Generación de un modelo para la extracción de la topología de una red compleja a partir de la evolución de las interacciones entre agentes sociales

Steven Fernando Rico Aparicio

Trabajo de grado para optar el título de Físico

Director
Johann Heinz Martínez Huartos

Codirector
Luis Alberto Núñez de Villavicencio Martínez

Universidad Industrial de Santander
Facultad de Ciencias
Escuela de Física
Bucaramanga
2021

Dedicatoria

A mi madre Subsy y a mi hermano Walter. El camino que hemos recorrido juntos lo tengo grabado en la mente.

Agradecimientos

Agradezco inmensamente a Johann Martínez y Mario Cosenza, por aceptarme como su estudiante y guiar el inicio de mi camino en los Sistemas Complejos.

A Luis Nuñez, por su ejemplo como excelente profesional y al grupo Halley, por la inmensa cantidad de oportunidades que me brindó.

A la Universidad Industrial de Santander, por el alto nivel de formación académica y a la Escuela de Física, especialmente a Amanda y Cecilia, por el ejemplo de formación personal.

A mis abuelos Vicente y Rosa, por el cálido amor que me han brindado desde el primer día, y a mi prima Dayana, por ser la imagen de excelencia que siempre persigo.

A mis amigos, y ahora colegas, Laura, Juan, Rolando y Lucía, por el apoyo mutuo y el constante crecimiento intelectual y personal que producen en mi.

A Danna y Daniela, por sus palabras de aliento incondicional y su sincera y longeva amistad.

Tabla de contenido

Introducción **9**

1 Materiales y métodos **12**
1.1 Datos y tratamiento 12
1.2 Definición del modelo físico para la captura de enlaces 14
1.3 Espacio de configuración, Matriz de adyacencia y visualización de red de pases . 16
1.4 Herramientas para medición de eficiencia 19
1.4.1 *Matriz de confusión* 19
1.4.2 *Creación de la matriz de verdaderos positivos y la matriz de falsos positivos* . . 20
1.4.3 *Herramientas de medición: Curvas ROC, Norma de Frobenius, divergencia Jensen-Shannon* 21
1.4.4 *Herramientas de medición: Coeficientes de red (Shortest Path, Clustering)* . . 24

2 Resultados **27**
2.1 Captura de sucesos 27
2.2 Medición de eficiencia 30
2.3 Criterios para la elección de la tupla óptima 33
2.4 Reconstrucción de la red compleja 36
2.5 Instancias individuales 40

3 Discusión **43**

4 Conclusiones **45**

Referencias **47**

5 Apéndice A **50**

Lista de figuras

1.1 *Esquema de contenido de datos y Esquema de distribución de instancias* . . . 13
1.2 *Vector resultante* $(\vec{r_{ib}})$ *para el marco de referencia no inercial (sobre el balón), y conjunto de magnitudes* $\{r_{ib}\}$ *para los jugadores i de un único "frame"* . . 14
1.3 *Ilustración del modelo físico que define el enlace (pase) en el sistema* 15
1.4 *Red de pases del F.C Barcelona. Encuentro contra el Real Madrid jugado en el estadio Bernabéu durante la temporada 2009/2010* 18
1.5 *Ilustración didáctica de las instancias de clasificación en la Matriz de Confusión* 20
1.6 *Ejemplo gráfico del Shortest Path y del Clustering* 26

2.1 *Matriz de Verdaderos Positivos (V) y Matriz de Falsos Positivos (F)* 29
2.2 *curva ROC individual* $\tau = 3.0(s)$ *y conjunto completo de curvas ROC* . . . 31
2.3 *Matriz de norma de Frobenius* (N) *y de divergencia Jensen-Shannon* (S) . . 33
2.4 *Matriz de Residuos R* 35
2.5 *Matriz de Adyacencia óptima* $W^{(0.3,3)}$ 36
2.6 *Reconstrucción de los coeficientes de red (Shortest Path y Clustering)* 38
2.7 *Visualización de la red compleja asociada al sistema agregado Macro* 39
2.8 *Reconstrucción de los coeficientes de red en las instancias Individuales* 41

5.1 *Conjunto de resultados para la primera instancia individual* 51
5.2 *Conjunto de resultados para la segunda instancia individual* 52
5.3 *Conjunto de resultados para la tercera instancia individual* 53
5.4 *Conjunto de resultados para la cuarta instancia individual* 54
5.5 *Conjunto de resultados para la quinta instancia individual* 55
5.6 *Conjunto de resultados para la sexta instancia individual* 56
5.7 *Conjunto de resultados para la séptima instancia individual* 57
5.8 *Conjunto de resultados para la octava instancia individual* 58

Lista de tablas

1.1 *Esquema de clasificación en una Matriz de Confusión* 19

2.1 *Resultados de instancias Individuales:* ρ',τ',vp,fp,*Frobenius,ROC* 40

Resumen

TÍTULO: Generación de un modelo para la extracción de la topología de una red compleja a partir de la evolución de las interacciones entre agentes sociales[*].

AUTOR: Steven Fernando Rico Aparicio[†]

PALABRAS CLAVE: Redes complejas, Análisis deportivo, Ciencia de datos.

DESCRIPCIÓN:

La Ciencia de Redes es una rama de la Física conocida por su transdisciplinariedad. Esta estudia sistemas provenientes desde la propia Física, hasta sistemas de la Biología, las Ciencias Sociales, la Economía, entre otros campos, y se enfoca en la colectividad de los agentes que componen el sistema, centrándose en sus interacciones. Este proyecto se basa en la perspectiva de red de un sistema social específico como sistema colectivo, el deporte del Fútbol. El estudio de deportes colectivos como este, es posible gracias al creciente surgimiento de modelos físicos y matemáticos que lo permiten. Sin embargo, existen dos aspectos relevantes que restringen la construcción de una red de pases. Por un lado, la creación de estas redes se basa en el conocimiento previo de los enlaces que conforman los sistemas deportivos. Por otro lado, existen vastas limitaciones en el acceso a estos datos de los sistemas futbolísticos.

Este proyecto propone una alternativa para la reconstrucción de una red compleja de fútbol sin conocimiento previo de los enlaces entre los jugadores. Para esto, se propone trabajar sobre los datos de rastreo (*Tracking*), datos en bruto que contiene información espacial (posición) y temporal (tiempo) de los agentes en la cancha. El proyecto pretende generar e identificar los enlaces que definen la red de pases a partir de estos datos alternos. Para esto, se desarrolla un modelo físico de *driven-data* que actúa bajo la restricción de una condición espacial (ρ) y una condición temporal (τ). Gracias al análisis de diferentes técnicas estadísticas sobre un espacio de configuración de los anteriores parámetros, se obtuvo como resultado una combinación de parámetros óptimos (ρ', τ') bajo la cual el modelo recupera los enlaces del sistema con una eficiencia mayor al $91\,\%$. Al mismo tiempo, se capturan los coeficientes de red del *Shortest Path* y del *Clustering*, con un error menor al $10\,\%$. Por lo que conocemos, esta es la primera aproximación en la reconstrucción directa de redes deportivas de fútbol, con base en la Física de sistemas complejos. El modelo propuesto se comporta con alta eficiencia, sin la necesidad de conocimiento previo de los enlaces del sistema. También se aporta a la aplicación de la ciencia de redes en sistemas sociales como los deportes colectivos.

[*]Trabajo de Grado

[†]Facultad de Ciencias, Escuela de Física, Johann Heinz Martínez Huartos (Director)

Abstract

TITLE: Generation of a model for the topology extraction of a complex network from the evolution of the interactions between social agents*.

AUTHOR: Steven Fernando Rico Aparicio†

KEYWORDS: Complex networks, Sports analytics, Data science.

DESCRIPTION:

Networks Science is a scientific branch of Physics known for its transdisciplinarity. It studies systems coming up from physics, biology, social sciences, economy, among other areas, and it focuses on the collectivity of agents forming a system, specially on the interactions among them. This project is based on the network aspect of a specific social system, the Football sport, as a collective system. The research of collective sports is possible thanks to the growing progress of physical and mathematical models. However, two facts are relevant when reconstructing a simple passing network. On one hand, the current field's literature about these networks is based on previous knowledge about how the links create those systems, which would present a major problem by generating some biases under the network reconstruction. On the other hand, it is noted the existence of vast limitations for accessing the data of Football systems.

This project proposes an alternative way to reconstruct a Football complex network without any prior knowledge. Hence, It is proposed to work with the tracking data. The well-known raw data that conglomerates the spatial (position) and temporal (times) information of these agents on the pitch. The project aims to generate and identify the links that define the passes network using such raw data. A physical and driven-data model is developed working under the constraints of a spatial condition (ρ) and temporal one (τ). Due to the analysis of different statistical techniques over a configuration space of previous parameters, it is obtained a combination of an optimal tuple (ρ', τ') that will recovers the system's links with an efficiency greater than $91\,\%$. At the same time, it captured the Shortest Path and the Clustering network coefficients, with an error lower than $10\,\%$. To the best of our knowledge, this is the first approximation around the direct reconstruction of soccer sports networks based on the Physics of complex systems. The model here proposed is well behaved with high efficiency, without the need for prior knowledge of the system links. It also shed light on the application of network science in social systems like collective sports.

*Bachelor Thesis

†Facultad de Ciencias, Escuela de Física, Johann Heinz Martínez Huartos (Director)

Introducción

La Ciencia de Redes (o Redes Complejas) es una rama académica que estudia cualquier sistema compuesto por una cantidad prominente de agentes, representados por nodos, tal que las interacciones entre ellos puedan ser identificadas y expresadas en forma de enlaces que conectan dichos nodos (Barabási y cols., 2016). En este campo, la investigación se centra en el estudio de las interacciones de la colectividad más que en el desarrollo de un único individuo. Aquí se observa que el intercambio de información entre nodos establece una estructura específica en el espacio (una red) que determina la dinámica del sistema, y es mediante el análisis de esta estructura que se hacen predicciones físicas sobre el comportamiento y el desarrollo del mismo (Boccaletti, Latora, Moreno, y cols., 2006). Este campo tiene como principales componentes la Teoría de Grafos desde la Matemática y la Mecánica Estadística desde la Física.

La Ciencia de Redes es una rama de investigación joven que ha presentado importantes avances y promete grandes resultados para el futuro. Apunta al mejor entendimiento de dinámicas climáticas para la prevención de eventos de calentamiento mundial extremos, o la temprana identificación de crisis sociales y el desarrollo de métodos de mitigación para esta clase de catástrofes (Havlin, Kenett, Ben-Jacob, y cols., 2012). Esta ciencia cuenta con un amplio abanico de acción, ya que incursiona con facilidad en el estudio transdisciplinar de otros campos como la Biología (Gosak, Markovič, Dolenšek, y cols., 2018), la Tecnología (van Raan, 2013), la Economía (Kenett y Havlin, 2015), el Arte (Fraiberger, Sinatra, Resch, Riedl, y cols., 2018), la Sociedad (Kadushin, 2012), entre muchos otros. En palabras de uno de los autores más prolíficos en el campo, M. Newman: "*Las redes constituyen un método, aunque general, muy poderoso de representar patrones de conexión e interacción entre las partes de un sistema*" (Newman, 2018). Estas redes se aplican sobre sistemas determinados "complejos", los cuales son definidos por las interacciones entre los agentes que lo componen, estos sistemas son estudiados al analizar los comportamientos que emergen de las interacciones, y no al observar individualmente los agentes que interaccionan. Desde la perspectiva de redes, la estructura de los enlaces describe las interacciones, y por consiguiente, el comportamiento emergente y la dinámica del sistema. A estos arreglos se les denominan "redes complejas" y los enlaces que la definen pueden tener características de peso y dirección, donde el primero indica el nivel de conectividad entre dos determinados nodos, y el segundo indica el sentido del intercambio de información entre ellos (Barabási y cols., 2016).

La Ciencia de Redes y el análisis deportivo parecieran ser dos entornos diferentes, pero actualmente estos ámbitos de estudio se enriquecen mutuamente (Grund, 2012; Wäsche, Dickson, Woll, y cols., 2017). En los últimos años, se ha podido describir el desarrollo y el desempeño global de los equipos de fútbol de alto nivel al evaluar el sistema deportivo desde el punto de vista matemáti-

co de jugadores y pases, lo que ha brindado grandes avances para ambos campos (Buldu, Busquets, Echegoyen, y cols., 2019). En el mundo deportivo, estos estudios implican utilidad en ámbitos desde la elección de estrategias de juego por parte de entrenadores, hasta bases de conocimientos para apuestas sobre equipos. Desde la perspectiva científica el estudio de estos sistemas decanta en un mejor entendimiento de interacciones no lineales, e implica la posibilidad de extrapolación de los análisis y los comportamientos a sistemas análogos de otras disciplinas.

Las redes complejas aplicadas a los sistemas futbolísticos actualmente aportan al mayor entendimiento del desempeño de los equipos (Clemente, Martins, Mendes, y cols., 2016) gracias a la creación de modelos físicos, matemáticos y geométricos de redes de pases (Buldú, Busquets, Martínez, y cols., 2018), y redes de campo (Herrera-Diestra, Echegoyen, Martínez, y cols., 2020). Trabajos de esta clase intentan recobrar la mayor información sobre la frecuencia de anotaciones, las distribuciones de los jugadores de defensa y ofensa, la probabilidad de victorias, las estadísticas de control sobre la cancha, etc. Todos estos modelos poseen algo en común: Sus resultados se basan en el análisis de datos de eventos (*events*), los cuales recuperan la información de todas las acciones que ocurren en un partido (pases, tiros al arco, tarjetas, entre otros) (Caicedo-Parada, Lago-Peñas, y Ortega-Toro, 2020). Normalmente se hace uso de estos datos para construir la red compleja del sistema deportivo, ya que los pases en los datos de eventos representan los enlaces entre los nodos (jugadores).

No obstante, existen algunas problemáticas que aún no se han abordado. Primero, la obtención de los datos de eventos representa un trabajo manual arduo realizado por extensos grupos de analistas (Pappalardo, Cintia, Rossi, y cols., 2019). Segundo, la industria privada del fútbol tiene poder sobre la gran mayoría de estos datos, lo que imposibilita el correcto desarrollo investigativo de estos sistemas sociales, ya que los datos no son de carácter libre. Para evitar la centralización de los datos y aportar a la automatización en la identificación de eventos, se han encontrado en la literatura modelos que intentan realizar predicciones sobre los equipos utilizando un conjunto de datos alternativo: los datos de rastreo o *Tracking* (Bialkowski, Lucey, Carr, y cols., 2014; Power, Ruiz, Wei, y cols., 2017). Estos corresponden a datos en bruto que almacenan solamente el seguimiento espacio-temporal de los jugadores en la cancha.

En esta línea, la ausencia de modelos para la identificación de interacciones entre los agentes de una red, es uno de los constantes retos de la Ciencia de Redes. Desde el punto de vista de las redes deportivas, se expresa la necesidad de reconstrucción de los pases del sistema sin conocimiento previo de los datos de eventos, y se cree firmemente que los datos en bruto de rastreo son una alternativa viable para este fin.

Por lo tanto, este proyecto hipotetiza la existencia de una estructura de red que puede ser reconstruida a partir de los datos de rastreo (*tracking*) en sistemas deportivos. Se apunta entonces a responder la pregunta ¿Es posible extraer la red compleja de un sistema aislado a partir del rastreo en bruto de sus jugadores en el tiempo? Finalmente, para resolver este interrogante, se propone como objetivo generar un modelo físico para extraer la topología de estas redes complejas a partir de la evolución de las interacciones entre los agentes, donde se establece como referencia teórica la información proporcionada en los datos de eventos.

El primer capítulo de este proyecto introduce los datos de rastreo (*tracking*), utilizados en el desarrollo del trabajo, junto con una descripción detallada del tratamiento al que se someten. Allí se presenta la propuesta del modelo físico creado, y a su vez, se describen las herramientas estadísticas empleadas para evaluar la eficiencia del modelo, y posteriormente avalarlo. El segundo capítulo expone los resultados obtenidos por las herramientas estadísticas utilizadas, y enuncia las condiciones óptimas bajo las cuales este modelo alcanza la mayor eficiencia posible en la reconstrucción de las redes de pases. En el tercer capítulo, se discute la validez del modelo con base en los resultados previamente mostrados, y se señalan algunos aspectos relevantes a tener en cuenta para la continuación de futuros trabajos afines. Por último, en el cuarto capítulo, se responde a la pregunta de investigación planteada y se indican las conclusiones obtenidas en el desarrollo de este proyecto.

1. Materiales y métodos

En este primer capítulo se presenta la información correspondiente a los datos utilizados para el desarrollo del proyecto, seguido de la depuración y el tratamiento que se les realizó. Se introduce también las consideraciones físicas que se establecieron para la construcción del modelo, el cual busca la reconstrucción de la red compleja. Finalmente, se muestran las herramientas matemáticas y estadísticas involucradas en la medición de la eficiencia para avalar el modelo.

1.1. Datos y tratamiento

Trabajar en el modelado de sistemas deportivos desde la rama de las redes complejas, específicamente en el fútbol, es una travesía con varias limitantes. La problemática principal radica en la adquisición de datos que sean de carácter público (Rein y Memmert, 2016). Estos datos se clasifican en dos clases. Por un lado, se encuentran los datos de **rastreo** o ***tracking***, los cuales recopilan información básica del partido: la ubicación espacial del balón y de los 22 jugadores en coordenadas cartesianas (Wei, Sha, Lucey, y cols., 2013). Esta información se almacena gracias a tecnología de monitoreo por visión artificial. El segundo paquete de datos corresponde a la información denominada **eventos**, los cuales son una construcción manual por analistas y expertos deportivos basados en los datos de rastreo (Pappalardo y cols., 2019). Aquí se clasifican las acciones que ocurren a lo largo del partido (pases, goles, tiros al arco, saques de banda, penaltis, entre otros). Actualmente, las redes complejas se construyen a partir de este ultimo *set* de datos, ya que los pases identificados corresponden a los enlaces de la red (Park y Yilmaz, 2010).

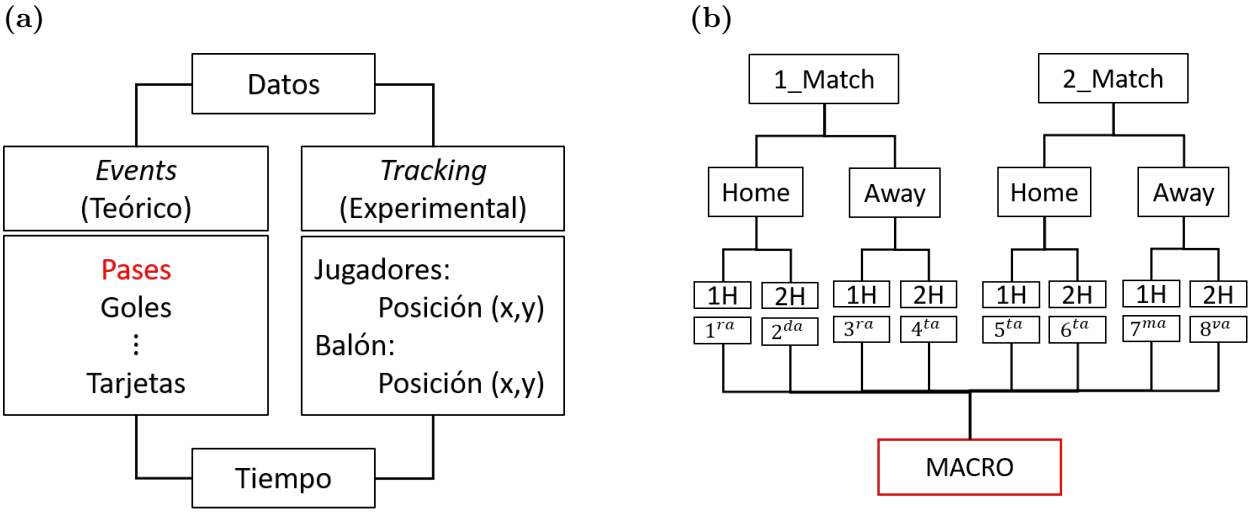
En abril de 2020, *Metrica Sport*, una plataforma privada de video análisis deportivo, abrió una ventana al modelamiento de estos sistemas. Esta empresa liberó al público un *set* de datos de rastreo y su equivalente de eventos sincronizados temporalmente (Fig. 1.1a) con una frecuencia de muestreo de $f = 25(fps)$. Una oportunidad sin precedentes. Ya que la red se construye generalmente con los pases de eventos, estos funcionarán como nuestros datos teóricos. Por otro lado, los datos en bruto de rastreo se usarán como nuestros datos experimentales con el fin de reconstruir a partir de ellos la red compleja. Ambos conjuntos de datos pueden ser recuperados de un repositorio en GitHub (Metrica, 2020).

Metrica Sport proporcionó información correspondiente a dos partidos anonimizados; la Figura (1.1b) presenta la estructura de división elegida para estos datos. Primero, en cada partido se almacena el desarrollo del equipo local y el equipo visitante, en la práctica cada equipo se describe

por una red individual (Cotta, Mora, Merelo-Molina, y cols., 2011). Segundo, para cada equipo se decidió reconstruir de forma separada la red del primer tiempo y del segundo tiempo debido a estudios que indican un desarrollo diferente en cada mitad (Rampinini, Impellizzeri, Castagna, y cols., 2009; Russell, Benton, y Kingsley, 2011). Con el fin de identificar cada mitad estudiada en este proyecto se les asoció números ordinales, es así que tenemos un total de ocho (8) instancias diferentes para trabajar, donde cada instancia posee información correspondiente al medio tiempo de un equipo anonimizado. Ahora bien, se debe tener en mente que un fin secundario de este proyecto es establecer un modelo físico que sea eventualmente aplicable a cualquier partido disponible. Con base en esto, se realizó un **"Macro"** partido definido como la agregada de las ocho (8) instancias. De esta forma se pueden tratar los cálculos posteriores como resultados generalizados del modelo.

Figura 1.1

Esquema de contenido de datos y Esquema de distribución de instancias



Nota. (a) Contenido de los set de datos sincronizados de *events* y *tracking*. (b) Segmentación de datos: Definición de las instancias a estudiar y del correspondiente sistema **"Macro"** agregado.

El sistema deportivo se sitúa espacialmente en una cancha de fútbol estándar de dimensión $68(m) \times 105(m)$, en este espacio se hace el seguimiento espacial y temporal de los jugadores y el balón. Los datos de rastreo entregan su información desde un punto de referencia en el centro de la cancha. Si bien lo anterior es bastante práctico, se decidió llevar el sistema a un marco de referencia no inercial sobre la pelota, tal que esta demarque el origen de coordenadas en todo momento. El balón representa el medio por el cual el sistema intercambia información entre los agentes y conocer las distancias relativas de todos los jugadores al balón en todo el momento facilita el seguimiento de las interacciones.

La Figura (1.2a) muestra como se define un primer vector $\vec{r_i}$ a un jugador i y un segundo vector $\vec{r_b}$ al balón. La magnitud del vector resultante representa la distancia relativa entre dicho jugador y la pelota, y se calcula como la resta de estos vectores $(\vec{r_b} - \vec{r_i})$ tal que su magnitud:

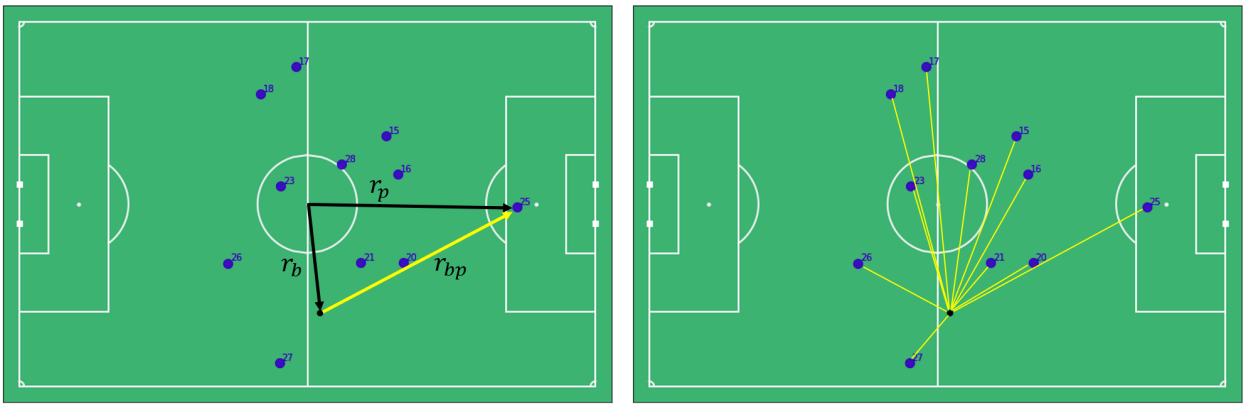
$$r_{ib} = \left| \sqrt{(x_i - x_b)^2 + (y_i - y_b)^2} \right| \quad . \tag{1.1}$$

Al aplicar este cambio de referencia en todos los agentes en juego, la magnitud r_{ib} permite conocer la proximidad del balón a cada uno de los jugadores (Fig. 1.2b), donde nos interesa identificar al jugador i con menor magnitud r_{ib}. Dicha magnitud representa al jugador con mayor cercanía al balón en un determinado momento, y con base en esta, es posible identificar si un jugador tiene posibilidades de contacto con el balón. El cambio de referencia se realiza en todos los frames de los noventa (90) minutos del partido.

Figura 1.2

Vector resultante $(\vec{r_{ib}})$ *para el marco de referencia no inercial (sobre el balón), y conjunto de magnitudes* $\{r_{ib}\}$ *para los jugadores* i *de un único "frame"*

(a) **(b)**



Nota. Los dos paneles identifican el mismo segundo de un partido. (a) Ilustración de un vector resultante en el cambio de referencia inercial desde el centro de la cancha a un marco de referencia no inercial con origen en el balón. (b) Ilustración de todas las magnitudes resultantes $\{r_{ib}\}$ para todos los jugadores i de un mismo equipo.

1.2. Definición del modelo físico para la captura de enlaces

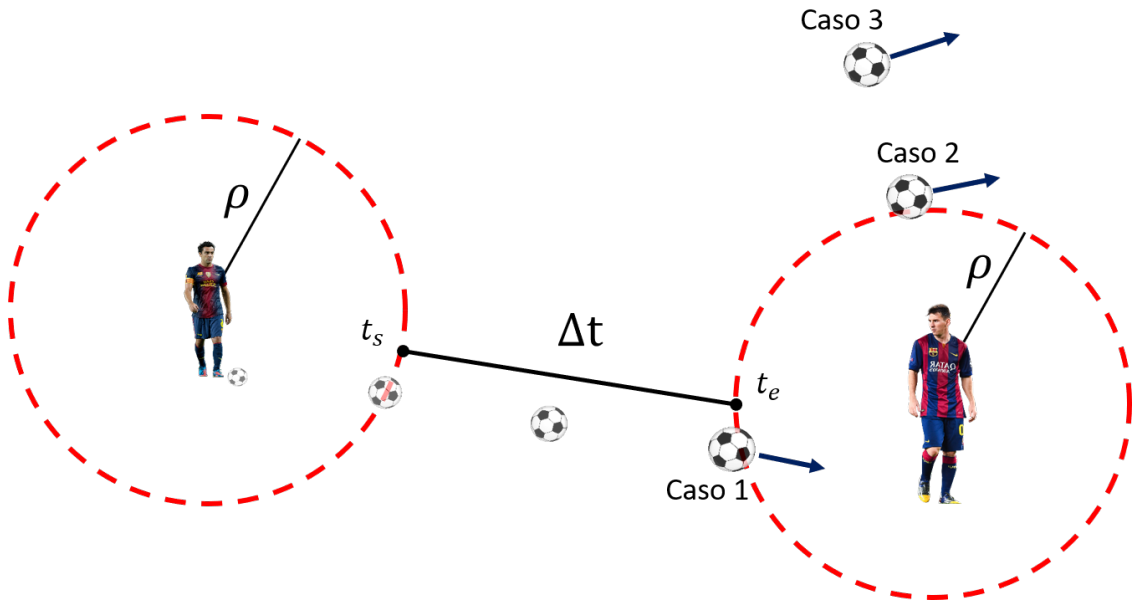
Nos interesa el punto de vista de la Física de Redes Complejas, donde cada jugador representa un nodo en una red, y se define una dinámica global con base en la estructura formada por los enlaces, al enviar información entre los nodos de la red (Clemente y cols., 2016). Estos enlaces son fundamentales para poder modelar el sistema deportivo, y para identificarlos nos centramos en las magnitudes fundamentales brindadas por los mismos datos de rastreo: posición y tiempo.

Con el fin de capturar los pases a partir de los datos de rastreo, se generó un modelo cuyo esquema puede ser visualizado en la Figura (1.3). El modelo parte de introducir dos variables físicas que funcionan como restricciones sobre el sistema, ρ y τ. La primera (ρ), es una consideración espacial que establece el radio de una circunferencia con centro en un jugador, dicho radio representa

el limite de una zona de contacto alrededor de los jugadores, donde es posible que haya contacto físico balón-jugador. La segunda (τ), es una consideración temporal que restringe el tiempo de vuelo del balón entre dos jugadores.

Figura 1.3

Ilustración del modelo físico que define el enlace (pase) en el sistema



Nota. El tiempo de vuelo (Δt) se establece desde el segundo (t_s) en que el balón sale de la frontera en la zona de contacto de un jugador i, hasta el segundo (t_e) en que ingresa a la frontera en la zona de contacto de un jugador j. Casos identificables por el modelo en la captura de pases: (Caso 1) Transición zona-zona, (Caso 2) Transición zona-frontera, (Caso 3) Transición zona-exterior.

La restricción (ρ) define una zona de contacto al rededor de cada uno de los jugadores, la cual representa una región de incertidumbre donde pueden haber interacciones balón-jugador. Si la distancia relativa r_{ib} entre un jugador i y el balón b es menor a la restricción ρ $(r_{ib} \leq \rho)$ el balón se encuentra dentro de la zona de contacto de dicho jugador i, de otra forma $(r_{ib} > \rho)$ el balón está fuera de la zona de contacto y no hay posibilidad de interacción balón-jugador. Esta zona se establece para todos los jugadores en el juego, y al hacer el seguimiento del balón sobre las diferentes zonas de contacto a las cuales ingresa el balón, se obtiene un conjunto de posibles pases (enlaces) que se establecen entre los jugadores dueños de dichas zonas.

Por otro lado, un pase real generalmente se realiza en un lapso corto de tiempo (Russell, Benton, y Kingsley, 2010) y en nuestro modelo esta condición se restringe con el parámetro temporal τ. Cuando el balón se traslada entre dos zonas de contacto (Fig. 1.3) se establece un posible pase (enlace) y este suceso se identifica por el tiempo de vuelo (Δt) que demora el balón en realizar dicha transición. El parámetro τ restringe la selección de pases al aplicar la siguiente condición:

$$0 < \Delta t \leqslant \tau \ . \tag{1.2}$$

Es decir, de todo el conjunto de posibles pases obtenido por el intercambio entre zonas de contacto (ρ), solo aquellos que cumplan con la condición (1.2) se seleccionarán como pases exitosos, de otra forma el suceso es descartado. Resumiendo, el modelo se centra principalmente en la identificación de los pases (enlaces) que definen la dinámica del sistema, para esto se implementan dos restricciones físicas (ρ, τ). Teniendo esto en cuenta se establece la definición de pase (enlace) bajo la cual trabaja nuestro modelo.

Definición 1 *Un pase (enlace) se define como una traslación del balón entre las zonas de contacto de dos jugadores del mismo equipo, definidas por (ρ), tal que el tiempo de vuelo del balón en su desplazamiento sea menor o igual al umbral de tiempo (τ).*

El modelo puede discernir tres posibles desenlaces (Fig. 1.3). *Caso 1*: transiciones de balón directas de una zona de contacto a otra. Este caso concuerda con la idea clásica de un pase. *Caso 2*: transiciones de una zona de contacto a una frontera de otra zona de contacto. Si bien aquí puede no haber contacto balón-jugador, según la definición propuesta este caso también se colecta como pase. *Caso 3*: transiciones desde una zona de contacto a una región exterior de otra zona de contacto. Este caso no se toma en cuenta ya que no cumple las condiciones necesarias establecidas.

Ahora, no se tiene conocimiento sobre cuales de los pases experimentales capturados por el modelo son reales, por lo tanto a estas capturas las denominamos como "sucesos capturados". Para evaluar la viabilidad del modelo, nos interesa conocer (dentro del conjunto de sucesos capturados) qué proporción fueron correctamente identificados, lo que es posible al compararlos con el conjunto de pases teóricos pertenecientes a los datos de eventos (Fig. 1.1b).

Un suceso capturado se clasifica como un pase real, si tanto los jugadores ij de las zonas de contacto involucradas en dicho suceso, como el tiempo en el que este fue identificado, coinciden con los jugadores y el tiempo de un pase perteneciente al conjunto de datos teóricos de eventos. Si no existe coincidencia alguna entre un enlace capturado y un pase teórico, el suceso corresponde a una falla sistemática del modelo. Con base en esto, es posible medir la eficiencia del modelo como la capacidad de capturar una alta cantidad de sucesos experimentales (enlaces) correspondientes a pases reales del sistema deportivo.

1.3. Espacio de configuración, Matriz de adyacencia y visualización de red de pases

La implementación de este modelo se realiza a través de un código desarrollado en *Python v3.7.7*, el cual solo requiere de una tupla (ρ, τ) para realizar el cambio al marco de referencia no inercial, e implementar las condiciones del modelo para la captura de sucesos experimentales.

Cada posible combinación de parámetros se traduce en un determinado número de pases recolectados, por lo que dicha tupla no puede ser de carácter arbitrario. Es necesario, entonces, encontrar la combinación de parámetros óptimos (ρ', τ') que maximice la eficiencia en la captura de los enlaces. Para esto se estableció un dominio para cada una de las restricciones, tal que se pueda obtener una amplia combinación de parámetros. Con el fin de no sobrestimar condiciones reales en las zonas de interacción balón-jugador, o en el tiempo de vuelo del balón, dichos dominios se definieron como:

$$\rho \in [0, 1](m), \quad \tau \in [0, 10](s) \ . \tag{1.3}$$

Estos rangos crean un espacio de configuración característico donde cada tupla posible (ρ, τ) captura una cantidad de sucesos específicos referentes a ese dúo de parámetros. Formalmente, la captura de enlaces tiene una representación matricial bien definida llamada la matriz de adyacencia $W^{(\rho,\tau)}$. Esta es una matriz cuadrada de $n \times n$ con diagonal nula, donde n es el numero de jugadores en el partido y sus elementos indican la existencia de enlaces entre dichos jugadores (Oliveira y Tyler, 2015). Ahora, los enlaces de la red que se quiere reconstruir son no dirigidos, es decir, no se discrimina un pase entre jugadores ij de un pase entre jugadores ji ya que ambos constituyen un mismo enlace entre los dos jugadores. Esto conlleva a que la matriz de adyacencia $W^{(\rho,\tau)}$ sea simétrica y que la totalidad de pases se almacenen en el triangulo superior o inferior del arreglo.

$$W^{(\rho,\tau)} = \begin{cases} \omega_{ij}^{(\rho,\tau)}, & \omega \text{ es la cantidad de pases entre jugadores } i \text{ y } j \text{ dada una} \\ & \text{configuración } (\rho, \tau), \\ 0 \ , & \text{si no hay pases entre estos.} \end{cases} \tag{1.4}$$

La matriz de adyacencia es la construcción matemática de la red de pases, y el proceso de construcción de la red finaliza al obtenerla. Al construir la matriz de adyacencia se recopila en sus filas/columnas la información de los nodos del sistema, y en sus elementos matriciales la cantidad de pases entre cada par de jugadores (enlaces). Es así que redes de pases, como la observada en la Figura (1.4) pueden ser visualizadas. Ahora bien, cada una de las posibles tuplas (ρ, τ) genera su propia y única matriz de adyacencia $W^{(\rho,\tau)}$, cada una con su respectiva representación gráfica.

Buscamos entonces encontrar la combinación de parámetros óptima (ρ', τ'), tal que su matriz de adyacencia $W^{(\rho',\tau')}$ reconstruya la red con mayor eficiencia. Para poder seleccionar cuál es la tupla óptima, es necesario medir la eficiencia del modelo entre las tuplas que conforman el espacio de configuración respecto a la matriz de adyacencia teórica T proveniente de los datos de eventos.

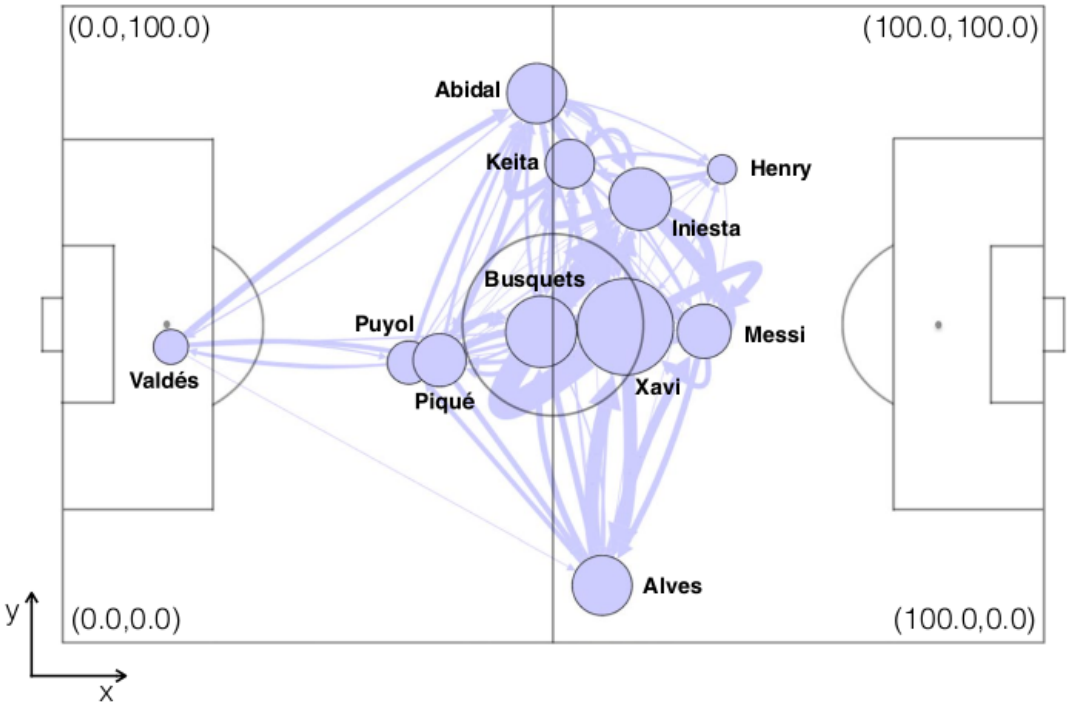
$$T = \begin{cases} t_{ij} \ , & t \text{ es la cantidad de pases teóricos entre jugadores } i \text{ y } j, \\ 0 \ , & \text{si no hay pases entre estos.} \end{cases} \tag{1.5}$$

En función de futuras operaciones cuantitativas entre los sucesos capturados experimentalmente y los pases reales, es necesario conocer cual es la cantidad total de pases reales dentro del sistema. La matriz de adyacencia T almacena esta información en sus elementos matriciales, por lo que al sumarlos se obtiene el total de pases realizados por todos los jugadores. Se define entonces la constante q, esta representa la cantidad total de pases reales en el sistema y funciona como referencia en posteriores comparaciones porcentuales de la cantidad de pases y la cantidad fallas sistemáticas capturadas por el modelo. Debido a la simetría de la matriz T, es necesaria la división de la sumatoria de pesos para tomar en cuenta únicamente el triangulo superior o inferior del arreglo. Entonces, se define q tal que:

$$q = \frac{\sum_{ij} t_{ij}}{2} \quad . \tag{1.6}$$

Figura 1.4

Red de pases del F.C Barcelona. Encuentro contra el Real Madrid jugado en el estadio Bernabéu durante la temporada 2009/2010



Nota. El radio de los nodos representa la magnitud relativa de centralidad de los jugadores, mientras que el ancho de los enlaces representa el peso de los pases establecidos en la matriz de adyacencia. La posición de los jugadores se establece como el promedio de las posiciones en donde cada jugador realizó los pases a lo largo del partido. Tomada de (Buldu y cols., 2019).

1.4. Herramientas para medición de eficiencia

1.4.1. *Matriz de confusión*

Luego de capturar los enlaces del modelo a través de las matrices de adyacencia $W^{(\rho',\tau')}$ es necesario identificar si aquellos sucesos corresponden, o no, a pases reales en el sistema. Para esto nos basamos en la matriz de confusión, una herramienta del campo de *Machine Learning* que cumple la función de clasificador binario de datos experimentales respecto a un conjunto de datos conocidos (Sammut y Webb, 2011). El conjunto de datos reales conocidos son identificados en dos clases (datos positivos y datos negativos), a la par y con ayuda de un umbral variable, el conjunto de datos experimentales se establecen dentro de estas dos mismas clases (datos positivos y datos negativos). Si se supone un conjunto de ratones de laboratorio bajo un estudio clínico de sobrepeso, las clases positiva y negativa se pueden establecer bajo el concepto de ratón obeso y ratón delgado, respectivamente. Al variar un umbral dentro de un determinado rango de peso en gramos, los diferentes ratones serán clasificados como obesos o delgados según superen, o no, dicho umbral. Las predicciones experimentales son ubicadas en el arreglo 2×2 que compone la matriz de confusión (Tb. 1.1), según sean correcta o incorrectamente clasificadas.

En nuestro caso, el conjunto de condiciones reales es compuesto por los datos de eventos, donde la clase positiva son eventos definidos como pases, mientras que la clase negativa son eventos que no son pases. Como umbral, se establece la tupla de restricciones (ρ, τ) ya que dependiendo de una combinación dada, se capturan por el modelo un conjunto de pases experimentales que serán clasificados en una clase positiva si representan un pase real o en una clase negativa si representa una falla sistemática.

La matriz de confusión sirve como base para el desarrollo de diferentes herramientas estadísticas que lleven a cabo el proceso de clasificación mencionado (Parker, 2001). La Tabla (1.1) muestra la estructura de clasificación binaria utilizada. Las fila representan las clases de los pases experimentales capturados, mientras que las columnas representan las clases del sistema real dado por los datos de eventos.

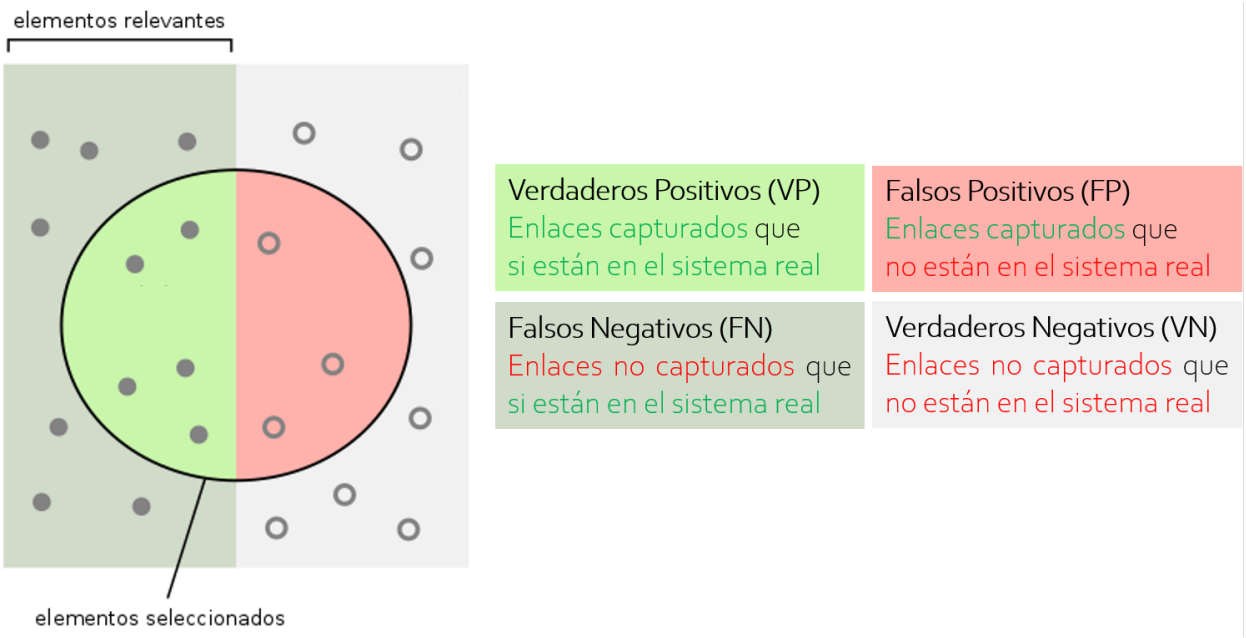
Tabla 1.1
Esquema de clasificación en una Matriz de Confusión

		Condición Real	
		Condición Positiva	Condición Negativa
Condición Predicha	Condición Positiva Predicha	Verdadero Positivo (vp)	Falso Positivo (fp)
	Condición Negativa Predicha	Falso Negativo (fn)	Verdadero Negativo (vn)

La instancia de **Verdaderos Positivos** (vp) corresponde a todos los enlaces capturados por el algoritmo que concuerdan con pases reales. Los **Falsos Positivos** (fp) encierran a todos los enlaces que fueron capturados pero que no existen en el sistema real. La instancia de **Falsos Negativos** (fn) corresponden a pases reales que no fueron capturados por el modelo. Finalmente, los **Verdaderos Negativos** (vn) son aquellos eventos que no son pases y no fueron capturados. La Figura (1.5) es una ilustración de estas instancias creada para facilitar su entendimiento. Para la medición de eficiencia dentro de nuestro modelo nos centramos en las instancias de Verdaderos Positivos (vp) y de Falsos Positivos (fp) al variar los parámetros (ρ, τ). Estos clasificadores representan la idea principal de capturas correctas e incorrectas dentro el modelo.

Figura 1.5

Ilustración didáctica de las instancias de clasificación en la Matriz de Confusión



Nota. Los elementos relevantes corresponden a los pases reales estipulados en los datos de eventos. Los elementos seleccionados identifican los sucesos capturados por el modelo. *modificado de* https://es.wikipedia.org/wiki/Sensibilidad_y_especificidad

1.4.2. *Creación de la matriz de verdaderos positivos y la matriz de falsos positivos*

Como es de esperarse, para cada combinación de parámetros en el espacio de configuración (ρ, τ), hay un número especifico de capturas correctas e incorrectas de pases. Con el fin de obtener un panorama global de la evolución de estas clasificaciones, se decidió crear dos matrices que representen el espacio de configuración al evaluar los resultados de vp y los resultados de fp de forma separada. Los elementos de estos arreglos evalúan el porcentaje de error de los pases capturados correctamente, e incorrectamente, respecto a la totalidad de pases reales q (Eq. 1.6) provenientes de los datos de eventos.

Primero se construye la matriz de Verdaderos Positivos (V), sus elementos presentan el error

relativo de la clasificación (vp) para cada tupla (ρ, τ) respecto a la totalidad de pases teóricos q. Esta matriz muestra el porcentaje de éxito en la captura de pases, es decir

$$V = [v_{\rho\tau}] \quad \text{donde} \quad v_{\rho\tau} = \frac{vp^{(\rho,\tau)}}{q} \cdot 100 \ . \tag{1.7}$$

Luego se construye la matriz de Falsos Positivos (F), ahora con elementos referentes al error relativo de la clasificación (fp) respecto a los pases totales q. Esta matriz muestra el porcentaje de fallo en la captura de pases y se define como

$$F = [f_{\rho\tau}] \quad \text{donde} \quad f_{\rho\tau} = \frac{fp^{(\rho,\tau)}}{q} \cdot 100 \ . \tag{1.8}$$

Se resalta que a diferencia de las matrices de adyacencia (Eq. 1.4 y 1.5) las columnas y filas de la matriz V y la matriz F no señalan jugadores, sino valores específicos de las restricciones ρ y τ del espacio de configuración definido en los dominios (1.3).

1.4.3. *Herramientas de medición: Curvas ROC, Norma de Frobenius, divergencia Jensen-Shannon*

Además de las matrices de verdaderos positivos (V) y falsos positivos (F), que permiten medir los porcentajes de éxito y fallo en la captura de pases, se mide la eficiencia del modelo a través de tres (3) métodos estadísticos más. El primero de ellos es **La curva ROC** (*Receiver Operating Characteristic*), un clasificador discreto propio de la teoría de señales, el cual se basa en las clasificaciones de la matriz de confusión dando una visión completa y directa del rendimiento del modelo (Fawcett, 2006). Esta herramienta ha sido utilizada recientemente como método para la elección de umbrales óptimos en modelos de predicción de victorias (Maimone y Yasseri, 2019), por ejemplo, sobre partidos de la copa mundial de fútbol 2018 (Hassan, Akl, Hassan, y cols., 2020).

La curva ROC consiste en una gráfica de dos dimensiones sobre un denominado espacio ROC, el cual contrasta dos indicadores: la "Sensibilidad" y "1-especificidad". El primero de ellos se relaciona con las condiciones positivas del sistema real (Tb. 1.1) y describe la proporción de éxito en la captura de enlaces. El segundo indicador se basa en las clasificaciones negativas del sistema real y representa la proporción de fallo en el modelo. La dinámica de este par de indicadores puede ser visto como una función análoga a la idea de costo/beneficio.

$$\text{Sensibilidad} = \frac{vp^{(\rho,\tau)}}{vp^{(\rho,\tau)} + fn^{(\rho,\tau)}}, \qquad \text{1-especificidad} = \frac{fp^{(\rho,\tau)}}{fp^{(\rho,\tau)} + vn^{(\rho,\tau)}} \tag{1.9}$$

El espacio ROC se compone de un eje horizontal (1-especificidad) y un eje vertical (sensibilidad), donde cada posible combinación de parámetros (ρ, τ) genera un único par (sensibilidad,1-especificidad) en este espacio. Dicho punto mide la proporción de capturas correctas contra la proporción de capturas incorrectas obtenida en el modelo al implementar esos valores (ρ, τ). La curva se crea al unir los puntos discretos resultantes al fijar el parámetro τ de la tupla, y variar su compañero ρ. El análisis de estas curvas permite conocer los umbrales (ρ, τ) bajo los cuales el modelo presenta una mejor relación de compromiso entre la capturas de pases correctos contra las fallas sistemáticas obtenidas del modelo. Dentro del espacio ROC la recta $x = y$ de las gráficas representa aleatoriedad en el modelo, todo punto ubicado en esta recta no posee relación entre las razones de acierto y fallo.

La segunda herramienta para ponderar la eficiencia del modelo y efectuar la búsqueda de la tupla óptima es la norma de Frobenius. Anteriormente se mencionó que la representación matemática de una red es la matriz de adyacencia, por lo tanto, para comparar dos redes, es necesario comparar las matrices de adyacencia que las define. **La norma de Frobenius** es una herramienta de geometría euclídea de alto orden que resta linealmente dos matrices de igual dimensión y posteriormente caracteriza la matriz resultante mediante el cuadrado de sus elementos (Golub y cols., 1996). Entre más cercano a cero (0) sea esta norma, menor distancia entre ellas. Es decir, mayor es el grado de similitud entre los elementos de estas matrices. Esta técnica se usa en análisis de dispersión sobre las formaciones de ataque y defensa en jugadores brasileños (Moura, Martins, y cols., 2012).

En el modelo nos remitirnos al espacio de configuración (ρ, τ), donde cada tupla genera una matriz de adyacencia $W^{(\rho,\tau)}$, las cuales deben ser comparadas respecto a la matriz de adyacencia teórica T. Con el fin de hallar la distancia entre estas matrices se definen las restas matriciales como $T - W^{(\rho,\tau)}$, donde la distancia de Frobenius se establece al realizar la norma de Frobenius, sobre los elementos sustraídos $t_{ij} - w_{ij}^{(\rho,\tau)}$. Como resultado del procedimiento se posee un conjunto de distancias matriciales correspondiente a todas las posibles combinaciones (ρ, τ).

Con el fin de poder visualizar el comportamiento de esta distancia al variar los umbrales de la tupla, se construye una nueva matriz con base en el espacio de configuración. Esta matriz se denomina Matriz de Norma de Frobenius (N) y sus elementos recopilan la distancia de Frobenius obtenida en cada una de las posibles tuplas. La implementación de la norma de Frobenius en el código desarrollado en *Python* se lleva a cabo gracias a la libreria *Numpy*, específicamente mediante la función *linalg.norm()*. La matriz de Frobenius se define como

$$N = [n_{\rho\tau}] \quad \text{donde} \quad n_{\rho\tau} = \|T - W^{(\rho,\tau)}\|_F = \sqrt{\sum_{ij} \left| t_{ij} - w_{ij}^{(\rho,\tau)} \right|^2} \ . \tag{1.10}$$

Como tercera herramienta utilizada para evaluar la similitud entre dos matrices se utiliza **la divergencia Jensen-Shannon** (JSD). Una medida proveniente de la Teoría de la información, donde las matrices de adyacencia son transformadas en distribuciones de probabilidad discretas (Fuglede y Topsoe, 2004). La medida de distancia se define como la diferencia de una distribución

de probabilidad respecto a una segunda distribución de referencia. Este método se ha usado, por ejemplo, en análisis de la dinámica de diferentes equipos de fútbol con base en la cantidad de goles anotados (Lopes y Tenreiro Machado, 2019).

En este caso, se calcula la distribución de probabilidad experimental (P) con base en los pesos de las matrices de adyacencia $W^{(\rho,\tau)}$, mientras que la distribución de probabilidad teórica se obtiene de los pesos correspondientes a la matriz de adyacencia teórica T. Se establecen las distribuciones como:

$$P^{(\rho,\tau)}(K^{(\rho,\tau)} = \{w^{(\rho,\tau)}\}), \qquad \text{y} \qquad Q(K = \{t_{ij}\}) \ . \tag{1.11}$$

Donde K es un espacio de probabilidad de números enteros $N = \{0, 1, 2, 3, ...\}$ que identifica el conjunto de cantidades de pases entre jugadores ij definido en la matriz de adyacencia teórica T. Con esto establecido, la divergencia entre las dos distribuciones se establece como

$$\text{JSD}(P^{(\rho,\tau)} \| Q) = \frac{1}{2} D(P^{(\rho,\tau)} \| M^{(\rho,\tau)}) + \frac{1}{2} D(Q \| M^{(\rho,\tau)}) \ , \tag{1.12}$$

donde la función D representa la entropía relativa de Kullback-Leiber (Fuglede y Topsoe, 2004). Al aplicarla, la divergencia en las distribuciones se reescribe tal que

$$\begin{aligned} \text{JSD}(P^{(\rho,\tau)} \| Q) = & \frac{1}{2} \sum_{ij} P^{(\rho,\tau)}(w_{ij}^{\rho,\tau}) \log \left(\frac{P^{(\rho,\tau)}(w_{ij}^{\rho,\tau})}{M^{(\rho,\tau)}(m_{ij}^{\rho,\tau})} \right) + \\ & \frac{1}{2} \sum_{ij} Q(t_{ij}) \log \left(\frac{Q(t_{ij})}{M^{(\rho,\tau)}(m_{ij}^{\rho,\tau})} \right) \ , \end{aligned} \tag{1.13}$$

donde

$$M^{(\rho,\tau)} = \frac{1}{2}(P^{(\rho,\tau)} + Q) \ . \tag{1.14}$$

Dentro del algoritmo construido, las matrices de adyacencia se transforman a distribuciones discretas mediante histogramas y el cálculo de JSD se realiza con la librería *Scipy*, específicamente a la función *stats.entropy()*. Como es de esperar, cada tupla de parámetros genera su propia distancia JSD respecto a la distribución Q teórica. Con base en esto, se genera nuevamente un arreglo matricial en el espacio de configuración (ρ, τ) llamada Matriz de JSD. Aquí se recopilan

las correspondientes distancias JSD para cada tupla con el fin de observar la diferencia entre las distribuciones de probabilidad de las matrices $W^{(\rho,\tau)}$ y T al variar los parámetros del modelo. Esta matriz se define entonces como:

$$S = [s_{\rho\tau}] \quad \text{donde} \quad s_{\rho\tau} = \mathrm{JSD}(P^{(\rho,\tau)} \| Q) \ . \tag{1.15}$$

1.4.4. *Herramientas de medición: Coeficientes de red (Shortest Path, Clustering)*

Adicional a los métodos presentados anteriormente, se decidió evaluar la eficiencia del modelo a través de dos coeficientes de red: el *Shortest Path* y el *Clustering*. Los coeficientes de red estudian la topología del sistema al cuantificar la organización de los enlaces y los nodos. Cada coeficiente tiene un fin determinado y en las redes de fútbol estos describen, desde propiedades de formación en los jugadores, hasta propiedades de sincronización entre ellos (Buldu y cols., 2019; Clemente y cols., 2015). Los diferentes coeficientes se calculan también mediante las matrices de adyacencia y en nuestro caso cumplen la función de contribuir a la medición de eficiencia del modelo al comparar los valores experimentales con los coeficientes teóricos.

El ***Shortest Path (sp)*** se presenta como la cuarta herramienta de comparación de este proyecto. Este coeficiente de red se usa en la teoría de grafos para hallar la menor distancia topológica entre dos nodos del grafo, siendo la distancia topológica entendida como la cantidad de enlaces intermedios que conectan dichos nodos (Watts y Strogatz, 1998). Para su aplicación se utiliza el algoritmo de Dijkstra, una serie de normas que permiten discernir el camino más corto dentro de todos los posibles (Dijkstra y cols., 1959) (Fig. 1.6). El *shortest path* es uno de los primeros coeficientes en ser calculados con el fin de describir la red compleja (Buldu y cols., 2019; Herrera-Diestra y cols., 2020) ya que representa una medida de integración, la cual indica qué tan fuerte son las conexiones entre los nodos del sistema, así mismo, el cómo diferentes regiones de una red se integran mediante estos *shortest path*. A nivel global, un *shortest path* promedio $\langle sp \rangle$ muy cercano a cero (0) implica mayor rapidez en el intercambio de información entre cualesquiera dos nodos de la red. Se define entonces:

$$\left\langle sp^{(\rho,\tau)} \right\rangle = \sum_{i,j} \frac{sp^{(\rho,\tau)}(i,j)}{h(h-1)} \ . \tag{1.16}$$

Donde $\langle sp \rangle$ es el promedio de todos los *shortest path* individuales $sp^{(\rho,\tau)}(i,j)$ desde un jugador "source" (i) a un jugador "target" (j), con h como el número total de nodos en grafo. Dentro del algoritmo en *Phyton* la matriz de adyacencia se transforma en un objeto grafo mediante la librería especializada *NetworkX* y se aplica las normas Dijsktra a través de la función *shortest_path_length()*, para obtener las $sp^{(\rho,\tau)}(i,j)$.

Por último, la quinta herramienta para la medición de la eficiencia en el modelo corresponde a un segundo coeficiente de red. El ***Clustering*** es un coeficiente de segregación que mide la capacidad de formación de grupos en la red, cuya relevancia se centra en la interconexión de tres (3) nodos (Fig. 1.6). El *clustering* tiene entonces la capacidad de cuantificar la eficiencia en la transmisión de información en el sistema en caso de que uno de los enlaces que lo conforman sea interrumpido (Onnela, Saramäki, y cols., 2005). Un jugador con alto *clustering* indica alta conexión de dicho jugador con sus compañeros más próximos y por tanto, mayor facilidad en el transporte del balón a lo largo de la cancha.

El cálculo de este parámetro varía según la red que se quiera describir, en este caso, el sistema deportivo que se estudia presenta una red compleja pesada no dirigida, por lo tanto el *clustering* para un único jugador en una red con tupla (ρ, τ) ya establecida se define como:

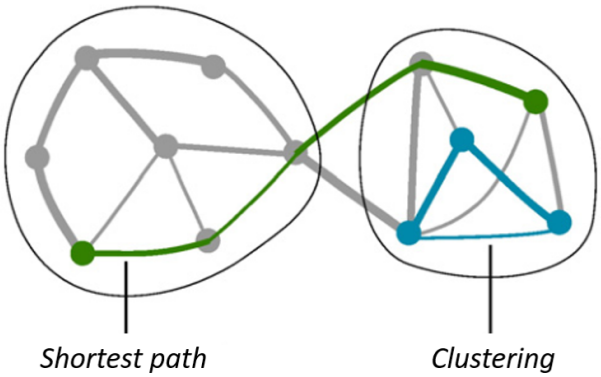
$$c_i^{(\rho,\tau)} = \frac{1}{k_i(k_i-1)} \sum_{jk} \left(\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk}\right)^{1/3} \quad . \tag{1.17}$$

Donde el grado (k_i) del nodo i corresponde a la suma de vecinos de dicho nodo, y los pesos ω_{ij} de los enlaces están normalizados por el peso máximo $\max(W^{\rho,\tau})$ de la red:

$$\hat{w}_{ij} = \frac{w_{ij}}{\max(W^{\rho,\tau})} \quad , \quad \hat{w}_{ik} = \frac{w_{ik}}{\max(W^{\rho,\tau})} \quad \text{y} \quad \hat{w}_{jk} = \frac{w_{jk}}{\max(W^{\rho,\tau})} \quad . \tag{1.18}$$

Para aplicar este coeficiente dentro del algoritmo en *Python*, se usó la librería *NetworX* para construir el grafo del sistema. La función *clustering()*, permite realizar los cálculos para cada uno de los jugadores de la red pesada no dirigida. Finalmente, cabe resaltar que tanto el *shortest path* como el *clustering* obtenidos para cada matriz de adyacencia experimental del espacio de configuración (ρ, τ) son contrastados con los correspondientes coeficientes teóricos definidos por la matriz de adyacencia original T.

Figura 1.6
Ejemplo gráfico del Shortest Path y del Clustering



Nota. En gris, los nodos y enlaces de una red. En verde, el camino más corto entre dos determinados nodos. En azul, la formación de tres (3) nodos interconectados. *Modificado de* (Rubinov y Sporns, 2010)

2. Resultados

En este capítulo se evalúa el modelo físico creado con el fin de encontrar una combinación de parámetros óptimos (ρ', τ') que maximice la eficiencia de la captura de pases, para esto se sigue la siguiente linea de procedimiento. Inicialmente se presenta el espacio de configuración como la captura de sucesos en los rangos de ρ y τ. Se expone, para cada posible tupla, tanto el porcentaje de capturas correctas, como el porcentaje de capturas incorrectas. Posteriormente, se hace un rastreo de eficiencia mediante las curvas ROC, la distancia de Frobenius y la divergencia de Jensen-Shannon. Con base en estos resultados, se presenta una propuesta para la elección de una tupla que represente la combinación óptima bajo la cual el modelo maximice la eficiencia de captura de pases. Al definir dicha combinación óptima, se extrae la correspondiente matriz de adyacencia de la red de pases y se calculan los coeficientes de red que se obtienen de ella.

Esta línea se lleva a cabo inicialmente con la representación "Macro" agregada de las ocho instancias, presentada en el diagrama de la Figura (1.1b), y posteriormente con sus respectivas ocho instancias individuales.

2.1. Captura de sucesos

Se estableció un rango de acción para las restricciones espaciales (ρ) y temporales (τ) (Eq. 1.3), bajo el cual se decidió rastrear los parámetros en once (11) pasos. De esta forma, la restricción espacial (ρ), que define la zona de contacto de los jugadores, puede tomar valores entre $[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0](m)$, mientras que la restricción (τ) para el tiempo del vuelo del balón puede variar entre $[1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0](s)$. Estos rangos de valores proporcionan un total de 121 tuplas (ρ, τ).

Cuando hablamos de captura de sucesos nos referimos al cambio del sistema de referencia de los datos de rastreo y, sobre ellos, la implementación de las restricciones establecidas en la definición (1). Este proceso se realiza para todas las combinaciones del espacio de configuración (ρ, τ), y para cada una de ellas, se obtiene como resultado un conjunto de sucesos. En otras palabras, un conjunto de posibles pases capturados por el modelo. Por tanto, es necesario la separación de los sucesos capturados entre correctos e incorrectos, para medir la eficiencia de captura del modelo, según una combinación (ρ, τ), y a su vez cuantificar las fallas sistemáticas que se puedan presentar.

Iniciamos este problema de clasificación binaria basados en la matriz de confusión (Tb. 1.1),

específicamente en la clase de verdaderos positivos (vp) y falsos positivos (fp). Al realizar la captura de pases, cada tupla (ρ, τ) arroja una cantidad especifica de estas clases, y con el fin de visualizar su comportamiento individual, se expresaron los resultados vp, fp en dos matrices independientes. Las matrices sitúan en su eje horizontal (columnas) los rangos temporales τ, y en su eje vertical (filas) los rangos espaciales ρ, de forma que cada elemento matricial represente uno de los 121 errores porcentuales de los sucesos capturados según las posibles combinaciones de parámetros.

De primera mano tenemos la matriz de Verdaderos Positivos (V), esta se presenta en la Figura (2.1a), donde sus elementos identifican el valor porcentual de los enlaces capturados que si están en el sistema real (Eq. 1.7). Dichos enlaces representan aquellos sucesos donde los jugadores involucrados en el pase y el tiempo en el cual se realizó la acción concuerdan con los jugadores y el tiempo de pases reales dentro de los datos teóricos de eventos. La intensidad del color rojo implica que dicha tupla capturó un número de pases correctos más cercano a la cantidad de pases expuestos en los datos de *events*. El patrón obtenido permite identificar un comportamiento creciente a medida que se incrementan los umbrales de las magnitudes (ρ, τ), llegando hasta valores superiores al $95\,\%$ en combinaciones extremas de las tuplas.

Se identifican regiones en donde tuplas con restricciones espaciales mayores a $0.2(m)$, combinadas con restricciones temporales mayores a $2.0(s)$, superan el $80\,\%$ en el porcentaje de capturas correctas. Analizando los casos extremos se aprecia que hay nulas capturas de pases en la columna correspondiente a $\tau = 0.0(s)$, lo que implica que el modelo no captura pases instantáneos entre dos zonas de contacto, algo que sería físicamente imposible, ya que para que exista una traslación, el balón requiere de un cambio en la posición a lo largo del tiempo. Por otro lado, al observar la fila correspondiente a $\rho = 0.0(m)$, existe un pequeño porcentaje de pases al incrementar la restricción en el tiempo de vuelo del balón. Una zona de contacto de $\rho = 0.0(m)$, y la existencia de estos sucesos, implican pases exactamente directos entre dos jugadores. Que el porcentaje de captura en este rango sea tan bajo implica que esta clase de suceso es muy esporádico. Generalmente los jugadores utilizan un espacio prudente a su alrededor para reacción y manejo de la pelota.

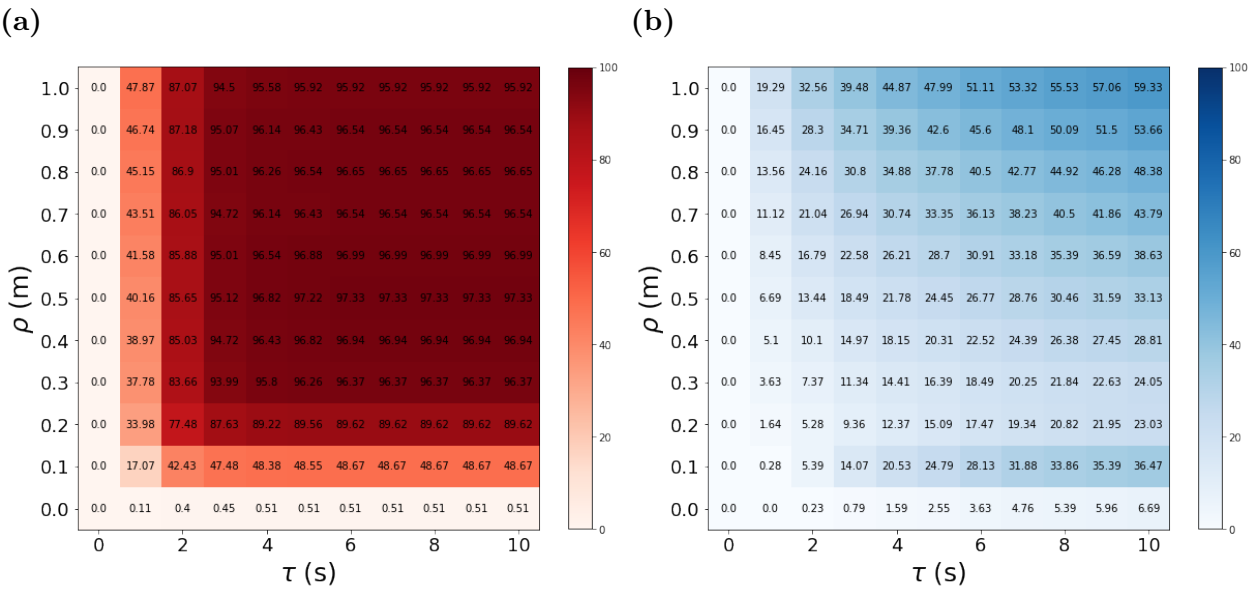
En el otro extremo de la matriz, observamos que con tuplas con una zona de contacto de $\rho = 1.0(m)$ y $\tau = 10.0(s)$ se captura casi la totalidad de pases teóricos, pero también que estos altos porcentajes de éxito se repiten desde combinaciones con magnitudes más bajas. Esto se debe a que los resultados son acumulativos, ya que las restricciones representan rangos y no valores fijos. Es decir, una zona de contacto de un metro incluye los posibles sucesos que ocurran desde 0 a 1 metro, y una restricción en el tiempo de vuelo de 10 segundos no implica que todos los pases duren 10 segundos, sino que se toman en cuenta todos los posibles pases con tiempo de vuelo entre 0 y 10 segundos. Debido a esto, los valores porcentuales altos en el extremo del espacio de configuración se deben a la acumulación de pases capturados en rangos inferiores, y no precisamente a valores realistas.

Posteriormente, se construyó la matriz de Falsos Positivos (F), apreciada en la Figura (2.1b). En esta matriz se identifican los valores porcentuales correspondientes a enlaces que fueron capturados pero que no existen en el sistema real (Eq. 1.8). Es decir, aquí se presenta el porcentaje de pases causados por fallas sistemáticas del modelo, siendo dichas fallas expresadas como sucesos

identificados como pases cuando los jugadores involucrados y el tiempo de la acción no concuerdan con ningún pase en los datos de eventos. Nuestra matriz de Falsos Positivos tiene un comportamiento creciente, es decir, a medida que se aumentan los rangos de los umbrales, la cantidad de sucesos erróneos capturados también aumenta. En este caso los valores máximos alcanzados se acercan al $60\,\%$, nuevamente en tuplas extremas. Se observa también un pequeño salto inesperado en el crecimiento del error porcentual al rededor de la fila correspondiente a $\rho = 0.1(m)$.

Figura 2.1

Matriz de Verdaderos Positivos (V) y Matriz de Falsos Positivos (F)



Nota. (a) Matriz de Verdaderos Positivos V: Espacio de configuración creado a partir de enlaces capturados que corresponden a enlaces reales del sistema. (b) Matriz de Falsos Positivos F: Espacio de configuración creado a partir de enlaces capturados que no concuerdan con enlaces reales del sistema.

El comportamiento creciente de la matriz F, se debe al carácter acumulativo de las restricciones ya mencionado, sin embargo se aprecia que, en general para cada elemento matricial, el porcentaje de falsos positivos es menor que su correspondiente valor porcentual de verdaderos positivos. Es decir al aplicar el modelo propuesto, en general se captura cuantitativamente una cantidad mayor de pases reales en cada tupla. Se observa que valores muy bajos en las restricciones (ρ, τ) llevan a bajas capturas erróneas las cuales pueden estar relacionadas con el bajo porcentaje de capturas que se aprecia para ese mismo rango en la matriz de verdaderos positivos (V). Por otro lado, rangos altos en las restricciones, presentan altos porcentajes de captura de sucesos erróneos.

Salta a la vista una región interesante en la fila $\rho = 0.1(m)$ donde hay un aumento inesperado en la captura de pases erróneos. Remitiéndonos a la literatura de métodos de medición para los datos de rastreo, encontramos que la resolución espacial de los instrumentos tecnológicos utilizados en el monitoreo de los partidos es de aproximadamente $0.1(m)$ (Linke, Link, y Lames, 2020). Con base en dicha concordancia, se cree que el modelo propuesto tiene la capacidad de reproducir condiciones reales del sistema que no han sido introducidas *a priori*. Estas condiciones afectan la clasificación de los pases capturados y facilita señalar zonas en el espacio de configuración que

deben ser evitadas.

La lectura de estas matrices no es trivial. Queda claro que en la búsqueda de una tupla óptima existe la tendencia a decantarse por combinaciones extremas de ρ y τ con base en su alto porcentaje de capturas acertadas (Fig. 2.1a). Pero estos mismos elementos matriciales presentan a su vez un alto porcentaje de capturas erróneas (Fig. 2.1b). Se debe hallar entonces una razón de compromiso entre ambos comportamientos con el fin de encontrar la mejor combinación (ρ, τ). La ventaja de construir estos espacios de configuración con base en la matriz de confusión es que existe una relación directa para evaluar el comportamiento de aciertos y desaciertos para todas las posibles combinaciones de parámetros al mismo tiempo.

2.2. Medición de eficiencia

Como se mencionó en la sección de métodos, las curvas ROC son herramientas útiles para llevar las clasificaciones de la matriz de confusión (de donde construimos las matrices V y F) a una gráfica en el espacio ROC, que expresa directamente la relación de compromiso entre la captura de pases reales y la captura de fallas sistemáticas mediante la clasificación de sensibilidad y 1-especificidad (Eq. 1.9).

La curva ROC funciona al variar un umbral, en este caso dicho umbral se estableció como la restricción espacial, es decir, para un τ fijo se varia su componente de tupla ρ, de esta forma, para un determinado τ, existen once (11) combinaciones con la restricción ρ y cada una de ellas posee un punto en el espacio ROC, como se ilustra en la Figura (2.2a). La curva ROC se forma al unir los puntos resultante de la variación ρ sobre un τ constante, esta curva muestra el impacto de las diferentes posibles combinaciones (ρ, τ) sobre la relación de compromiso de capturas correctas contra fallas sistemáticas. Para $\tau = 3(s)$ (Fig. 2.2a), umbrales ρ de $0.3(m)$, en adelante, poseen altos porcentajes de sensibilidad pero a la vez aumenta su porcentaje de capturas incorrectas, mientras que si nos deslizamos a umbrales bajos, tanto capturas correctas como incorrectas disminuyen; lo que indica realmente que en esta zona de parámetros no hay captura alguna de sucesos.

Cada valor de τ posee su propia curva ROC y para poder observar el comportamiento global de la variación de parámetros sobre el modelo se presenta el aglomerado de estas en la Figura (2.2b). Esta figura recopila la relación de compromiso entre la capturas de pases correctos contra las fallas sistemáticas de las 121 tuplas del espacio de configuración (ρ, τ). Encontramos aquí varios resultados interesantes.

Primero, si bien en las curvas ROC obtenidas, hay una pequeña cantidad de puntos que se acercan a la recta $x = y$ que representan aleatoriedad en el modelo, la mayoría de tuplas se ubican firmemente en el triangulo superior de la gráfica, mostrando que nuestro modelo captura una razón de acierto, para cada tupla, superior a su porcentaje de fallo en la captura de pases.

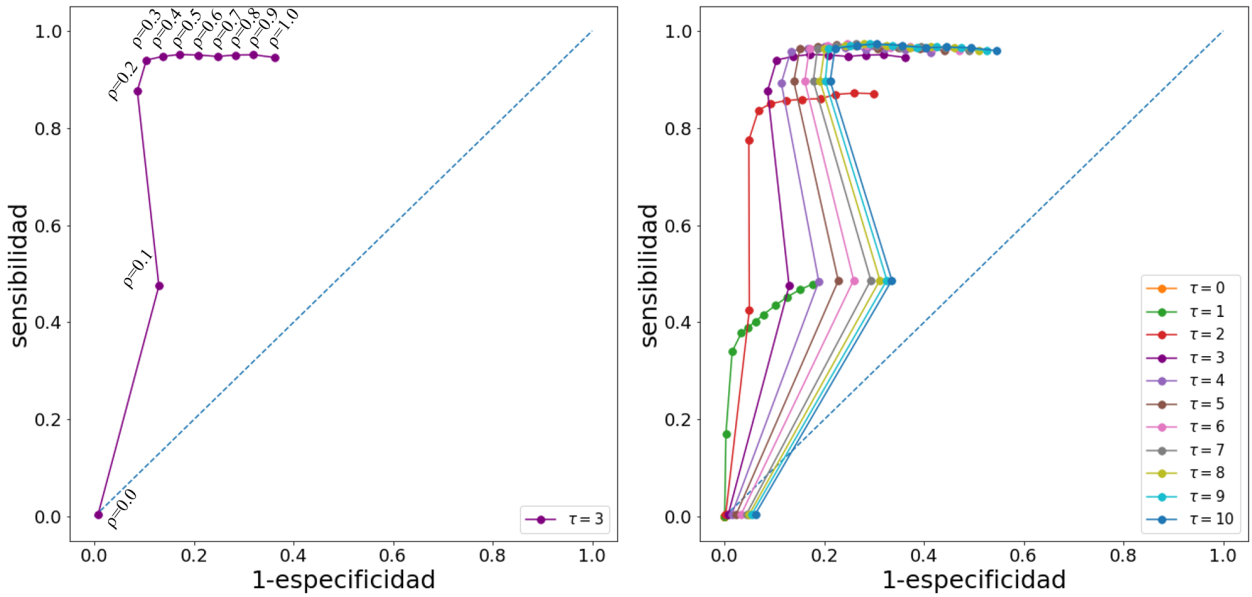
Segundo, en general se obtiene alta sensibilidad en las curvas ROC, pero en la búsqueda de la combinación de parámetros que maximice la eficiencia debemos centrarnos en aquellos puntos ROC que presenten alta sensibilidad y a la vez baja especificidad, tuplas que se acerquen a la es-

quina superior izquierda del espacio ROC. Vemos en esta zona un conjunto de posibles candidatos, los cuales se componen de combinaciones asociadas a los valores centrales de los rangos de los parámetros (ρ, τ). Aunque no es directo discernir que tupla es la mejor, poseemos ahora un rango de interés para focalizar la búsqueda.

Tercero, las tuplas cuyos resultados se acercan a la recta de aleatoriedad, debido a su incremento abrupto en la especificidad, concuerdan con el crecimiento inesperado en la matriz F (Fig. 2.1b). Esto nos indica una zona dentro de los rangos espaciales de $\rho = 0.1(m)$ que se recomienda evitar en la selección de tuplas bajo la precaución de que aquí existe predilección a capturas erróneas indeseadas, dadas por la resolución espacial en la toma de los datos de rastreo.

Figura 2.2
curva ROC individual $\tau = 3.0(s)$ y conjunto completo de curvas ROC

(a) **(b)**



Nota. (a) curva ROC individual: Cada punto de la curva identifica la relación de compromiso entre la sensibilidad y 1-especificidad para los diferentes umbrales ρ, con $\tau = 3(s)$ constante. (b) Curvas ROC: Comparación entre todas las curvas ROC del sistema.

La representación matemática fundamental de las redes es la matriz de adyacencia, por lo tanto una forma diferente pero complementaria de evaluar la eficiencia de captura de enlaces se expresa al construir la matriz de adyacencia $W^{(\rho,\tau)}$ y compararla respecto a la matriz de adyacencia teórica T. La matriz de adyacencia, a diferencia de las matrices V y F construidas sobre el espacio de configuración (ρ, τ), identifica en sus filas y columnas los jugadores de un equipo en el partido, y sus elementos matriciales expresan la cantidad de pases entre todos los pares de agentes.

La norma de Frobenius se basa en las matrices de adyacencia y funciona como método adicional en la medición de la eficiencia del modelo. Recordemos que la distancia en la norma de

Frobenius (Eq 1.10) es un número positivo que mide la diferencia entre los elementos de dos matrices, en este caso, la matriz de adyacencia experimental $W^{(\rho,\tau)}$ y la teórica T. Una distancia en la norma de Frobenius de cero (0) identifica dos arreglos con valores de elementos matriciales idénticos. En nuestro caso cada una de las 121 posibles combinaciones de parámetros posee su propia matriz de adyacencia $W^{(\rho,\tau)}$, y por lo tanto, su propia distancia de Frobenius.

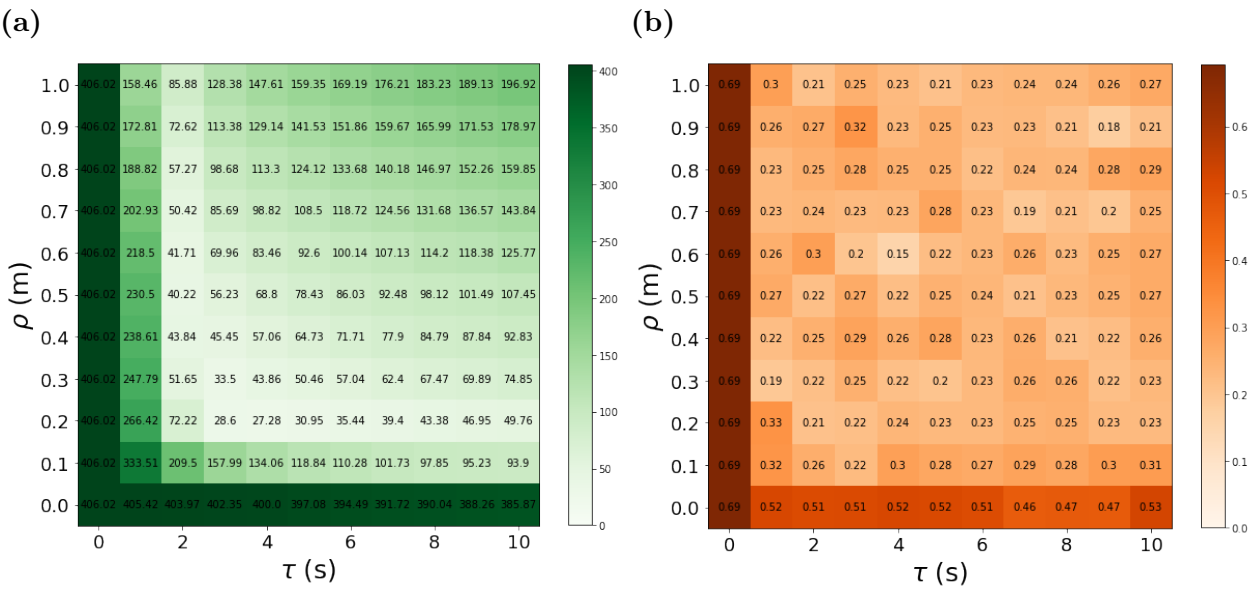
Para observar el comportamiento de estas distancias al variar los parámetros se construyó una matriz sobre el espacio de configuración (ρ, τ) semejante a las matrices V, F. En este caso cada elemento matricial presenta la distancia de Frobenius correspondiente a cada tupla (Fig. 2.3a), este arreglo matricial se denota como la matriz de distancias en la norma de Frobenius (N).

Se obtiene un comportamiento peculiar. En la matriz N se observa una región de mínimos globales correspondiente a valores centrales bajos de los parámetros (ρ, τ), esta zona de alto interés concuerda con la zona marcada por los resultados de las curvas ROC. Al centrarnos en la fila del caso extremo $\tau = 0.0(s)$, donde ya habíamos visto que no se captura ningún suceso, hay una distancia máxima de Frobenius de 406.02. Si expresamos esto como un límite superior de referencia, se aprecia que las magnitudes de la región de mínimos tienen un orden de magnitud menor, si bien estas distancias de Frobenius no son nulas, las medidas son bajas. Esto señala una alta concordancia entre el peso de los pases asociados a cada uno de los jugadores entre $W^{(\rho,\tau)}$ y la matriz de referencia T.

La tercera herramienta basada en la Teoría de la información, toma como base las matrices de adyacencia. La divergencia de Jensen-Shannon (Eq. 1.12) compara las matrices de adyacencia $W^{(\rho,\tau)}$ y T al convertir los elementos de $W^{(\rho,\tau)}$ en la distribución de probabilidad P y establece la distancia no euclídea como el desplazamiento de esta respecto a la distribución original Q obtenida de la matriz T (Eq. 1.11). Para observar el comportamiento global de estas mediciones se expresan los resultados para cada combinación de parámetros en un arreglo matricial sobre el espacio de configuración (ρ, τ) denominada la matriz de Jensen-Shannon (S) (Fig.2.3b).

Las distribuciones de probabilidad que obtenemos siguen una distribución con ley de potencia. Esta clase es común observarla en fenómenos de interacción de "mundo-real" sin embargo no pueden ser caracterizadas fácilmente como otras distribuciones también de "mundo-real", como por ejemplo, las acampanadas (Clauset, Shalizi, y Newman, 2009). En el comportamiento de S, no se observa un patrón claro que brinde información relevante sobre la clasificación del sistema. Claramente no existen mínimos globales, sino que las relaciones de distancia parecieran no tener un patrón marcado. Esta herramienta de medición mediante distribuciones de probabilidad no captura los posibles patrones emergentes del sistema deportivo, por lo que se cree, consecuentemente, que el método JSD puede no ser una buena elección al medir la semejanza de estas matrices de adyacencia en la reconstrucción de estas redes complejas.

Figura 2.3
Matriz de norma de Frobenius (N) *y de divergencia Jensen-Shannon* (S)



Nota. (a) Matriz de distancias de Frobenius N: Cada elemento se obtiene al aplicar la norma de Frobenius sobre la resta entre la matriz de adyacencia teórica y la correspondiente experimental, nótese la emergencia de cierta región de mínimos. (b) Matriz S: Cada elemento representa la comparación entre dos distribuciones de probabilidad. Toda distribución fue segmentada en 50 Bins.

Ahora, los resultados de las matrices V, F, N y las curvas ROC, nos muestran una zona de acción eficaz donde se recomienda trabajar; un rango de parámetros aptos para recuperar la mayor eficiencia en la captura de enlaces. Esta zona se define espacialmente desde $\rho > 0.1(m)$ y temporalmente desde $\tau > 2(s)$. Dicha zona, omite potenciales capturas erróneas debido a la resolución espacial de los datos de rastreo utilizados (Linke y cols., 2020), y concuerda con estudios donde el tiempo promedio de los pases entre jugadores se establece al rededor de $2(s)$ (Goes, Kempe, y cols., 2019).

Para cada equipo que participe en un determinado partido, existiría una combinación de parámetros (ρ, τ) que describiría mejor su respectiva red de pases. Sin embargo, al estar trabajando principalmente con un estado Macro generado por la agregada de ocho instancias deportivas, se considera que es posible establecer un valor fijo de parámetros de forma que el modelo pueda ser utilizado sobre un partido de fútbol sin conocimiento previo de los eventos que lo definen. Es por esto que se propone la siguiente metodología que planea concretar la búsqueda de la tupla óptima utilizando las matrices creadas sobre el espacio de configuración (ρ, τ).

2.3. Criterios para la elección de la tupla óptima

De los resultados anteriores, se obtuvo un rango de interés demarcado por los mínimos de la matriz de la norma de Frobenius N (Fig. 2.3a), dentro de ella se ubica la tupla óptima (ρ', τ') que maximiza la eficiencia en la captura de pases. Para hallarla nos centramos en la matriz de verdaderos positivos V y la matriz de falsos positivos F, las cuales están directamente relacionadas con

las curvas ROC y representan la eficiencia y errores porcentuales en la identificación de pases, respectivamente.

El criterio principal para la elección se centra en maximizar el porcentaje de capturas correctas a la vez que se minimiza el porcentaje de capturas incorrectas. En otras palabras, necesitamos obtener la mayor cantidad de vp a la vez que se obtiene la menor cantidad posible de fp. Recurrimos entonces a restar las matrices V y F obteniendo una nueva matriz residual R (Fig. 2.4) donde se expresa la resta de las 121 tuplas (ρ, τ).

La magnitud en los elementos de la matriz R representan la diferencia matemática entre el clasificador binario de verdaderos positivos y el clasificador de falsos positivos. Esto significa que entre mayor sea la brecha entre ellos, se conserva una relación de mayor cantidad de capturas correctas, a la par de una menor cantidad de capturas incorrectas.

$$R = V - F \quad \text{donde} \quad r_{\rho\tau} = vp_{\rho\tau} - fp_{\rho\tau} \tag{2.1}$$

En esta matriz R nos interesan los valores (ρ, τ) que maximicen la diferencia $r_{\rho\tau}$. Se obtuvo una región de máximos que coincide con la región de valores óptimos en la matriz de Frobenius. Se observa que en la región extrema de $\rho = 0.0(m)$ se hallan diferencias negativas la cuales expresan que en dicha zona la cantidad de pases erróneos capturados superan aquellos correctamente identificados. Dentro del rango de interés encontramos un elemento que representa el máximo global ubicado en la combinación de parámetros $\rho = 0.3(m), \tau = 3.0(s)$. Esta tupla se establece entonces como el par de restricciones (ρ', τ') que maximiza la eficiencia en la captura de enlaces dentro de nuestro modelo de reconstrucción.

La matriz de residuo R cumple una función muy importante en el desarrollo de este proyecto. Esta cuantifica de forma directa la relación de compromiso entre el alto porcentaje de capturas correctas y el bajo porcentaje de capturas incorrectas que hemos venido buscando. Dentro de R, se identifica la combinación de parámetros óptimos (ρ', τ'), ya que esta garantiza el máximo funcionamiento del modelo creado, se hace explicita su definición, con el fin de ubicarla inequívocamente dentro del espacio de configuración (ρ, τ).

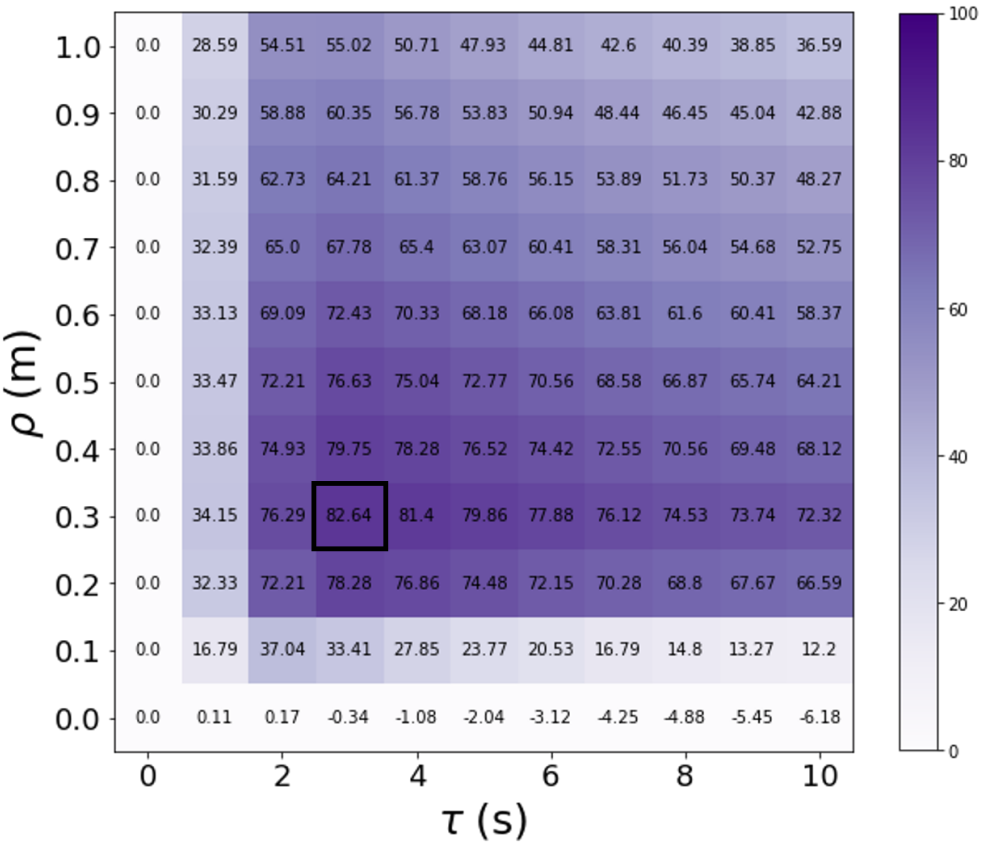
Definición 2 *La tupla óptima (ρ', τ'), es un par ordenado cuyos elementos representan la restricción espacial ρ y la restricción temporal τ que identifican el elemento de la matriz residual R de mayor magnitud.*

$$(\rho', \tau') := (\rho, \tau) \quad | \quad r_{\rho\tau} = \max(R)$$

Pueden haber casos en los cuales más de un elemento de la la matriz R cumplan con la condi-

ción de tupla óptima, en estos casos se propone dirigirse a la matriz de Frobenius (N) y tomarla como segundo filtro. Es decir, en caso de existir varios elementos óptimos con la misma magnitud en la diferencia entre capturas correctas e incorrectas, aquel que posea la menor distancia en la norma de Frobenius se alzará como tupla óptima.

Figura 2.4
Matriz de Residuos R



Nota. La matriz de Residuos representa el primer filtro para la elección de la combinación óptima de parámetros. Se obtiene al restar la matriz de Verdaderos Positivos y la matriz de Falsos Positivos. En el recuadro negro se presenta el elemento óptimo (ρ, τ) que maximiza la eficiencia en la captura de enlaces.

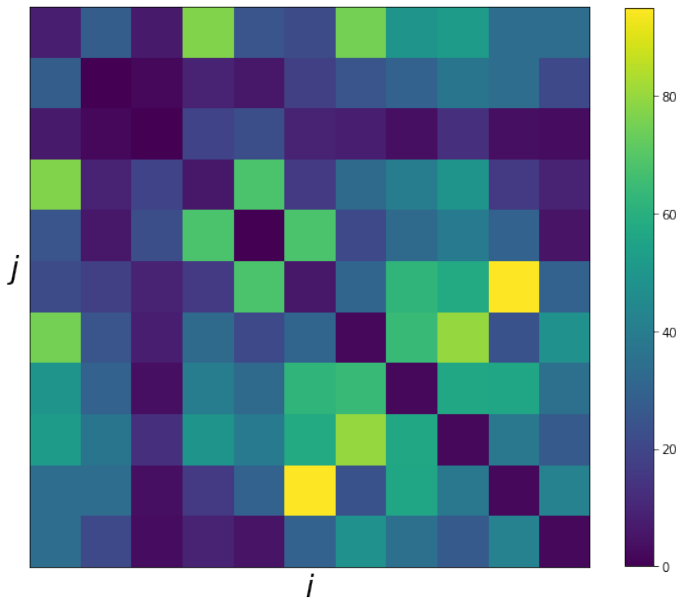
Para el arreglo Macro, la combinación de parámetros óptimos que maximiza la eficiencia en nuestro modelo es $\rho' = 0.3(m), \tau' = 3(s)$. Con base en que este sistema se conforma como la agregada de ocho instancias individuales, la tupla óptima obtenida representa un resultado general. Se recomienda utilizarla como parámetros predeterminados a la hora de aplicar el modelo físico propuesto, en caso de querer construir una red de pases sin conocimiento previo de datos de eventos como referencia teórica. Se debe tener en cuenta que al aumentar el banco de datos de rastreo y de eventos sincronizados estos valores podrían variar. La tupla óptima $\rho' = 0.3(m), \tau' = 3.0(s)$ presenta un porcentaje de verdaderos Positivos de $93.99\,\%$, y un porcentaje de Falsos Positivos de $11.34\,\%$. Su posición en el espacio ROC la conforma una sensibilidad de 0.940 y un 1-especificidad de 0.104, mientras que su distancia en la norma de Frobenius corresponde a 33.5, la cual corresponde al mínimo global en el espacio de configuración.

Los enlaces capturados conforman la red de pases que caracteriza la dinámica del sistema, por lo tanto la eficiencia definitiva del modelo se establece mediante comparaciones de los coeficientes de red. En este orden de ideas, se presenta la red de pases con base en las capturas de la tupla óptima y el correspondiente cálculo de los coeficientes de red, en contraste con sus equivalentes teóricos obtenidos de la matriz T.

2.4. Reconstrucción de la red compleja

A la tupla óptima $(\rho' = 0.3(m), \tau' = 3(s))$ le corresponde una única matriz de adyacencia $W^{(0.3,3)}$ de una red compleja (Fig. 2.5). Esta matriz representa la mejor reconstrucción experimental de la red de pases, respecto a la original, que nuestro modelo puede reproducir con los datos disponibles. Aquí, cada elemento matricial no nulo representa la existencia de un enlace en la red, en otras palabras un pase entre dos jugadores ij, mientras que el peso del enlace identifica la cantidad de pases entre dichos dos jugadores ij. En el sistema Macro estudiado se capturó un total de 3714 pases.

Figura 2.5
Matriz de Adyacencia óptima $W^{(0.3,3)}$



Nota. Matriz de adyacencia óptima $W^{(0.3,3)}$: Todo elemento matricial representa la existencia de enlaces (pases) entre dos jugadores ij. Su peso identifica la cantidad de pases realizados entre ellos.

Como es de esperarse, nuestra matriz de adyacencia experimental es simétrica debido a que la red en consideración es no direccionada. Sin embargo, esta no posee una diagonal nula. Se entiende, entonces, que el modelo acepta *self-loops* entre agentes. En otras palabras, el modelo creado acepta auto-pases de jugadores, siendo estos definidos como una recuperación del balón por un

mismo jugador. Si bien la cantidad de enlaces capturados bajo esta categoría es muy baja y no afecta significativamente en la dinámica de la red, esto representa un fenómeno digno de estudiar en trabajos posteriores, agregando restricciones adicionales como por ejemplo: el tiempo de posesión de la pelota dentro de las zonas de contacto por los jugadores.

Con base en esta matriz de adyacencia $W^{(0.3,3)}$ se calculan los coeficientes de red, como se expresa en los métodos de medición de eficiencia presentados en secciones anteriores. Desde el punto de vista de la Física de Redes Complejas, los coeficientes de red identifican las propiedades emergentes generadas por las interacciones entre los agentes del sistema. La forma de medir la eficiencia con estos coeficientes es compararlos con los respectivos coeficientes extraídos de la red de pases teórica T.

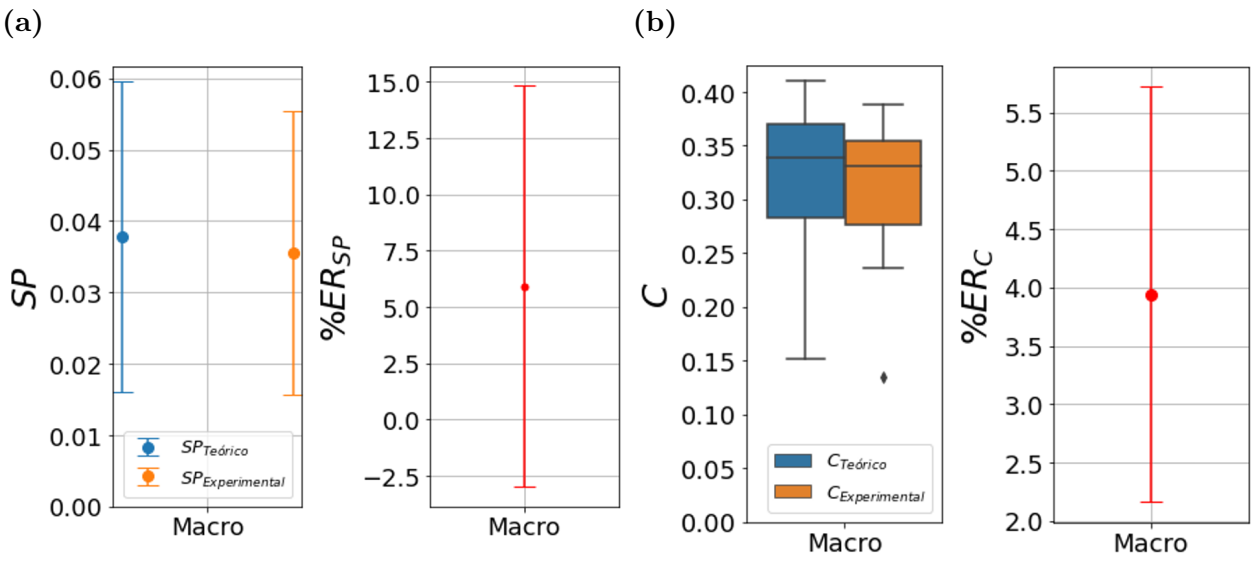
El panel izquierdo de la Figura (2.6) muestra el resultado del primer coeficiente de red calculado: el *Shortest Path* (sp). El sp calcula el camino más corto para la transmisión de información dentro de la red, un coeficiente bajo significa una menor cantidad nodos por los que pasa la información en camino a su destino. En los resultados, tanto la magnitud del promedio teórico como la magnitud del promedio experimental son cercanas a cero, lo que implica que esta red presenta una buena integración entre sus jugadores, es decir que con un bajo número de pases el balón puede desplazarse entre cualesquiera dos jugadores del equipo. Más allá de la cifra obtenida en sí, nos interesa la comparación entre estos valores. La Figura (2.6a) muestra cómo el algoritmo reproduce de forma certera, no solo el $\langle sp \rangle$ de la red (Eq. 1.16), sino también su desviación estándar σ_{sp}. Se puede señalar que el modelo captura el comportamiento a nivel global de la red, pero para comprobarlo nos debemos remitir también a la comparación del coeficiente sp a nivel de individuos (nodos).

Tanto el *shortest path* promedio teórico como el experimental son el resultado de los *shortest path* de todos los posibles pares de jugadores ij (Eq. 1.16). Como las dos matrices de adyacencia contienen el mismo número de nodos, la cantidad de pares (ij) es la misma para ambas redes lo que permite hacer una comparación biyectiva entre pares de nodos. El panel derecho de la Figura (2.6a) presenta el porcentaje de error relativo ($\%ER_{sp}$) en la reconstrucción de estos *shortest path* individuales.

$$\%ER_{sp_{ij}} = \frac{sp_{ij} - sp_{ij}^{(\rho',\tau')}}{sp_{ij}} \tag{2.2}$$

Donde sp_{ij} representa el *shortest path* entre dos nodos ij dentro de la red original. Lo que se observa es que el modelo, bajo la aplicación de la tupla óptima, reproduce certeramente el *shortest path* promedio de la red y también lo hace a nivel individual con un error relativo promedio de $5.922 \pm 8.899\,\%$. La baja magnitud de $\%ER_{sp}$ confirma que se puede recuperar con alta exactitud el coeficiente de integración de la red original, incluso con el limitado *set* de datos disponible. La magnitud de $\sigma_{\%ER_{sp}}$ implica la necesidad de *set* de datos con más partidos.

Figura 2.6
Reconstrucción de los coeficientes de red (Shortest Path y Clustering)



Nota. (*a*) Comparación de *Shortest Path*: El *shortest path* promedio y su desviación estándar son calculados mediante los *shortest path* individuales de todos los agentes en el sistema. (*b*) Comparacion de *Clustering*: Distribución de valores de *clustering* compuesta por los coeficientes de los 11 jugadores. En ambos coeficientes el valor teórico es definido mediante la red construida por los pases de *events*. Los valores experimentales corresponden a los resultados de la matriz de adyacencia óptima.

En el cálculo del coeficiente de *Clustering* ocurre algo similar. En este caso no se promediaron los valores de los pares ya que al ser una medida de segregación, cada jugador tiene un valor específico que mide la facilidad de formar triángulos (Fig. 1.6). En el panel izquierdo de la Figura (2.6b), se muestran los *boxplots* que definen la distribución del *clustering* de los 11 jugadores para la red teórica y experimental. Se observa una alta semejanza desde el comportamiento de las distribuciones hasta el valor de la media recuperada. Nuevamente, la forma de medir dicha semejanza se hizo mediante el cálculo del error relativo jugador por jugador, en este caso definida por la forma $\%ER_C$.

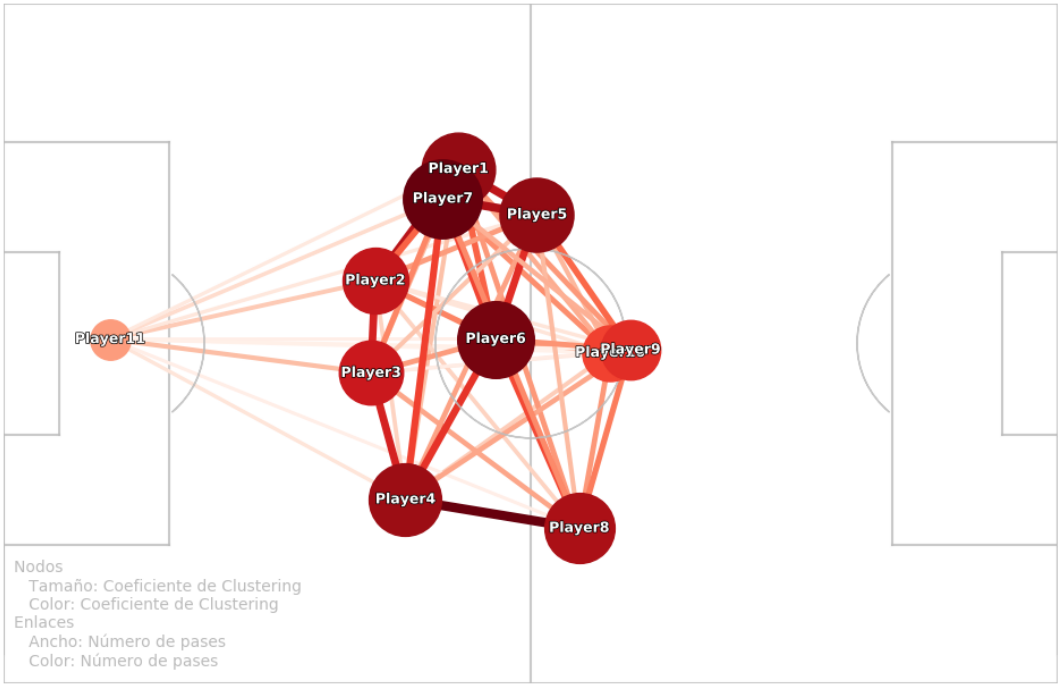
$$\%ER_{C_i} = \frac{c_i - c_i^{(\rho',\tau')}}{c_i} \tag{2.3}$$

El error obtenido en la reconstrucción del coeficiente de segregación es de $3.942 \pm 1.777\,\%$. Con base en estos resultados, se observa la buena reproducción en las características que definen la estructura del sistema real. Si bien los resultados son alentadores, hay que tener en cuenta dos elementos. El primero es que la eficiencia del modelo podría mejorar al incrementar la muestra estadística. El segundo se relaciona con la disminución de los errores relativos de los coeficientes. Estos pueden ser disminuidos al implementar en el modelo condiciones adicionales cómo el tiempo de contacto entre los jugadores y el balón.

Resumiendo el proceso, es necesario obtener las matrices de clasificación binaria de verdaderos positivos V, y de falsos positivos F, a la par de la matriz de distancia en la norma de Frobenius N para poder establecer una combinación de parámetros óptimos (ρ', τ') bajo la cual se usa la matriz de adyacencia correspondiente para calcular los coeficientes de red. Estos resultados se recopilan en la visualización de la red (Fig. 2.7).

Figura 2.7

Visualización de la red compleja asociada al sistema agregado Macro



Nota. La red de pases fue construida bajo los resultados de la tupla óptima. Existen un total de once nodos cuyo color y diámetro se relaciona con su correspondiente coeficiente de *clustering* mientras que el ancho y color de los enlaces indican el peso de los pases entre cuales quiera dos jugadores.

Para la visualización, los nodos son ubicados en el espacio real de la cancha con dimensiones de $65(m) \times 108(m)$, en donde la posición de cada uno corresponde al promedio de las posiciones cuando un pase fue realizado. Esto permite identificar, en diferentes estudios, las posiciones de: arqueros, defensas, medio campistas y demás. Si bien los nodos se representan como una circunferencia, estas poseen diferentes diámetros y colores que representan los coeficientes de *clustering* para cada uno de los jugadores, brindándonos información visual sobre el nivel de segregación de los equipos. Para el caso de los enlaces, estos identifican los pases existentes entre todos los nodos, información que proviene de los elementos en la matriz de adyacencia; en este sentido su color y el ancho de la linea representa la cantidad de pases que se realizaron entre los jugadores, mostrando así cuáles son las interacciones predominantes en el partido, y que jugadores son más relevantes a la hora de distribuir la pelota en el campo.

2.5. Instancias individuales

Los resultados recién mostrados corresponden al sistema "Macro", sin embargo es necesario evaluar el comportamiento del modelo a menor escala. Los siguientes resultados corresponden a las ocho instancias individuales que se presentaron en secciones anteriores (Fig. 1.1b). Para cada instancia, se implementó la metodología de selección de tupla óptima. Primero se construye el espacio de configuración (ρ, τ) en la matriz de Verdaderos Positivos V, la matriz de Falsos Positivos F y las curvas ROC. Se genera la matriz de residuos R y se selecciona la tupla que maximice la diferencia (ρ', τ'). Si hay varios máximos globales, se hace uso de las distancias de la Norma de Frobenius para discernir entre ellos. Luego, se obtiene la matriz de adyacencia experimental correspondiente a la tupla óptima, con base en esta, se reconstruyen los coeficientes de red (sp y c).

En la Tabla (2.1), se presenta un resumen de resultados finales para las ocho instancias. Se exponen en primera mano la combinación de parámetros óptima, y su correspondiente porcentaje de Verdaderos Positivos, el porcentaje de Falsos Positivos, la norma de Frobenius y las relaciones de las curvas ROC. Las matrices V, F, R, N del espacio de configuración del cual se obtuvo la tupla óptima de cada instancia se visualizan en el Apéndice (A).

Tabla 2.1

Instancia	ρ' **(m)**	τ' **(s)**	$vp\,\%$	$fp\,\%$	**Frobenius**	**ROC**	
						FPR	**VPR**
1^{ra}	0.3	3	91.98	16.04	10.198	0.113	0.920
2^{da}	0.3	3	94.40	8.00	8.124	0.093	0.944
3^{ra}	0.3	3	93.96	13.74	10.296	0.097	0.940
4^{ta}	0.3	3	95.00	9.44	7.746	0.082	0.950
5^{ta}	0.3	3	97.63	9.83	9.487	0.115	0.976
6^{ta}	0.3	3	97.58	10.08	8.718	0.100	0.976
7^{ma}	0.4	3	97.12	10.70	11.045	0.121	0.951
8^{va}	0.3	3	87.64	12.36	7.483	0.061	0.921

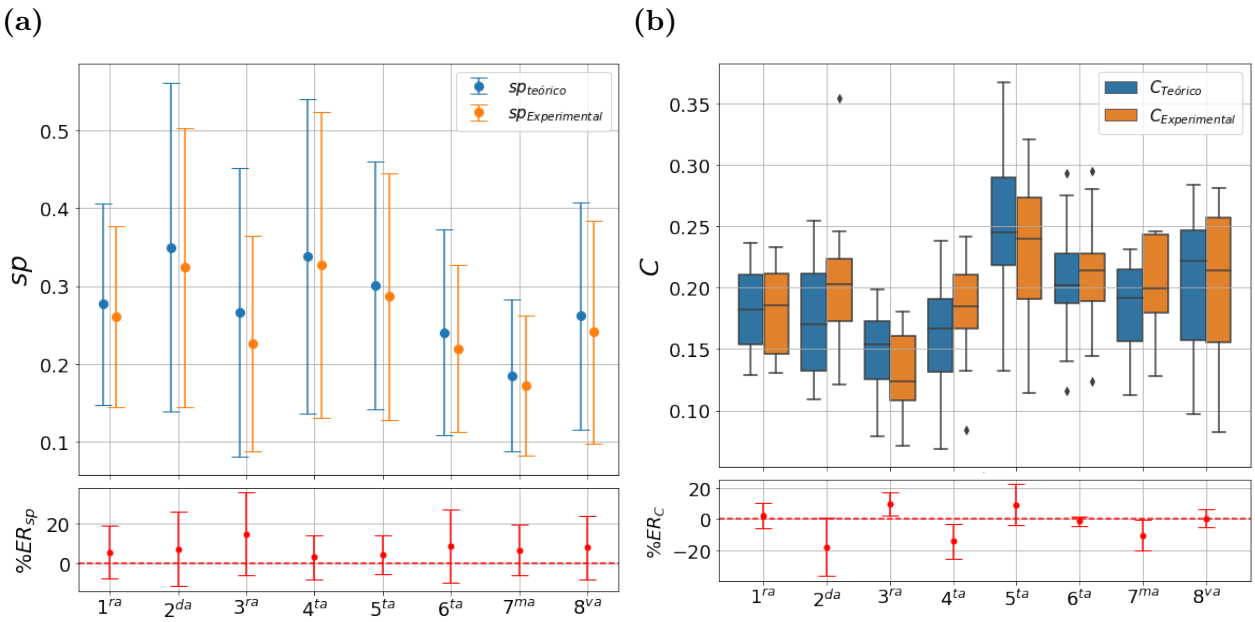
Esta tabla muestra el panorama global de la elección de parámetros óptimos a nivel de instancias. Es de resaltar que la mayoría de las tuplas elegidas por la metodología concuerdan con los valores de $\rho = 0.3(m), \tau = 3.0(s)$. Desde la perspectiva de un sistema físico, cada equipo tiene varias estrategias y desempeños dependiendo de sus entrenamientos. Esto implica que el tiempo de vuelo en los pases, y el alcance de los jugadores al balón varían también de partido a partido. El resultado de la constante selección de $\rho = 0.3(m), \tau = 3.0(s)$ como tupla óptima, da un indicio de que bajo nuestro modelo, esta tupla específica puede reconstruir la red de pases con alta relación de compromiso entre sensibilidad y 1-especificidad, independientemente del estilo de juego de los equipos.

Vemos cómo por lo general, el porcentaje de enlaces capturados que sí están en el sistema real (vp) supera el $92\,\%$. Por otro lado, el porcentaje de enlaces capturados que no existen en el sistema real (fp) se mantiene por lo general por debajo del $14\,\%$. Las distancias en la norma de Frobenius distan poco del cero, como es de esperarse al comparar dos sistemas matriciales semejantes. Y fi-

nalmente las relaciones del espacio ROC demarcan siempre alta sensibilidad y baja especificidad, denotando así la relación de compromiso entre capturas correctas e incorrectas que hemos venido mencionando. Luego de obtener las combinaciones de parámetros óptima podemos generar la matriz de adyacencia para cada instancia y con base en ellas los respectivos coeficientes de red. Estos valores se encuentran visibles en la Figura (2.8).

Figura 2.8

Reconstrucción de los coeficientes de red en las instancias Individuales



Nota. (*a*) Comparación del coeficiente *Shortest Path* (ocho instancias): El valor promedio y la desviación es calculado a partir de los coeficientes individuales de *shortest path* de todos los jugadores para cada una de las instancias. (*b*) Comparación del coeficiente *Clustering* (ocho instancias): La distribución se hace con base en los coeficientes de *clustering* individuales de los 11 jugadores, para cada instancia.

La aplicación del modelo sobre las instancias individuales nos facilita la observación del funcionamiento de este. En la Figura (2.8a) se observan varios elementos. De primera mano, se aprecia cómo el modelo reproduce el *shortest path* promedio $\langle sp \rangle$ en las ocho redes estudiadas, con un error relativo promedio de $7.46 \pm 14.94\,\%$. La totalidad de estos coeficientes poseen una magnitud menor a su par teórico, lo que puede indicar que nuestro modelo posee una tendencia a subestimar el valor promedio del sp de red, comportamiento que no se deducía en el resultado "Macro". Por otro lado, en el caso de los resultados de *Clustering*, las distribuciones de los coeficientes parecen corresponder con los coeficientes de segregación esperados, los cuales se reconstruyen con un error relativo promedio de $8.13 \pm 9.58\,\%$.

Se puede afirmar entonces que siendo una primera aproximación a la reconstrucción de redes complejas en este tipo de sistemas sociales, el modelo creado es altamente eficiente. Usando una combinación óptima de parámetros, que pudiera ser $\rho = 0.3(m), \tau = 3.0(s)$ de forma general, se puede reconstruir tanto la matriz de adyacencia de los partidos a estudiar como los coeficientes de red que definen las características de integración y segregación en esta; teniendo en cuenta que se

puede incurrir en una leve subestimación de dichos coeficientes.

3. Discusión

La creación de modelos para la identificación de interacciones entre los agentes de una red es uno de los grandes retos en la Ciencia de Redes. La incursión en esta problemática puede ser llevada a cabo especialmente en la red social de sistemas deportivos, como el fútbol, donde las interacciones son mediadas por el balón en juego y este puede ser rastreado. Con base en esto, se hipotetiza la existencia de una red que puede ser reconstruida a partir de los datos de rastreo (*tracking*) de estos sistemas. En este proyecto se apunta a responder si es posible extraer dicha red compleja desde el rastreo en bruto de los agentes en el tiempo. Por lo tanto, se introduce un nuevo modelo de "captura de pases" con base en consideraciones físicas para identificar los enlaces entre los jugadores y mediante estos reconstruir la red deportiva.

El modelo propuesto trabaja bajo la aplicación de una intuitiva condición espacial (ρ), y temporal (τ), donde variaciones en las combinaciones de estos conllevan a diferentes niveles de eficiencia en la estructura de la red reconstruida. Los resultados obtenidos en este proyecto son alentadores. Primero, desde un punto de vista computacional, el modelo se desarrolló en el lenguaje de programación *Python v.3.7.7* donde consume un tiempo de computo de tres (3) horas para el análisis de una instancia monitoreada a 25 *fps*. Al finalizar, el algoritmo devuelve la elección de una tupla óptima (ρ', τ'), bajo la cual se obtiene la mayor eficiencia en la reconstrucción de la red. También captura los respectivos valores asociados del porcentaje de pases capturados correcta, e incorrectamente, asociados a la matriz de adyacencia de la red compleja, y sus coeficientes de red: *shortest path* y *clustering*, que la definen.

Segundo, tanto para el sistema Macro agregado, como para las ocho instancias individuales de los encuentros, la selección de una tupla óptima se decantó por la combinación $\rho' = 0.3(m)$ y $\tau' = 3.0(s)$. Estos resultados concuerdan con indicadores en la literatura asociados justamente a la naturaleza de nuestros parámetros. Por un lado, se encuentran estudios donde la duración promedio de los pases en partidos de fútbol de alto nivel es de dos (2) segundos (Goes y cols., 2019). Que nuestro modelo indique como resultado un umbral de tiempo de vuelo de tres (3) segundos, muestra que este tiene potencial de recuperar condiciones reales del sistema que estudiamos al no sobrestimar el tiempo prudente del pase, aún teniendo en cuenta la escasez de datos actual. Por otro lado, en el ámbito espacial, las mediciones de *tracking* incurren en un error máximo en el seguimiento posicional del balón y de los jugadores, correspondiente a aproximadamente $10(cm)$, debido a la precisión de los instrumentos de visión artificial (Linke y cols., 2020). En nuestro modelo, esto corresponde con el rango $\rho = 0.1(m)$, el cual sobresale notoriamente en los resultados

del espacio de configuración de capturas incorrectas F, y en el aumento abrupto de 1-especificidad en las curvas ROC. Esta concordancia llama la atención porque se señala sutilmente que el modelo toma en cuenta de forma automática las consideraciones en la precisión de los datos utilizados, mostrando así, regiones de alto fallo que deben ser evitadas.

Tercero, la importancia del modelo creado se centra en la capacidad de reproducción de una red compleja mediante la aplicación de condiciones físicas intuitivas y elementales (posición y tiempo). Cuarto, la eficiencia del proceso de captura de pases también se basa en la comparación de los coeficientes de red, ya que estos valores definen la estructura y el transporte de información dentro de ella. Ahora, que los errores relativos obtenidos en esta reconstrucción tanto para el *shortest path*, como para el *clustering*, sean menores al $10\,\%$ señalan un gran éxito en esta primera aproximación, estableciendo un trabajo novedoso e innovador en el contexto de la socio-física (en particular aplicada a sistemas de competición) desde la perspectiva de la Ciencia de Redes.

Finalmente, se confirma que el modelo físico funciona y cumple el fin de reconstruir una red social bajo evolución temporal con alto nivel de eficiencia. Cabe resaltar que un objetivo a largo plazo de este proyecto es desarrollar la capacidad de recuperar una red de pases sin el conocimiento previo de los datos teóricos de eventos. Para esto es necesario encontrar rangos de una tupla de parámetros canónicos que maximicen la efectividad del modelo sobre cualquier partido arbitrario de fútbol. Si bien en estos casos se alienta a usar el modelo bajo las condiciones óptimas (ρ', τ') ya encontradas, queda abierta la ventana a trabajos posteriores enfocados a mejorar la precisión sobre la elección de esta tupla. Esto sería posible mediante: el incremento en la estadística de datos proporcionados, la implementación de parámetros adicionales sobre el modelo, como por ejemplo el tiempo de contacto entre balón-jugador, y/o re-evaluar los resultados sobre el mismo espacio de configuración haciendo uso de una resolución más fina.

4. Conclusiones

Como respuesta a la pregunta de investigación de este proyecto se concluye que sí es posible extraer la red de un sistema aislado a partir del rastreo en bruto de las interacciones de los agentes en el tiempo. Específicamente, es posible reconstruir la red de pases de los sistemas de fútbol a partir de los datos de rastreo por visión artificial (*tracking*). Este proyecto representa un trabajo innovador y de alto impacto a nivel internacional, que cuenta con el potencial para abrir puertas a investigaciones futuras en la creación automatizada de enlaces en sistemas complejos.

Se creó un modelo funcional cuya potencia se basa en la capacidad de reconstrucción de las redes de pases con alta efectividad haciendo uso, únicamente, de dos condiciones físicas (ρ, τ). Donde ρ representa una restricción espacial en la zona de contacto de los jugadores con el balón, y τ una restricción temporal en el tiempo de vuelo del balón entre dos jugadores. La aplicación del modelo, bajo una combinación de parámetros óptimos, presenta una alta fiabilidad con un porcentaje de capturas correctas de enlaces de hasta $93.99\,\%$ y un rango bajo de error con un porcentaje de capturas incorrectas de hasta $11.34\,\%$. El modelo también reproduce de forma acertada la dinámica de la red compleja a nivel global e individual de los jugadores del sistema. Esto es apreciable en los bajos errores relativos de los coeficientes de red de integración (*Shortest Path*) con $7.46\,\%$ y de segregación (*Clustering*) con $8.13\,\%$.

Se extrae de los resultados una combinación de parámetros óptimos correspondientes a $\rho = 0.3(m)$ y $\tau = 3.0(s)$, los cuales concuerdan con criterios propios de las practicas deportivas. Para futuros análisis de encuentros futbolísticos, en los que no se tengan conocimientos de los eventos en el juego, se recomienda usar el modelo bajo esta combinación encontrada en aras de crear la red de pases del sistema.

Finalmente, en el caso de los métodos para el cálculo de eficiencia se indica que la divergencia de Jensen-Shannon puede no ser una buena herramienta para medir la semejanza de las matrices de adyacencia. Se aprecia que este método no captura la sensibilidad en la variación (ρ, τ) del espacio de configuración debido a sus resultados homogéneos. Se señala, entonces, la apertura a futuros trabajos de investigación con foco en la comprobación del modelo mediante diferentes técnicas estadísticas de alto orden.

Este proyecto cumplió con el objetivo de extender los campos de estudio actuales de la física en la Universidad Industrial de Santander, y en Colombia, siendo, además, el primer trabajo en

identificación y reconstrucción de enlaces en la Física de sistemas complejos a nivel internacional, al generar un modelo con alta eficiencia en la captura de pases. Adicionalmente, ha captado el interés para futuras colaboraciones con grupos de investigación internacionales activos en Física de sistemas complejos, como el Grupo Interdisciplinar de Sistemas Complejos (GISC) y el *Complex Systems* & *Sport Analytics* de Madrid-España, quienes se encuentran a la espera de los resultados para una eventual replicación de la metodología con datos oficiales de la liga española.

Referencias

Barabási, A.-L., y cols. (2016). *Network science*. Cambridge university press.

Bialkowski, A., Lucey, P., Carr, P., y cols. (2014). Large-scale analysis of soccer matches using spatiotemporal tracking data. En *2014 ieee international conference on data mining* (pp. 725–730).

Boccaletti, S., Latora, V., Moreno, Y., y cols. (2006). Complex networks: Structure and dynamics. *Physics reports*, *424*(4-5), 175–308.

Buldu, J. M., Busquets, J., Echegoyen, I., y cols. (2019). Defining a historic football team: Using network science to analyze guardiola's fc barcelona. *Scientific reports*, *9*(1), 1–14.

Buldú, J. M., Busquets, J., Martínez, J. H., y cols. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in psychology*, *9*, 1900.

Caicedo-Parada, S., Lago-Peñas, C., y Ortega-Toro, E. (2020). Passing networks and tactical action in football: a systematic review. *International Journal of Environmental Research and Public Health*, *17*(18), 6649.

Clauset, A., Shalizi, C. R., y Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Clemente, F. M., Martins, F. M. L., Mendes, R. S., y cols. (2016). *Social network analysis applied to team sports analysis*. Springer.

Clemente, F. M., y cols. (2015). Using network metrics in soccer: a macro-analysis. *Journal of human kinetics*, *45*(1), 123–134.

Cotta, C., Mora, A. M., Merelo-Molina, C., y cols. (2011). Fifa world cup 2010: A network analysis of the champion team play. *arXiv preprint arXiv:1108.0261*.

Dijkstra, E. W., y cols. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C., y cols. (2018). Quantifying reputation and success in art. *Science*, *362*(6416), 825–829.

Fuglede, B., y Topsoe, F. (2004). Jensen-shannon divergence and hilbert space embedding. En *International symposium oninformation theory, 2004. isit 2004. proceedings.* (p. 31).

Goes, F. R., Kempe, M., y cols. (2019). Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big data*, *7*(1), 57–70.

Golub, G. H., y cols. (1996). Cf vanloan, matrix computations. *The Johns Hopkins*.

Gosak, M., Markovič, R., Dolenšek, J., y cols. (2018). Network science of biological systems at different scales: A review. *Physics of life reviews*, *24*, 118–135.

Grund, T. U. (2012). Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, *34*(4), 682–690.

Hassan, A., Akl, A.-R., Hassan, I., y cols. (2020). Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. *Sensors*, *20*(11), 3213.

Havlin, S., Kenett, D. Y., Ben-Jacob, E., y cols. (2012). Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics*, *214*(1), 273–293.

Herrera-Diestra, J. L., Echegoyen, I., Martínez, J. H., y cols. (2020). Pitch networks reveal organizational and spatial patterns of guardiola's fc barcelona. *Chaos, Solitons & Fractals*, *138*, 109934.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oup Usa.

Kenett, D. Y., y Havlin, S. (2015). Network science: a useful tool in economics and finance. *Mind & Society*, *14*(2), 155–167.

Linke, D., Link, D., y Lames, M. (2020). Football-specific validity of tracab's optical video tracking systems. *Plos one*, *15*(3), e0230179.

Lopes, A. M., y Tenreiro Machado, J. (2019). Entropy analysis of soccer dynamics. *Entropy*, *21*(2), 187.

Maimone, V. M., y Yasseri, T. (2019). Football is becoming boring; network analysis of 88 thousands matches in 11 major leagues. *arXiv preprint arXiv:1908.08991*.

Metrica. (2020). https://github.com/metrica-sports/sample-data.

Moura, F. A., Martins, L. E. B., y cols. (2012). Quantitative analysis of brazilian football players' organisation on the pitch. *Sports biomechanics*, *11*(1), 85–96.

Newman, M. (2018). *Networks*. Oxford university press.

Oliveira, A. P., y Tyler, H. R. (2015). Measurement and comparison of passing networks in collegiate soccer. *Minnesota Journal of Undergraduate Mathematics*, *1*(1).

Onnela, J.-P., Saramäki, J., y cols. (2005). Intensity and coherence of motifs in weighted complex

networks. *Physical Review E*, *71*(6), 065103.

Pappalardo, L., Cintia, P., Rossi, A., y cols. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, *6*(1), 1–15.

Park, K.-J., y Yilmaz, A. (2010). Social network approach to analysis of soccer game. En *2010 20th international conference on pattern recognition* (pp. 3935–3938).

Parker, J. R. (2001). Rank and response combination from confusion matrix data. *Information fusion*, *2*(2), 113–120.

Power, P., Ruiz, H., Wei, X., y cols. (2017). Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. En *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1605–1613).

Rampinini, E., Impellizzeri, F. M., Castagna, C., y cols. (2009). Technical performance during soccer matches of the italian serie a league: Effect of fatigue and competitive level. *Journal of science and medicine in sport*, *12*(1), 227–233.

Rein, R., y Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1–13.

Rubinov, M., y Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, *52*(3), 1059–1069.

Russell, M., Benton, D., y Kingsley, M. (2010). Reliability and construct validity of soccer skills tests that measure passing, shooting, and dribbling. *Journal of sports sciences*, *28*(13), 1399–1408.

Russell, M., Benton, D., y Kingsley, M. (2011). The effects of fatigue on soccer skills performed during a soccer match simulation. *International journal of sports physiology and performance*, *6*(2), 221–233.

Sammut, C., y Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

van Raan, A. F. J. (2013). *Handbook of quantitative studies of science and technology*. Elsevier.

Wäsche, H., Dickson, G., Woll, A., y cols. (2017). Social network analysis in sport research: an emerging paradigm. *European Journal for Sport and Society*, *14*(2), 138–165.

Watts, D. J., y Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.

Wei, X., Sha, L., Lucey, P., y cols. (2013). Large-scale analysis of formations in soccer. En *2013 international conference on digital image computing: techniques and applications (dicta)* (pp. 1–8).

Apéndice A

Resultados de las Instancias Individuales

El modelo creado y la metodología propuesta fueron aplicados a las ocho instancias de forma individual con el fin de conocer en ellas el comportamiento de la eficiencia y a partir de esta construir la red de pases experimental con mayor semejanza a las correspondientes teóricas. Cada instancia posee información correspondiente al medio tiempo de un equipo anonimizado y sobre estas se aplica el modelo y las correspondientes restricciones espaciales y temporales del espacio de configuración (ρ, τ). En la sección de resultados se presenta la tabla (2.1) con el resultado final de las mediciones de eficiencia, aquí en cambio presentamos el espacio de configuración completo del cual se obtuvieron dichas tuplas óptimas.

A continuación se presenta, para cada instancia individual, la matriz de verdaderos positivos V y la matriz de falsos negativos F necesarias para construir la matriz residual R, cuyo elemento de mayor magnitud identifica la combinación óptima (ρ', τ'). Se presenta también la matriz de distancia de Frobenius N, que funciona como segundo filtro en caso de la existencia de varias posibles combinaciones óptimas y adicionalmente se expone la visualización de la red de pases experimental que recupera la mayor eficiencia posible al aplicar el modelo creado.

Figura 5.1
Conjunto de resultados para la primera instancia individual

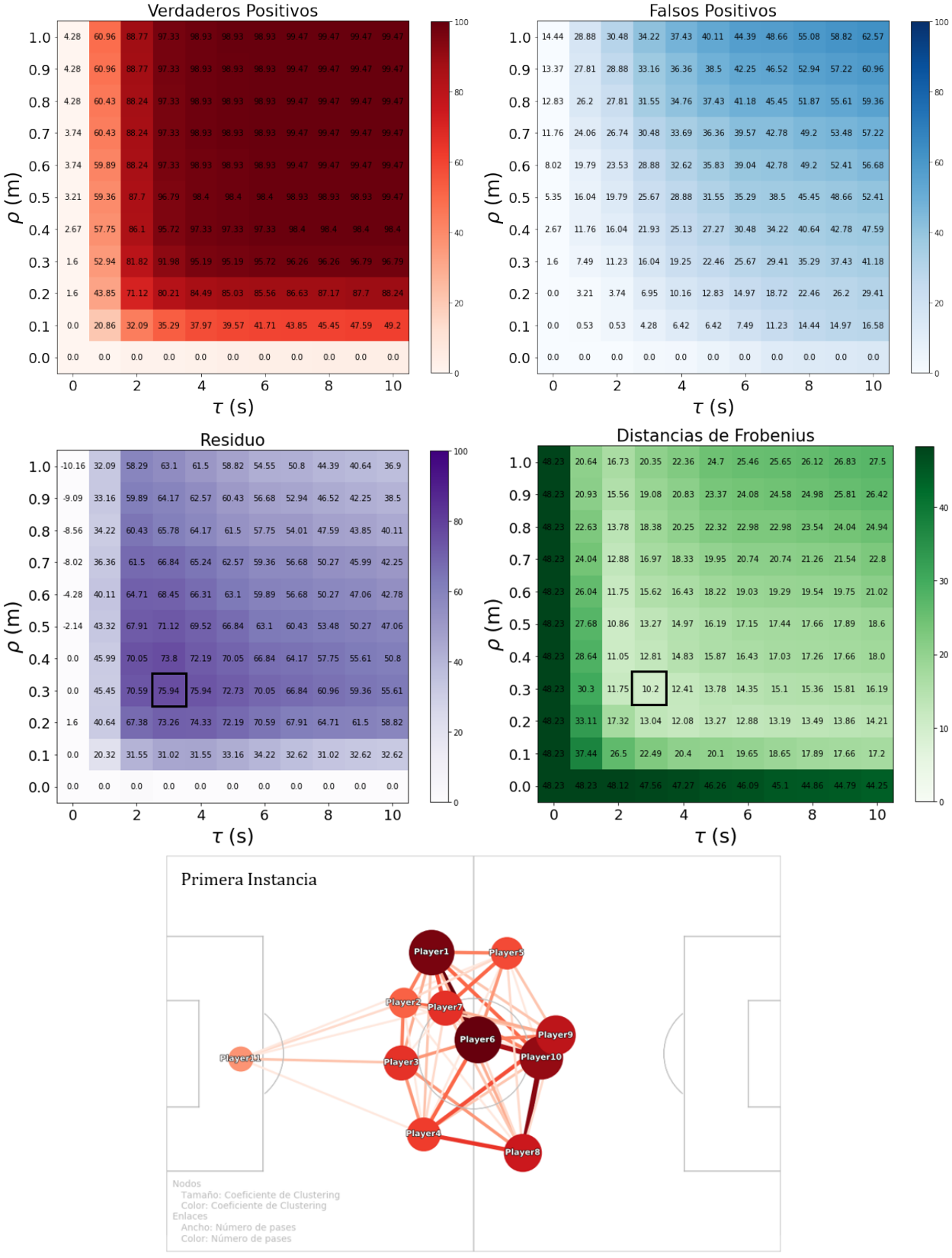


Figura 5.2
Conjunto de resultados para la segunda instancia individual

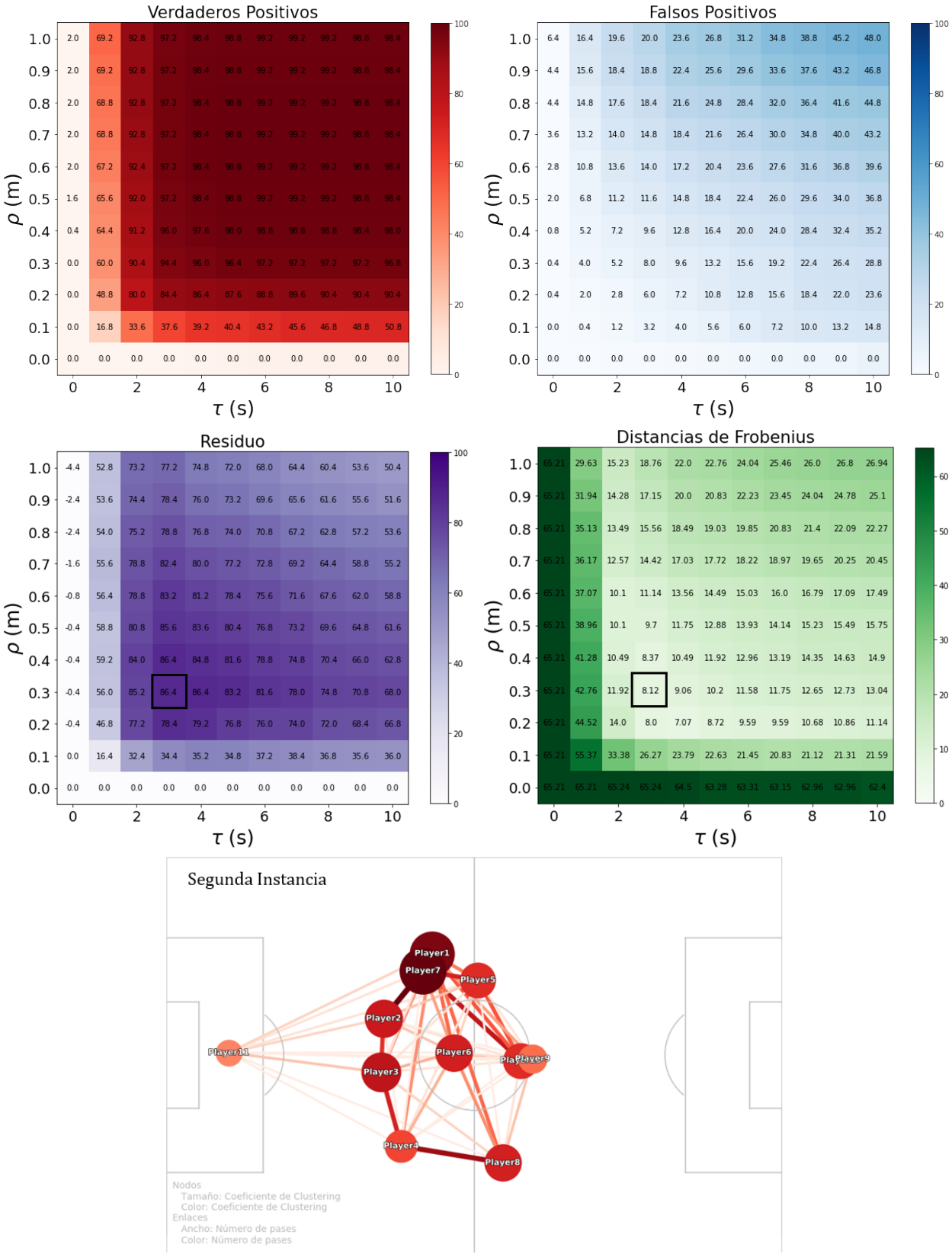


Figura 5.3
Conjunto de resultados para la tercera instancia individual

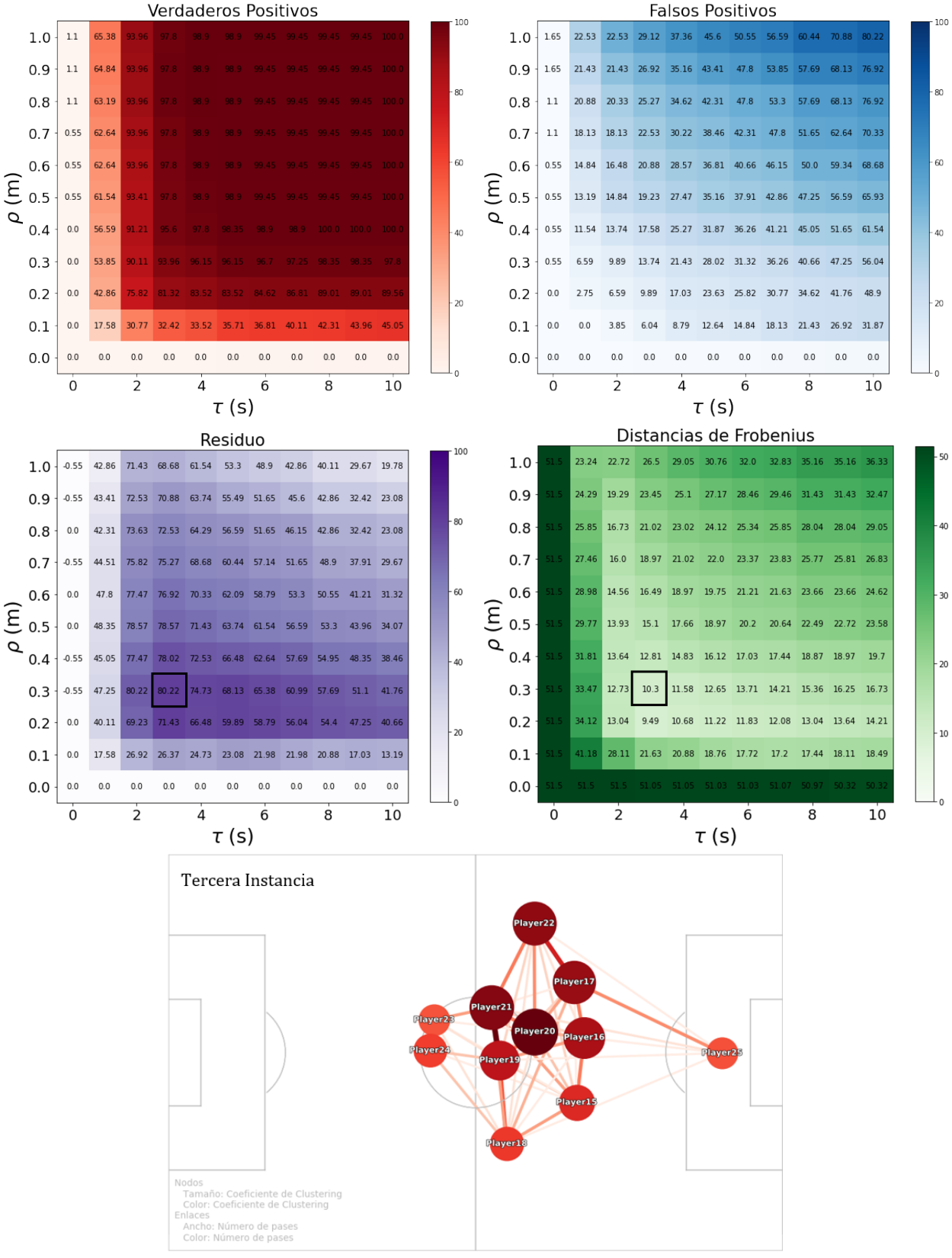


Figura 5.4
Conjunto de resultados para la cuarta instancia individual

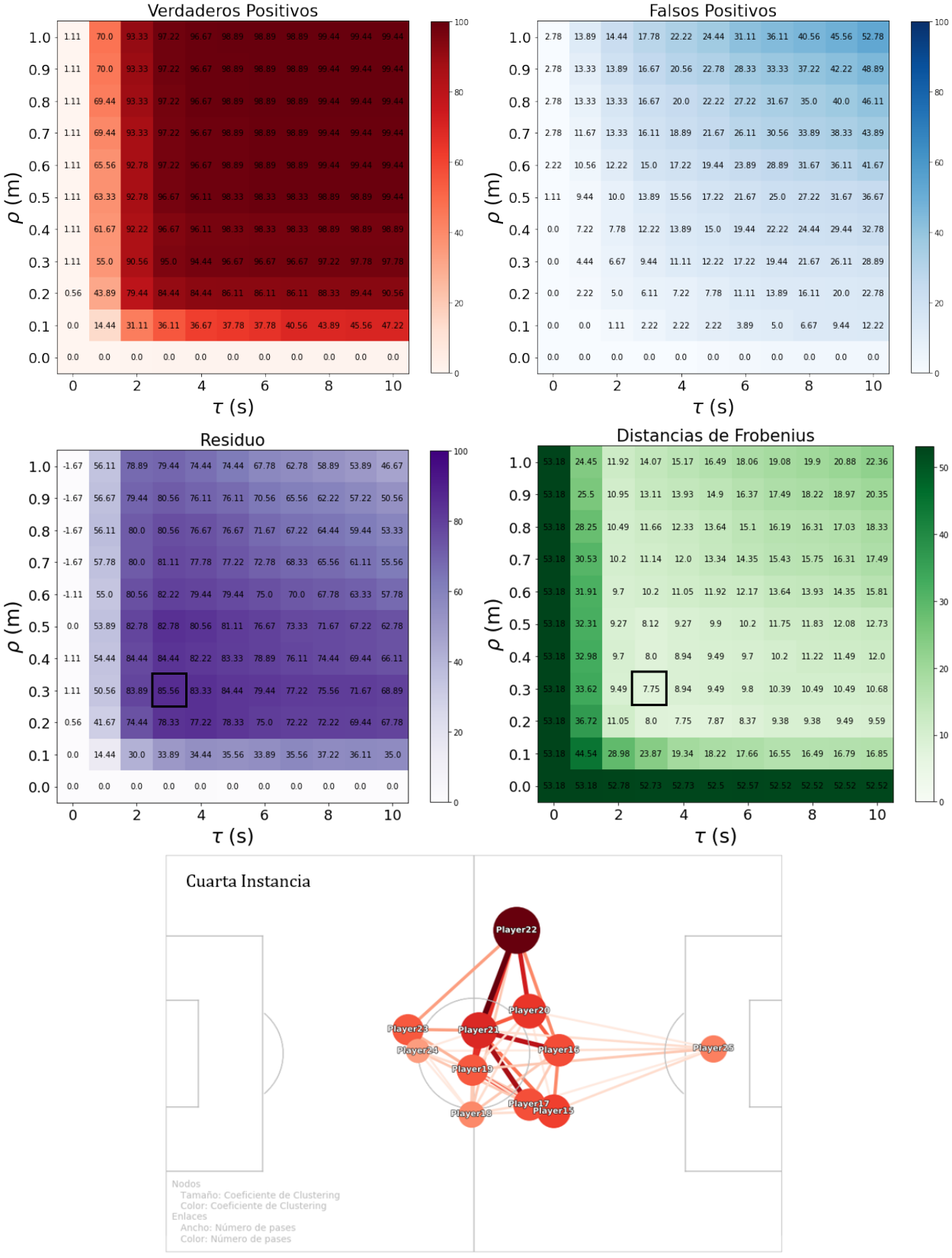


Figura 5.5
Conjunto de resultados para la quinta instancia individual

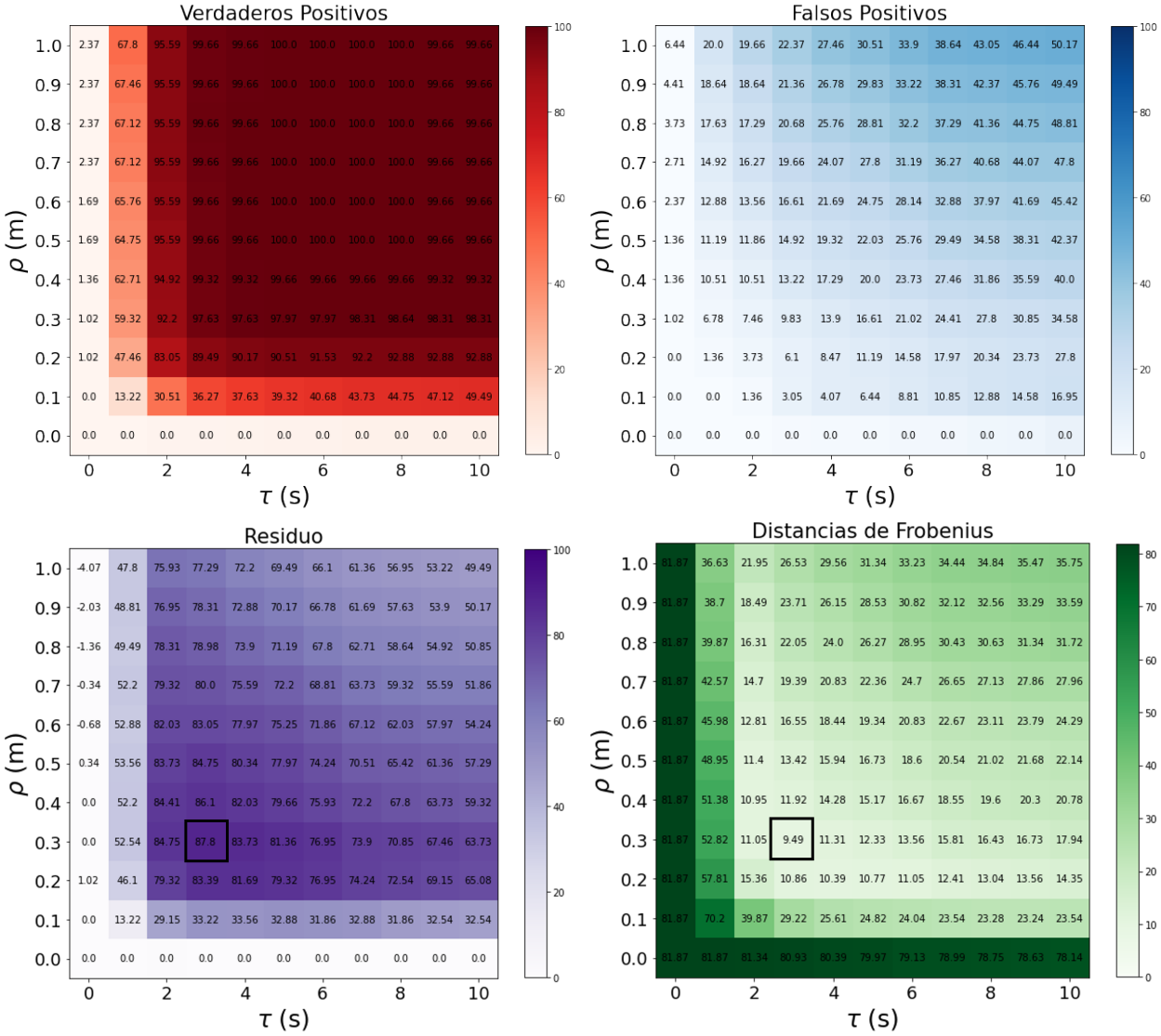


Figura 5.6
Conjunto de resultados para la sexta instancia individual

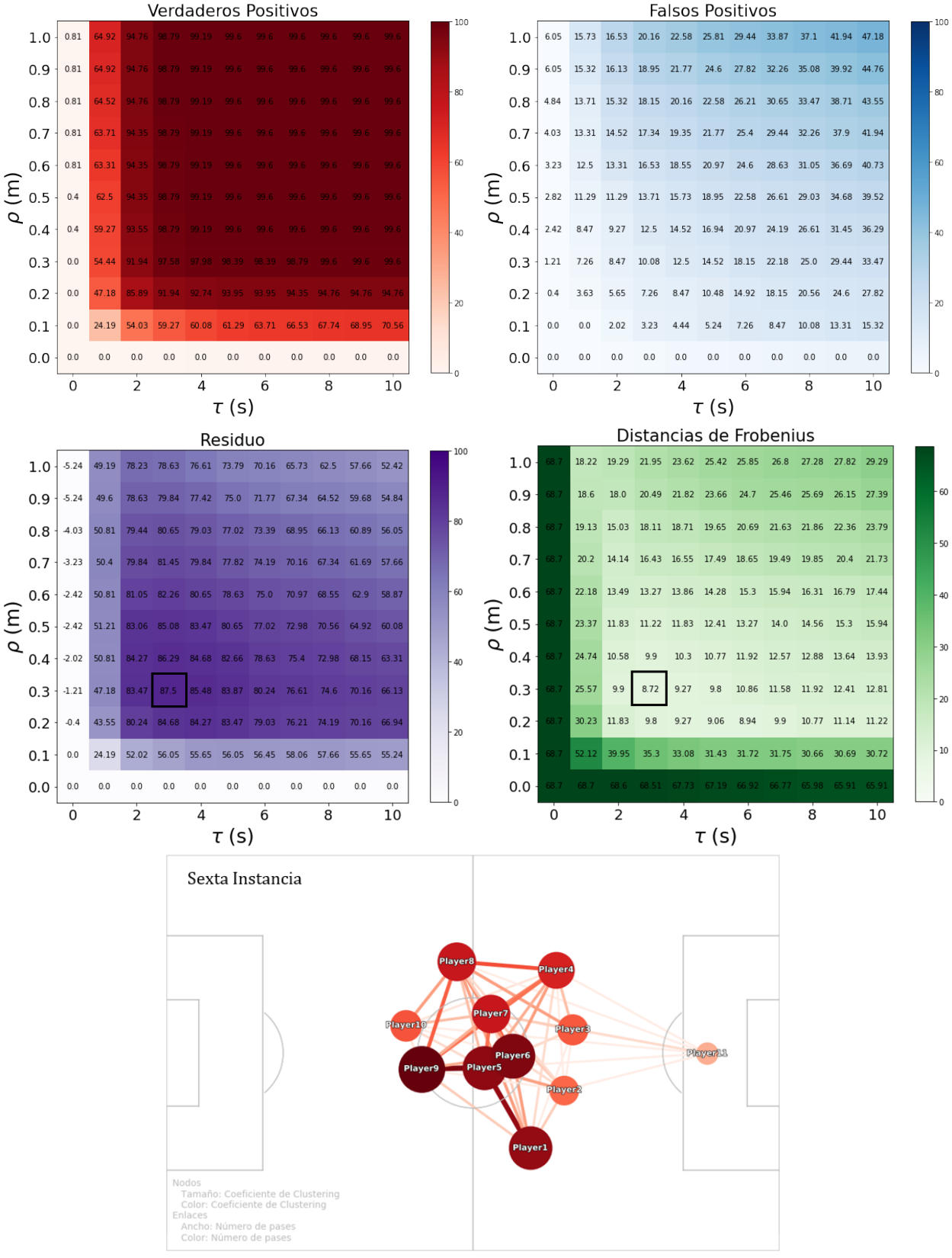
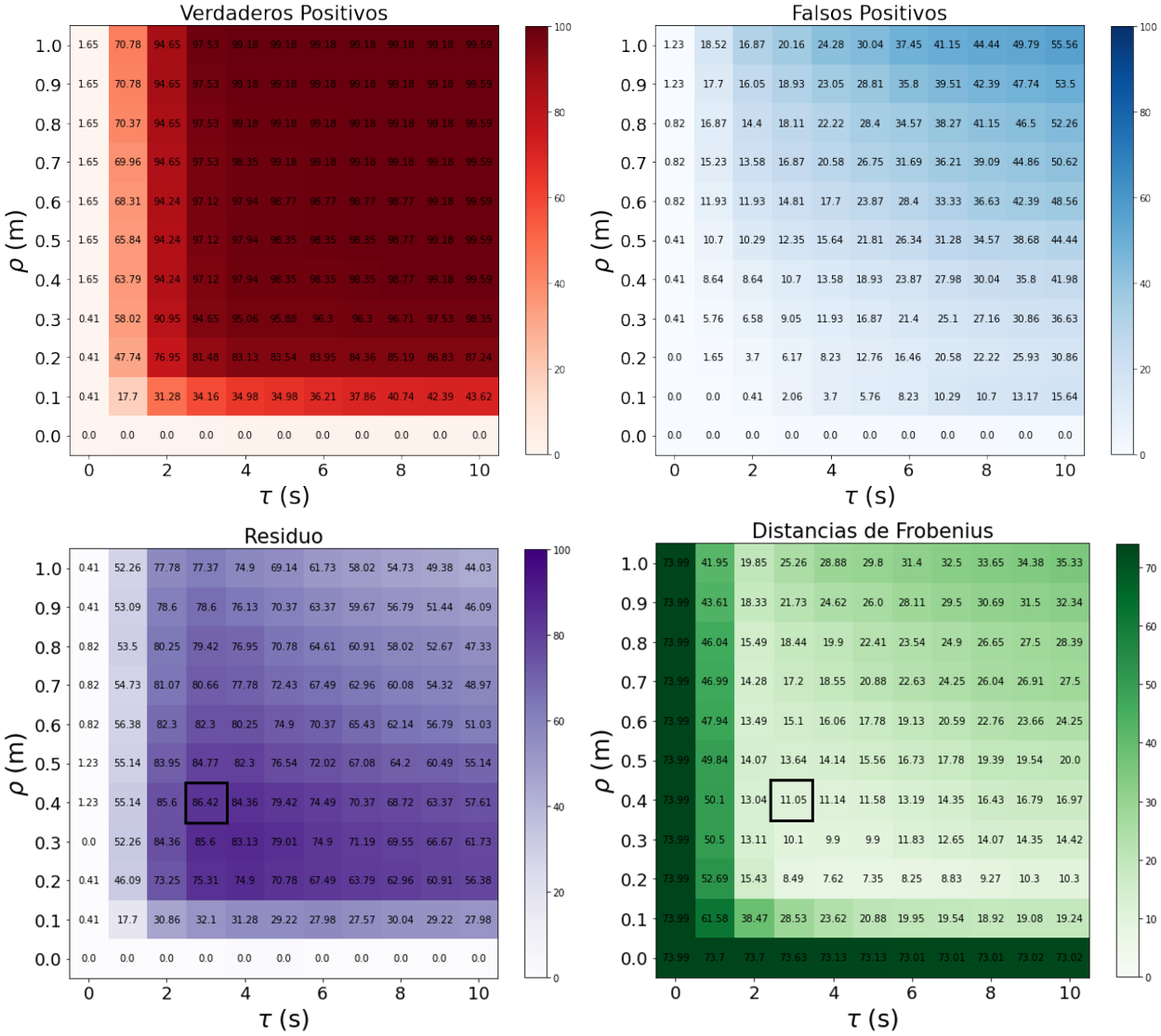


Figura 5.7
Conjunto de resultados para la séptima instancia individual



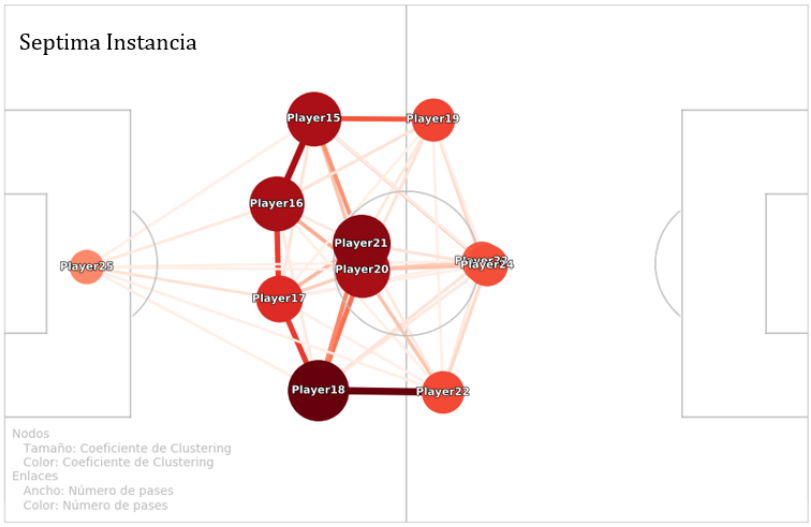
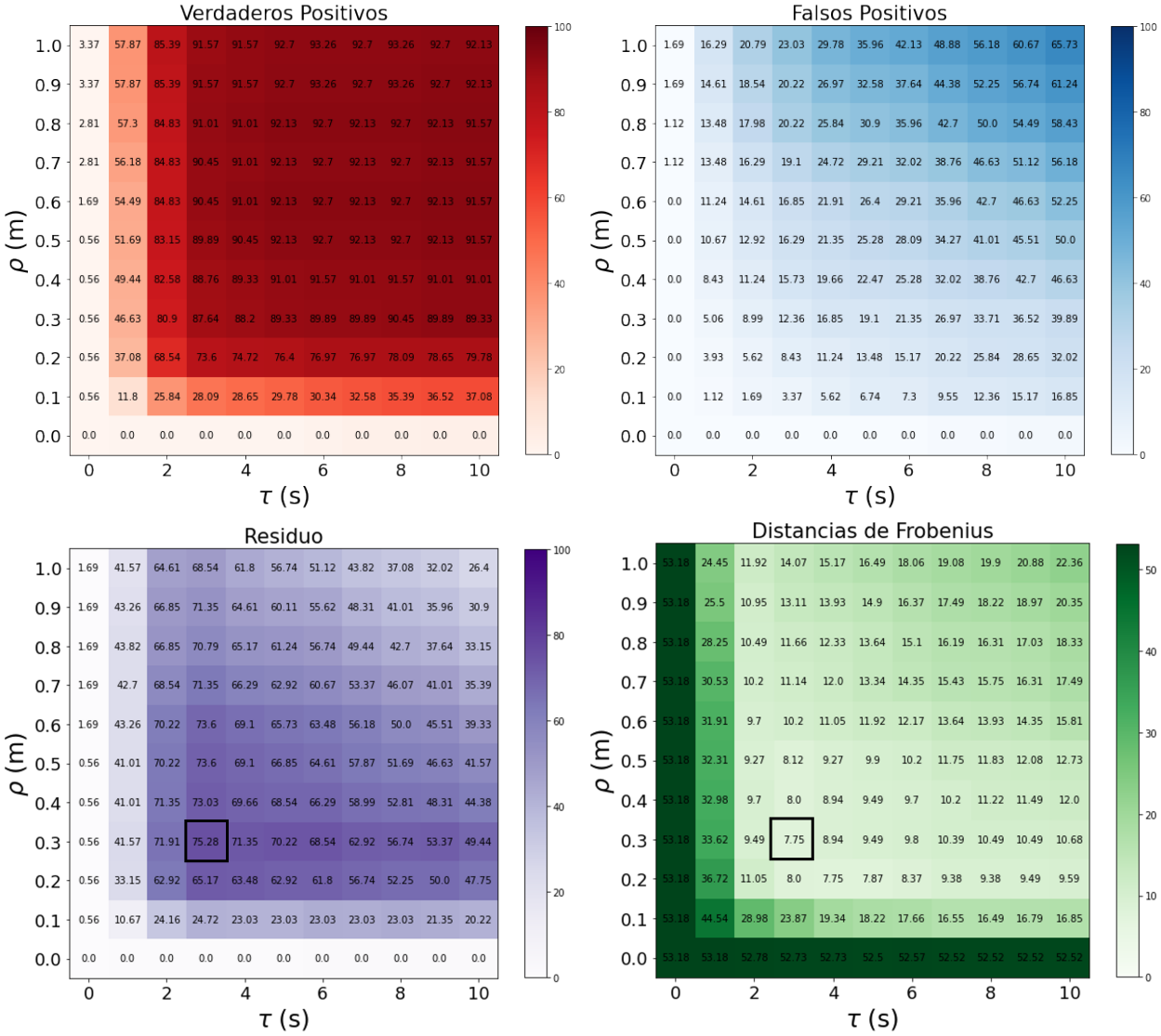
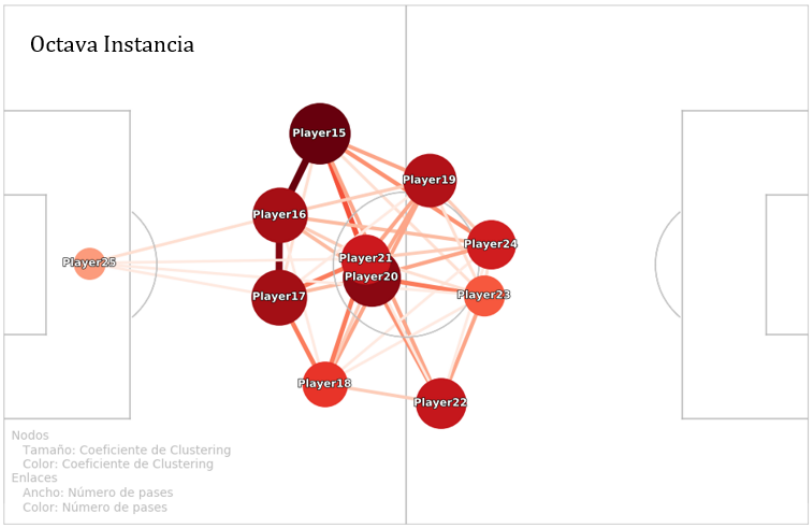



Figura 5.8
Conjunto de resultados para la octava instancia individual