titulo

Andres Chacoma*

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física. Buenos Aires, Argentina and CONICET - Universidad de Buenos Aires, Instituto de Física Interdisciplinaria y Aplicada. Buenos Aires, Argentina.

Juan I. Perotti

Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía, Física y Computación. Córdoba, Argentina and CONICET- Universidad Nacional de Córdoba, Instituto de Física Enrique Gaviola (IFEG). Córdoba, Argentina.

Orlando V. Billoni

Universidad Nacional de Córdoba, Facultad de Matemática, Astronomía, Física y Computación. Córdoba, Argentina and CONICET- Universidad Nacional de Córdoba, Instituto de Física Enrique Gaviola. Córdoba, Argentina.

Abstract

We ...

I. INTRODUCTION

II. DATA

A. Recopilación de métricas

En este trabajo utilizamos la base de datos de evento provista por L. Pappalardo et al en [1]. En ese articulo, los autores visualizan todos los partidos de la temporada 2017-2018 de las principales ligas de fútbol europeas: La Liga (España), Premier league (Inglaterra), Seria A (Italia), Bundesliga (Alemania), Ligue 1 (Francia). Por cada partido detectan, clasifican, y localizan en tiempo y espacio todos los eventos: goles, tiros al arco, pases, saques de esquina, faltas, etc. En el sistema de referencia que utilizan para ubicar los eventos en tiempo y espacio, t expresa el tiempo transcurridos desde el inicio del partido, la coordenada x expresa la distancia respecto del arco del equipo creador del evento, y la coordenada y la distancia respecto a la banda lateral derecha. Las unidades de las coordenadas espaciales están dadas en porcentajes del campo juego, siendo por ejemplo $x = 0$, $x = 50$ y $x = 100$ las posiciones de la linea de meta propia, la linea de centro del campo y la linea de meta del rival, respectivamente.

En este marco, definimos intervalo de posesión de pelota (BPI) como al conjunto dado por una secuencia continua de eventos generados por un equipo. Note que cada BPI contiene información de un solo equipo. Recopilamos todos los BPI de todos los equipos de cada liga, y sobre esos datos extrajimos métricas que nos permiten detectar algunos de los recursos tácticos que están utilizando los equipos en esa ventana temporal del partido. Las métricas recopiladas para nuestro análisis están basadas en las propuestas por J. Fernandez-Navarro en [2]. El estudio de estas metricas fue utilizado tambien en un trabajo anterior para determinar estilos futbolisticos caracterisiticos [3]. A continuación describimos en detalle a cada una de estas,

* achacoma@df.uba.ar

1. *Direct play.* Cada vez que hay un pase o un tiro libre en un BPI, medimos la velocidad media en la dirección de ataque, dada por el cociente entre la distancia recorrida por la pelota en el eje x y el tiempo transcurrido. De cada BPI tomamos el valor máximo. Esto nos permite detectar que tan directo hacia la portería rival es el movimiento de la pelota en el equipo.

2. *Counterattack.* Dado dos eventos consecutivos en un BPI, si el primero se observa en $x_1 < 40$ y el segundo en $x_2 > 60$ a una diferencia temporal Δt, se informa la velocidad como $v = \frac{x_2 - x_1}{\Delta t}$. En otro caso se informa 0. Esto es una medida de que tan rápido un equipo pasa de una posición defensiva en su campo a una ofensiva en campo rival.

3. *Maintenance.* Dado un BPI, se calcula el promedio de las posiciones en el eje x, donde se generaron todos los eventos. Si se cumple $\bar{x} < 40$, es decir si los eventos se desarrollaron mayoritariamente en la zona defensiva del equipo, se informa el tiempo total de la posesión. En otro caso se informa 0. Esto nos permite detectar que tanto un equipo decide mantener y construir su juego desde su propio campo.

4. *Build up.* Si en un BPI se verifica que $\bar{x} > 60$, es decir la posesión se desarrolla mayoritariamente en campo rival, se informa el tiempo total de la posesión. En otro caso se informa 0. Esta métrica informa el tiempo de posesión en situaciones donde el equipo invade fuertemente el campo rival.

5. *Midfield play.* Si en un BPI se observa que $\bar{x} \leq 60$ y $\bar{x} \geq 40$, es decir la posesión de desarrolla mayoritariamente por el centro del campo de juego, se informa el tiempo total de la posesión, en otro caso se informa 0.La idea de esta variable es medir el tiempo que el equipo pasa en el sector medio del campo de juego.

6. *Flow rate.* En cada BPI donde $\bar{x} \geq 50$ se toma la diferencia temporal entre todos los eventos, y se calcula el valor medio, $\bar{dt}$. Luego se define la métrica como $1/\bar{dt}$. De esta manera se tiene una medida de que tan rápido el equipo mueve la pelota en el campo rival.

7. *Crossing.* Si en un BPI se observa un evento centro, se informa 1, en otro caso se informa 0. Esta métrica sirve para contabilizar los intentos de llegada por vía aérea.

8. *Pressure point.* De cada BPI se toma el primer evento, y se extrae la posición en la variable x, es decir donde el equipo comienza su posición. Esto nos permite medir si el equipo esta recuperando la pelota en su campo, en la zona media o en el campo rival.

9. *Pressure loss.* Si un BPI comienza en un evento donde $x > 40$, se informa el tiempo total de la posesión del adversario del BPI inmediatamente anterior. Este métrica es útil para observar si el equipo esta relajando o aumentando el nivel de presión que ejerce sobre el juego del rival en las zonas medias y altas del campo de juego.

10. *Shots.* Si en un BPI se registra un evento "Shot", se informa 1, en otro caso se informa 0. Esta métrica permite contabilizar los tiros al arco de cada equipo.

Para nuestro análisis,se descartaron todas los BPI con menos de 3 eventos y con tiempo total menor a 2 segundos. La idea de esto es descartar pequeñas recuperaciones pasajeras y quedarnos con posesiones consolidadas. Del proceso de recopilación se obtuvieron 215681 BPI. Luego de calcular los valores de las métricas en cada uno de estos, estudiamos la distribución de los datos. Observamos que las métricas parecen seguir una distribución tipo log-normal, por lo tanto decidimos transformar los datos como $x \to log(1 + x)$ para trabajar con distribuciones aproximadamente normales. Posteriormente, agrupamos la información por partido y por equipo, y sumamos los valores obtenidos en cada métrica. De esta manera por ejemplo el feature *Shots* cuantifica la cantidad de tiros al arco ejecutados por el equipo en ese partido. Asimismo el feature *Build-up* cuantifica la cantidad de tiempo neto en la cual un equipo sostuvo una posición de ataque frente al rival en ese partido. Note que en la temporada 2017/2018, en las ligas Española, Inglesa, Francesa e Italiana los equipos jugaron 38 partidos. Por lo tanto, al tomar los datos de los primeros 4 equipos cada liga aporta un total de $38 \times 4 = 152$ muestras al archivo de datos. Asimismo, en la liga alemana, al ver menos equipos, se jugaron 34 partidos, por lo tanto esta liga aporta 136 muestras. En consecuencia, la matriz de datos consta de 744 filas y 10 columnas. Por ultimo, en un dataset aparte recopilamos meta-data asociada a cada muestra, útil luego para realizar el análisis: a que equipo pertenece esa muestra, cual es la liga de pertenencia, y el resultado final en la tabla de posiciones.

B. Representación en redes complejas

En lo que sigue, presentamos nuestra propuesta para representar las métricas de rendimiento en términos de redes complejas. Definimos $M(i,\ j,\ g)$ como la métrica de rendimiento correspondiente al equipo i cuando enfrenta al equipo j en el partido g. Por ejemplo, puede representar la cantidad de tiros al arco realizados por el FC Barcelona al jugar contra el Real Madrid en el primer encuentro del torneo español *La Liga*. En nuestro conjunto de datos, todos los equipos participaron en un formato de liga todos contra todos, enfrentándose dos veces: un partido de ida (g_1) y otro de vuelta (g_2). Utilizando la información de ambos encuentros, definimos una métrica agregada que resume el desempeño observado entre esos dos equipos a lo largo del torneo:

$$M(i,\ j) = \sum_{g_1, g_2} M(i,\ j,\ g).$$

En el ejemplo anterior, $M(i,\ j)$ representa la cantidad total de tiros al arco realizados por el FC Barcelona contra el Real Madrid en ambos partidos del torneo. Calculando $M(i,\ j)$ para cada par de equipos en una liga L, es posible representar estas relaciones de desempeño mediante un grafo dirigido y ponderado $G(L,\ M)$, cuyos pesos se definen como

$$f_{ij} = M(j,\ i) \ -\ M(i,\ j).$$

Nótese que, en esta representación, $f_{ij} < 0$ indica que el equipo i superó al equipo j en la métrica considerada. En este marco, construimos 50 grafos en total, diez grafos por liga. Cada uno asociado a una métrica de rendimiento distinta.

III. THEORY

IV. RESULTADOS

A. Estadística del rating real

La idea de esta sección es definir un modelo estadistico para la distribucion de probabilidad del rating real, R_T, definido como la cantidad de puntos totales obtenidos por un equipo durante la liga. R_T es una variable estocastica que depende de la cantidad de par-

tidos ganados, empatados y perdidos, por los equipos de la liga. En las ligas de futbol, un equipo obtiene 3 puntos cuando gana, 1 cuando empata y 0 cuanto pierde. En este marco, R_T puede expresarse en terminos de las variables estocasticas W y D , que representan el numero de partidos ganados y empatados, respectivamente,

$$R_T = 3W + D.$$

Sea n el número total de partidos en la temporada. Modelamos W con una distribución binomial,

$$W \sim Bin(n|p_w)$$

donde p_w es la probabilidad de ganar un partido. Condicional al número de victorias $W = w$, los empates siguen,

$$D|W = w \sim Bin(n - w|\frac{p_d}{1 - p_w}),$$

con p_d probabilidad de empatar. Dado que los equipos tienen distinto nivel competitivo, cada equipo posee probabilidades intrínsecas p_w y p_d de ganar y empatar, respectivamente. Modelamos estas probabilidades mediante un enfoque jerárquico: cada equipo tiene un par de habilidades latentes $\boldsymbol{\eta} = (\eta_w, \eta_d)^T$, de ganar o empatar, que siguen una distribución normal bivariada,

$$\boldsymbol{\eta} \sim \mathcal{N}\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})\ ,$$
$$\boldsymbol{\mu} = (\mu_w, \mu_d)^T, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_w^2 & \rho\sigma_w\sigma_d \\ \rho\sigma_w\sigma_d & \sigma_d^2 \end{pmatrix},$$

donde $\boldsymbol{\mu}$ es el centro de la distribucion, $\boldsymbol{\Sigma}$ la matriz de covarianza. En este marco, ρ representa la correlación entre las habilidades para victoria y empate. Las probabilidades se obtienen mediante la transformación softmax,

$$p_w = \frac{e^{\eta_w}}{1 + e^{\eta_w} + e^{\eta_d}}, \quad p_d = \frac{e^{\eta_d}}{1 + e^{\eta_w} + e^{\eta_d}}.$$

Esta transformación garantiza que $p_w + p_d \leq 1$ y que el modelo sea identificable al fijar implícitamente $\eta_l = 0$ para las derrotas. Con estos elementos, podemos escribir la distribucion conjunta de victorias y empates marginalizando sobre los efectos aleatorios $\boldsymbol{\eta}$,

$$P(W = w, D = d) = \int_{\mathbb{R}^2} Bin(n|p_w(\boldsymbol{\eta}))\ Bin(n - w|\frac{p_d(\boldsymbol{\eta})}{1 - p_w(\boldsymbol{\eta})})\ \phi(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\boldsymbol{\eta}, \quad (1)$$

Notar que la integral en 1 no tiene forma cerrada, para resolverla utilizamos el metodo de monte carlo. Con estos elementos, podemos finalmente escribir la distribucion de probabilidad teorica para el true rating como,

$$P(R_T = r_T) = \sum_X P(W = w,\, D = d), \tag{2}$$

$$X = \{w,\, d > 0,\, 3w + d = r_T,\, w + d \leq n\}.$$

Los grados de libertad que otorgan los parametros μ_w, μ_d, σ_w, σ_d y ρ, permiten ajustar la curva teorica 2 a la curva empirica. El ajuste fue realizado sobre el conjunto de datos asociado a las ligas Inglesa, Francesa, Italiana y Española. Para este analisis, decidimos no utilizar los datos de la liga Alemana, ya que su torneo se disputa entre un numero menor de equipos (18), y por lo tanto se genera una distribucion de valores de true rating menor respecto de las otras ligas, en las cuales participa un total de 20 equipos por liga. Para hacer el ajuste, utilizamos el algoritmo Nelder-Mead minimizando el RMSE entre las CDF empírica y teórica. Con los parámetros óptimos, realizamos bootstrap paramétrico (1000 réplicas) para estimar la incertidumbre. Para la probabilidad de victoria obtuvimos $\bar{p}_w = 0.366$ ($SD = 0.010$) con un intervalo de confianza del 95% de [0.3527, 0.3798]. Asimismo, para la probabilidad de empate obtuvimos $\bar{p}_d = 0.280$ ($SD = 0.020$) con un intervalo de confianza del 95% de [0.2548, 0.3074]. Para contrastar estos resultados con los datos empiricos, utilizando la informacion de los resultados de los partidos, calculamos para cada equipo i las probabilidades de victoria y empate empiricas, q_w y q_d,

$$q_w^{(i)} = \frac{\#wins^{(i)}}{\#matchs^{(i)}}, \quad q_d^{(i)} = \frac{\#draws^{(i)}}{\#matchs^{(i)}}.$$

Calculando el valor medio y el desvio estandar sobre el conjunto de todos los equipos, obtuvimos para la probabilidad de victoria $\bar{q}_w = 0.379$ ($SD = 0.168$) con un intervalo de confianza del 95% de [0.1572, 0.7638], y para la probabilidad de empate $\bar{q}_d = 0.243$ ($SD = 0.081$) con un intervalo de confianza del 95% de [0.1046, 0.3954]. Podemos observar, (1) Las medias son similares (0.366 vs 0.379 para victorias, 0.280 vs 0.243 para empates), lo que valida el centro de la distribución. (2) las desviaciones estándar empíricas (0.168, 0.081) son mucho mayores que las implícitas en el modelo (0.01, 0.02), indicando que el modelo subestima la heterogeneidad entre equipos.

Con respecto al parametro de correlacion, obtuvimos $\bar{\rho} = -0.130$ ($SD = 0.07$) con un intervalo de confianza del 95% de $[-0.3040, -0.0371]$. Esto indica la presencia de una

correlacion estructural negativa debil entre las habilidades de victoria, η_w, y empate, η_d, de los equipos. Esto sugiere que equipos con mayor habilidad para ganar tienden ligeramente a tener menor habilidad para empatar.

En la Fig. 1 (a), mostramos la CDF del true rating junto con el ajuste. Podemos ver que el modelo propuesto captura muy bien el comportamiento de la curva en todo el soporte. En Fig. 1 (b) mostramos un grafico con la relacion entre los quantiles teoricos y de los datos. A modo de referencia, mostramos tambien la relacion entre los quantiles de los datos y los asociados a una distribucion gausiana con media y desvio estandar igual a los valores muestrales obtenidos de los datos, $\bar{R}_T = 52.45$ ($SD = 18.56$). En primer lugar, podemos ver que el modelo propuesto captura bien el comportamiento en la zona central izquierda y de manera aceptable el comportamiento de la cola de la derecha. La comparación con el modelo gaussiano nos permite ver que la distribucion real presenta colas mas livianas a la izquierda y mas pesadas a la derecha, respecto del comportamiento gaussiano, develando así la presencia de una asimetria en la distribuicion. La concabidad hacia abajo en el centro del grafico, indica tambien que la mediana de los datos esta corrida hacia la izquierda.

B. Comparacion entre el rating real y el rating obtenido de las metricas

Definimos como true ranking de un equipo, al puesto final alcanzado en la tabla de posiciones de su liga. Por ejemplo, el true ranking del equipo FC Barcelona en la liga Española es 1, por que salio campeon esa temporada. Asimismo, definimos como metric rating, R_M, de un equipo de una liga, en una metrica dada, al valor de rating calculado a partir del metodo HodgeRank para la red asociada a esa metrica.

Nuestra idea es comparar el true rating con el rating obtenido a partir del metodo de HodgeRank. Naturalmente, estas cantidades estan en escalas distintas, por lo tanto para hacer la comparación es necesario llevar los valores de ambas variables a la misma escala. Por tal motivo, decidimos estandarizar las variables y hacer la comparacion sobre el *z-score* de cada una de estas. En lo siguiente discutimos algunas comparaciones intersantes entre true raiting y el metric rating obtenido a partir de alguna de las metricas. En la Fig. 2 (a) comparamos las CDF del true rating asociado a los datos de todos los equipos y todas las ligas, con la del metric rating asociado a los datos de todos los equipos de todas las ligas y de las diez metricas estudiadas. A ambas curvas se les estrajeron los valores outliers

considerados como aquellos casos que se encuentran por encima del quantile $Q_{99.7}$ Podemos observar una coincidencia notable en casi todo el rango, observando algunas diferencias en las colas: el metric rating muestra mas valores positivos mas alejados de la media, y el true rating muestra mas valores negativos levemente mas alejados de la media. En la Fig. 2 (b) comparamos las curva R_T y R_M en función del true rnking, para el caso de la metrica *Presure Point* en la liga Inglesa. En primer lugar podemos notar que R_T decrece a medida que crece el ranking, esto es trivial por defincion: el primer equipo tiene un true raiting mayor o igual que el segundo, el segundo tiene un true raiting mayor o igual que el tercero, y asi hasta el ultimo puesto. En la curva de R_M observamos una tendencia tambien decreciente, pero no monotona. En este caso, podemos decir que en terminos generales el metric rating aproxima relativamente bien la tendencia del true rating. En la Fig. 2 (c) mostramos una comparacion analoga a la anterior para el caso de la metrica *Maintenance* en la liga Italiana. En este caso vemos que la curva R_M no sigue una tendencia decreciente y se obverva una total descorrelacion con el true rating. Vemos entonces que el true rating de una dada liga puede diferir poco o mucho del metric rating dependiendo de la metrica que utilizamos para inferir el rating. Una pregunta que surge naturalmente es cual es la estadistica asociada a esas diferencias. Para estudiar esto, definimos la cantidad ΔR como la diferencia entre el true rating de un equipo i de una liga l y el metric rating asociado a una metrica m,

$$\Delta R = R_T(i,\, l) \,-\, R_M(i,\, l,\, m)$$

Note que en las ligas Inglesa, Francesa, Italiana y Española, liga obtuvimos 20×10 muestras de ΔR, ya que tenemos 20 equipos y 10 metricas distintas. No obstante, en la liga Alemana obtuvimos 18×10 muestras. Nos interesa veriricar si las distribuciones de esas diferencias tienen comportamiento normal y si difieren entre una liga y otra. Para esto, utilizamos el grafico quantil quantile que mostramos en la Fig. 2 (d). Para este grafico, tomamos los valores de ΔR de cada liga y los estandarizados al valor medio y la desviacion estandar. Luego calculamos los quantiles y los comparamos contra los quantiles de una distribucion normal. En general se observa un comportamiento bastante cercano al normal. La liga inglesa, podemos ver que presenta colas ligeramente pesadas tanto a izquierda como a derecha. Por el contrario, en la liga alemana podemos ver una asimetria, mostrando colas levente livianas a la iquierda y levemente pesadas a la derecha.

Nuestro objetivo ahora es cuantificar la diferencia entre la curva de true rating, R_T, y las

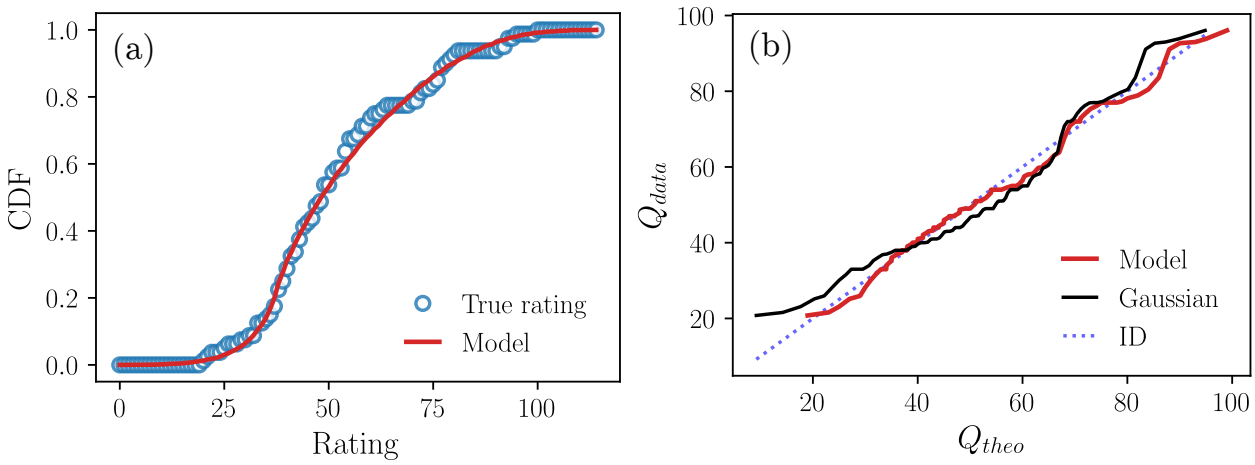

FIG. 1. Estadística del rating real. (a) Comparación de la Cumuative distribution function (CDF) de los valores de rating real con lo obtenido a partir del modelo. (b) Relacion entre los Quantile asociado a los datos y los obtenidos a partir del modelo. A modo de referencia, se agrega tambien una comparación con una distribucion gaussiana de media y desvio estandar igual al de los datos.

curvas de metric rating, R_M, asociadas a cada metrica de cada liga. Nuestra idea es verificar de metrica se obtiene el rating mas similar al real, y si eso depende de la liga. Para esto, definimos la distancia D, como el valor medio de las diferencias cuadraticas entre dos curvas (RMSE), esto cuantifica que tan diference es R_T de R_M en cada caso. Ademas, calculamos el coeficiente de Person entre ambas curvas, para estudiar el comportamiento de la tendencia.

C. Comparacion entre el ranking real y el ranking obtenido de las metricas

D. Propuesta de un raiting multivariado

Para cada liga, definimos la matriz de rating multivariado, $\boldsymbol{R_{MV}}$. En este objeto, las fila representan los equipos, y las columnas los metric ratings calculados con el metodo de HodgeRank. La dimension de esta matriz, en el caso de Espana por ejemplo, esta matriz sera 10

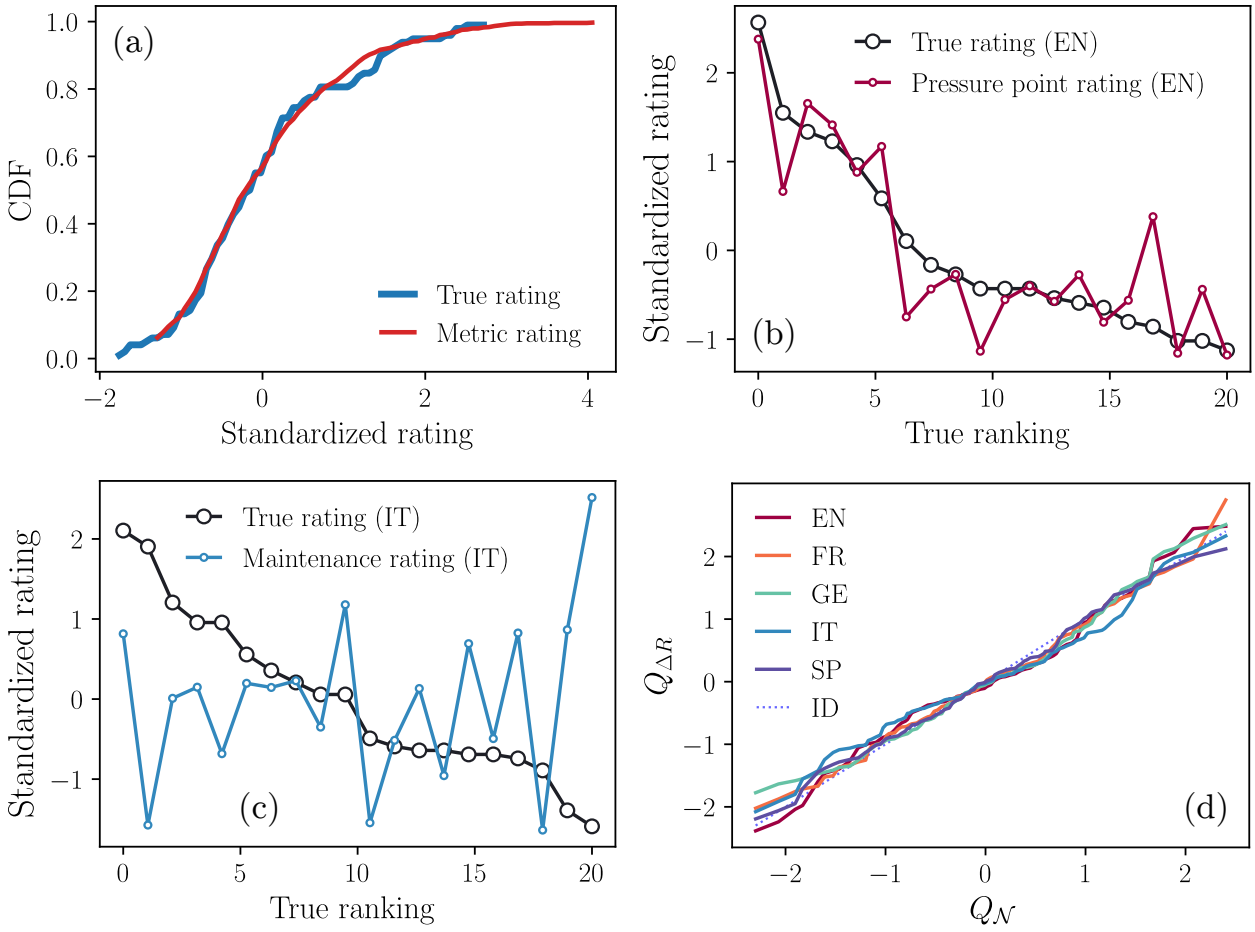

FIG. 2. Comparacion entre true rating y metric rating. Debido a que estas cantidades estan en unidades distintas, para comparar se usan los datos estandarizados. (a) Comparacion de las cumulative distribution functions asociadas al true ranking y al metric rating. (b) Comparación en el caso Presure point en la liga inglesa. (c) Comparación en el caso Maintenance en la liga italiana.. (d) Cumulative distribution function asociadas a las diferencias punto a punto entre el true rating y el metric raiting. A modeo de referencia, en lineas punteadas, se muestra la CDF de una densidad gaussiana de media y desvio igual al conjunto de todos los datos.

V. DISCUSION Y CONCLUSION

[1] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer

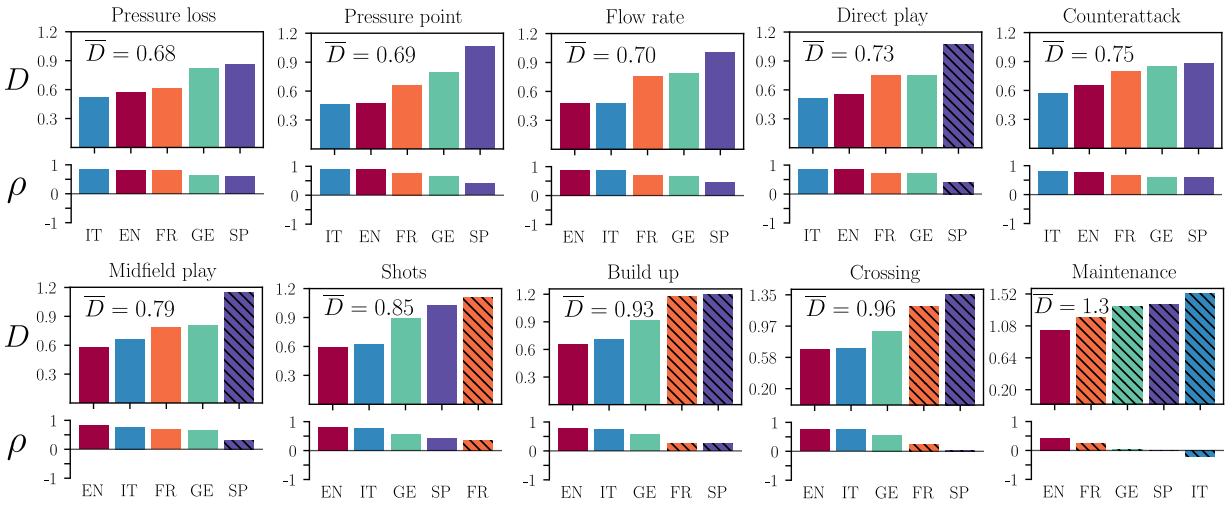

FIG. 3. Distancia D y coeficiente de person ρ, entre el rating verdadero y el rating dado por las metricas. Cada panel contiene la informacion de una metrica y las barras muestran el valor de D obtenido en cada liga. Las barras en cada panel estan ordenadas en orden creciente respecto del valor de D. Asimismo, los paneles estan ordenado en orden creciente respecto del valor promedio en cada metrica, $\overline{D}$. Las barras con hatch muestran los casos donde el coeficiente de Pearson indica una correlacion debil, $|\rho| < 0.4$.

competitions. Scientific data, 6(1):236, 2019.

[2] Javier Fernandez-Navarro, Luis Fradua, Asier Zubillaga, and Allistair P. McRobert. Influence of contextual variables on styles of play in soccer. International Journal of Performance Analysis in Sport, 18(3):423–436, 2018.

[3] Andres Chacoma and Orlando V Billoni. Data-driven approach to defining football styles in major leagues. Chaos, Solitons & Fractals, 200:116926, 2025.

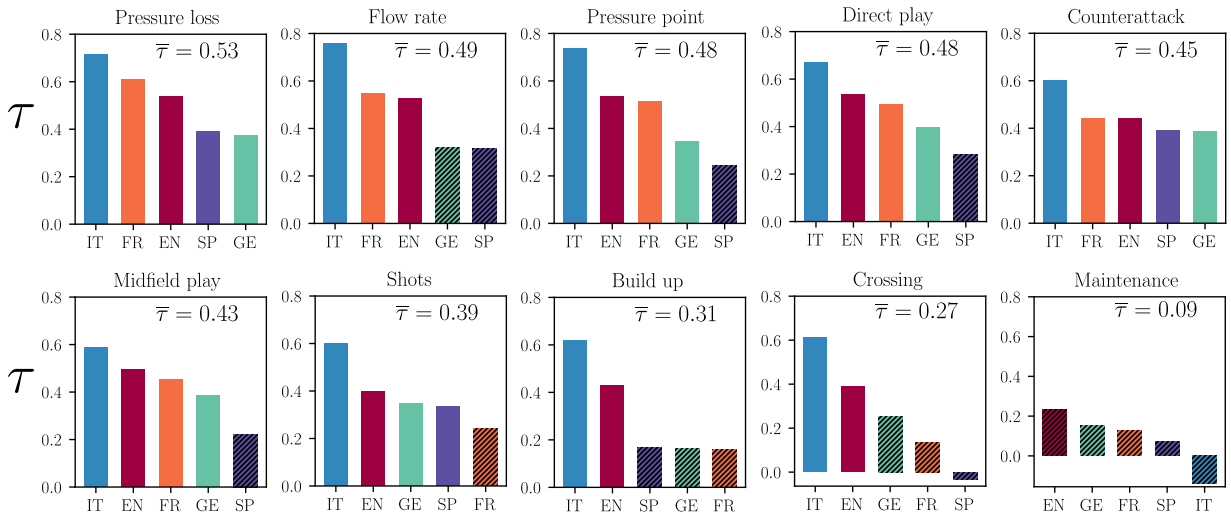

FIG. 4. Coeficiente de Kendall, τ, entre los rankings verdaderos y los rankings dados por las metricas. Cada panel contiene la informacion de una metrica y las barras muestran el valor de τ obtenido en cada liga. Las barras en cada panel estan ordenadas en orden decreciente respecto del valor de τ. Asimismo, los paneles estan ordenado en orden decreciente respecto del valor promedio en cada metrica, $\overline{\tau}$. Las barras con hatch indica los casos donde el $p - value < 0.05$. En esos casos no hay evidencia suficiente para afirmar que existe correlacion significativa.

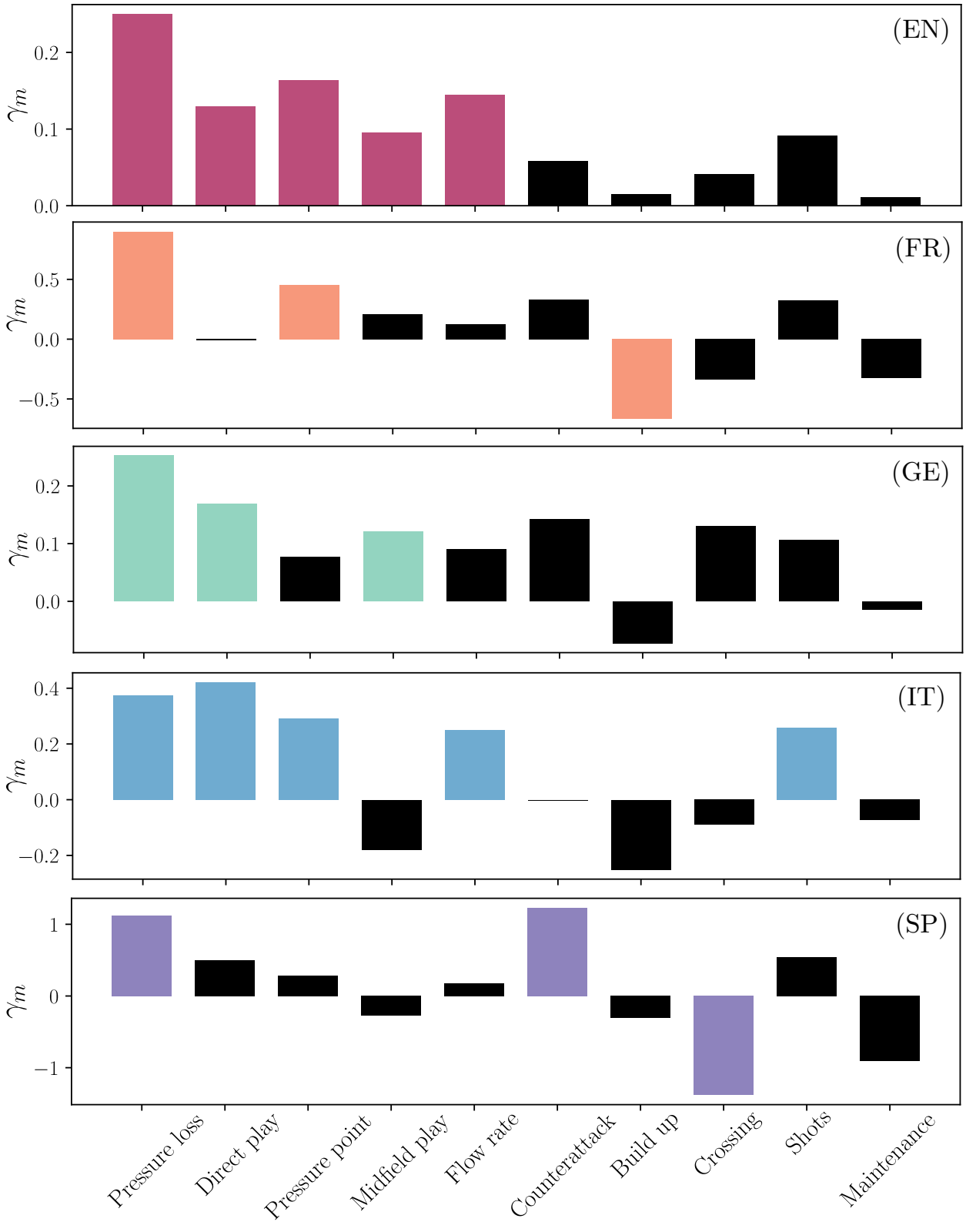

FIG. 5. Relevancia de las metricas en el rating multivariado. Cada panel muestra los valores de γ_i en cada metrica para una liga dada. Las barras negras muestran los casos donde no hay significancia estadistica.

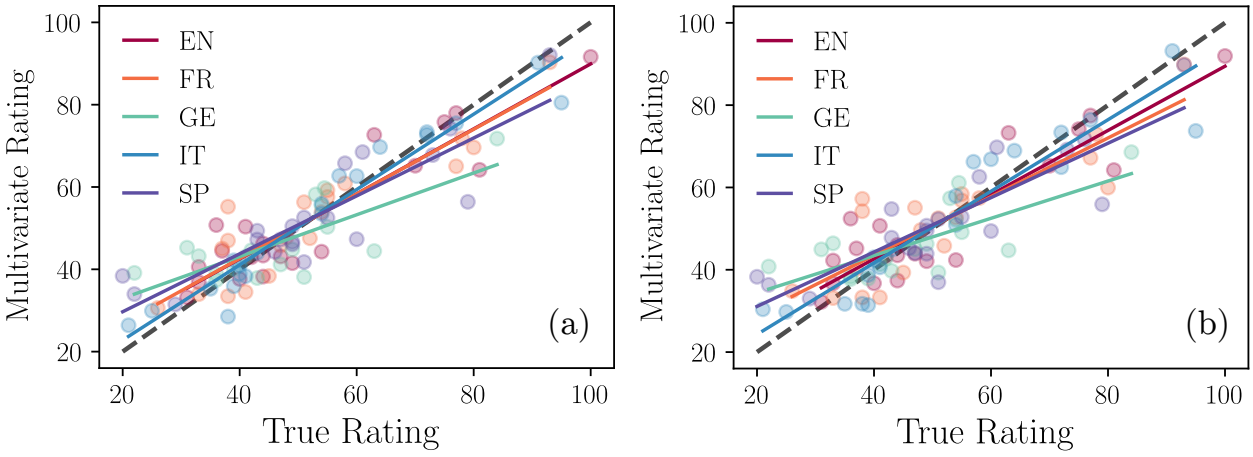

FIG. 6. Comparacion el true rating y el multivariate rating. (a) Utilizando todas las variables para calcular el raiting multivariado. (b) Utilizando solo las variables relevantes de cada liga (ver Fig. 5). En ambos paneles, los scatters muestran los valores exactos para cada equipo de cada liga, las rectas fueron calculadas a partir un fit lineal con el objetivo de mostrar la tendencia que siguen las relaciones graficadas. La linea negra punteada muestra la funcion identidad.