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 (INFINA). Buenos Aires, Argentina.

Juan I. Perotti and 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 (IFEG). 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. 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 partidos 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 G, E y P, que representan la cantidad de partidos ganados, empatados y perdidos por un equipo,

$$R_T = 3G + 1E.$$

Definiendo $\mathbf{p} = (p_g,\ p_e,\ p_p)$ como el vector que contiene las probabilidades de ganar, empatar o perder un partido, para el equipo dado, podemos modelar la cantidad de partidos ganados, empatados y perdidos por ese equipo con una distribucion multinomial,

$$(G,\ E,\ P) \sim Multinomial(n,\ \mathbf{p}),$$

Donde n es la cantidad total de partidos jugados, lo cual en una liga es igual para todos los equipos. Proponemos que cada equipo tiene asociado un vector $\mathbf{p}$ diferente, dependiendo de su nivel competitivo. Por lo tanto $\mathbf{p}$ tambien puede ser considerada una variable estocastica. Podemos modelar este parametro a partir de una distribución de Dirichlet,

$$\mathbf{p}_i \sim Dirichlet(\boldsymbol{\alpha}),$$
$$\boldsymbol{\alpha} = (\alpha_g,\ \alpha_e,\ \alpha_p).$$

Esta distribución, en función del conjunto de parámetros $\boldsymbol{\alpha}$, genera vectores de tres componentes, $\mathbf{p}$, que cumplen con las propiedades requeridas para una terna de probabilidades:

$$p_{g_i} > 0,\ p_{ei} > 0,\ p_{p_i} > 0, \tag{1}$$
$$p_{g_i} + p_{ei} + p_{p_i} = 1.$$

Con estos elementos, podemos escribir la probabilidad de que un equipo gane $G = g$ partidos y empate $E = e$, durante la liga,

$$P(G = g,\ E = e) = \int_S f_M(g,\ e;\ n|\mathbf{p}) f_D(\mathbf{p},\ \boldsymbol{\alpha}) d\mathbf{p}. \tag{2}$$

Las densidades de probabilidad dentro de la integral son la densidad multinomial condidicionada y la densidad de Dirichlet. Estas pueden escribirse como,

$$f_M(g,\ e;\ n|\mathbf{p}) = \frac{n!}{g!e!(n\ -\ g\ -\ e)!}\ p_g^g\ p_e^e\ p_p^{(n-g-e)},$$
$$f_D(\mathbf{p},\ \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} p_g^{(\alpha_g - 1)}\ p_e^{(\alpha_e - 1)}\ p_p^{(\alpha_p - 1)}.$$

Ademas, el espacio de integración, S, es el simplex generado por las eqs. 1. La integral 2 admite solucion analitica. Resolviendo, obtenemos la probabilidad de ganar $G = g$ partidos y empatar $E = e$, en funcion de los parametros α_g, α_e y α_p,

$$P(G = g,\, E = e) = \frac{n!}{g!e!(n\, -\, g\, -\, e)!} \frac{B(\alpha_g\, +\, g,\, \alpha_e\, +\, e,\, \alpha_p\, +\, n\, -\, g\, -\, e)}{B(\alpha_g,\, \alpha_e,\, \alpha_p)}. \qquad (3)$$

A partir de la eq. 3, escribimos la distribucion de probabilidad teorica para el true rating,

$$P(R_T = r_T) = \sum_X P(G = g,\, E = e)$$

$$X = \{g,\, e > 0,\, 3g\, +\, 1e = r_T,\, g\, +\, e \geq n\}$$

Los grados de libertad que otorgan los parametros $\boldsymbol{\alpha}$, permiten ajustar la curva teorica a los datos. Lo hicimos a partir de un procedimiento tipo boostraping para 1000 muestreos con remplazo, utilizando los datos de la liga Inglesa, Francesa, Italiana y Epañola. En la Fig. 1 (a), mostramos la cumulative distribution function (CDF) del true rating junto con el ajuste. El grafico incluye también una densidad gaussiana de valor medio y desviacion estandar igual a los valores muestrales $\bar{R}_T = 52.45$ ($s_{R_T} = 18.56$), a modo de referencia. Los valores obtenidos del ajuste para los parametros son $\alpha_g = 0.85 \pm 0.03$, $\alpha_e = 2.60 \pm 0.07$, y $\alpha_p = 0.46 \pm 0.05$. Respecto a los resultados obtenidos, en primer lugar, podemos observar que el Multinomial-Dirichlet (M-D) model ajusta bien los datos empiricos en toda la curva, notandose un destacable rendimiento en la zona de raiting bajos. Comaprando estos resultados con la curva gaussiana, vemos que los true rating muestran colas mas livianas y un exceso de concentración en la zona de la mediana. En la Fig. 1 (b), la realción entre los quantiles resalta el buen renidmiento del modelo y las diferencias claras repecto del modelo gaussiano en la zona de las colas y de la mediana.

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

Definimos como true ranking, r, 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 Espanola es 1, por que salio campeon esa temporada. Definimos tambien, true rating, R_T, de un equipo a la cantidad de puntos obtenidos por ese equipo en su correspondiente liga. Por ejemplo, el true rating del FC Barcelona en esa temporada fue 93 puntos. 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.

Por otro lado, como nuestra idea es comparar el true rating con el rating obtenido a partir del metodo de HodgeRank, es necesario llevar los valores de ambas variables a la misma escala. Por tal motivo, decidimos estandarizar las variables utilizando la media y la desviacion estandar, de la siguiente manera,

$$R_T \rightarrow \frac{R_T - \langle R_T \rangle}{\sigma(R_T)},$$
$$R_M \rightarrow \frac{R_M - \langle R_M \rangle}{\sigma(R_M)},$$

donde $\langle * \rangle$ indica el valor medio y $\sigma(*)$ la desviacion estandar. Por ultimo, definimos como $\Delta_{i,m}$ a la diferencia cuadratica de rating de una metrica m, de un equipo i, como,

$$\Delta_{i,m} = \big(R_T(i) - R_M(i,\, m)\big)^2. \tag{4}$$

Esta cantidad cuantifica que tan distinto es el true rating de un equipo respecto al metric rating calculado a partir de HodgeRank para ese mismo equipo en una metrica dada.

En lo siguiente dsicutimos en orma cualitativa algunas comparaciones intersantes entre true raiting y metric rating. En la Fig. 2 (a) mostramos la curva R_T vs. r y R_M vs. r 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 r, esto es trivial por defincion. Observamos tambien que R_M tiene tambien una tendencia decreciente, pero no monotona. En terminos generales podemos decir que, en este caso, el metric rating se aproxima bastante bien al true rating. En la Fig. 2 (b) mostramos una comparacion analoga a la anterior para el caso de la metrica *Direct play* en la liga Espanola. En este caso podemos ver que R_M parece seguir una tendencia levemente decreciente, pero difiere bastante de la curva R_T. En la Fig. 2 (c) mostramos lo mismo que antes pero 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. Note, que la diferencia punto a punto entre la curvas R_T y R_M, en cualquiera de estos casos, representa la diferencia entre el true rating y el metric rating para cada equipo de la liga. En estos ejemplos, podemos ver que esta diferencia puede ser muy variada dependiendo del caso. Para estudiar en mas detalle estas variaciones, utilizamos la eq. 1 para calcular la diferencia cuadratica $\Delta_{i,m}$ en todos los equipos de todas las ligas y en todas las metricas. La distribucion de esos valores se muestra en la Fig. 2 (d).

Se observa una distribucion tipo power law con un cut off alrdedor de $\Delta_{i,m} \approx 1$, esto indica que rara vez las diferencias superan 1 desviacion estandar. En el rango $10^{-6} < \Delta_{i,m} < 1$ ajustamos el modelo $P(\Delta_{i,m}) \propto \Delta_{i,m}^{-\alpha}$ para obtener el valor del exponente. Obtuvimos $\alpha = 0.61 \pm 0.02$, lo que indica una distribucion de valores extremadamente heterogeneos. En este caso el valor medio no es un estadisitico informativo, no obstante si lo es la mediana, la cual arrojo un valor de 0.26. Podemos concluir entonces, que si bien la distribucion de diferencias cuadraticas exhibe un alto grado de complejidad los valores observados son en terminos generales relativamente pequenos.

La idea a continuacion es cuantificar la diferencia entre la curva de true rating, R_T, y la curva metric rating, R_M, asociada a una metrica m en una liga l. Para eso, en primer lugar definimos $\{\Delta_{i,m,l}\}$ como al conjunto de las diferencias cuadraticas que se observan punto a punto entre las curvas R_T y R_M asociadas a una metrica m en una liga l. A partir de esto, definimos la distancia $D(m,\, l)$ como la raiz cuadrada de la mediana de ese conjunto,

$$D(m,\, l) = \sqrt{M_d(\{\Delta_{i,m,l}\})}. \tag{5}$$

Notar, que $D(m,\, l)$ esta definida de manera analoga a la cantidad Root Mean Square Error (RMSE), la cual se define como el mean de las diferencias cuadraticas entre dos curvas. En nuestro caso decidimos tomar la mediana, debido a que la heterogeneidad de la distribucion $P(\Delta_{i,m})$ indica que el estadisitico mean podria ser poco informativo en este caso.

En este marco utilizamos la eq. 5 para calcular el valor de D para las 10 metricas en las 5 ligas. En resultado se muestra en la Fig. 3, la cual esta separada en 10 paneles uno por cada metrica, indicada en el parte superior del grafico. En cada panel se encuentran 5 barras que indican el valor de D para cada liga. Cada panel fue ordenado en orden ascendente, mostrando primero las ligas que muestran una menor distancia. En la parte inferior de cada panel mostramos el coeficiente de correlacion de pearson, ρ, el cual mide la correlacion entre las curvas R_T y R_M en cada caso. Los colores de las barras estan asociados a las ligas para una mejor visualizacion. Asimismo, las barras negras indican que en ese caso la correlacion de Person es debil, $|\rho| < 0.4$.

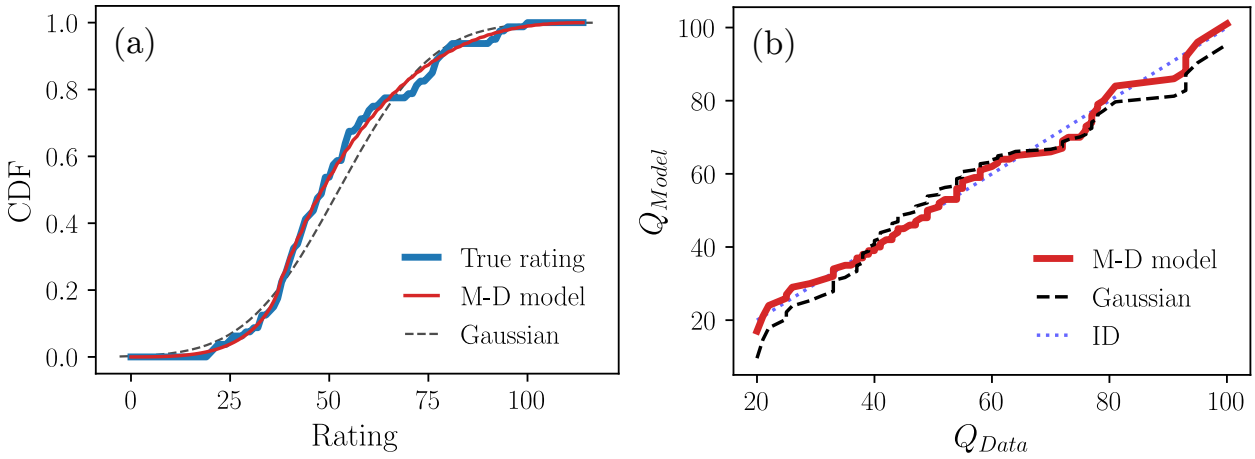

FIG. 1. Estadística del rating real. (a) Comparación de la Cumuative distribution function (CDF) de los valores de rating real con lo obtenido a partir del M-D model y el modelo gaussiano. (b) Relacion entre los Quantile asociado a los datos y los obtenidos a partir del M-D model y el modelo gaussiano.

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

IV. 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 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

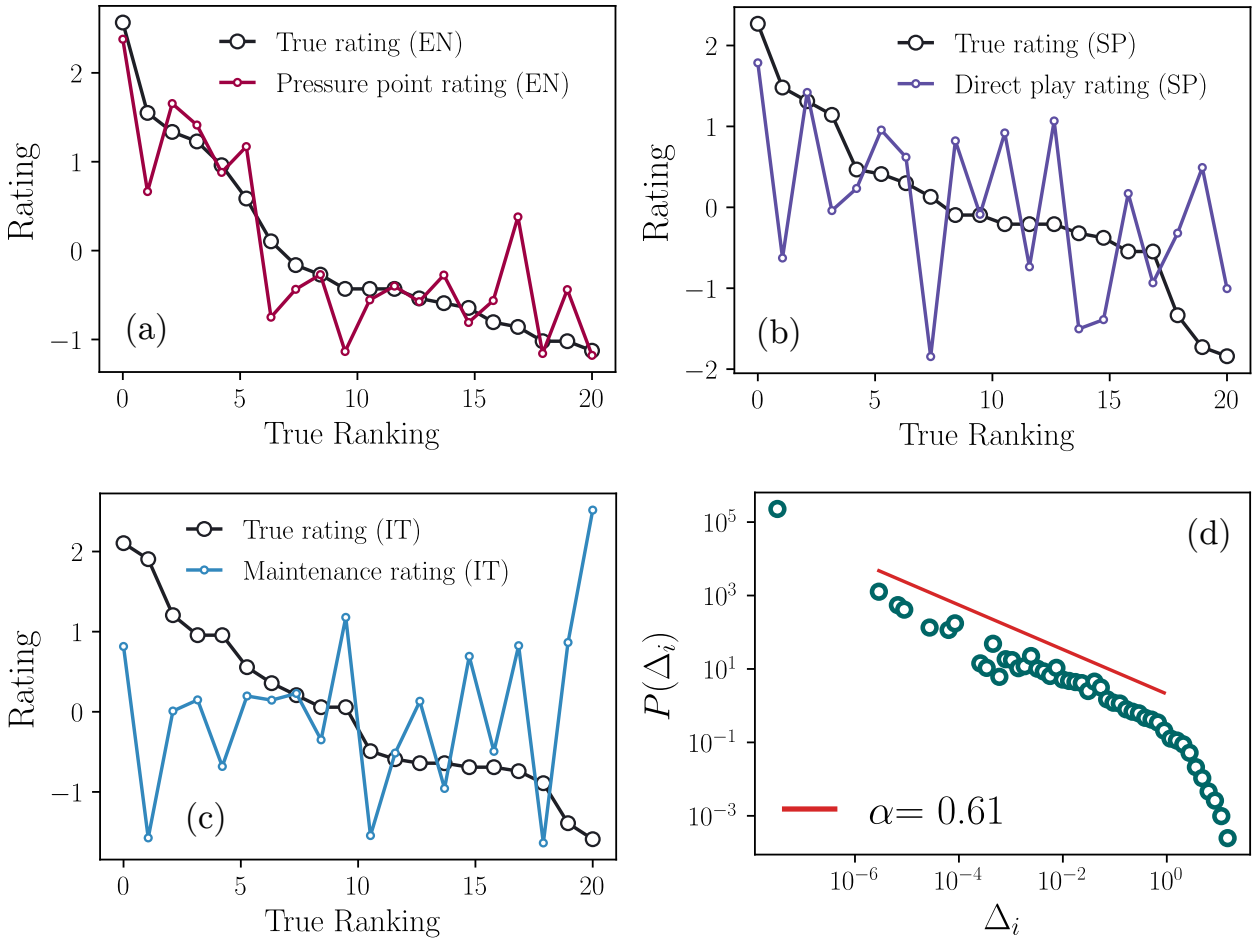

FIG. 2. Comparacion entre el rating verdadero y el rating obtenido de las metricas (algunos casos ejemplares). (a) Caso Presure point en la liga inglesa. (b) Caso Direct play en la liga espanola. (c) Caso Maintenance en la liga italiana. (d) Distribucion de diferencias cuadraticas calculada sobre el conjunto de todas las metricas de todas las ligas.

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.

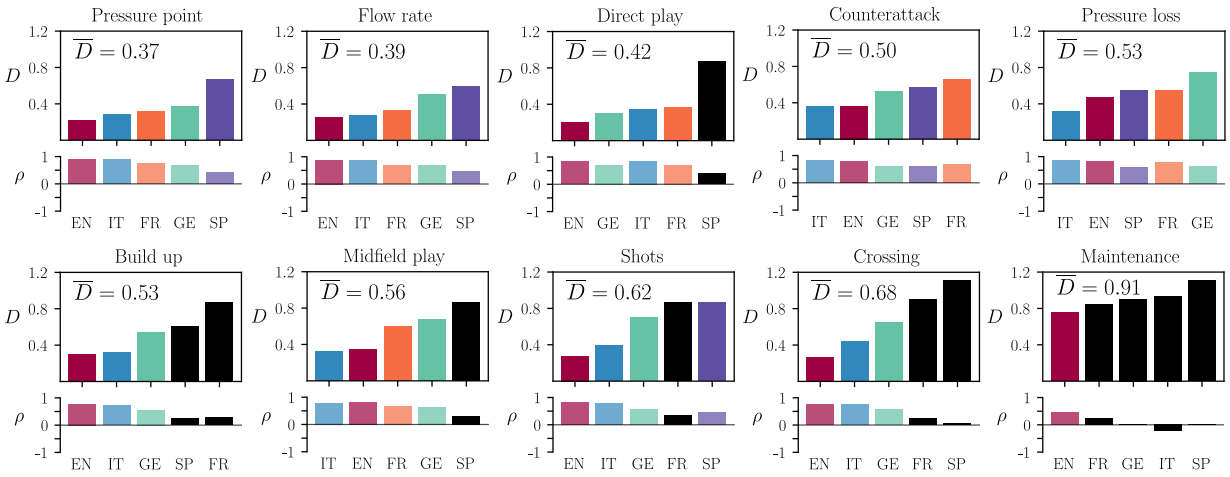

FIG. 3. Distancia D y coeficiente de person ρ, entre el rating verdadero y el rating dado por las metricas. Cada panel contiene la informacion de una metrica y las barras muestran el valor de D obtenido en cada liga. Las barras en cada panel estan ordenadas en orden creciente respecto del valor de D.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

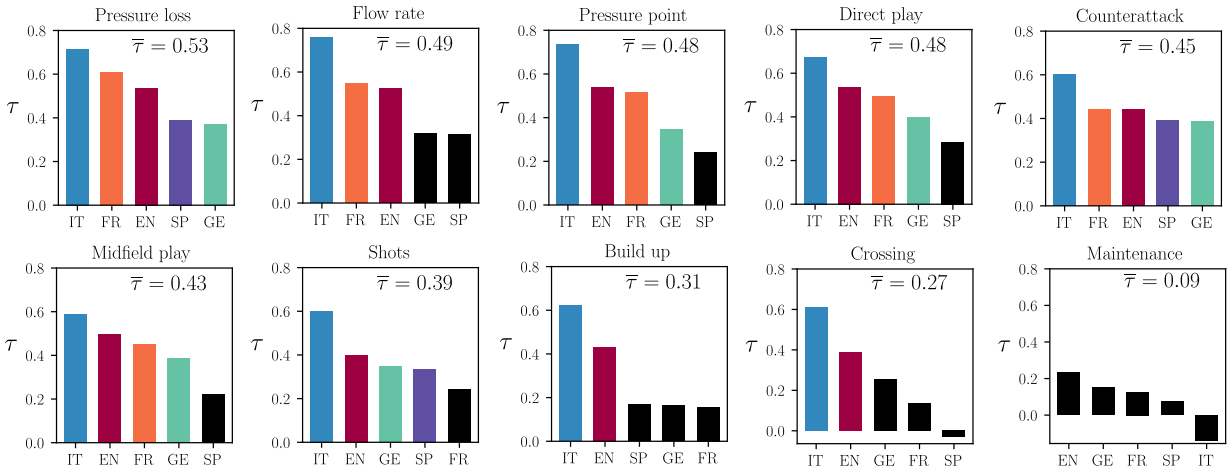

FIG. 4. Coeficiente de Kendall, τ, entre los rankings verdaderos y los rankings dados por las metricas. Cada panel contiene la informacion de una metrica y las barras muestran el valor de τ obtenido en cada liga. Las barras en cada panel estan ordenadas en orden decreciente respecto del valor de τ. Asimismo,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

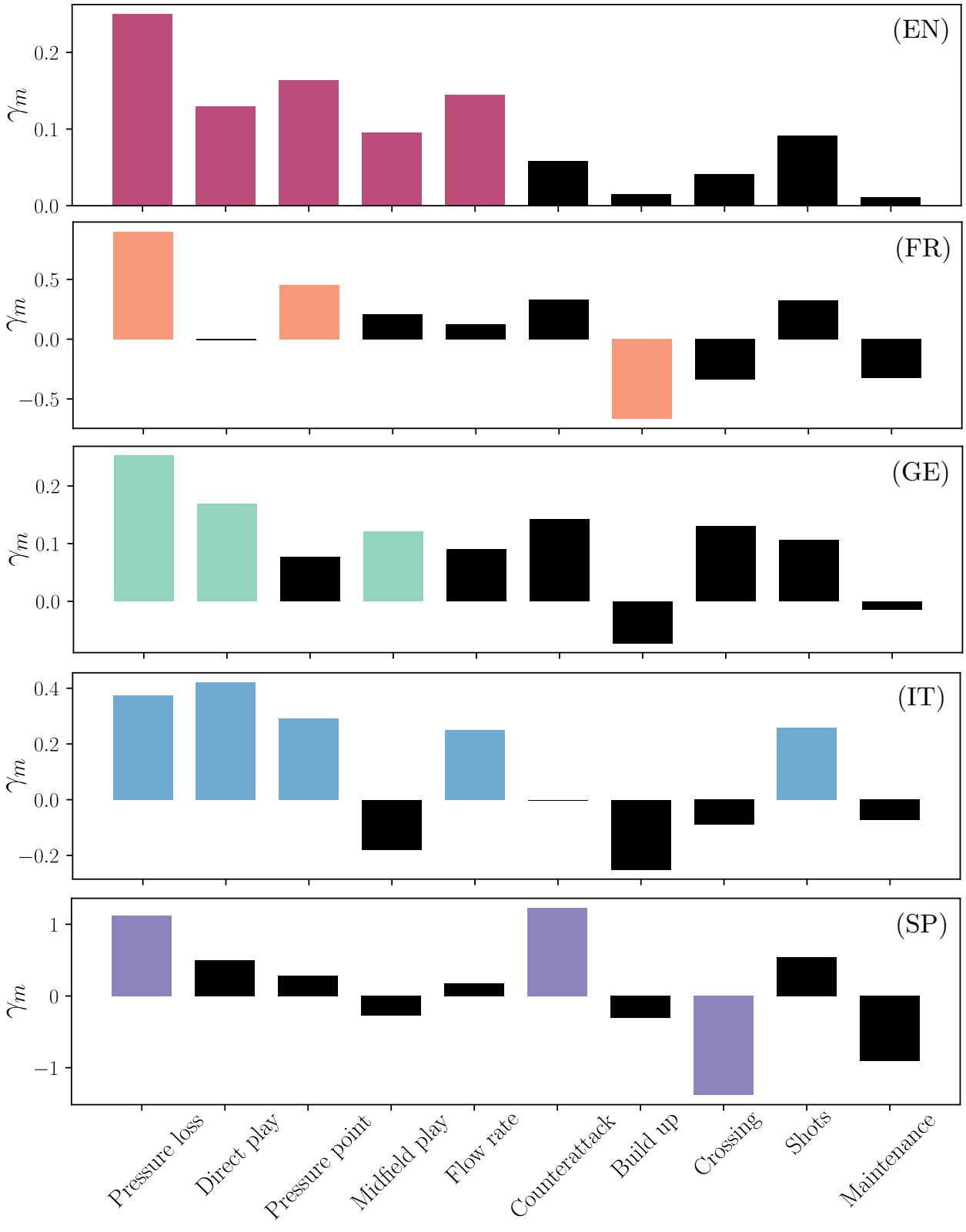

FIG. 5. Relevancia de las metricas en el rating multivariado. Cada panel muestra los valores de γ_i en cada metrica para una liga dada. Las barras negras muestran los casos donde no hay significancia estadistica.

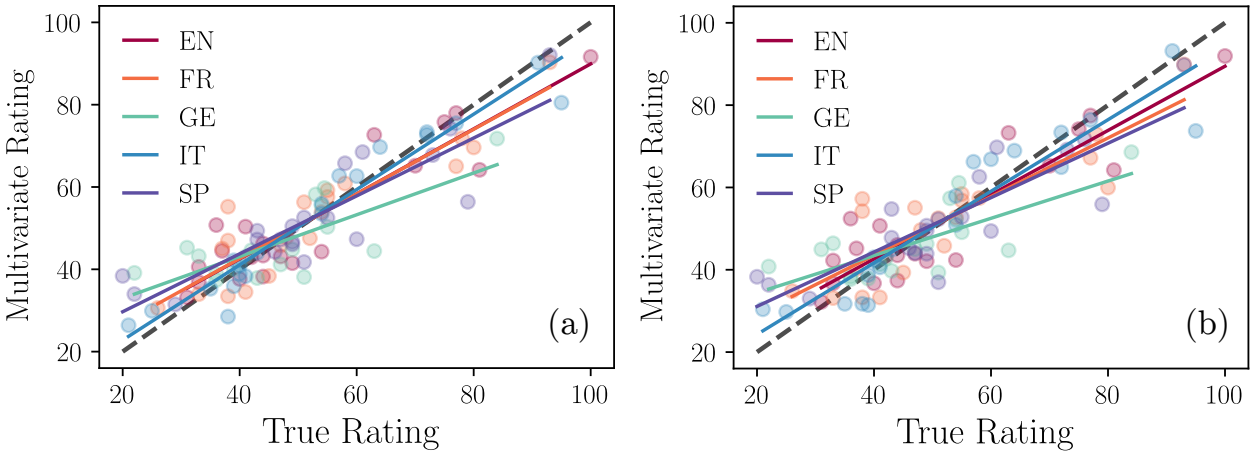

FIG. 6. Comparacion el true rating y el multivariate rating. (a) Utilizando todas las variables para calcular el raiting multivariado. (b) Utilizando solo las variables relevantes de cada liga (ver Fig. 5). En ambos paneles, los scatters muestran los valores exactos para cada equipo de cada liga, las 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.