lo concluido por Reep parece ser un tanto superficial dado que, por lo general, la mayoría de los goles están precedidos por secuencias de pases más bien cortas.

Contrariamente a lo concluido por Reep, la posesión sí es uno de los aspectos más relevantes en lo que refiere al análisis del juego. El análisis de pases en el contexto de un partido de fútbol representa un campo interesante para estudiar, debido a que el conjunto de éstos representa, en gran parte, la idea de juego de los equipos. En ese sentido, la estructura de pases (*flow motifs*) entre compañeros es relevante a la hora de considerar el estilo de juego de un equipo según el director técnico y los jugadores con los que cuente (Håland, Salte, Hvattum, y Stålhane, 2020; Bekkers y Dabadghao, 2019) ya sea para analizar puntos fuertes y debilidades de los rivales o según las formaciones iniciales (Beernaerts, De Baets, Lenoir, y Van de Weghe, 2022). Asimismo, resulta de interés estudiar cómo pueden variar estos, por ejemplo, según el rival o el resultado en cierto momento del partido mediante distintas técnicas y modelos estadísticos: *clustering* de posesiones de la pelota según características de las mismas, así como clasificación supervisada para predecir el éxito o no de dichas secuencias.

Es por eso que en este trabajo se propone analizar las posesiones (y sus estructuras de pases) de los equipos en los mundiales de fútbol de Catar 2022 (masculino) y Australia/Nueva Zelanda 2023 (femenino) de modo de agruparlas (análisis de clúster) según sus características comunes. Posteriormente, se buscará caracterizar la estructura de grupos resultante de modo de extraer patrones de juego derivados de esos clústeres. Para ello, se cuenta con bases de datos de *eventing* de cada uno de los 64 partidos de ambas competiciones provenientes de la empresa proveedora de datos deportivos *StatsBomb*.

Capítulo 2

Métodos

2.1. Aprendizaje Estadístico No Supervisado

El aprendizaje estadístico es un área de la Estadística que consiste, a grandes rasgos, en el estudio de una serie o conjunto de datos de modo de poder extraer la información y los patrones que de estos subyacen. La selección de las técnicas apropiadas depende fundamentalmente de la naturaleza del problema, el tipo de variables involucradas y los objetivos definidos en el análisis.

De manera general, el aprendizaje estadístico puede clasificarse en dos grandes categorías: supervisado y no supervisado. En el aprendizaje supervisado, el interés radica en estimar o predecir una variable de respuesta (denominada *etiqueta*), a partir de un conjunto de variables explicativas u observables (regresores). En este contexto, el objetivo radica en ajustar modelos que describan la relación funcional entre ellas como por ejemplo los modelos de regresión lineal y logística, los árboles de decisión y los métodos basados en máquinas de soporte vectorial.Por otro lado, el aprendizaje no supervisado se centra en el análisis de estructuras inherentes a los datos pero con la diferencia de que no se cuenta con una variable de respuesta. En este caso, el objetivo es descubrir agrupaciones naturales, reducir la dimensionalidad o identificar relaciones latentes entre observaciones, exclusivamente a partir de las variables disponibles. Técnicas clásicas en este ámbito incluyen el análisis de componentes principales (PCA), los métodos de agrupamiento (*clustering*), entre otros. En resumen, podría decirse que el aprendizaje no supervisado consiste en describir y caracterizar la estructura interna de los datos (Hastie, Tibshirani, y Friedman, 2001).

El presente trabajo se enmarca en el campo del aprendizaje estadístico no supervisado, con foco en el análisis de clúster. Este enfoque permite segmentar un conjunto de observaciones en grupos lo más homogéneos posibles internamente y lo más heterogéneos externamente, basándose en métricas de similitud o distancia definidas sobre el espacio de las variables explicativas.

2.1.1. Análisis de Clúster

El análisis de clúster consiste en el estudio de las características de un conjunto de individuos u observaciones con el objetivo de agruparlos en conjuntos homogéneos según un criterio o medida de similitud, generalmente definida a partir de una función de distancia. Esta distancia determina cuán similares son dos individuos, y por tanto, si deberían pertenecer al mismo clúster. En términos generales, el objetivo puede resumirse como: *"Dado un conjunto de puntos de datos, dividirlos en un conjunto de grupos que sean lo más similares posible"* (Aggarwal y Reddy, 2014). Las aplicaciones de las técnicas de clustering abarcan una amplia variedad de dominios, adaptándose a diferentes tipos de datos y contextos: segmentación de clientes (Benítez, Quijano, Díez, y Delgado, 2014), agrupación de fotógrafos mediante sus fotos (u otros archivos multimedia) (Jang, Yoon, y Cho, 2010; Platt, 2000), estilos de juego en fútbol (Bekkers y Dabadghao, 2019) o jugadores según sus rendimientos (Akhanli y Hennig, 2023), entre otras.

Desde el punto de vista metodológico, el análisis de clúster se puede dividir en dos grandes enfoques: *clustering* jerárquico y *clustering* no jerárquico. Por un lado, del *clustering* jerárquico se desprende que existe una estructura jerárquica en los datos de modo que los grupos resultantes reflejen este aspecto. Para ello, existen algoritmos para agrupar las observaciones ya sea de forma aglomerativa (1) o de forma divisiva (2): en (1) se comienza considerando cada observación como un clúster individual y, en cada iteración, se fusionan los pares de clústeres *más similares*, según la métrica previamente definida, hasta obtener un único grupo que contiene a todas las observaciones, mientras que en (2) funciona a la inversa partiendo de un único clúster que contiene a todas las observaciones y se procede a dividirlo de manera recursiva en subgrupos cada vez más pequeños, hasta llegar a la situación en la que cada observación constituye su propio clúster. La elección de la cantidad *óptima* de grupos suele apoyarse en la interpretación del dendrograma, que es un diagrama en forma de árbol que resume el proceso de agrupación y cuya altura de corte define la estructura final de clústeres. Por otro lado, el *clustering* no jerárquico se orienta en la búsqueda de una estructura latente en las observaciones y en la estructura de grupos resultante. Este enfoque se basa en la partición directa de las observaciones (*k-means*, DBSCAN, PAM) en un número predeterminado o estimado de clústeres, optimizando algún criterio de homogeneidad interna y heterogeneidad externa. Un aspecto central en estos métodos es la elección de la métrica de distancia utilizada para evaluar la similitud entre observaciones, la cual puede variar según el tipo de dato.

En términos generales, el procedimiento del análisis de clúster y su conformación puede resumirse, más allá del tipo de agrupación de la siguiente manera:

1. **Selección de características:** elegir o transformar las variables relevantes. En muchos casos, es necesario normalizar o estandarizar las variables para llevar a cabo el análisis.

2. **Cálculo de distancias o similitudes:** definir una métrica para poder agrupar aquellas observaciones próximas entre sí. Para ello existen una vasta cantidad de métricas según el tipo de datos con los que se trabaje. En ese sentido, en lo que refiere a variables cuantitativas ,la más utilizado es la métrica euclídea que mide la separación entre dos puntos en $\mathbb{R}^n$. Además existen la distancia de Manhattan definida por sumatoria de la diferencia absoluta de las coordenadas de los puntos, de Gower para combinar variables de tipo mixto, índice de Jaccard que mide el grado de similitud entre 2 conjuntos, de Mahalanobis entre otras. Se obtiene una matriz $D_{n\times n}$ que refleja las distancias entre cada una de los n individuos teniendo en cuenta cada una de las variables relevadas.

3. **Aplicación del algoritmo** de modo de obtener la partición y asignación de cada observación a un clúster.

4. **Evaluación de la asignación:** existen varios índices de modo de definir la partición *óptima* de los datos siendo el Índice de silueta y el Método del codo (*Elbow Method*) los más conocidos y aplicados. Sin embargo, existen otros índices como el ARI o el índice de Dunn. En caso de contar con la información, se puede comparar los resultados hallados luego del clustering con las etiquetas reales de esas observaciones.

En el caso de contar con variables con diferentes unidades de medida, es razonable transformar las características de las secuencias para que éstas puedan ser comparables entre sí. Más precisamente, se estandarizan las variables mediante la normalización estándar:

$$z = \frac{x - \bar{x}}{s}.$$

También es necesario definir una métrica que cuantifique la proximidad entre ellos de modo de calcular el grado de similitud entre individuos. Una opción comúnmente utilizada es la **distancia euclídea**, que mide la separación entre dos puntos en un espacio $\mathbb{R}^n$. Por defecto, utilizaremos esta métrica, cuya expresión está dada por:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

donde $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ y $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ representan las coordenadas de dos individuos en un espacio n-dimensional. Sin embargo, existen distintos tipos de distancias variando en cada una de ellas el grado de similitud entre 2 observaciones. Por lo general, algunas de estas métricas funcionan *mejor* bajo algún tipo de variable. De manera análoga, se estudia la disimilaridad de estos

individuos de modo de maximizar esa distancia entre los distintos grupos resultantes. Esto resulta un aspecto fundamental a la hora de la partición en grupos ya que distintas métricas, incluso para un mismo conjunto de datos y aplicando el mismo procedimiento de *clustering*, pueden derivar en diferentes resultados.

En este caso, se elige trabajar con el algoritmo *k-means* (k-medias) que consiste en la optimización de la estructura de grupos minimizando la distancia entre individuos pertenecientes al mismo grupo así como maximizando la distancia con los restantes. Para ello, deben definirse previamente el número de clústeres k. El objetivo del algoritmo es optimizar la estructura de los grupos minimizando la distancia entre los individuos pertenecientes al mismo grupo y maximizando la separación entre los distintos grupos. El procedimiento del algoritmo *k-means* se desarrolla en los siguientes pasos:

1. Se eligen k centroides iniciales, que pueden seleccionarse de manera aleatoria o siguiendo algún criterio específico.

2. Cada individuo del conjunto de datos se asigna al clúster cuyo centroide esté más cercano, utilizando una métrica de distancia, generalmente la distancia euclídea:

$$d(\mathbf{x}, \mathbf{c}_j) = \sqrt{\sum_{i=1}^{n} (x_i - c_{j,i})^2}$$

   donde $\mathbf{x}$ representa un punto de datos y $\mathbf{c}_j$ es el centroide del clúster j.

3. Una vez asignados todos los individuos a los k clústeres, se recalculan los centroides como el promedio de los puntos asignados a cada grupo:

$$\mathbf{c}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

   donde C_j es el conjunto de puntos asignados al clúster j y $|C_j|$ es el número de elementos en dicho grupo.

4. Se repiten los pasos de asignación de individuos y actualización de centroides hasta que la asignación de los individuos no cambie significativamente o se alcance un número máximo de iteraciones.

El algoritmo *k-means* es ampliamente utilizado por su simplicidad y eficiencia computacional, aunque presenta ciertas limitaciones, como la necesidad de especificar el número de clústeres k de antemano y su sensibilidad a la elección de los centroides iniciales. A los efectos de encontrar una solución más robusta, en la función `kmeans` permite ajustar a estos efectos los parámetros `iter.max` y `nstart`. Por un lado, el primer argumento indica el número máximo de iteraciones para que el algoritmo converja en cada ejecución, es decir, que valores

altos pueden contribuir a la convergencia del procedimiento en caso de datos complejos para los cuales pueda resultar difícil alcanzar dicho punto de convergencia. Por otra parte, el segundo argumento fija la cantidad de veces que se reinicia el algoritmo con inicializaciones aleatorias distintas en cada instancia para cada uno de los K clústeres. Es decir, el algoritmo se ejecuta `nstart` veces, cada vez con una inicialización distinta de los centroides $\mu_1^{(j)}, \ldots, \mu_K^{(j)}$, para $j = 1, \ldots, nstart$. Finalmente, *k-means* conserva la asignación que presenta menor variabilidad intra grupo:

$$\min_{j=1,\ldots,nstart} \left(\sum_{k=1}^{K} \sum_{x_i \in C_k^{(j)}} \|x_i - \mu_k^{(j)}\|^2 \right),$$

donde μ_k es el *centroide* (media) del clúster C_k. Esto resulta relevante dada la sensibilidad del algoritmo a la elección inicial de dichos centros. En este caso, para robustecer el proceso, se fija `iter.max=100` y `nstart=40`.

Finalmente, se estudia la cantidad *óptima* de clústeres mediante diferentes métodos que resumen la composición final de los grupos en una medida que se toma como métrica. Si bien la estructura *óptima* de grupos no es cerrada ya que puede depender en los objetivos del problema analizado, existen métodos para medir cada una de las particiones obtenidas. En el presente trabajo nos centraremos en (1) el Método del codo (*Elbow Method*), (2) el Índice de silueta y (3) el Índice Rand Ajustado (*Adjusted Rand Index*, ARI).

(1) Método del codo

Una forma de comprender cómo se descompone la variabilidad total en el algoritmo k-means es a través de la relación entre la **suma total de cuadrados** (Total Sum of Squares, TSS), la **suma intra-clúster** (Within-Cluster Sum of Square, WCSS) y la **suma entre clústeres** (Between Clusters Sum of Squares, BCSS):

$$TSS = WCSS + BCSS,$$

donde:

- $TSS = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$: variación total respecto al centroide global.
- $WCSS = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$: variación dentro de cada clúster.
- $BCSS = \sum_{k=1}^{K} |C_k| \cdot \|\boldsymbol{\mu}_k - \bar{\mathbf{x}}\|^2$: variación explicada por la separación entre clústeres.

Este desglose ayuda a evaluar cuánta varianza es capturada por la estructura de clústeres. Por lo tanto, se calcula la suma de distancias cuadráticas dentro de clústeres (WCSS) de modo de elegir la *mejor* partición de grupos. Al

graficar $WCSS(k)$ y k, el punto de inflexión (*codo*) indica la elección recomendada de cantidad de grupos. A medida que k aumenta, WCSS disminuye, pero eventualmente las mejoras marginales se vuelven insignificantes. Ese punto es considerado el valor *óptimo*.

(2) Índice de silueta

Para cada observación i, se define

$$a(i) = \frac{1}{|C(i)| - 1} \sum_{\mathbf{x}_j \in C(i), j \neq i} d(\mathbf{x}_i, \mathbf{x}_j),$$
$$b(i) = \min_{\ell \neq C(i)} \frac{1}{|C_\ell|} \sum_{\mathbf{x}_j \in C_\ell} d(\mathbf{x}_i, \mathbf{x}_j),$$

siendo $C(i)$ el clúster de la observación i. Para calcular las distancias se toma la métrica euclidiana. Luego, la silueta de i es:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]. \tag{2.1}$$

El valor promedio $\bar{s} = \frac{1}{n} \sum_i s(i)$ evalúa globalmente la calidad de la partición.

Es importante examinar la distribución de los valores $s(i)$ por clúster para detectar estructuras internas. Valores cercanos a 1 indican que la observación está bien asignada; valores cercanos a 0 indican que está en el límite entre dos clústeres, y valores negativos implican una asignación incorrecta: la observación está más cercana al centroide de otro clúster que al suyo.

(3) ARI

El ARI es una métrica utilizada en el contexto de análisis de clúster de modo de elegir una partición en función de otra de referencia. Dicho de otro modo, compara todas las combinaciones posibles de observaciones entre las 2 particiones y mide las coincidencias entre ambas de modo de establecer un índice de similitud entre los individuos agrupados y los clústeres resultantes.

$$\mathrm{ARI}(P^*, P) = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j^*}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j^*}{2}\right] - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j^*}{2}\right] / \binom{n}{2}} \in [-1, 1], \tag{2.2}$$

con n siendo el total de observaciones, n_{ij} la cantidad de observaciones que simultáneamente pertenecen al clúster i-ésimo de la partición P y al clúster j-ésimo de la partición P^*, n_i el total de observaciones a $k_i \in P$ y n_j^* el total de observaciones $k_j^* \in P^*$. Un valor del índice igual a 1 nos indica que las particiones son idénticas, mientras que el valor 0 nos indica particiones aleatorias; es decir, cuanto más cercano a 1 sea dicho índice, más similares serán las

conformaciones de grupos resultantes. Un ARI < 0 nos indica que la partición resultante es peor que una asignación al azar de grupos. De esta manera, el ARI nos permite comparar asignaciones de clústeres que fueron calculadas con distintas variables y para distinta cantidad de grupos. Utilizamos dicho índice así como la interpretación futbolística de los grupos resultantes para elegir la partición *óptima* entre todas las calculadas.

2.1.2. Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica descriptiva para la reducción de dimensiones. Se utiliza para transformar un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables incorrelacionadas llamadas *componentes principales*, la cual se enmarca dentro del Análisis Factorial (AF). Su objetivo es resumir la variabilidad total de los datos utilizando la menor cantidad de componentes posibles, preservando la mayor proporción de información. Dicho de otro modo, el ACP busca transcribir los datos en una menor cantidad de dimensiones, construyendo unos nuevos ejes ortogonales para así facilitar, entre otros, su visualización y así poder trabajar en $\mathbb{R}^J$ con $J < P$, siendo P la cantidad de variables del conjunto de datos.

Planteamiento algebraico

Sea $\mathbf{X}_{n\times p}$ la matriz de datos centrados por columnas, donde n es el número de individuos y p el número de variables. El ACP busca encontrar un sistema de variables $\mathbf{Z}_{n\times p}$ de la forma:

$$\mathbf{Z} = \mathbf{XU}$$

donde:

- $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_p]$ es la matriz ortogonal de vectores propios (autovectores) de la matriz de covarianzas muestral $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$,
- Cada vector $\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$ representa la j-ésima componente principal.

Realizando la descomposición espectral de la matriz $\mathbf{S}$ se obtiene:

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

donde $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ contiene los autovalores ordenados descendentemente: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Los vectores propios $\mathbf{u}_j$ definen las direcciones principales, y cada componente principal es:

$$\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$$

La varianza de $\mathbf{z}_j$ es el autovalor correspondiente: $\text{Var}(\mathbf{z}_j) = \lambda_j$.

Por otra parte, desde un punto de vista geométrico, el ACP proyecta los datos sobre una nueva base ortonormal tal que el primer eje (componente) es la dirección de máxima dispersión de la nube de puntos y cada uno de los ejes siguientes es ortogonal a los anteriores a la vez que maximiza la varianza residual.Cada observación es proyectada en este nuevo sistema de coordenadas, facilitando su visualización en dos o tres dimensiones.

Cabe destacar que el enfoque del ACP puede hacerse desde el punto de vista de la nube de individuos, cada observación está en un espacio $\mathbb{R}^p$, o desde la nube de las variables, cada una siendo representada por un vector en el espacio $\mathbb{R}^n$.

Calidad de representación en el ACP

Para evaluar los resultados del ACP, no basta con observar la proyección de individuos y variables en los planos principales; también es fundamental analizar la calidad de representación de cada elemento en esos planos.En el caso de las variables, la calidad de representación indica qué proporción de la varianza total de cada variable queda explicada por las primeras componentes principales consideradas. Un valor elevado sugiere que la variable está bien capturada por el plano seleccionado, reflejando adecuadamente su contribución en la estructura global. Para los individuos (observaciones), la calidad de representación mide la proporción de varianza de cada punto que se conserva en el plano principal. Valores próximos a uno implican que el individuo está bien proyectado y que su posición en el gráfico es fiable para la interpretación.

Gráficamente, en el caso de las variables, esta calidad se asocia a la longitud del vector: cuanto mayor es, mejor es la representación mientras que en el enfoque de los individuos, se relaciona con la distancia al origen: los puntos más alejados del centro están mejor explicados por el plano.

2.2. Visualización estadística de datos

La visualización estadística es una herramienta fundamental en el análisis de datos, tanto en sus etapas exploratorias como en la comunicación de los resultados encontrados. Permite identificar anomalías, relaciones entre variables y estructuras de dependencia que podrían no detectarse necesariamente mediante resúmenes numéricos. Precisamente, los gráficos son esenciales para resumir información y describir relaciones entre variables o individuos, y métodos como el *clustering* (por ejemplo, mediante dendrogramas en la agrupación jerárquica) o el ACP, con el círculo de correlaciones, no son la excepción. Existen múltiples herramientas visuales más o menos complejas, cuya elección depende de la naturaleza de los datos y del objetivo del análisis. Algunas herramientas visuales un poco más complejas pueden ser las curvas ICE (*Individual Conditional Expectation*) o el PDP (*Partial Dependent Plot*) en el contexto del Aprendizaje

Supervisado Interpretable.

Además de su rol descriptivo y exploratorio, la visualización puede utilizarse también como instrumento de inferencia. En contextos de contraste de hipótesis o evaluación de supuestos estadísticos al momento de llevar a cabo un modelo, los gráficos pueden complementar —e incluso reemplazar— a los tests estadísticos *tradicionales*. Por ejemplo, el uso del protocolo de *lineup*, mediante el cual se grafica en una grilla el residuo de un modelo ajustado contra sus valores predichos y se compara con otras 19 visualizaciones generadas mediante simulaciones de datos con distribución bajo la hipótesis nula. Este enfoque permite evaluar visualmente la presencia de no linealidad, heterocedasticidad o desviaciones respecto a la normalidad de los errores. Los resultados obtenidos con este método han demostrado ser comparables, e incluso más robustos en ciertos casos, que los de pruebas estadísticas como Shapiro–Wilk (no normalidad), Breusch–Pagan (heterocedasticidad) y RESET (no linealidad) (Li, Cook, Tanaka, y VanderPlas, 2024).

Este tipo de enfoques, que piensan a la visualización en un plano similar al de los métodos inferenciales (tests estadísticos), destacan la importancia de contar con herramientas gráficas en todo el proceso de análisis: desde la descripción de los datos hasta su validación. Para ello, resulta fundamental comprender no solo qué información se desea representar, sino también cómo construir esas visualizaciones y, de alguna manera, poder unificarlas con un criterio común. En ese sentido, la *gramática de los gráficos*, popularizada por Hadley Wickham mediante el paquete `ggplot2`, ofrece un marco conceptual sólido para componer gráficos mediante capas, escalas, transformaciones y temas (Wickham, 2010).

2.2.1. Gramática de capas

En esencia, lo que propone Wickham es un criterio común al momento de graficar. Más importante aún, la posibilidad de modificar cada uno de los aspectos que componen las visualizaciones de modo de poder transcribir visiblemente las posibles relaciones entre los objetos de estudio, así como los resultados hallados. Precisamente, cada uno de estos aspectos se puede hacer mediante cada una de las capas que lo componen.

Si bien a lo largo del análisis se exploran y utilizan una gran variedad de visualizaciones, se propone presentar dos herramientas gráficas que nos permiten resumir los resultados finales hallados. Precisamente, se propone describir en detalle la construcción del gráfico de jugadas, cuya definición se presenta en la Sección 3.1 así como el gráfico de coordenadas paralelas analizado en la Sección 4.2. Ambas visualizaciones nos permiten explicar los resultados obtenidos con mayor profundidad de análisis.

2.2.2. Gráfico de jugadas

Para graficar las jugadas definidas en la Sección 3.1 nos serviremos de la *gramática de capas* para poder resumir las distintas secuencias en el plano (x, y) de una gran variedad de eventos (3.1) por lo cual utilizaremos varias capas de modo de resumir esa información.

En este análisis se diferencian 4 capas: (1) las medidas de la cancha con sus respectivas delimitaciones, (2) los eventos *estáticos*, (3) los eventos que implican un movimiento en la trayectoria de la secuencia y (4) la dirección de los tiros al arco.

En el punto (1) definimos los límites del terreno de juego según las medidas que utiliza *StatsBomb* en sus datos, tal como se detalla en la Figura 2.1. De esta manera, la primera capa de la visualización queda definida según estas medidas. Además, el equipo poseedor de la pelota, siempre defiende el arco de la *izquierda*, coordenada $x = 0$, mientras que ataca el de la *derecha*, coordenada $x = 120$. De igual manera se define la coordenada $y = 0$ como la línea lateral izquierda de la cancha y $y = 80$ como la línea derecha (en el sentido de ataque).

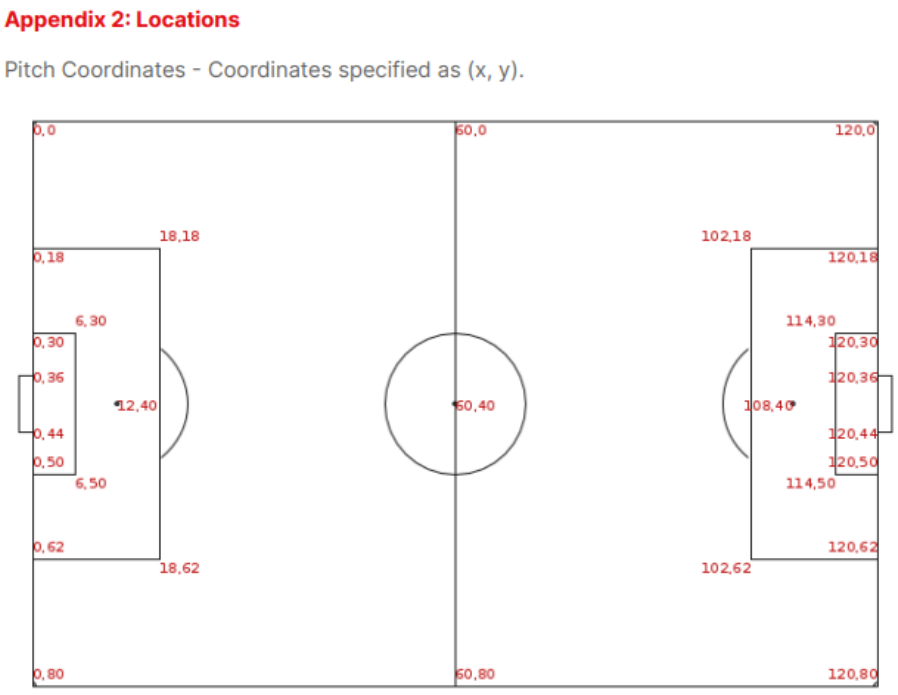
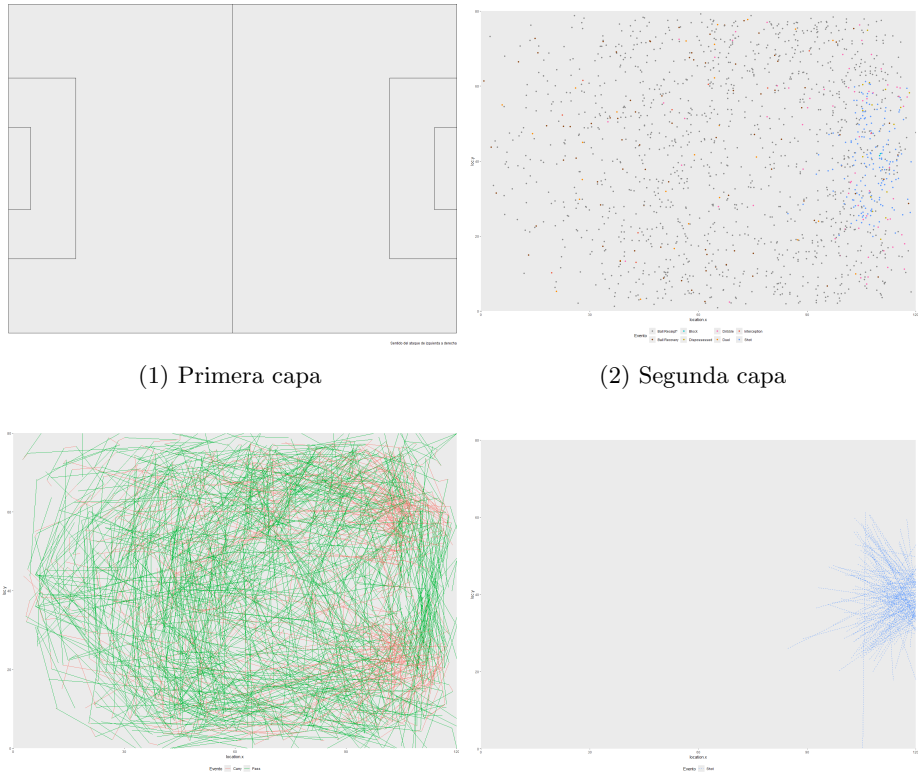


Figura 2.1: Coordenadas (x, y) de cada una de las delimitaciones de la cancha de *StatsBomb*

Al empezar a graficar cada posesión, se separan las acciones en 3 tipos distintos. En el punto (2) tomamos eventos *estáticos* en lo que refiere a la pelota. Concretamente, hay acciones en los que la trayectoria del balón no se modifica de manera considerable por lo que podemos representarlos únicamente mediante puntos (`ggplot2::geom_point`). La mayoría de estas dan inicio o culminan las jugadas ya que refieren a intercepciones, duelos, recuperaciones y robos de pelota, o, en menor medida, regates a rivales (*dribble*). Además, de esta misma manera, se visualizan las recepciones de balón que representan una parte importante de los datos ya que están asociados a los pases que reciben los/as futbolistas de sus compañeros/as. De esta manera, de modo de identificar cada tipo de evento, se asocia un color distinto a cada uno.



(1) Primera capa

(2) Segunda capa

(3) Tercera capa

(4) Cuarta capa

Figura 2.2: Visualización de las 4 capas del gráfico por separado.

En el punto (3), se tiene en cuenta las acciones más frecuentes en cada posesión: los pases y los traslados de pelota (*carry*). Éstos representan en gran medida los estilos de juego de los equipos y la manera de intentar llegar al arco rival al momento de la posesión. A diferencia de la capa (2), estas acciones

representan una trayectoria de la pelota por lo que se busca trasladar ese movimiento a la visualización. En ese sentido, se cuenta con las coordenadas (x, y) de inicio y final tanto de los pases como de los traslados, y utilizamos segmentos de línea para graficarlos (`ggplot2::geom_segment`) con colores específicos para cada tipo de evento.

En el punto (4) se toma en cuenta la dirección de los tiros al arco. Se decide graficarlos por separado debido a que esta no forma parte de la secuencia en cuestión, pese a que pueda representarse de igual manera que los pases y traslados mediante segmentos en función de su dirección. De esta manera, se utilizan las coordenadas (x, y) del remate para incluirlo mediante un punto, pero a su vez graficaremos la dirección del mismo con una línea punteada de modo de dejar en claro que la misma no forma parte en sí de la jugada a analizar (simplemente, si fue gol o no). Al igual que con las capas anteriores, se utiliza el mismo color específico para estos remates.

En la Figura 3.4 se puede ver una secuencia completa graficada según lo descrito anteriormente. Finalmente, se pueden *complejizar* estas visualizaciones, utilizando `ggplot2:: facet_grid` separando las secuencias ya sea según la competición, del clúster al que corresponden, del éxito (o no) de la jugada, o cualquiera de las variables restantes.

2.2.3. Coordenadas paralelas

En términos generales, el gráfico de coordenadas paralelas (*Parallel Coordinates Plot*, PCP) constituye una de las herramientas más utilizadas para la visualización de datos multivariantes, especialmente en el caso de variables numéricas. Este tipo de representación permite analizar simultáneamente múltiples dimensiones y explorar posibles patrones, asociaciones o agrupamientos en un conjunto de datos de alta dimensionalidad. Además, existen algunas alternativas para incluir también variables categóricas (Ge y Hofmann, 2020).

El PCP se construye representando cada variable como un eje vertical dispuesto en paralelo, distribuidos de manera equidistante sobre el eje horizontal. Cada observación individual se representa mediante un segmento que conecta, de manera secuencial, los valores que adopta en cada una de las variables. De este modo, la trayectoria de cada línea codifica la información multivariante de una observación en el plano (x, y), facilitando la comparación visual entre individuos. Asimismo, se puede agregar información adicional sobre clases o grupos mediante el uso de atributos visuales, como el color o la transparencia de las líneas. Al asignar un color específico a cada clase, se pueden identificar patrones comunes, tendencias o posibles separaciones entre grupos previamente definidos (por ejemplo, clústeres).

Desde un punto de vista formal, el PCP proyecta un conjunto de observaciones en $\mathbb{R}^n$, a un plano bidimensional (x, y), preservando la información de cada variable a través de los ejes verticales. Sin embargo, para que la comparación entre variables sea representativa, resulta indispensable estandarizar o normalizar previamente las escalas, ya que cada variable puede estar medida en distintas unidades. Usualmente se recurre a transformaciones lineales para asegurar que todas las variables queden representadas en un rango comparable. Otro aspecto crítico en la construcción de un PCP es el ordenamiento del eje horizontal. El orden en que se dispongan las variables influye directamente en la percepción visual. Por lo tanto, además del método de escalado de las variables, resulta indispensable definir su ordenamiento en el eje X. De hecho, existen varios métodos los cuales modifican la visualización final.

En síntesis, el gráfico de coordenadas paralelas constituye una herramienta versátil y una alternativa visual interesante para describir cada una de las secuencias representadas anteriormente en función de sus características y destacando según el clúster al que pertenecen. Para graficar estos resultados, se utiliza la función `GGally::ggparcoord` de modo que cada segmento representa una posesión distinta en función de los valores de cada una de sus características. No obstante, su correcta interpretación requiere una cuidadosa elección del escalado y del orden de los ejes. En el presente análisis se trabajará con el método de ordenamiento *allClass* y de escalado *uniminmax*. En base a este ordenamiento y método de escalado se pueden observar ciertos patrones de cada grupo los cuales se profundizan en la Sección 4.2:

- *allClass*: se realiza un ANOVA para cada una de las variables (variable de respuesta) en función de la clase, el clúster al que pertenece (variable de entrada), y se ordenan de manera decreciente según los valores del estadístico F resultante
- *uniminmax*: escala los valores de cada variable de manera que el rango sea [0,1]. Es decir, el valor estandarizado $x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$

Capítulo 3

Datos

3.1. Base de *eventing*

En el presente capítulo se propone describir los datos utilizados para el análisis así como el proceso realizado para obtener la base de datos a ser analizada. Es decir, se busca resaltar la importancia del procedimiento realizado en cada instancia hasta obtener la matriz utilizada para el estudio de las jugadas, es decir, para el *clustering*. En el contexto de Analítica del Deporte y estudio de los estilos de juego de los distintos equipos, existen principalmente 2 tipos de datos diferenciados: datos de *eventing* y de *tracking* provistos por las empresas de datos deportivos. Para el presente trabajo se utilizará únicamente la base que detalla cada uno de los eventos que se dieron con el balón en el transcurso de los 64 partidos del Mundial de Fútbol FIFA disputado en Catar en 2022 y del disputado en Australia y Nueva Zelanda en 2023. Estos datos se pueden obtener de manera libre y gratuita a través del paquete `StatsBombR`[1]. Por lo general, estos datos no son de fácil acceso ya que es parte de los servicios que estas empresas ofrecen a los clubes. Esta base detalla cada una de las acciones que se realizan a través de la pelota, denotándose éstas como **eventos**, y de variables relacionadas a cada uno de estos eventos según corresponda. Estos datos son recabados por la empresa *StatsBomb*, especializada en este tipo de datos. En este caso, la base del mundial masculino consta de 233821 observaciones, mientras que la del mundial femenino tiene 226146 filas y ambas bases cuentan con 184 variables (columnas). Precisamente, estas variables engloban una vasta cantidad de aspectos: identificadores de los jugadores, equipos, partidos y las posesiones, así como las distintas acciones que realizan los jugadores con la pelota. Además del tipo de acción realizada, se cuenta con otra cantidad de características asociadas a cada acción. En resumen, cada una de estas acciones se denotan como **eventos** y distintas filas de la base de datos refieren a distintas situaciones de cada uno de los partidos registradas mediante esos eventos. De hecho, *StatsBomb* registra una amplia variedad de estas acciones de modo de

[1]El link del respositorio de github es: https://github.com/statsbomb/StatsBombR

50/50	*Bad Behaviour*	*Ball Receipt**
Ball Recovery	*Block*	*Carry*
Clearance	*Dispossessed*	*Dribbled Past*
Dribble	*Duel*	*Error*
Foul Committed	*Foul Won*	*Goal Keeper*
Half End	*Half Start*	*Injury Stoppage*
Interception	*Miscontrol*	*Offside*
Own Goal Against	*Own Goal For*	*Pass*
Player Off	*Player On*	*Pressure*
Referee Ball-Drop	*Shield*	*Shot*
Starting XI	*Substitution*	*Tactical Shift*

Tabla 3.1: Lista de los 33 tipos de eventos.

trasladar lo más fielmente posible lo que sucede realmente dentro del campo de juego.

3.2. Variables

El valor de cada variable depende del tipo de acción registrada, ya que cada variable está vinculada a un evento específico. A excepción de la variable que indica a qué tipo de evento se refiere cada acción, se cuenta con otras 180 variables referidas a cada una de esas acciones, tomando valores únicamente en aquellas que corresponda según el tipo de acción descrita. Dentro de ese grupo de variables se cuenta, por una parte, con algunas relativas a identificadores, ya sea del evento en sí mismo, del partido, tiempo y minuto de partido, de la posesión, entre otras. Si bien estas variables no refieren a aspectos específicos del juego, nos serán de utilidad a la hora de caracterizar las distintas secuencias. Por otra parte, se dispone de información precisa sobre dónde ocurre cada acción de juego (coordenadas de la cancha), si la acción se realiza bajo presión de un rival, el tiempo que dura la posesión o el tiempo que llevan con el dominio de la misma.

3.2.1. Eventos

Concretamente, la empresa proveedora de los datos registra 33 tipos de eventos distintos (StatsBomb, 2022). Algunos refieren a eventos relativos a los encuentros, ya sea su inicio y su final, sustituciones de jugadores, formaciones iniciales, cambios de formación, sueltas neutrales de balón o incluso pausas en el partido, ya sea por lesión u otra situación excepcional, mientras que otros refieren a acciones específicas del partido en si mismo.

Además, dependiendo del tipo de evento en cuestión, se cuenta con varias variables relacionadas según corresponda. Asimismo, se cuenta con información

detallada adicional relativa a éstas, como a qué partido corresponden, el minuto en que se dieron y en qué sectores de la cancha. En términos generales, la base de datos no presenta grandes errores más allá de alguna cuestión puntual relativa al tiempo en el que ocurren y, por lo tanto, de la duración de la posesión. Más precisamente, se encontró, únicamente para algunas posesiones puntuales, que el momento del tiempo en el que se efectuó la última acción de la secuencia era previo a la penúltima acción. Al observarlo, se corrigió asignando manualmente un segundo más a esos últimos eventos respecto del anterior. De esta manera, se calcula entonces la duración de la posesión como el tiempo (en segundos) que pasa entre la primera y la última acción de la secuencia. De hecho, al contar con una variable que indica esa duración, se puede verificar el cálculo realizado y éste coincide.

Para este trabajo se excluirán del análisis los eventos: inicio y final de cada tiempo, formaciones iniciales, sustituciones, tarjetas mostradas por el juez, salida e ingreso de un jugador del terreno de juego sin que se haya realizado una sustitución, cambios de formación, pausas del partido por lesiones y sueltas neutrales del árbitro, ya que se entiende que no aportan información respecto al análisis del juego en sí mismo (*Bad Behaviour, Half End, Half Start, Injury Stoppage, Player Off, Player On, Referee Ball-Drop, Starting XI, Substitution, Tactical Shift* respectivamente). Finalmente, se trabajará únicamente con 23 eventos relacionados con la posesión de la pelota de alguno de los 2 equipos a excepción de las acciones de presión (*Pressure*) que son realizadas sin la pelota y los duelos (*Duel* y *50/50*) en los cuales el balón se encuentra en disputa. Las variables (características) asociadas a cada tipo de evento varían según la acción relevada. Además, cabe destacar que para aquellas variables que reportan características de otro evento que el reportado, estas columnas figuran vacías para esos registros. La distribución de dichas características según el tipo de evento se puede observar en la Tabla A.2. En la Figura 3.1 se puede observar la distribución de las principales acciones registradas según su frecuencia por partido diferenciando entre la competición masculina y femenina. Para la visualización se tuvo en cuenta únicamente aquellos eventos cuya mediana (incluyendo ambas competiciones) superaba las 25 ocurrencias por partido. En lo que refiere a Pases (y Recepciones de pelota, *Ball Receipt*) y Traslados de balón (*Carry*) se ve mayor cantidad de estas acciones por partido a nivel masculino lo que es consistente con una mayor cantidad de acciones defensivas a nivel femenino como son las acciones de presión (*Pressure*), Recuperaciones de pelota (*Ball Recovery*), Duelos (*Duel*), Despejes (*Clearance*), Bloqueos (*Block*).

Aunque el gol es el factor que determina de manera directa el desarrollo y, especialmente, el resultado de un partido, las secuencias de pases representan un objeto de estudio clave para comprender las estrategias con las que los equipos buscan alterar el marcador a su favor. En el contexto del análisis de los partidos mediante los datos de *eventing* se puede acceder a información detallada de modo de profundizar en el estudio de los partidos más allá de si consiguen anotar (o no) esos goles. Tal como se mencionó anteriormente, se registra una

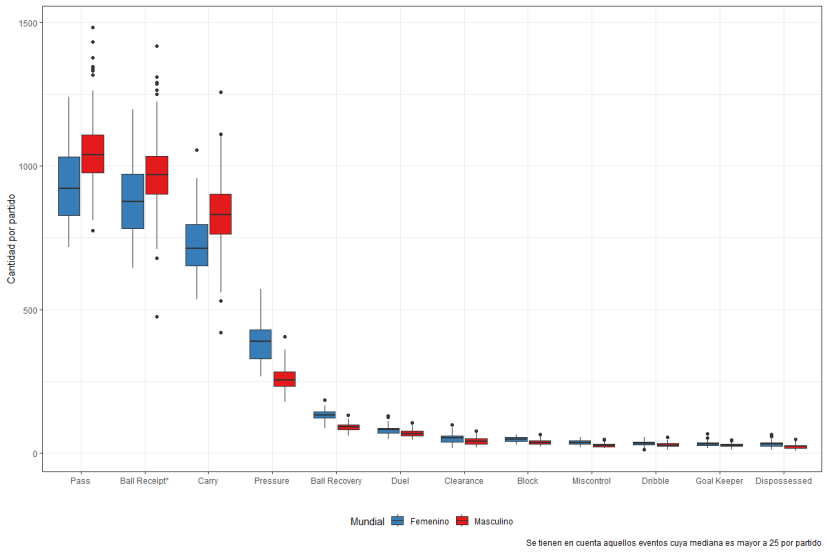


Figura 3.1: Distribución de la cantidad de eventos por partido según competición.

importante variedad de eventos pero se propone prestar particular atención en los pases debido a que éstos representan no solo la mayoría de las acciones del juego sino que además nos permite analizar y categorizar la forma de jugar de un equipo.

Pases

El análisis de las secuencias con la pelota es muy relevante, y en ese contexto los pases (cómo, cuándo y dónde se realizan) y sus características adquieren una relevancia notable para identificar los comportamientos ofensivos y defensivos de los equipos. En total, contamos con 68514 pases efectuados en los partidos del mundial masculino y 59837 en los partidos del mundial femenino, de los cuales 56346 y 44290 fueron exitosos, respectivamente. La base de datos cuenta con información específica asociada a cada pase que se dio en cada uno de los partidos. Más precisamente, se cuenta con 31 variables relacionadas con sus características, detalladas en la Tabla A.3. Estas variables refieren a aspectos de cada uno de los pases que se realizan en los partidos. Se puede profundizar para ver cómo éstas pueden resultar de utilidad a la hora de analizar el comportamiento de los jugadores y de los equipos. La variable *pass.angle* nos da el ángulo (en radianes) del pase en función de su dirección. Esta variable puede ser útil para categorizar cada uno de los pases según hayan sido para atrás, para adelante o hacia alguno de los laterales (izquierda o derecha), tomando como referencia el sentido de ataque. Por lo general, en lo que refiere a posibles estilos de juego, aquellos equipos que se caracterizan por dominar la posesión del balón

durante la mayor parte del encuentro, suelen jugar más en corto y mayoritariamente hacia los costados o hacia atrás en el entendido de que muchas veces los pases hacia adelante suponen un mayor riesgo de pérdida de la pelota. En ese sentido, podemos analizar esta cuestión para las competiciones de estudio normalizando la cantidad de pases realizados por cada equipo a 90 minutos en función del tiempo que cada conjunto efectivamente tuvo la pelota en su poder y no de la duración total del partido. De esta manera, se busca reducir el sesgo para los equipos con menor posesión y, además, visualizar de una manera más clara los estilos de juego.

Tiros al arco

En el análisis de los tiros al arco, la base cuenta con 1494 remates al arco efectuados en el mundial masculino y 1680 en el femenino, de los cuales 195 y 184 terminaron en gol, respectivamente. Se cuenta asimismo con 24 variables relacionadas con las características de cada uno de esos tiros detalladas en la Tabla A.4. En detalle, si se profundiza en el *resultado* de cada uno de esos remates se obtienen las variables descritas en la Tabla 3.2.

Valor	**Definición**
Blocked	Un disparo que fue detenido por un defensor antes de que pudiera continuar su trayectoria.
Goal	El disparo terminó en gol.
Off T	Un disparo cuya trayectoria inicial terminó fuera del arco.
Post	Un disparo que impactó uno de los tres palos (verticales o travesaño), sin entrar al arco.
Saved	Un disparo que fue detenido por el portero rival y que iba a puerta.
Wayward	Un disparo muy desviado o mal ejecutado, sin potencia o dirección. Puede incluir errores de contacto.
Saved Off Target	Un disparo que no iba al arco pero fue detenido por el portero igualmente.
Saved To Post	Un disparo que el portero ataja y en el rebote el balón pega en el palo.

Tabla 3.2: Definiciones de los valores posibles de *shot.outcome.name* en StatsBomb.

Tal como se mencionó anteriormente, los avances científicos y de la tecnología han permitido el registro y medición de todas las acciones de un partido de fútbol (con y sin pelota), dando lugar a múltiples análisis y es en esa línea que, de modo de medir la calidad de las situaciones de gol, se creó la métrica de Goles Esperados (xG, por *Expected Goals*). Esta es una de las métricas *avanzadas* más expandidas actualmente. Concretamente, el xG mide, en el instante previo a que el remate se lleve a cabo, la probabilidad de que éste termine en gol teniendo en cuenta: la ubicación del pateador, de los jugadores de campo rivales, del arquero rival, si el remate se da de primera/de cabeza o con tiempo para definir, entre otros aspectos (Lucey, Bialkowski, Carr, Yue, y Matthews, 2014). En resumen, se trata de un modelo estadístico que estima, según la ca-

racterística de la situación **previa** al tiro y un histórico de situaciones similares, la probabilidad de que termine en gol[2]. Esta métrica se utiliza para medir la calidad de las situaciones en las que se efectuó un tiro, es decir que cuanto mayor sea esa probabilidad, se puede afirmar que fue una *mejor* situación lograda por el equipo. Uno de los aspectos importantes a tener en cuenta al momento de analizar los Goles Esperados, es su poca robustez en caso de que se consideren dentro de los disparos los penales ya que estos inflan el valor de la métrica y no necesariamente resumen correctamente la calidad de la generación de situaciones del equipo. Si bien *StatsBomb* le asigna una probabilidad de 0,7835 a los penales, el xG varía según proveedor, pero siempre se encuentra en el entorno del $[0,76; 0,80]$. Es por esta razón que se suelen excluir los penales al momento de analizar la capacidad y calidad ofensiva o defensiva de los distintos conjuntos. En los datos, el xG promedio es de 0,090 en 2022 y de 0,096 en 2023 (excluyendo penales). Asimismo, si contamos únicamente los tiros desde dentro del área el promedio de gol esperado es de 0,135 en el mundial masculino y 0,121 en el mundial femenino (sobre 883 y 1055 remates totales, respectivamente). De manera similar se miden los goles esperados por el equipo rival, de modo de analizar la faceta defensiva de estos equipos. Esta métrica se define como el xGA (*Expected Goals Against*) y es la sumatoria de los xG de cada uno de los tiros al arco realizados por el rival. Dado que hay disparidad en los partidos (y minutos) disputados por los equipos en cada uno de los 2 mundiales analizados, se normalizan el xG y el xGA a 90 minutos de modo de poder comparar los rendimientos de los distintos equipos.

Por otra parte, es posible analizar la ubicación de los tiros en función de sus probabilidades, así como el desenlace de éstos de la Tabla 3.2. A modo de ejemplo, seleccionamos los 2 jugadores con mayor cantidad de disparos en el mundial 2022 y las 2 jugadoras en el mundial 2023 (en ambos casos no se tiene en cuenta los penales). De hecho, se cuenta con la ubicación de la cancha en dónde se efectuó el tiro (así como de todas las acciones) mediante las coordenadas (x, y), sino que además se registran las coordenadas del destino del remate, es decir, la dirección del remate tal como se puede observar en la Figura 3.2. En el caso de que el tiro no vaya en dirección al arco (entre los 3 palos), la coordenada z de dicho registro figura vacía. Esta coordenada es de suma importancia a la hora de obtener la métrica de xGOT (*Expected Goals On Target*) que mide la probabilidad de gol únicamente de aquellos tiros que van al arco teniendo en cuenta su dirección.

Además de estos eventos detallados, se cuenta con variables asociadas a cada una de las acciones registradas en los partidos con las cuales se pueden construir métricas para estudiar la dominancia de un equipo por sobre otro en un partido. Por ejemplo, la métrica *Field Tilt* se define como el cociente entre el total de pases realizados por el equipo en el último tercio sobre el total de pases realizados en ese mismo sector del campo, o el Tiempo Efectivo de los partidos que consis-

[2]Breve descripción: https://www.youtube.com/watch?v=GYib8tyExUg

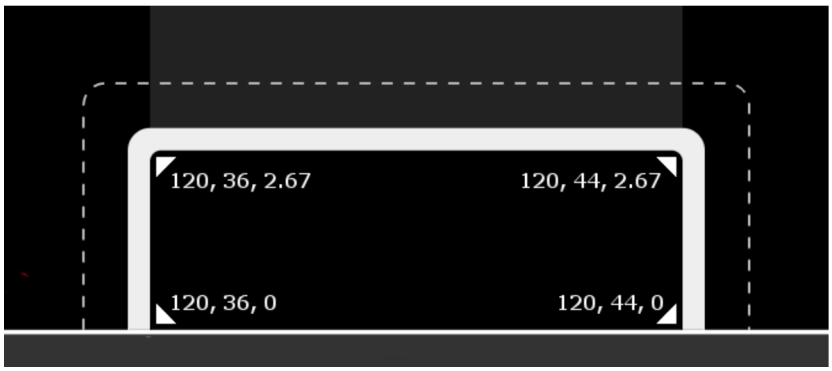


Figura 3.2: Coordenadas (x, y, z) del arco.

te en medir cuánto tiempo está *efectivamente* en disputa la posesión de la pelota.

3.2.2. Variables creadas

Con el objetivo de analizar en profundidad las secuencias de pases, consideramos crear nuevas variables que nos permitan complementar mejor el contexto en el que se dan cada una de las situaciones relevadas. Por ejemplo, en la base original no se cuenta con la información del resultado del partido de cada uno de los eventos detallados. Esto podría ser un aspecto relevante debido a que hay ciertos equipos que pueden variar su forma o estilo de juego según se encuentren o no en ventaja. Así, para cada evento, tendremos si el equipo que realiza la acción va ganando, perdiendo o empatando en ese momento del partido y ver de alguna manera si dicha variable influye en la manera de jugar del equipo o si hay un cambio de estilo a raíz de un resultado adverso. Por lo tanto, una de las características de las secuencias a analizar será el resultado del equipo dueño de la posesión en ese instante del partido.

Asimismo, se calcula el tiempo efectivo de posesión tanto a nivel general como para cada equipo en cada partido. Este tiempo se define como el período en el que el balón está realmente en juego, es decir, bajo el control de alguno de los equipos. Por lo general, esta medida difiere significativamente de los 90 minutos reglamentarios del encuentro, lo que permite evaluar qué tan interrumpido estuvo el partido y cómo se distribuyó la posesión entre los participantes. Los datos de *eventing* nos permiten calcular esta métrica sumando la duración de las secuencias en las que cada equipo tuvo participación. Esta métrica será tenida en cuenta más adelante al momento de caracterizar las secuencias, viendo cuánto representó cada una de ellas en el tiempo efectivo de posesión del partido en cuestión.

Por último, para complementar la información relativa a la ubicación exacta

de cada acción mediante sus coordenadas, se busca agrupar todos aquellos eventos que se dan en sectores similares de la cancha y así tenerlo en consideración para el posterior análisis. En ese sentido, se divide la cancha en 30 zonas tal como se puede observar en la Figura 3.3. De esta manera, todas las acciones de alguna secuencia que pasen por sectores cercanos de la cancha estarán agrupadas en la misma zona. Esto nos permite estudiar si hay alguna preferencia particular a nivel espacial en cuanto a los sectores de juego de los distintos equipos. La elección de la división en esas 30 zonas se definió en función de poder asignar características espaciales a las secuencias de una manera más general, de modo de asignarles valor en términos futbolísticos.

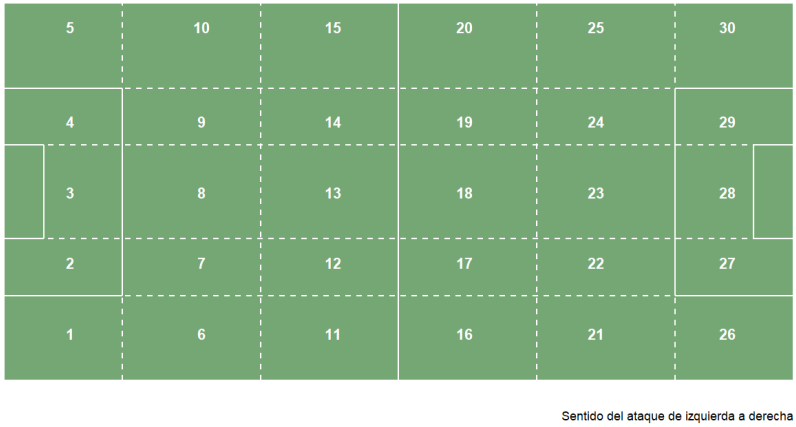


Figura 3.3: División de la cancha en 30 zonas.

3.3. Base de posesiones

Como el objeto de estudio del presente trabajo son las posesiones o secuencias de los equipos, se define la posesión de la pelota como todas aquellas acciones que se realizan una vez asegurado el control de la pelota por al menos 3 segundos. De hecho, una de las variables de la base de datos refiere a un identificador de cada una de estas posesiones que se dan a lo largo de los partidos, es decir, una vez que un equipo logra asegurar la posesión del balón, dicho contador se reinicia indicando una nueva secuencia. Sin embargo, si el equipo rival intercede en la disputa de la pelota pero no logra asegurar la posesión de la misma, puede darse que dentro de una misma secuencia se registren acciones con la pelota de ambos equipos. Dichas secuencias no serán incluidas en este análisis. Asimismo, el indicador de la jugada se reinicia cuando la pelota sale del terreno de juego o el partido se pausa por una falta, pese a que el balón le siga perteneciendo al mismo equipo previo a la interrupción. Por simplicidad nos quedaremos únicamente con aquellas posesiones en las cuales las acciones de las mismas son realizadas por el equipo poseedor del balón o, en su defecto, que las acciones

del rival no involucren directamente la pelota: acciones de presión, faltas cometidas o recibidas, atajadas del golero y/o jugadores *dribleados* del conjunto que está defendiendo. De esta manera, la trayectoria de la pelota se ve únicamente afectada por las acciones realizadas por el equipo que tiene la posesión y no por el rival. En resumen, trabajaremos con aquellas posesiones en las cuales solo se registran acciones del equipo poseedor de la jugada y cuya trayectoria puede realizarse de manera *aproximadamente* continua. Al aplicar este filtro, nos quedamos finalmente con 4961 posesiones del mundial masculino y 4409 del femenino, de las cuales estudiaremos la trayectoria de la pelota en esas secuencias (aproximadamente son 40 % de las secuencias totales). Esta decisión nos permite analizar secuencias y no eventos puntuales. En este trabajo, a partir de los datos de *eventing* construiremos una nueva base que contenga características de las 9370 secuencias a analizar de modo de agruparlas en función de éstas. Se entiende que son variables y aspectos relevantes para el análisis del juego y consideramos que deben ser relevantes para los entrenadores con el objetivo de definir un estilo de juego.

Dentro de una posesión se puede encontrar una gran variedad de acciones y eventos tanto del equipo que tiene la posesión como del rival. A modo de ejemplo, se puede ver en la Figura 3.4 el ejemplo de qué se considera una secuencia. Esta posesión es la más larga de las consideradas en el análisis, logrando el equipo de Argentina mantener la posesión durante más de 2 minutos y llegando a rematar al arco en el final de la jugada (sin terminar en gol).

Esta secuencia cuenta con 51 pases del equipo poseedor de la pelota y se registraron acciones en 20 zonas distintas del campo. Estas características y el resto definidas en el párrafo anterior están asociadas a dicha jugada en particular y es en base a éstas que se procede a agrupar las distintas posesiones. De modo de ilustrar el procedimiento, en la Figura 3.5 se puede observar cómo se construye cada posesión en función de la base de eventos. En la jugada del ejemplo, se registraron 3 pases y 2 traslados, a la vez que comenzó en las coordenadas (24,7; 51,4) y finalizó en (66,2; 44,1) como se ve en la primera y última fila de la Figura 3.5(1). El detalle de todas las variables creadas y su descripción de esta nueva base de datos que contiene las 9370 posesiones se puede observar en la Tabla A.1.

Asimismo, de modo de profundizar en el análisis, se busca estudiar el *éxito* de las secuencias. Para ello, según nuestro criterio, se definen como *exitosas* aquellas jugadas que cumplan alguna de las siguientes cinco condiciones:

(1) La posesión termina en un tiro al arco.

(2) La posesión termina en un córner a favor.

(3) La secuencia logra llegar hasta el área rival, iniciando desde, al menos, detrás del *último cuarto ofensivo*.

(4) El rival no logra recuperar la posesión y comete falta.

Figura 3.4: Secuencia de Argentina en el partido contra Polonia (122.91 segundos de duración).

(5) La posesión termina en un lateral a favor.

3.4. Análisis Exploratorio de Datos

Como se visualiza en la Figura 3.1, la mayor cantidad de pases y traslados observados a nivel masculino en contraposición a la mayor cantidad de acciones defensivas que dan a nivel femenino, parece indicar mayor duración en las posesiones durante Catar 2022 con la pelota en su poder. De hecho, el tiempo promedio de las posesiones para los varones es de 19,73 segundos, mientras que el de las mujeres ronda los 16,89. Los tiempos de cada jugada representan un aspecto fundamental en lo que refiere al dominio del partido y, por lo tanto, de imponer las ideas de juego de un equipo por sobre las de sus rivales. En la Figura 3.6 se observa la dispersión de dichos tiempos de posesión por partido y, si bien la distribución entre ambas competiciones es similar, en 2022 se registraron posesiones más largas en promedio. Precisamente, de modo de calificar a los equipos *dominantes* respecto de sus rivales, podemos utilizar la métrica del xG (ver Sección 3.2.1). Es posible afirmar que aquellos equipos que generaron mayor cantidad de goles esperados (situaciones de mayor riesgo), sean equipos

match_id	possession	minute	second	location.x	location.y	type.name	team.name	resultado
3893787	62	26	1	24.7	51.4	Pass	Norway	empate
3893787	62	26	2	25.7	62.4	Ball Receipt*	Norway	empate
3893787	62	26	2	25.7	62.4	Carry	Norway	empate
3893787	62	26	3	24.4	61.8	Pass	Norway	empate
3893787	62	26	4	35.0	64.0	Ball Receipt*	Norway	empate
3893787	62	26	4	35.0	64.0	Carry	Norway	empate
3893787	62	26	5	37.5	73.0	Pass	Norway	empate
3893787	62	26	6	66.2	44.1	Ball Receipt*	Norway	empate

(1) Detalle de la posesión (base de *eventing*)

match_id	possession	team.name	tiempo	resultado	n_pases	n_traslados	x_inicio	x_fin	y_inicio	y_fin
3893787	62	Norway	5.504	empate	3	2	24.7	66.2	51.4	44.1

(2) Resumen de la posesión (base de posesiones)

Figura 3.5: Ejemplo de una posesión resumida en una fila según sus características.

más dominantes. Para su análisis, normalizaremos a 90 minutos los xG y los xGA generados por cada equipo en función de los minutos disputados en su respectiva competición, sin tener en cuenta los tiros provenientes de penales (*npxG_90*). Además, si se compara el valor de esta métrica con los goles efectivamente realizados o recibidos, se observa la efectividad tanto ofensiva como defensiva de los distintos equipos.

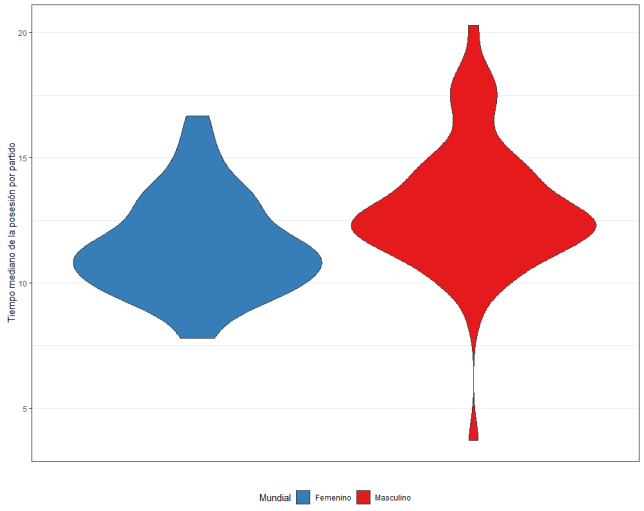


Figura 3.6: Distribución de tiempos de posesión de los equipos.

En la Figura 3.7 se observan los distintos equipos según competición or-

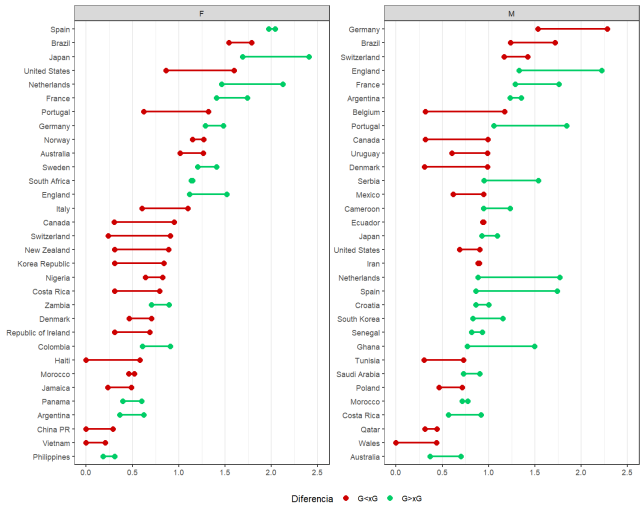


Figura 3.7: Diferencia entre xG y Goles por equipo normalizado a 90 minutos (excluyendo penales).

denados de forma decreciente según el *npxG_90* destacando selecciones como España a nivel femenino, Alemania a nivel masculino, así como Brasil en ambas competiciones. Esto es destacable, ya que tanto Brasil femenino como Alemania masculino quedaron eliminados en fase de grupos y, pese a ello, fueron los equipos de mayor generación de xG. El hecho de que hayan anotado menos goles de los que generaron ($G < xG$) contribuye a explicar sus posiciones finales en los torneos. En adición, el equipo femenino de Brasil, según la Figura A.2 es uno de los equipos a los cuales le generaron mayor xGA (además de la diferencia entre xGA y goles recibidos) lo que pauta su rendimiento defensivo. Para describir estilos de juego se puede analizar además el ángulo, largo y parte del cuerpo con la que fueron realizados los pases y comparar esto en función de la situación de gol que lograron generar (medido por el xG de la posesión). Por lo general, aquellos equipos que generan mayor xG (equipos *más dominantes*) son aquellos equipos que juegan mayoritariamente por bajo y en corto. Precisamente, las variables *pass.height.name* y *pass.length* nos permiten analizar a grandes rasgos el comportamiento en la posesión de la pelota de los equipos. En la Figura A.4 se puede ver cómo el equipo masculino de España es el que más pases por bajo por posesión realiza, teniendo en cuenta las dos competiciones, con un alto porcentaje de éxito en dichas acciones (95.5 % de efectividad) seguido por Inglaterra. A nivel femenino, si bien el margen es menor, destacan Alemania en primer lugar y luego España e Inglaterra nuevamente, lo que habla de una idea de juego entre las selecciones. Además, resulta relevante destacar el caso de Australia, Polonia y Corea del Sur a nivel masculino, y el caso de Filipinas, Países Bajos, Jamaica y Noruega a nivel femenino, como aquellas de las pocas selecciones que en promedio realizan más de un pase largo (*High Pass*) por cada

posesión en la que tienen la pelota. Además, a grandes rasgos, es posible afirmar que los equipos *más ofensivos* son aquellos equipos que en promedio realizan pases *más cortos*. A excepción de China y Argentina en 2023 y España en 2022, parece existir una predominancia de equipos del cuartil 4 entre los equipos con menor distancia y del primer cuartil entre aquellos equipos que juegan más con pases largos.

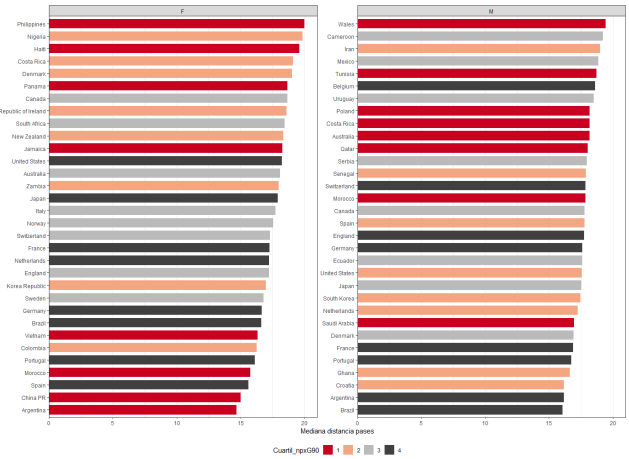


Figura 3.8: Mediana de la distancia de los pases por equipo según cuartil del *npxG90*.

En este análisis, se puede observar que equipos como España (a nivel masculino y femenino) y Bélgica masculino juegan secuencias de pases hacia alguno de los laterales mientras que otros como es el caso de Vietnam o Marruecos femenino, o como Australia o Túnez a nivel masculino la mayoría de sus pases son hacia adelante, y en largo como se ve en la Figura 3.8. Además de medir características de juego del equipo poseedor de la pelota, también podemos analizar la intensidad de la defensa y de la presión que ejerce desde la óptica de cuántos pases le permite realizar al equipo rival cuando no tiene la posesión. De hecho, esta métrica se conoce como PPDA (*Passes Per Defensive Action*) y se calcula mediante el cociente entre la cantidad de pases que realiza el equipo que tiene la pelota en los tres quintos más cercanos a su propio arco sobre la cantidad de acciones defensivas realizadas por el equipo en situación defensiva en ese mismo sector del campo (Trainor, 2014). Es decir, se utiliza como medida de la intensidad de la presión del equipo que no tiene la posesión en las zonas cercanas al arco rival. Es decir, cuanto menor sea su valor, mayor será el indicio de una alta presión dado que el conjunto en fase defensiva permite pocos pases del rival en función de sus acciones defensivas. Dado que la base de *eventing* no cuenta con el valor de esta métrica, se calcula teniendo en cuanto las acciones defensivas y pases de los equipos durante los partidos teniendo en cuenta la cantidad total de minutos disputados y el tiempo de posesión efectivo durante esos partidos. En

la Figura 3.9 se puede observar que, en términos generales, aquellos que suelen *dominar* los partidos desde la posesión, son aquellos que menos pases permiten hacer a su rival en su propio campo, es decir, de mayor intensidad de presión. Asimismo, se observa que la diferencia de en los tiempos de posesión entre las selecciones femeninas parece ser mayor que las masculinas.

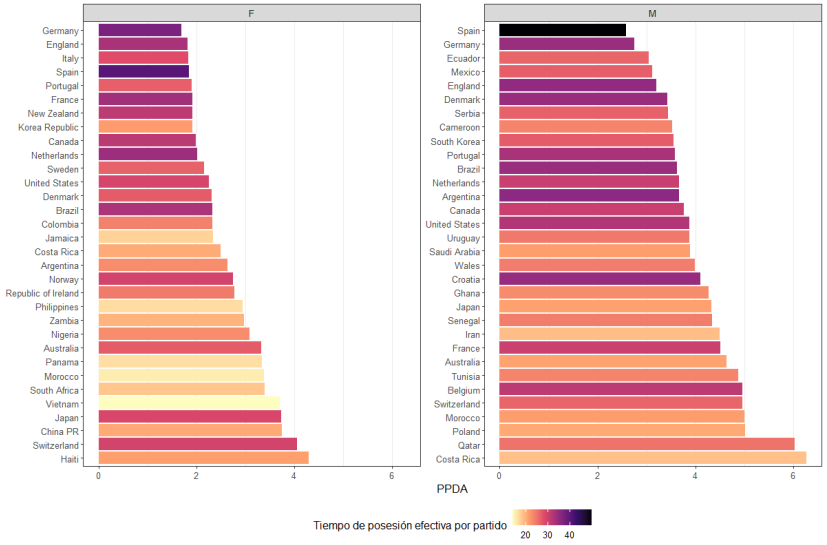


Figura 3.9: PPDA de cada equipo según tiempo efectivo total de posesión.

Para realizar el análisis de las posesiones se propone un análisis exploratorio de las secuencias y sus características. En ese sentido, en la Figura A.5 se observa que la distribución de pases y traslados por posesión parece ser similar. Separando por competición, se observa una mayor cantidad de estas acciones a nivel masculino respecto del femenino, así como una mayor duración de estas posesiones, tal como se mencionó anteriormente. Sin embargo, si estudiamos la cantidad de futbolistas involucrados en esas jugadas, se observan más jugadoras que participan en cada posesión.

Siguiendo con la descripción de las secuencias a analizar, podemos observar que a medida que la finalización de las mismas se aleja de la mitad de la cancha (en ambas direcciones), más se diferencian los valores de la verticalidad entre campo rival y campo propio. Se propone, tal como se observa en la Figura 3.10 podemos agrupar las 30 zonas (Figura 3.3) en pares según sean en campo propio o campo rival, representando la media $\pm$ 1 desvío estándar de las secuencias tomando como eje de simetría la mitad de la cancha. Esto es, a medida que las secuencias finalizan más cerca del arco rival, en promedio, recorrieron mayor distancia hacia adelante (valores altos de *vert_tot*). Lo inverso ocurre a medida que nos acercamos al arco propio. Resulta interesante destacar cómo, en

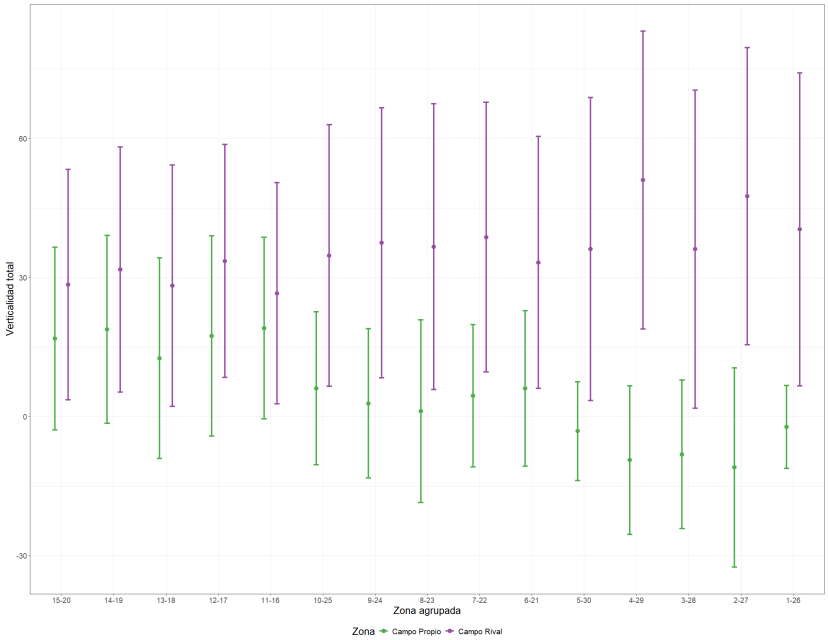


Figura 3.10: Comparación de la verticalidad total (*vert_tot*) según las zonas agrupadas de finalización (*zona_fin*) diferenciando campo propio y campo rival.

términos generales, a medida que las jugadas finalizan más cerca del arco rival, mayor verticalidad total. Estas jugadas que finalizan más próximas al área rival pueden darse habiendo iniciado tanto en campo propio como en campo rival. De hecho, si observamos los segmentos violetas se observa una mayor variabilidad en términos generales. Además, se observa que las secuencias con verticalidad total negativa, es decir, que arrancan más adelante en el campo respecto de dónde finalizan, se dan en zonas en el campo propio del equipo poseedor de la pelota.

A continuación se presenta el análisis y proceso realizado para construir los clústeres de las 9370 posesiones descritas en función de las variables creadas para la posterior caracterización de la partición resultante mediante *k-means*.

Capítulo 4

Resultados

4.1. Evaluación y elección de la partición

La base de posesiones de la Sección 3.3 se construyó según los aspectos que pueden aportar *valor* a los estilos de juego de los diferentes equipos. Sin embargo, tal como se puede ver en la Figura A.6 varias de estas variables están altamente correlacionadas, por lo que su inclusión podría afectar negativamente los resultados así como su interpretación. De modo de paliar este efecto, se propone realizar el análisis con **algunas** de las variables que presentan una alta correlación. De esta manera, en una primera instancia, se procede a realizar el análisis de clúster solo con algunas variables de la base. Como se puede observar en la Tabla A.5, se consideran 4 sets de variables distintos y se aplica el *k-means* para $k = \{2, 3, 4, 5, 6\}$. Si bien el criterio principal para la selección los sets de variables es el de la alta correlación, también se tiene en cuenta la importancia que puedan tener en explicar alguna de las facetas del juego, por lo que puede que, de todos modos, se tenga en cuenta algún par de variables correlacionadas de la Figura A.6.

En esta Sección se analizan los resultados hallados en la *clusterización* realizada mediante *k-means*, dividiendo el conjunto de variables en 4 sets tal como se presenta en la Tabla A.5. Asimismo, dado que los grupos resultantes con cada

	Set 1	**Set 2**	**Set 3**	**Set 4**
1	5174	787	592	552
2	2852	6526	2878	3810
3	1005	2057	251	250
4	339		2096	432
5			2550	3075
6			1003	1251

Tabla 4.1: Tamaño de cada clúster según partición *óptima* elegida.

uno de esos 4 análisis fueron realizados con variables distintas, se propone utilizar el ARI (ver 2.1.1) para definir la partición *óptima*. A esos efectos, se define la partición P_1^* (ecuación 2.2) como aquella obtenida a partir de los componentes resultantes del ACP calculado con la totalidad de las variables cuantitativas. Se realiza este Análisis Factorial (AF) con el fin utilizar el conjunto entero de características creadas sin tener los problemas de alta correlación mencionados, no para generar una reducción de dimensiones. Dicho de otro modo, se toma como referencia las particiones obtenidas a partir de los componentes calculados, para luego quedarnos con aquella/s partición/es más similares según el criterio del ARI, es decir, para aquellos sets de variables de mayor valor del índice presentado. Tal como se observa en la Figura A.7, los valores altos del índice ARI se dan entre la agrupación de los componentes del ACP (para todos los valores de k). Sin embargo, si tomamos como referencia alguna de las particiones del AF, vemos que aquellas de mayor ARI son el Set 1 para $k = 4$, el Set 2 para $k = 3$, el Set 3 para $k = 6$ y el Set 4 para $k = 6$. Estudiando la silueta promedio para los distintos valores de k en cada una de las particiones obtenidas para cada uno de los 4 sets de variables, se obtiene que los dos valores máximos de dicha silueta se dan para $k = 2$ y $k = 3$ en el Set 1 seguido por los valores del Set 2 tal como se observa en la Figura 4.1. Sin embargo, de modo de poder caracterizar con mayor precisión las secuencias se opta por una mayor cantidad de grupos ya que se entiende que hay gran variedad en las posesiones, así como en los estilos de juego, y es precisamente esto último lo que se busca capturar. En ese sentido, es el Set 4 el que presenta mayor silueta promedio. Además, comparando con los restantes sets de variables, parece ser el de menor variabilidad en lo que refiere a la silueta promedio de los grupos resultantes, para cada uno de los k propuestos. Asimismo, en la Figura 4.1 se observa que para $k = 6$ en el Set 4, en comparación con los restantes *óptimos* de cada Set, es el único que representa el valor de mayor silueta promedio de todos los k analizados. Se decide entonces seguir adelante con los 6 grupos derivados de la *clusterización* realizada con las 16 variables presentadas en la cuarta columna de la Tabla A.5.

A través de dicha partición, en este capítulo se busca encontrar dentro de los grupos resultantes, los patrones generales en los estilos de las posesiones de los equipos y, a grandes rasgos, definir estrategias de juego entre equipos.

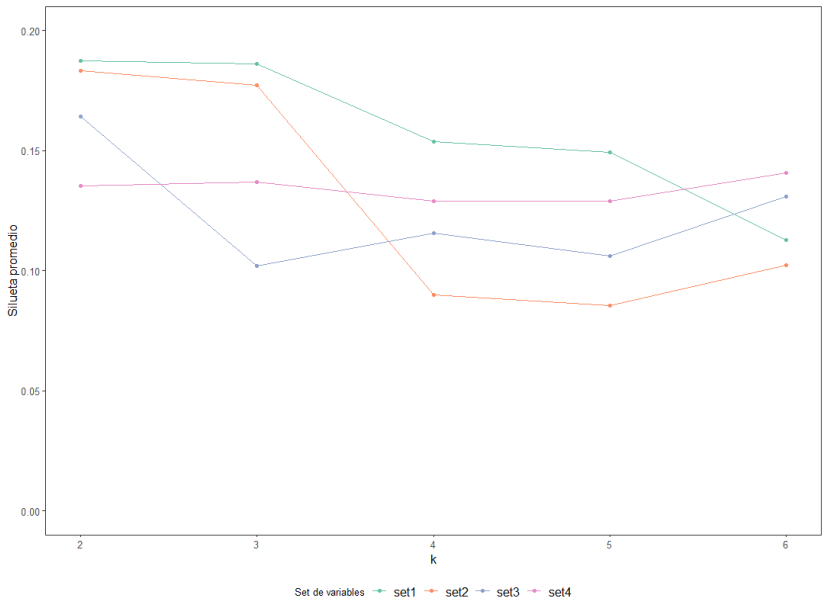


Figura 4.1: Silueta promedio para cada k según set de variables.

4.2. Caracterización de los grupos

En la Tabla 4.1, se observa el cardinal de cada uno de los clústeres calculados en esta instancia para el Set 4 y se obtiene que entre el clúster 2 y el clúster 5 acumulan casi el 74 % de las posesiones analizadas. En esa misma línea, se puede analizar la conformación de esos grupos diferenciando según las secuencias sean de equipos femeninos o masculinos. Concretamente, la proporción es similar en cada uno de los clústeres con valores que varían entre el 42 % y 49 % para las jugadas femeninas, y entre el 51 % y 58 % para las masculinas, siendo el grupo 6 el único con mayoría de jugadas femeninas, con un 51 % y 49 % respectivamente.

A los efectos de caracterizar la estructura de grupos resultante, se propone estudiar su composición desde el punto de vista de los equipos que los integran. En la Figura 4.2 se representa la proporción de jugadas de cada equipo (diferenciando según la competición femenina y masculina) que integran cada clúster, representados estos según el cuartil de *npxG90* generado (Sección 3.2.1). Asimismo, el segmento punteada representa la situación en la que las jugadas estén idénticamente distribuidas al interior de cada clúster. Es decir, que la proporción de cada selección sea igual $\frac{1}{64}$. Según el análisis gráfico, podemos notar, por un lado cómo la mayoría de secuencias del grupo 3 corresponden a equipos *"poderosos"*, tales como Croacia, Francia, Argentina (los 3 semifinalistas), Alemania, Portugal a nivel masculino y Estados Unidos, España y Australia a nivel femenino. De hecho, la mayoría de los equipos cuya proporción se encuen-

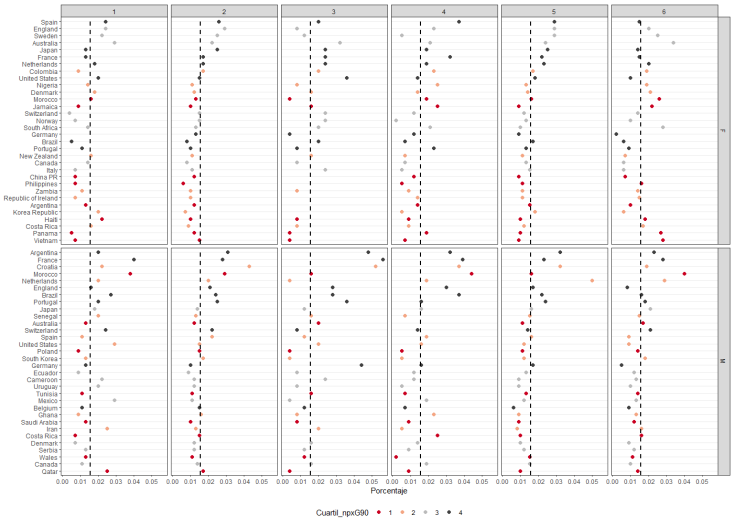


Figura 4.2: Proporción de jugadas de cada equipo por clúster según cuartiles de *npxG90*.

tra por encima de la mediana del clúster son los del último cuartil en cuanto a generación de xG por partido (excluyendo penales). De igual manera, es posible observar que la mayoría de secuencias pertenecientes a equipos femeninos del clúster 6 pertenecen a los de la mitad inferior en cuanto a generación de xG por partido (cuartiles 1 y 2). Además, observando las variables cuantitativas, podemos comparar sus medias estandarizadas para cada uno de esos grupos de modo de profundizar lo comentado anteriormente. Concretamente, si observamos nuevamente el grupo 3, se observa que las posesiones que lo componen representan mayoritariamente los tiros al arco (valores más altos de xG) e ingresos al área rival (mediante pases, conducciones o centros) lo cual sustenta lo

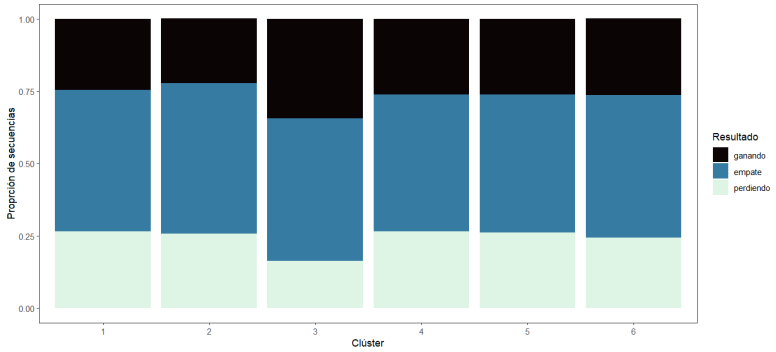


Figura 4.3: Proporción de secuencias de cada clúster según resultado.

comentado respecto a los equipos de mayor poderío ofensivo. Más en detalle, según la Figura 4.3 se observa que menos del 20 % de las jugadas de dicho grupo se dan cuando el equipo poseedor de la pelota se encuentra ganando el partido en ese momento. De hecho, estas posesiones son en promedio más largas, lo cual se puede interpretar que la manera de *sostener el resultado* por parte de estos equipos es, no solo mantener la posesión, sino que atacando el arco rival.

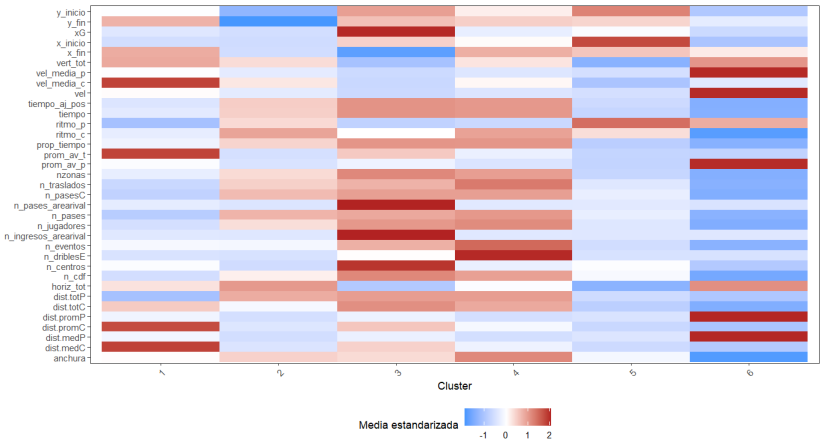


Figura 4.4: Medias estandarizadas por clúster.

De modo de profundizar en la caracterización se calcula la media estandarizada de cada una de las variables según cada clúster, de modo de poder comparar entre características con distinta unidad de medida. Esto se ve en el *heatmap* de la Figura 4.4. Se puede destacar que los promedios más elevados de las variables asociadas a los traslados de pelota (*type.event='Carry'*) se dan en el primer grupo. Más precisamente, la distancia promedio de avance de dichos traslados en el clúster 1 es 7 veces más grande que el promedio de dicha variable si se mira la totalidad de las 9370 posesiones analizadas. Algo similar sucede con la distancia promedio y mediana de los traslados de esas jugadas. Por otra parte, en dicho grupo los pases no presentan una particularidad, ya sea por muy largos o muy cortos. Las posesiones del clúster 6 representan en gran medida aquellas iniciadas por el golero, en general a través de saques de arco, en particular con saques largos, lo que se visualiza a través de una mayor distancia en los pases y velocidad de dichas secuencias: dicho de otro modo, pases que recorren una mayor distancia del terreno de juego pero en un menor tiempo. Esto se observa en la Figura 4.5. En esa misma línea, se observan posesiones más verticales (*vert_tot*) y menor tiempo promedio así como menores valores en la variable *anchura* la cual se define como la diferencia entre el máximo y mínimo valor registrado en el eje y, es decir en el ancho del terreno. Este aspecto se traduce en que estas secuencias no presentan grandes variaciones en el ancho de la cancha dada su juego más directo. A grandes rasgos, se puede definir este

grupo como el opuesto al clúster 1. Por otra parte, las secuencias del clúster 5 representan la mayoría de las jugadas de pelota quieta en campo rival como es el caso, por ejemplo, de los tiros de esquina. Esto se visualiza a través de valores mayores, en promedio, en lo que refiere a la ubicación espacial del comienzo y finalización de la jugada (coordenada x del inicio y fin). Por lo tanto, al arrancar muchas de estas jugadas cerca del arco rival (en posición ofensiva), se pueden observar menores valores promedio de verticalidad. Asimismo, el clúster 4 se compone de jugadas de las cuales en una gran proporción se observan regates (*dribles*) exitosos lo que habla de una cierta complejidad de esas secuencias. En esa misma línea, en ese clúster se observan jugadas muy variadas, es decir, que presentan una importante cantidad de eventos distintos (variable *n_eventos*). Finalmente, sobre el clúster 2 y 4 destacan, por ejemplo, las jugadas más bien horizontales, dado el número elevado de pases llevando la pelota de un costado a otro (terminando principalmente en el sector derecho del ataque, dado los valores bajos de *y_fin*).

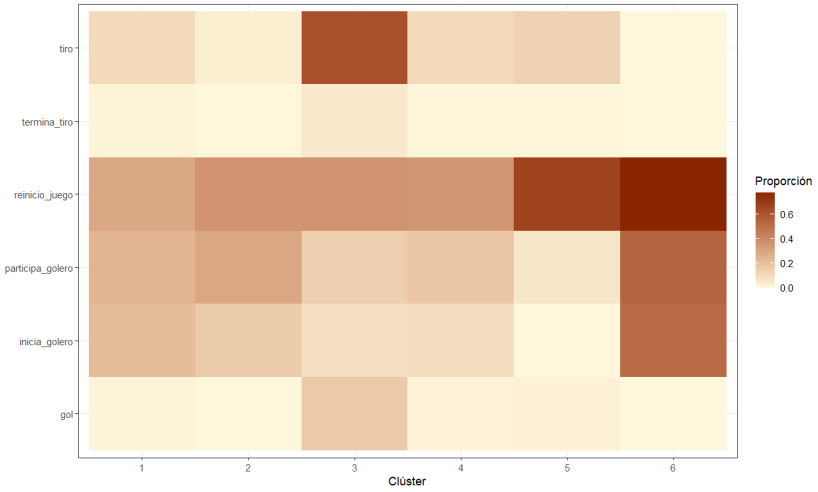


Figura 4.5: Proporción de jugadas por clúster según las variables binarias.

Por otra parte, si se observan en la Figura 4.5 las variables cualitativas creadas, podemos obtener la proporción que éstas representan en cada uno de los grupos hallados. Más precisamente, en concordancia con lo descrito anteriormente, el grupo 3 es el que presenta mayor cantidad de jugadas en las que se registra un remate al arco, mientras que en el 5 y 6 se ve gran cantidad de secuencias que corresponden a un reinicio de juego, jugadas de pelota quieta y saques de arco, respectivamente. De hecho, en la mayoría de las secuencias del grupo 6, las inicia (o participa) el arquero del equipo poseedor de la pelota. En cuanto a la visualización de las secuencias tal como la de la Figura 3.4, pero diferenciando según el clúster al que pertenecen, se pueden observar algunos de

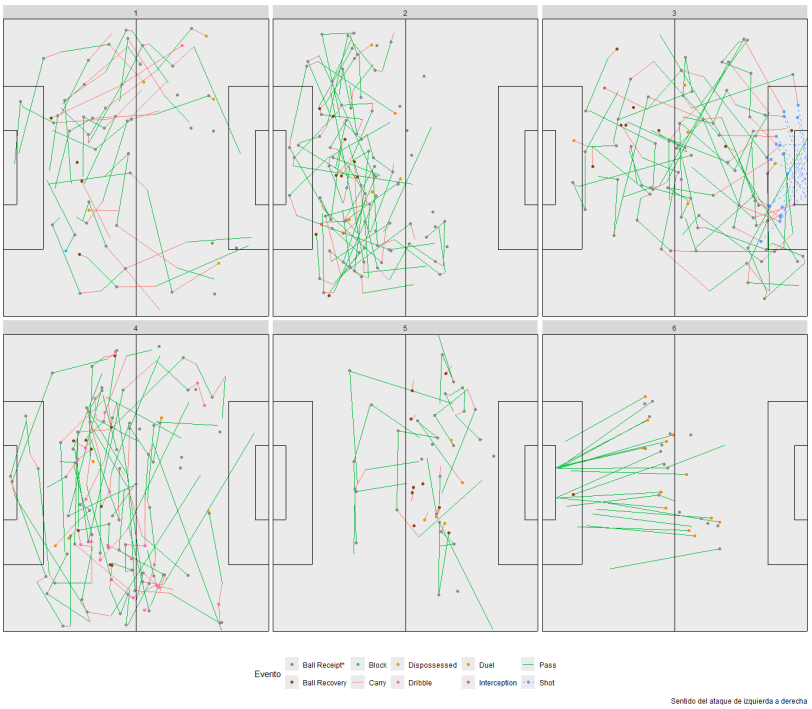


Figura 4.6: Las 20 jugadas *más representativas* de cada clúster.

los patrones ya mencionados anteriormente. Una manera de resumir las jugadas pertenecientes a cada grupo puede ser estudiando las jugadas representativas de cada clúster calculando la distancia euclidiana de cada secuencia a las restantes pertenecientes a su mismo grupo. Luego, se suman las distancias calculadas y se ordenan de menor a mayor. Se definen entonces las jugadas *más representativas* de cada clúster como aquellas que minimizan esta suma de distancias. Dicho de otra manera, estas jugadas corresponden a los elementos *más centrales* del clúster en términos de distancia, y resultan útiles para representar dichos patrones de secuencias en cada uno de los grupos calculados. En este caso, se toman las 20 jugadas que más caracterizan los clústeres anteriormente descritos. En la Figura 4.6 se pueden vislumbrar de mejor manera aspectos ya comentados previamente tales como jugadas más bien por los laterales y con una importante cantidad de traslados (y de larga distancia) en el clúster 1, jugadas que comienzan y por lo general terminan también en el campo propio del equipo poseedor, jugadas con una rebuscada elaboración que finalizan en el área y con remates al arco en el clúster 3, y secuencias en largo, desde el arco propio y por el carril central caracterizando el clúster 6.

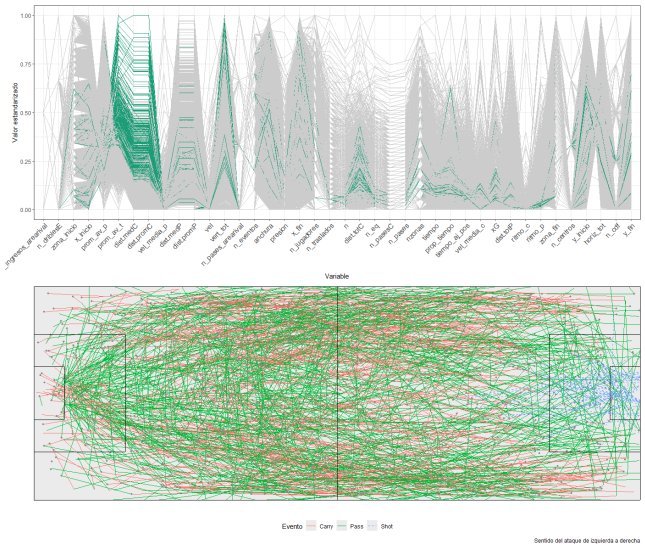


Figura 4.7: Secuencias del clúster 1.

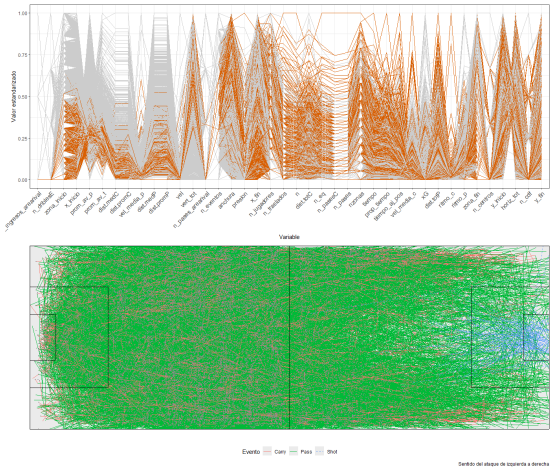


Figura 4.8: Secuencias del clúster 2.

De modo de profundizar en la caracterización de los grupos, nos resulta de interés ver todas las secuencias de manera conjunta. A estos efectos, por un lado, el *parallel plot* es una alternativa interesante para observar cada jugada según sus valores en cada una de las variables estandarizadas de tal manera que el mínimo y el máximo de cada una sean 0 y 1 respectivamente (ver 2.2.3). Concretamente, en dicha visualización se comparan los valores que toma cada una de las secuencias de cada clúster para todas sus características con el fin

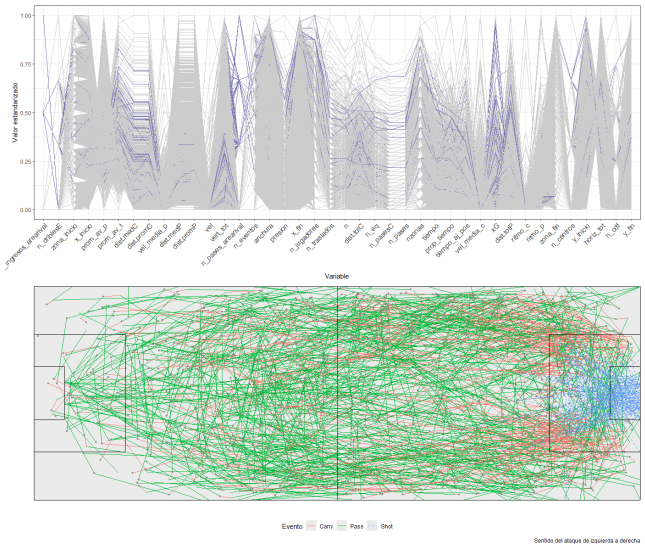


Figura 4.9: Secuencias del clúster 3.

de comparar con la representación de dichas jugadas en el campo de juego, graficando únicamente los pases, traslados y tiros al arco de modo de facilitar su lectura.

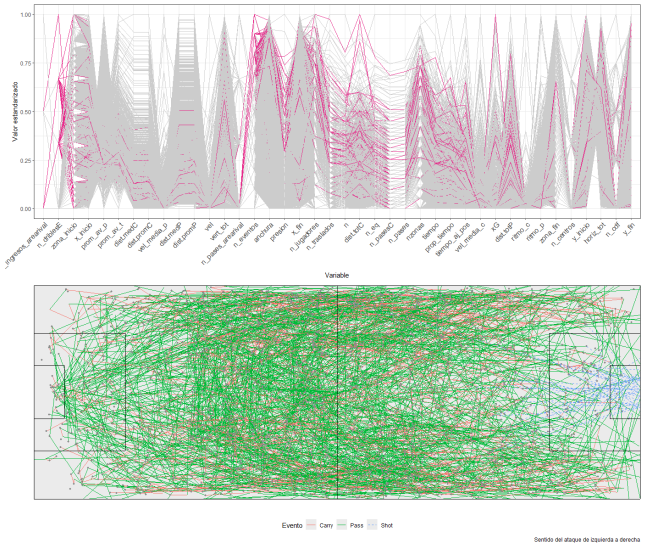


Figura 4.10: Secuencias del clúster 4.

En la Figura 4.7, en primera instancia, es posible reforzar lo comentado anteriormente, sobre la jugadas del clúster 1 caracterizadas por valores altos en lo que refiere al avance promedio por los traslados (*prom_av_t*), así como distancia media y mediana de esos traslados (*dist.promC* y *dist.medC*). Se observan también algunas jugadas muy verticales.

Luego, en cuanto a los sectores de la cancha dónde se desarrollan estas 552 posesiones, se puede observar que la predominancia de traslados, representados por los segmentos naranjas, se dan mayoritariamente por los laterales de la cancha. Esto es, por lo general, en el centro del campo suele haber mayor densidad de jugadores lo que dificulta trasladar la pelota, sobre todo de larga distancia. De hecho, alrededor del 46 % de las secuencias finalizan en alguna de las zonas laterales representadas en la Figura 3.3.

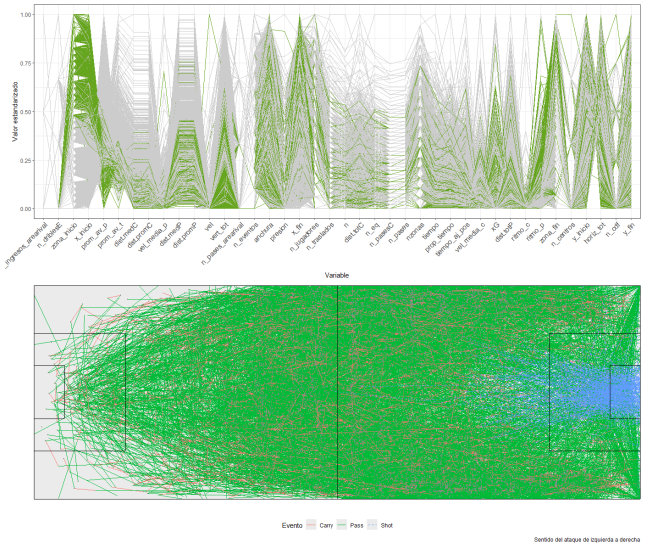


Figura 4.11: Secuencias del clúster 5.

Por otra parte, las secuencias del clúster 2, la mayoría suelen iniciar en campo propio en contrapartida con aquellas del clúster 5, las que suelen iniciar mayoritariamente en campo rival (Figura A.10). Las secuencias del clúster 3, dado que la gran mayoría finalizan con un remate al arco, finalizan en campo rival, pero más precisamente en las zonas que refieren al área del arco del equipo rival. De hecho, en el PCP se observa como estas secuencias son las de valores más elevados en lo que refiere a situaciones dentro del área, ya sea por ingresos mediante traslados o pases (*n_ingresos_arearival*, *n_pases_arearival* y *n_centros*). Además, notoriamente, sucede lo mismo con los valores de goles esperados que se generaron en estas posesiones. Asimismo, el clúster 3 y 4 contienen una elevada proporción de jugadas elaboradas al utilizar diversas zonas del campo en comparación con el resto de los grupos (Figura A.9). En esa misma línea, las secuencias del grupo 4 presentan valores elevados en lo que refiere a la variedad

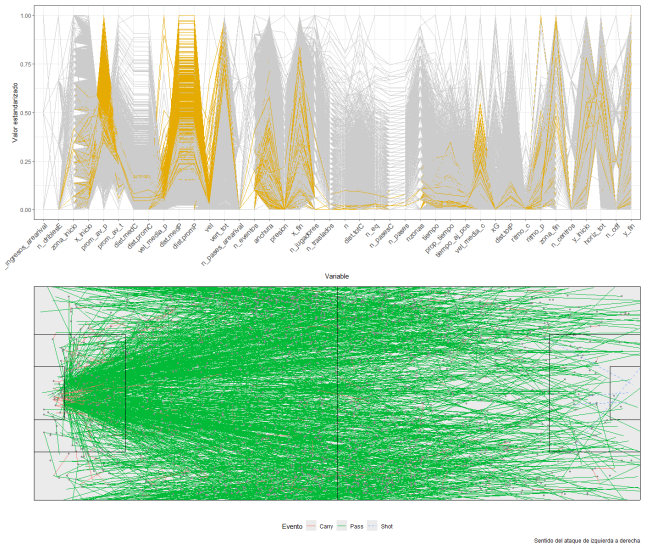


Figura 4.12: Secuencias del clúster 6.

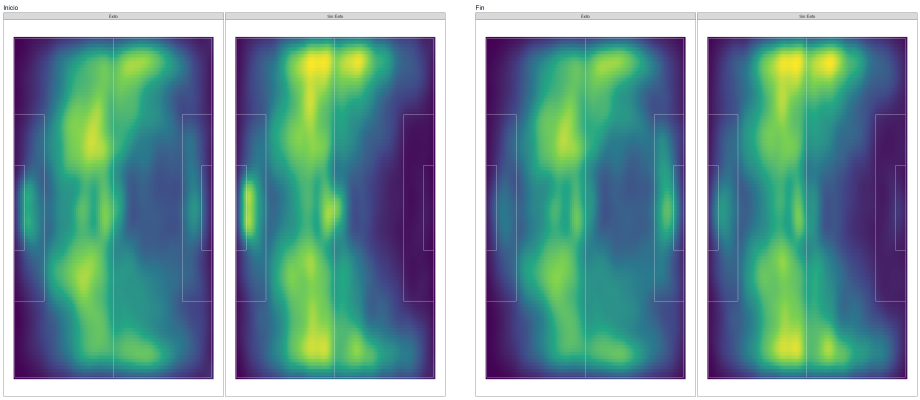
de eventos que se dan en la jugada (*n_eventos*). Por ejemplo, en este clúster se encuentra la posesión con mayor cantidad de regates exitosos (*n_driblesE*) así como de distancia total recorrida mediante traslados de pelota (*dist.totC*).

Complementando el análisis del clúster 5, se puede ver que la mayoría de secuencias inician en campo rival, pero más precisamente, ese en este clúster que lo que refiere a acciones a balón parado como tiros de esquina están en este grupo tal como se observa en la Figura 4.11. De hecho, los segmentos saliendo de las esquinas, la mayoría hacia el área de manera directa, representan este tipo de jugadas. A diferencia del grupo 5, en lo que refiere a zonas de inicio, el clúster 6 engloba aquellas jugadas iniciadas en campo propio pero más precisamente aquellas que surgen de saques de arco y de laterales en zona defensiva. En ese sentido, también representan secuencias de *reinicio de juego*. Más precisamente, en lo que refiere a los saques de arco, se ve una predominancia de pases más largos y hacia los costados (Figura 4.12) lo que deriva en eventos de tipo *duelos*, por lo que mayor riesgo de perder la posesión, y, por ende, menor proporción de secuencias en la que se registran tiros al arco.

Por otra parte, con el objetivo de resumir cada secuencia, puede resultar de interés estudiar el éxito de las secuencias según las condiciones descritas en la Sección 3.3. Podemos analizar la densidad de cada una de las acciones que componen cada secuencia, según haya finalizado de manera *exitosa* o no. Previo a diferenciar el análisis por cada uno de los clústeres realizados, podemos estudiar esta densidad para el conjunto de las 9370 secuencias tal como se observa en la Figura 4.13. En términos generales, podemos ver cómo hay una densidad im-

portante en las jugadas *exitosas* que finalizan en el área rival (posiblemente, en remates al arco) o en las zonas cercanas a la misma. Asimismo, cabe remarcar una mayor densidad de las jugadas *no exitosas* que arrancan en el arco propio (clúster 6) y en los sectores laterales del campo, sobre todo del sector izquierdo tomando como referencia el sentido del ataque.

En primera instancia, cabe destacar que en el grupo 3, dada la presencia en gran medida de remates al arco, la mayoría de esas jugadas se definen como *exitosas*. Más precisamente, de las 250 jugadas que lo componen, 94 terminaron de manera *exitosa* para el equipo poseedor de la pelota, ya sea en remates al arco o en ingresos al área (77 y 17 jugadas, respectivamente). El resto, si bien no fueron exitosas, alcanzaron a llegar al área ya sea en conducción o mediante pases. De hecho, en la Figura A.14 se observa con mayor detalle dónde inician y culminan estas jugadas *exitosas*. Asimismo, se puede ver, dentro del total de secuencias *no exitosas* de ese grupo, como la mayoría culminan en las inmediaciones del arco rival. Luego, para los clústeres restantes, la mayoría de las jugadas son *no exitosas* lo que es consistente con que aproximadamente el 28 % de las 9370 secuencias analizadas son *exitosas*. Sin embargo, cabe destacar que en el primer grupo se reparten prácticamente a partes iguales las jugadas finalizadas de manera exitosa de las que no, lo cual tiene sentido dada las características de esas secuencias: compuestas de conducciones de pelota (y de larga distancia) lo que permite asegurar el control de la misma y, por consiguiente, mayor dificultad para el rival de recuperarla (Figura A.12). A la inversa, el clúster 6 es el que registra menor proporción de jugadas *exitosas* lo cual es consistente con lo descrito anteriormente: secuencias con pases en largo (sobretodo de saque de arco) lo que dificulta el control de la posesión facilitando así la recuperación del rival. En la Figura A.17 se pueden observar estas jugadas según su ubicación de



(1) Inicio de secuencias *exitosas* y *no exitosas*.

(2) Finalización de secuencias *exitosas* y *no exitosas*.

Figura 4.13: Densidad de las 9370 secuencias según su *éxito*.

inicio y finalización. Precisamente, si observamos las 20 jugadas características de cada grupo, a grandes rasgos se mantiene la proporción de éxito en cada clúster: mayoría de jugadas *sin éxito* en los grupos 2, 4, 5 y 6, todas jugadas exitosas en el grupo 3, mientras que en el clúster 1 se reparte a partes iguales las *exitosas* y las que no.

En términos generales, podemos observar dónde empiezan y dónde culminan las jugadas según hayan sido exitosas o no, diferenciando según el grupo. Cabe aclarar, como se mencionó anteriormente, que la densidad de las jugadas *sin éxito* del clúster 3 ya que la cantidad de secuencias no es representativa. Además, la densidad de las Figuras A.121,A.131,A.141,A.151,A.161,A.171 fue calculada únicamente con las jugadas del clúster correspondiente, por lo que la comparación entre grupos no tiene sentido.

4.3. Jugadas consecutivas

Cabe recordar que al filtrar algunas de las posesiones de la base de datos (ver 3.3), no todas las jugadas analizadas se dan de manera consecutiva en el transcurso del partido. Precisamente, podría resultar de interés estudiar las

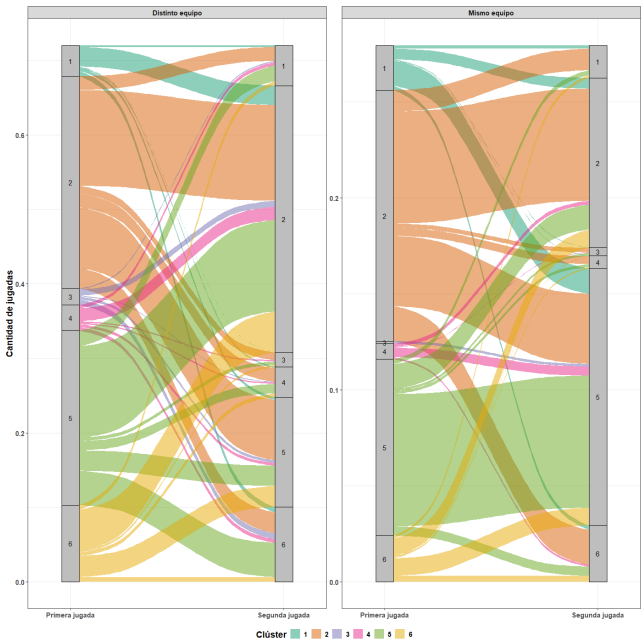
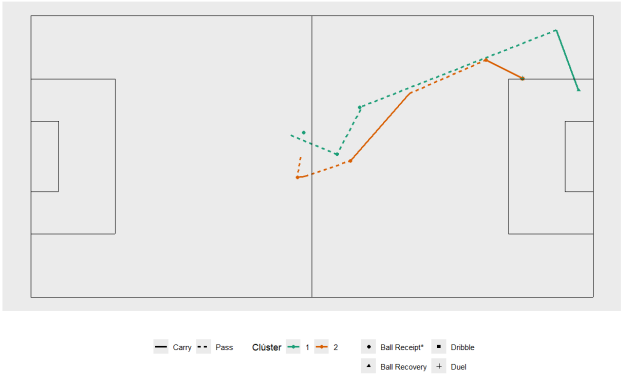
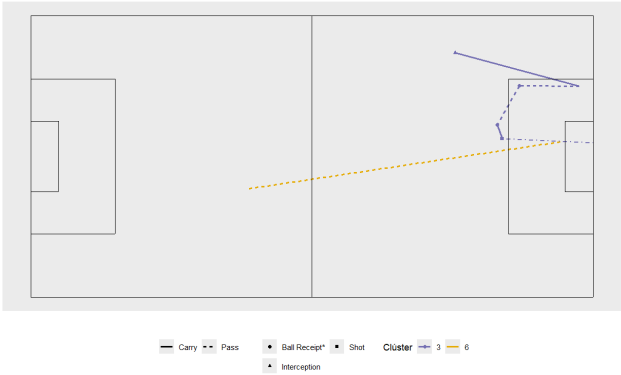


Figura 4.14: Proporción de pares de jugadas consecutivas según clúster ya sea si la posesión corresponde o no al mismo equipo.

jugadas consecutivas en el sentido de si existe alguna relación entre los clústeres formados y lo que sucede en estas jugadas que ocurren seguidas en el partido. En las posesiones analizadas, se cuentan con 6066 jugadas (de las 9370) que forman parte de alguna secuencia de jugadas consecutivas, ya sea de largo 2 o más. Más precisamente, se cuenta con 3938 secuencias compuestas por pares de jugadas consecutivas (según el id de posesión de cada partido), las cuales podemos dividir en 2 grupos: si la posesión corresponde al mismo equipo (2837 pares) o si ésta se alterna (1101 pares). En primer lugar, cabe mencionar que el hecho de haber notoriamente menos secuencias consecutivas que correspondan al mismo equipo puede deberse al filtro realizado en el momento del análisis de las posesiones. Luego, podemos observar en la Figura 4.14 que cuando la primera jugada pertenece al clúster 3, y cuando pertenecen a distintos equipos, la segunda jugada se divide mayoritariamente entre el grupo 2 y 6.



(1) Secuencias consecutivas pertenecientes al clúster 2 (España) y al clúster 1 (Alemania).

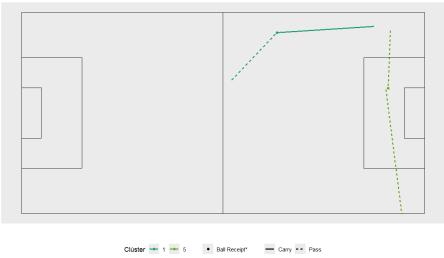


(2) Secuencias consecutivas pertenecientes al clúster 3 (Portugal) y al clúster 1 (Uruguay).

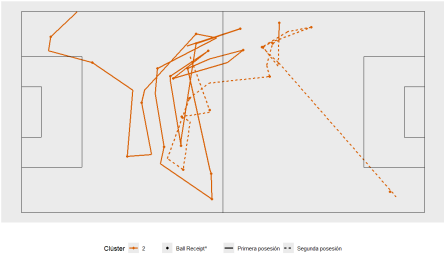
Figura 4.15: Ejemplos de secuencias consecutivas pertenecientes a distintos equipos.

Esto tiene *sentido futbolístico* por el hecho de que las jugadas del clúster 2 inician mayoritariamente en campo propio, mientras que las del 6 pueden estar haciendo referencia en gran medida a saques de arco. Es decir, esto último puede decirse que se dio a raíz de que el remate derivó en un saque de arco para el equipo rival.

Asimismo, se puede observar la trayectoria de estas secuencias de modo de observar los sectores en la cancha y cómo los clústeres de dichas posesiones varían, lo cual refuerza la caracterización anteriormente descrita. En la Figura 4.15 se observan jugadas cuya posesión no corresponde al mismo equipo ni al mismo grupo. En la Figura 4.151 se puede ver cómo la jugada de España comienza en mitad de cancha mientras que Alemania logra recuperar la pelota (luego de un regate sin éxito del jugador español) a la entrada de su propia área, y logrando salir por la zona lateral derecha tal como se describió en las secuencias del clúster 1.



(1) Secuencias consecutivas del clúster 1 y 5 pertenecientes a Brasil.



(2) Secuencias consecutivas pertenecientes al clúster 2 y a Inglaterra.

Figura 4.16: Ejemplos de secuencias consecutivas pertenecientes al mismo equipo.

A diferencia de esta, la Figura 4.152 nos muestra cómo la jugada del equipo portugués finaliza con un remate al arco, y agrupada en el clúster 3, mientras que el saque de arco en largo que se da a continuación pertenece al clúster 6. Luego en la Figura 4.16 se observan pares de posesiones consecutivas correspondientes

al mismo equipo: mientras que las secuencias de la Figura 4.161 corresponden a grupos distintos (la secuencia del clúster 5 se da a raíz de un tiro libre a favor), las de la de la Figura 4.162 corresponden a dos jugadas elaboradas del grupo 2. Cabe mencionar que en la Figura 4.162 se representan 2 secuencias en el mismo gráfico diferenciando ambas posesiones según si el trazo de la secuencia es continua (posesión 116 del partido) o punteado (posesión 117 del partido). De hecho, la primera empieza en campo propio y cuenta con 17 pases y 15 traslados mientras que la segunda comienza en campo propio terminándose de desarrollar en campo rival. Asimismo, cuenta con 11 pases y 9 traslados. Pese a contar con menos pases en esta segunda jugada, el avance promedio por cada uno de esos pases fue mayor al de la posesión anterior (5 veces más alto).

El estudio de las jugadas consecutivas representa un aspecto fundamental en la dinámica de juego de los partidos así como de estilo de juego de los equipos. Además, la alternancia (o no) de los clústeres en jugadas consecutivas pueden indicar distintos aspectos del juego o las ideas de los entrenadores.

Capítulo 5

Conclusiones

El presente análisis permitió identificar y caracterizar patrones comunes en las posesiones de los equipos participantes en las Copas del Mundo de la FIFA 2022 (masculina) y 2023 (femenina), a partir de la base de *eventing* de *StatsBomb*. Para ello, se reconstruyó y resumió la información de eventos en una matriz que describe, para cada posesión, características relevantes relacionadas con la dinámica y el desarrollo de la secuencia. Mediante la aplicación de técnicas de Aprendizaje Estadístico No Supervisado, en particular el algoritmo *k-means*, se obtuvieron seis clústeres representativos de distintos tipos de secuencias. Estos incluyen: posesiones seguras basadas en traslados, secuencias con alta cantidad de pases tanto en campo propio como en campo rival, jugadas elaboradas que alcanzan el área y el arco rival, jugadas de pelota quieta ofensiva y secuencias caracterizadas por envíos largos desde el propio campo. Asimismo, cada clúster fue caracterizado según el sector de la cancha en que inicia y finaliza, y se evaluó su nivel de *éxito* a través de métricas y visualizaciones creadas especialmente a esos efectos.

Como extensión futura, precisamente, resulta de interés profundizar el análisis del *éxito* descrito en la Sección 3.3 de las secuencias mediante técnicas de Aprendizaje Supervisado, con el objetivo de predecir la probabilidad de éxito en función de las características de la posesión y de factores contextuales (por ejemplo, el marcador parcial, el género o el rival). De igual manera, puede resultar una alternativa interesante seguir profundizando el estudio de las jugadas que se dan de forma consecutiva, tal como se introdujo en la Sección 4.3, extendiendo el largo y no quedándonos únicamente con aquellas de largo 2. Además, se sugiere considerar aspectos adicionales que podrían influir en el desarrollo de las posesiones, como la existencia de expulsiones, lesiones u otras situaciones que condicionen la dinámica (por ejemplo, inferioridad numérica). Complementar este estudio con datos de tipo *tracking* permitiría un análisis aún más detallado del posicionamiento y los movimientos de los jugadores. Otra línea interesante sería extender el análisis a la totalidad de secuencias de ambos mundiales, no centrarse únicamente en aquellas donde solo el equipo que tiene la posesión

genera acciones sobre la pelota. Finalmente, se propone aplicar la metodología desarrollada a nuevas competiciones, como la Copa América masculina 2024 o las Eurocopas 2024 y 2025, masculina y femenina respectivamente, con el fin de validar y comparar los patrones encontrados en contextos competitivos diferentes.

En conjunto, los resultados de este estudio ofrecen un aporte relevante para el entendimiento cuantitativo de las secuencias de juego en el fútbol de élite y abren múltiples caminos para investigaciones futuras que profundicen en la relación entre dinámica de posesión, contexto y rendimiento.

Referencias

Aggarwal, C. C., y Reddy, C. K. (2014). *Data clustering: Algorithm and applications*. Chapman&Hall/CRC.

Akhanli, S. E., y Hennig, C. (2023). Clustering of football players based on performance data and aggregated clustering validity indexes. *Journal of Quantitative Analysis in Sports*, *19*(2), 103–123. doi:10.1515/jqas-2022-0037

Beernaerts, J., De Baets, B., Lenoir, M., y Van de Weghe, N. (2022). Qualitative team formation analysis in football: A case study of the 2018 fifa world cup. *Frontiers in Psychology*, *13*, 863216. doi:10.3389/fpsyg.2022.863216

Bekkers, J., y Dabadghao, S. (2019). Flow motifs in soccer: What can passing behavior tell us? *Journal of Sports Analytics*, *16*(5), 299–311. doi:10.3233/JSA-190290

Benítez, I., Quijano, A., Díez, J.-L., y Delgado, I. (2014). Dynamic clustering segmentation applied to load profiles of energy consumption from spanish customers. *International Journal of Electrical Power & Energy Systems*, *55*, 437–448. doi:10.1016/j.ijepes.2013.09.022

Cao, S. (2024). Passing path predicts shooting outcome in football. *Scientific reports*, *14*(1), 9572. doi:10.1038/s41598-024-60183-7

Fujii, K. (2021). Data-driven analysis for understanding team sports behaviors. *Journal of Robotics and Mechatronics*, *33*(3), 505—514. doi:10.48550/arXiv.2102.07545

Garrido, D., Burriel, B., Resta, R., del Campo, R. L., y Buldú, J. M. (2022). Heatmaps in soccer: Event vs tracking datasets. *Chaos, Solitons & Fractals*, *165*(18), 112827. doi:10.48550/arXiv.2106.04558

Ge, Y., y Hofmann, H. (2020). A grammar of graphics framework for generalized parallel coordinate plots. *arXiv preprint arXiv:2009.12933*. doi:10.48550/arXiv.2009.12933

Gilovich, T., Vallone, R., y Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295-314. doi:10.1016/0010-0285(85)90010-6

Hastie, T., Tibshirani, R., y Friedman, J. (2001). *The elements of statistical learning*. Springer.

Håland, E. M., Salte, A. W., Hvattum, L. M., y Stålhane, M. (2020). Evaluating the effectiveness of different network flow motifs in association

football. *Journal of Quantitative Analysis in Sports*, *16*(4), 311–323. doi:10.1515/jqas-2019-0097

Jang, C., Yoon, T., y Cho, H.-G. (2010). Digital photo classification methodology for groups of photographers. *Multimedia Tools and Applications*, *50*, 441–463. doi:10.1007/s11042-010-0485-3

Li, W., Cook, D., Tanaka, E., y VanderPlas, S. (2024). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *Journal of Computational and Graphical Statistics*, *33*(4), 1497–1511. doi:10.1080/10618600.2024.2344612

Llana, S., Madrero, P., y Fernández, J. (2020). The right place at the right time: Advanced off-ball metrics for exploiting an opponent's spatial weaknesses in soccer. En *Proceedings of the 14th MIT Sloan Sports Analytics Conference*.

Lucey, P., Bialkowski, A., Carr, P., Yue, Y., y Matthews, I. (2014). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. *MIT Sloan Sports Analytics Conference*.

Martínez Arastey, G. (2019). *Historia del análisis del rendimiento deportivo: el controvertido pionero Charles Reep.* Descargado de https://www.sportperformanceanalysis.com/article/history-of-performance-analysis-the-controversial-pioneer-charles-reep (Accedido el 27 de junio de 2025)

Morgans, R., Ju, W., Radnor, J., Zmijewski, P., Ryan, B., Haslam, C., ... Oliveira, R. (2025). The positional demands of explosive actions in elite soccer: Comparison of english premier league and french ligue 1. *Biology of Sport*, *42*(1), 81–87. doi:10.5114/biolsport.2025.139083

Mortensen, J., y Bornn, L. (2019). From markov models to poisson point processes: modeling movement in the nba. En *Proceedings of the 13th MIT Sloan Sports Analytics Conference* (Vol. 10).

Otero-Saborido, F. M., Aguado-Méndez, R. D., Torreblanca-Martínez, V. M., y González-Jurado, J. A. (2021). Technical-tactical performance from data providers: A systematic review in regular football leagues. *Sustainability*, *13*(18), 10167. doi:10.3390/su131810167

Parmar, N., James, N., Hearne, G., y Jones, B. (2018). Using principal component analysis to develop performance indicators in professional rugby league. *International Journal of Performance Analysis in Sport*, *18*(6), 938–949. doi:10.1080/24748668.2018.1528525

Plakias, S., Moustakidis, E., Mitrotasios, M., Kokkotis, C., Tsatalas, T., Papalexi, M., ... Tsaopoulos, D. (2023). A multivariate and cluster analysis of diverse playing styles across european football leagues. *Journal of Physical Education and Sport*, *23*(7), 1631–1641. doi:10.7752/jpes.2023.07200

Platt, J. (2000). Autoalbum: Clustering digital photographs using probabilistic model merging. En *2000 Proceedings Workshop on Content-based Access of Image and Video Libraries* (pp. 96–100). doi:10.1109/IVL.2000.853847

Pollard, R. (2002). Charles reep (1904-2002): pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, *20*(10), 853–855. doi:10.1080/026404102320675684

Rico-González, M., Pino-Ortega, J., Méndez, A., Clemente, F., y Baca, A. (2023). Machine learning application in soccer: a systematic review. *Biology of sport*, *40*(1), 249–263. doi:10.5114/biolsport.2023.112970

Soroka, A., Duda, H., Stuła, A., Ambroży, T., Kromke, C., y Te Poel, H.-D. (2023). Ball possession as an indicator identifying differences in the efficient operation of football teams during the World Cup-Qatar 2022. *Journal of Kinesiology and Exercise Sciences*, *33*(102), 9–20. doi:10.5604/01.3001.0053.5968

StatsBomb, H. (2022). *Statsbomb open events structure and data specification.* Descargado de https://github.com/statsbomb/open-data/blob/master/doc/Open%20Data%20Events%20v4.0.0.pdf (Accedido el 01 de julio de 2025)

Trainor, C. (2014). *Defensive activity and passes allowed per defensive action (ppda).* Descargado de https://statsbomb.com/articles/soccer/defensive-activity-and-passes-allowed-per-defensive-action-ppda/ (Accedido el 25 de junio de 2025)

Wickham, H. (2010). A layered grammar of graphics. *Journal of computational and graphical statistics*, *19*(1), 3–28. doi:10.1198/jcgs.2009.07098

Yu, Y. Y., Wu, P. P.-Y., Mengersen, K., y Hobbs, W. (2022). Classifying ball trajectories in invasion sports using dynamic time warping: A basketball case study. *PLOS ONE*, *17*(10), 1-15. doi:10.1371/journal.pone.0272848

Anexo A

Anexo

A.1. Tablas

Variable	Descripción
match_id	Identificador del partido
possession	Identificador de la posesión
period	Tiempo al que corresponde la posesión
tiempo	Duración de la posesión en segundos
resultado	Resultado del partido al momento que se registra la secuencia, según el equipo dueño de la posesión
possession_team.id	Identificador del equipo dueño de la posesión
possession_team.name	Equipo dueño de la posesión
team.name	Equipo que realiza las acciones de la posesión
n	Cantidad de acciones registradas en la posesión
x_inicio	Coordenada x de la acción inicial de la secuencia (largo de la cancha)
y_inicio	Coordenada y de la acción inicial de la secuencia (ancho de la cancha)
x_fin	Coordenada x de la acción final de la secuencia (largo de la cancha)
y_fin	Coordenada y de la acción final de la secuencia (ancho de la cancha)
n_eventos	Cantidad de tipos de eventos registrados en la posesión
n_pases	Cantidad de pases registrados en la posesión
n_pasesC	Cantidad de pases completados registrados en la posesión
n_traslados	Cantidad de traslados registrados en la posesión
n_driblesE	Cantidad de regates exitosos registrados en la posesión
prom_av_p	Avance promedio mediante pases
prom_av_t	Avance promedio mediante traslados de pelota
n_jugadores	Cantidad de jugadores involucrados en la posesión
n_centros	Cantidad de centros registrados en la posesión
n_cdf	Cantidad de cambios de frente registrados en la posesión
n_pases_arearival	Cantidad de pases dentro del área rival registrados en la posesión
n_ingresos_arearival	Cantidad de ingresos al área rival mediante conducciones de pelota registrados en la posesión

Variable	Descripción
inicia_golero	Si la posesión es iniciada por el golero
participa_golero	Si el golero participa en la posesión
termina_tiro	Si la posesión termina en un tiro al arco
tiro	Si se registra un tiro al arco en la posesión
gol	Si la posesión termina en gol
reinicio_juego	Si la posesión proviene de un reinicio de juego (tiro libre, saque de esquina, saque de arco o lateral)
zona_inicio	Zona de la acción inicial de la secuencia
zona_fin	Zona de la acción final de la secuencia
nzonas	Cantidad de zonas por las que se registra alguna acción en lo que dura la posesión
xG	Gol esperado promedio de los tiros al arco registrados en la secuencia
vel_media_p	Velocidad promedio (en m/s) de los pases registrados en la secuencia
vel_media_c	Velocidad promedio (en m/s) de los traslados de pelota registrados en la secuencia
vert_tot	Verticalidad total de la posesión ($x\_fin - x\_inicio$)
horiz_tot	Horizontalidad total de la posesión ($y\_fin - y\_inicio$)
dist.promP	Distancia promedio (en metros) de los pases de la secuencia
dist.promC	Distancia promedio (en metros) de los traslados de la secuencia
dist.medP	Distancia mediana de los pases de la secuencia
dist.medC	Distancia mediana de los traslados de la secuencia
presion	Cantidad de acciones realizadas bajo presión (*under pressure*) del rival
n_eq	Cantidad de acciones realizadas por el equipo que tiene la posesión
éxito	Si la posesión fue exitosa o no
termina	En qué termina la jugada (categorías de clasificación de éxito)
mundial	Si la posesión corresponde a un partido del mundial masculino o femenino
prop_tiempo	Proporción de tiempo que duró la secuencia entre el tiempo efectivo de posesión del equipo en el partido
anchura	Ancho de la secuencia (diferencia entre máximo y mínimo valor en la coordenada y)
ritmo_p	Ritmo de los pases en la secuencia ($\frac{n\_pases}{tiempo}$)
ritmo_c	Ritmo de los traslados en la secuencia ($\frac{n\_traslados}{tiempo}$)

Tabla A.1: Descripción de las variables utilizadas en el análisis de posesiones.

Evento	Variables asociadas
Pass	31
Shot	25
Goalkeeper	16
Clearance	7
Foul Commited	7
Dribble	5
Duel	4
Carry	3
Block	3
Foul Won	3
Interception	2
Ball Receipt	2
Ball Recovery	2
50-50	2
Miscontrol	1
Pressure	1

Tabla A.2: Cantidad de variables asociadas a cada tipo de evento.

Variable	Descripción
pass.aerial_won	Indicadora si el pase se da a partir de un duelo aéreo ganado o no
pass.angle	Àngulo (en radianes) del pase
pass.assisted_shot_id	Id del pase que generó el tiro
pass.body_part.id	Id de la parte del cuerpo con la que se realizó el pase
pass.body_part.name	Parte del cuerpo con la que se realizó el pase
pass.cross	Indicadora si el pase es un centro o no
pass.cut_back	Indicadora si el pase es un centro o no
pass.deflected	Indicadora si el pase fue bloqueado por un jugador rival
pass.end location.x	Coordenada x donde termina el pase
pass.end location.y	Coordenada y donde termina el pase
pass.goal_assist	Indicadora si el pase resulta ser asistencia de un gol o no
pass.height.id	Id de la altura del pase
pass.height.name	Altura del pase
pass.inswinging	Indicadora si el pase es un córner con efecto hacia adentro (curvado hacia el área)
pass.length	Largo del pase (en yardas)
pass.miscommunication	Indicadora si el pase no se logra por una mala comunicación entre compañeros
pass.no_touch	Indicadora si el deja pasar deliberadamente la pelota para otra compañero en vez de controlar y pasar
pass.outcome.id	Id del desenlace del pase
pass.outcome.name	Desenlace del pase
pass.outswinging	Indicadora si el pase es un córner con efecto hacia afuera (curvado alejándose del área)
pass.recipient.id	Id del jugador que recibe el pase
pass.recipient.name	Jugador que recibe el pase
pass.shot_assist	Indicadora si el pase resulta ser asistencia de un tiro o no
pass.straight	Indicadora si el córner no tiene efecto
pass.switch	Indicadora si el pase es un cambio de frente o no
pass.technique.id	Id de la técnica empleada en el pase
pass.technique.name	Técnica empleada en el pase
pass.through_ball	Indicadora si el pase atraviesa la última línea defensiva
pass.type.id	Id del tipo de jugada de donde proviene el pase
pass.type.name	Tipo de jugada de donde proviene el pase

Tabla A.3: Variables asociadas a los pases.

Variable	Descripción
shot.aerial_won	Indicadora si el remate se da a partir de un duelo aéreo ganado o no
shot.body_part.id	Id de la parte del cuerpo con la que se realizó el remate
shot.body_part.name	Parte del cuerpo con la que se realizó el remate
shot.deflected	Indicadora si el remate fue bloquedo por un jugador rival
shot.end_location.x	Coordenada x dónde se realizó el remate
shot.end_location.y	Coordenada y dónde se realizó el remate
shot.end_location.z	Coordenada z hacia dónde fue el remate
shot.first_time	Indicadora si el remate se dio de primera o no (sin control previo)
shot.follows_dribble	Indicadora si el remate se da luego de un regate
shot.freeze_frame	*Snapshot* (congelado) del contexto posicional en el momento del disparo. Se utiliza para los datos de *tracking*
shot.key_pass_id	Id del pase clave asociado al tiro
shot.one_on_one	Indicadora si el remate se da a partir de un mano a mano con el golero rival
shot.open_goal	Indicadora si el remate se da con el arco libre
shot.outcome.id	Id del desenlace del remate
shot.outcome.name	Desenlace del remate
shot.redirect	Indicadora si el remate proviene de un desvío de la pelota (sin control previo)
shot.saved_off_target	Indicadora si el remate no iba en dirección al arco y fue atajado por el golero rival
shot.saved_to_post	Indicadora si el remate fue atajado por el golero rival y luego pega en el palo
shot.statsbomb_xg	Gol esperado (xG) del remate
shot.technique.id	Id del tipo de remate (técnica de tiro)
shot.technique.name	Tipo de remate (técnica de tiro)
shot.type.id	Id del tipo de jugada de la cual proviene el remate
shot.type.name	Tipo de jugada de la cual proviene el remate
shot_impact_height	Altura de la pelota al momento del tiro

Tabla A.4: Variables asociadas a los tiros al arco.

Set 1	Set 2	Set 3	Set 4
dist.promC	*dist.medC*	*dist.medP*	*dist.medC*
dist.promP	*horiz_tot*	*dist.promC*	*horiz_tot*
horiz_tot	*n_cdf*	*horiz_tot*	*n_driblesE*
n_cdf	*n_centros*	*n_ingresos_arearival*	*n_eventos*
n_centros	*n_driblesE*	*n_pases_arearival*	*n_ingresos_arearival*
n_driblesE	*n_eventos*	*presion*	*n_pases_arearival*
n_eventos	*n_ingresos_arearival*	*prom_av_t*	*n_traslados*
n_ingresos_arearival	*n_pases_arearival*	*tiempo*	*prom_av_p*
n_pases	*n_traslados*	*vel_media_c*	*prom_av_t*
n_pases_arearival	*presion*	*vel_media_p*	*vel_media_c*
presion	*prom_av_p*	*vert_tot*	*vert_tot*
prom_av_p	*prom_av_t*	*xG*	*x_inicio*
prom_av_t	*tiempo_aj_pos*	*zona_fin*	*xG*
tiempo_aj_pos	*vel_media_c*	*zona_inicio*	*y_inicio*
vel_media_c	*vert_tot*		*zona_fin*
vert_tot	*x_fin*		*zona_inicio*
xG	*x_inicio*		
zona_fin	*xG*		
zona_inicio	*y_fin*		
	y_inicio		

Tabla A.5: Sets de variables utilizados para la conformación de clústeres.

Clúster	Exitosas	No exitosas
1	269	283
2	1098	2712
3	233	17
4	169	263
5	688	2387
6	197	1054

Tabla A.6: Éxito de las secuencias de cada clúster.

A.2. Gráficos

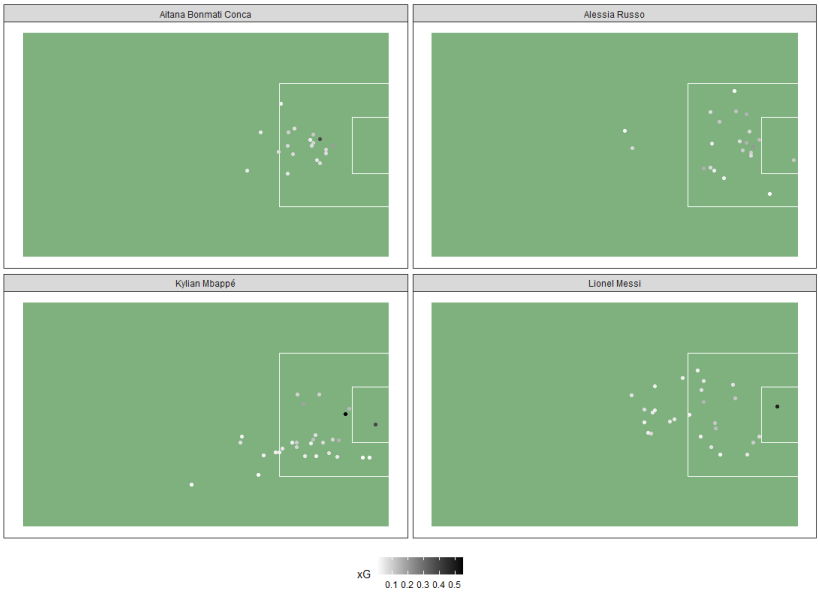


Figura A.1: Ubicación de todos los tiros por competición (excluyendo penales) según valores del xG.

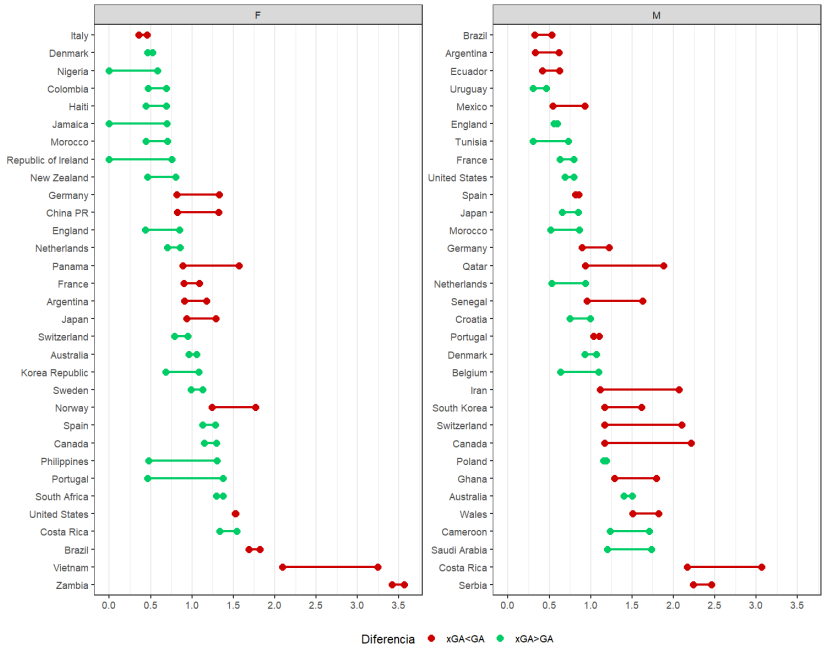


Figura A.2: Diferencia entre xGA y Goles Recibidos por equipo normalizado a 90 minutos (excluyendo penales).

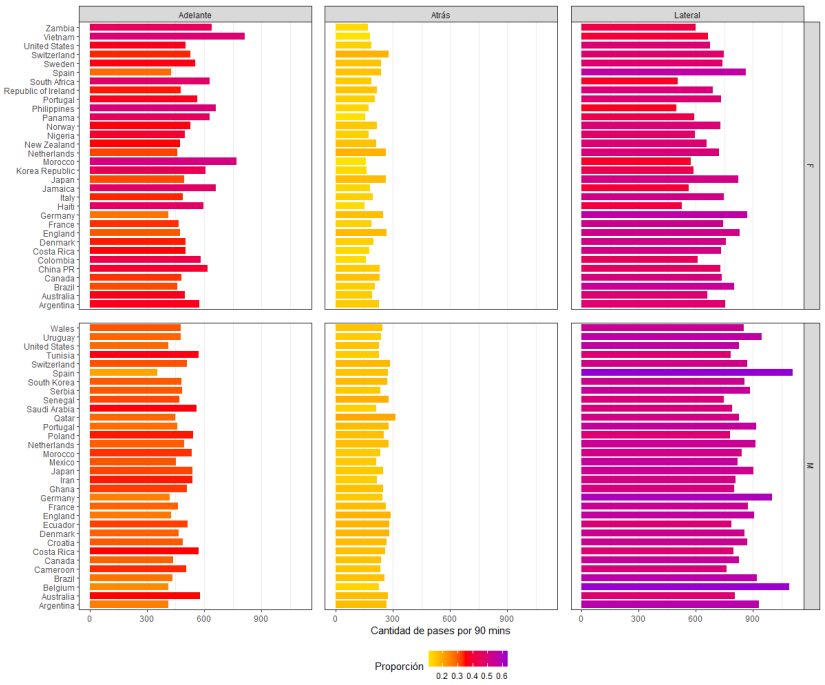


Figura A.3: Cantidad de pases normalizados según tiempo de posesión por equipo según su dirección y la proporción sobre el total de pases.

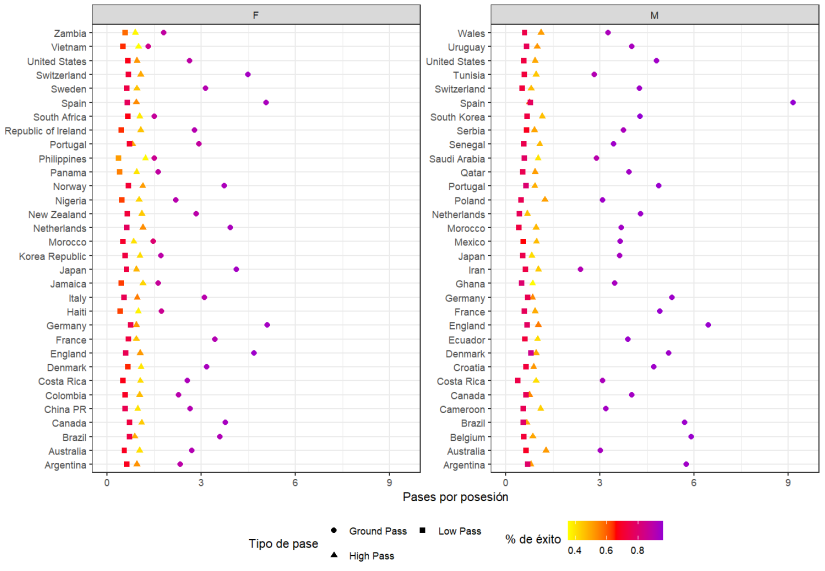


Figura A.4: Cantidad y porcentaje de acierto según cada tipos de pase.

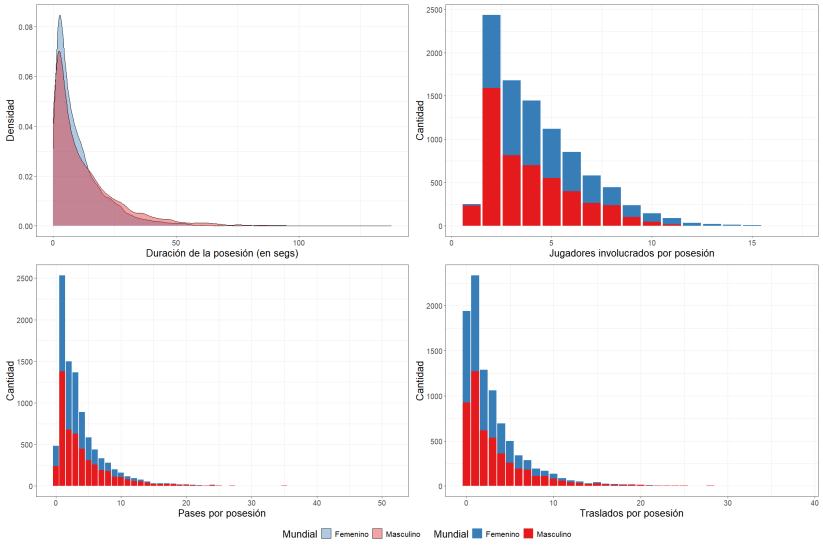


Figura A.5: Distribución de la duración (arriba, izquierda), futbolistas involucrados (arriba, derecha), pases (abajo, izquierda) y traslados (abajo, derecha) por posesión.

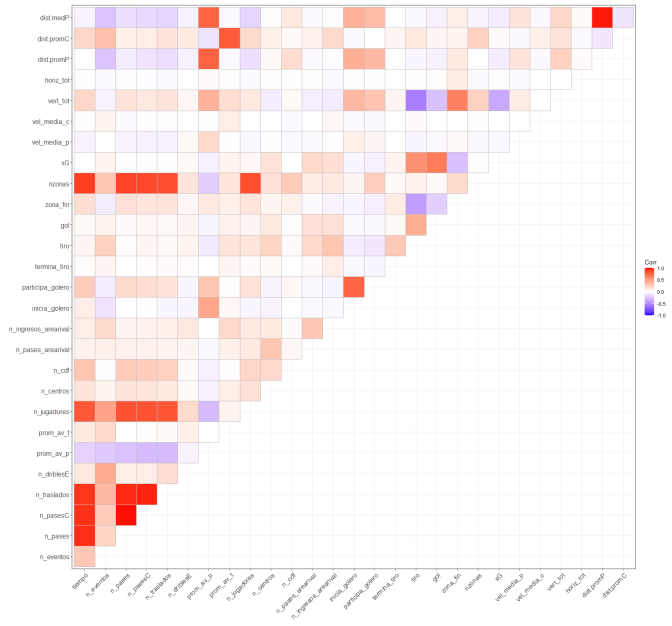


Figura A.6: Correlaciones entre variables creadas.

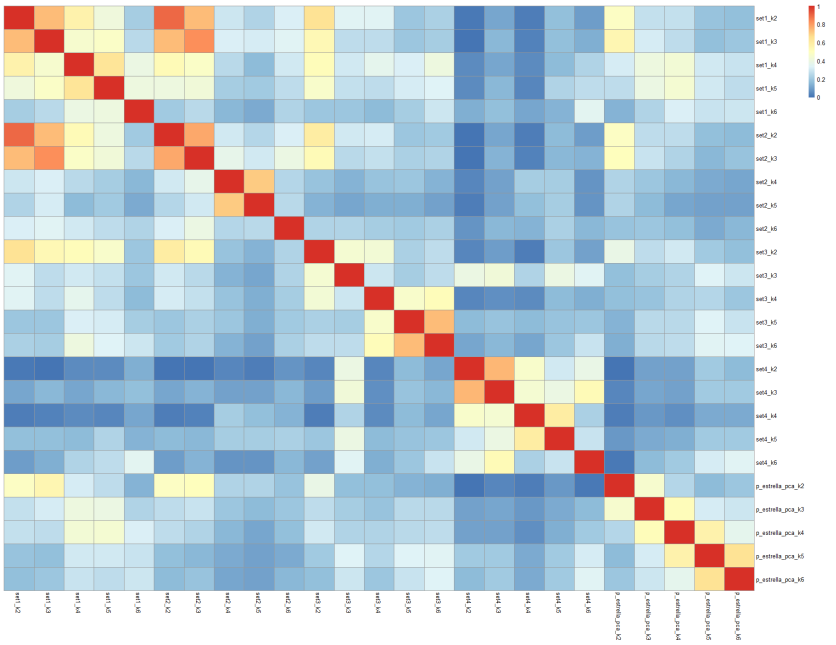


Figura A.7: Matriz con los ARI para cada par de particiones generadas.

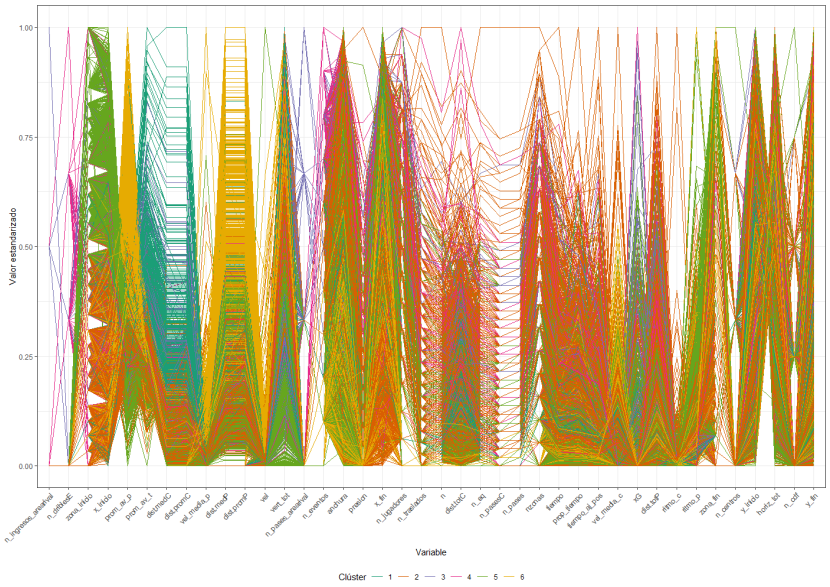


Figura A.8: Parallel plot de las secuencias según clúster.

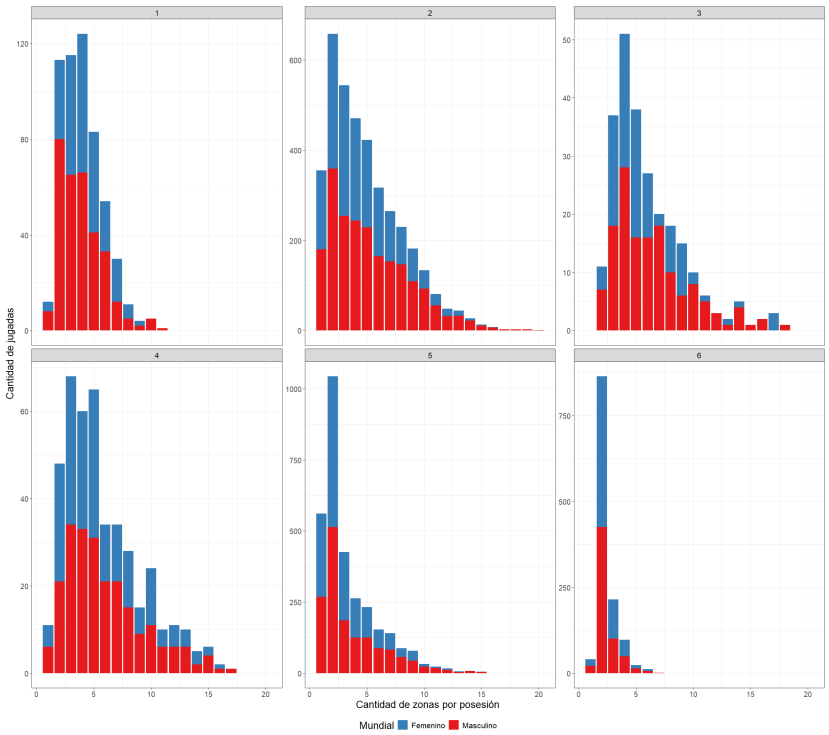


Figura A.9: Distribución de la cantidad de zonas por las que pasan las jugadas según clúster.

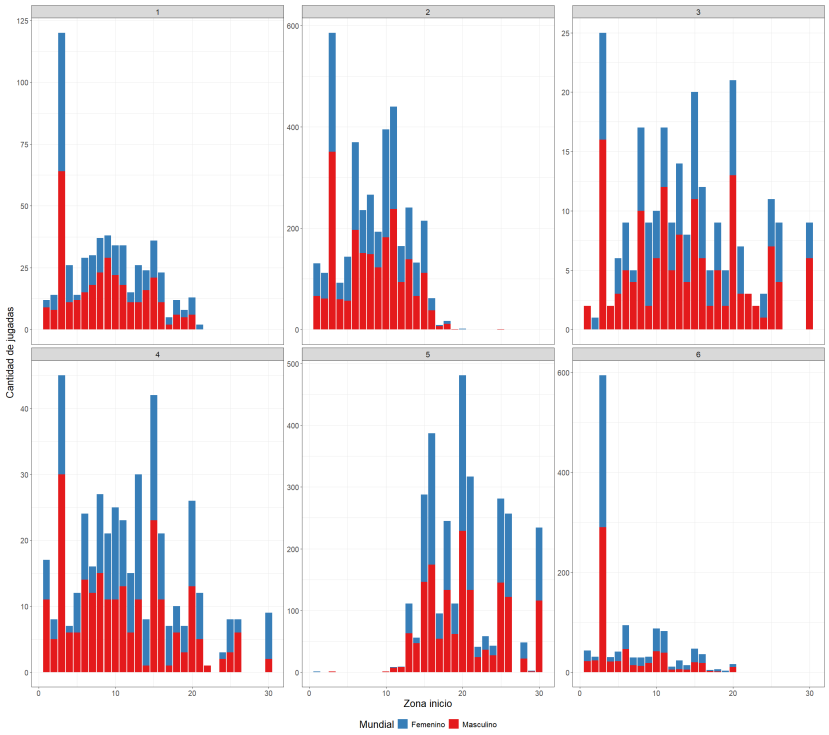


Figura A.10: Distribución de las zonas de inicio de las jugadas según clúster.

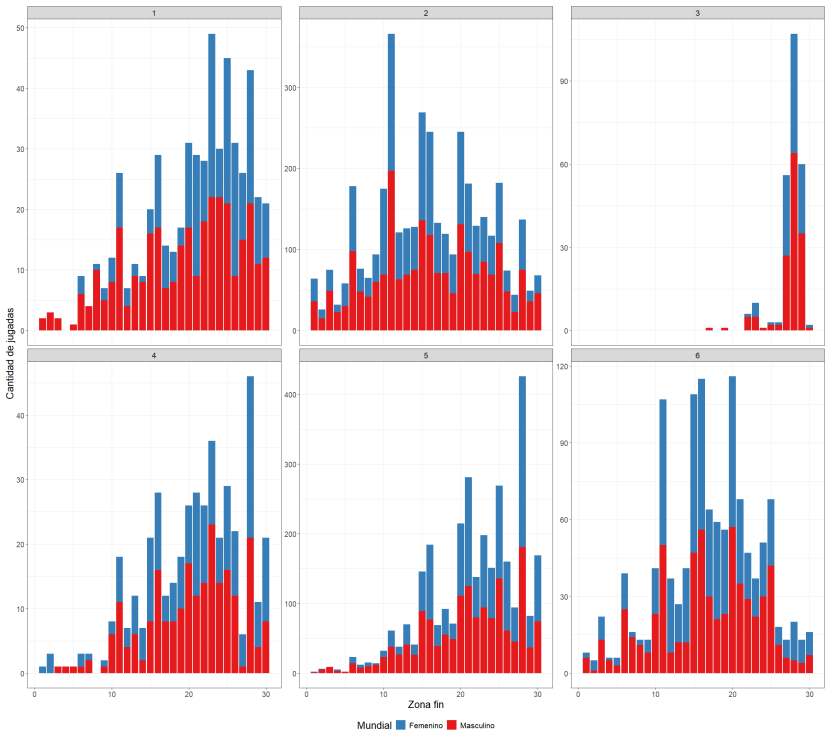
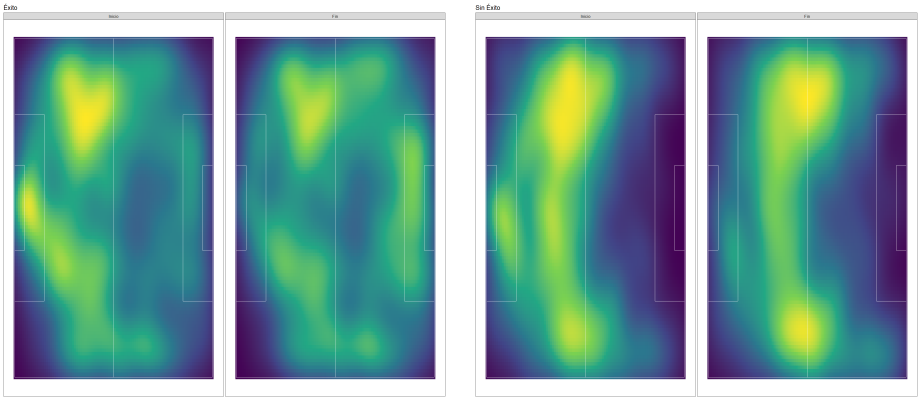
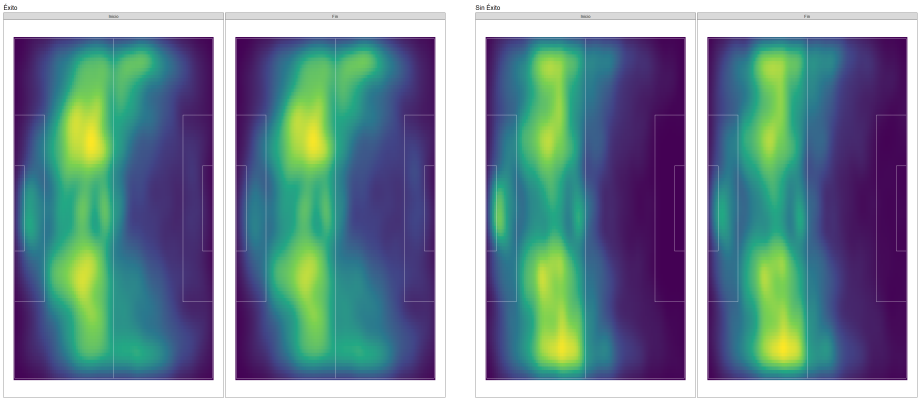


Figura A.11: Distribución de las zonas de finalización de las jugadas según clúster.



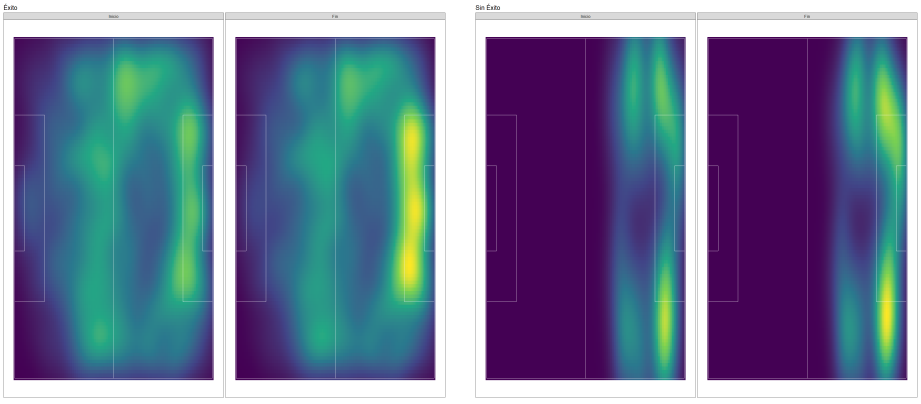
(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.12: Densidad de las secuencias del clúster 1 según su *éxito*.



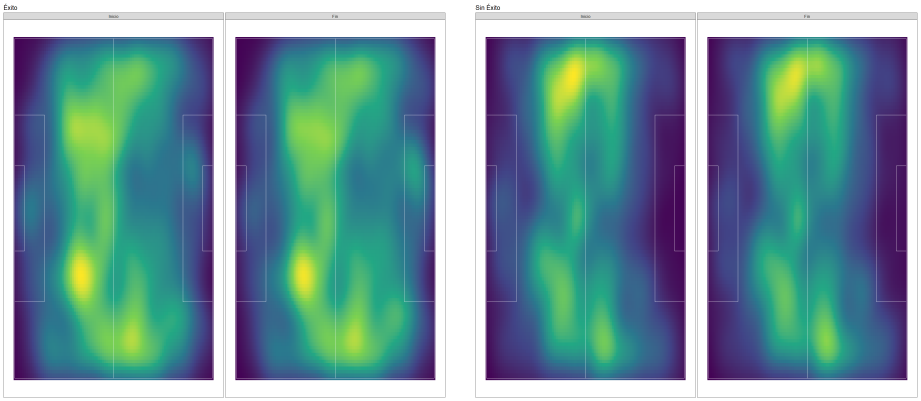
(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.13: Densidad de las secuencias del clúster 2 según su *éxito*.



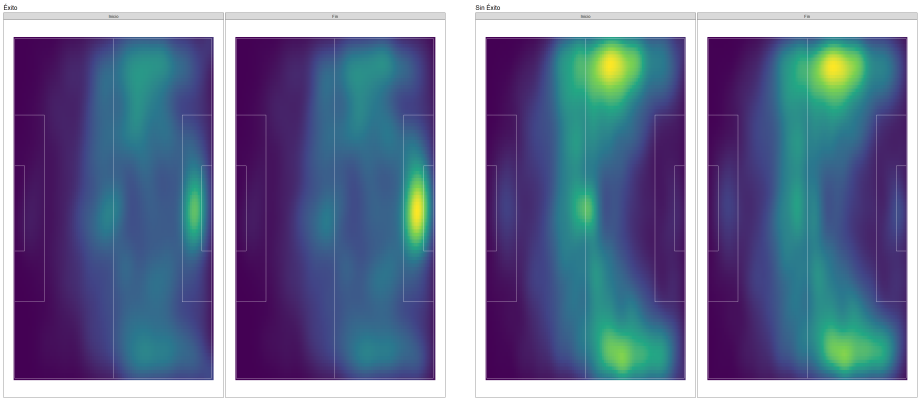
(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.14: Densidad de las secuencias del clúster 3 según su *éxito*.



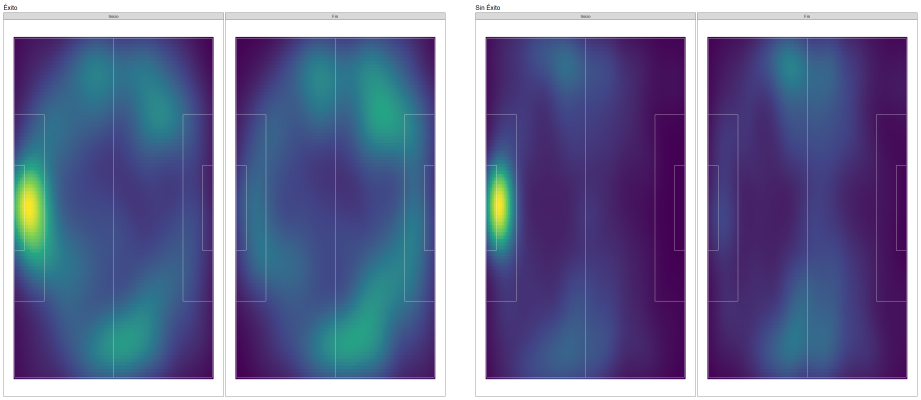
(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.15: Densidad de las secuencias del clúster 4 según su *éxito*.



(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.16: Densidad de las secuencias del clúster 5 según su *éxito*.



(1) Inicio y final de secuencias *exitosas*. (2) Inicio y final de secuencias *no exitosas*.

Figura A.17: Densidad de las secuencias del clúster 6 según su *éxito*.