



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

DINÁMICA DE RANGO EN DEPORTES Y JUEGOS

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Físico

PRESENTA:

José Antonio Morales Álvarez

TUTOR:

Dr. Carlos Gershenson García

Ciudad de México, 2018





Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Hoja de datos del jurado

Datos del alumno: *Morales*

Álvarez

José Antonio

5525264401

Universidad Nacional Autónoma de México

Facultad de Ciencias

Física

106003222

Datos del tutor: *Dr.*

Carlos

Gershenson

García

Datos del sinodal 1: *Dr.*

Adonis Germinal

Cocho

Gil

Datos del sinodal 2: *Dr.*

Jorge Andrés

Flores

Valdés

Datos del sinodal 3: *Dr.*

Ruben Yvan Maarten

Fossion

Datos del sinodal 4: *Dr.*

Ricardo Atahualpa

Solórzano

Kraemer

Datos del trabajo escrito: *Dinámica de rango en deportes y juegos*

148 p

2018

A mis padres que desde siempre me han dado todos su amor y han hecho todo para impulsarme a alcanzar mis sueños.

A mi hermana, quien además de ser pariente sanguíneo, es mi mejor amiga y siempre me ha acompañado en los momentos buenos y difíciles.

A la Facultad de Ciencias y a la Universidad Nacional Autónoma de México, por la formación y oportunidades que me han dado.

Reconocimientos

Expreso mi más grande agradecimiento a la *Universidad Nacional Autónoma de México*, que desde los 12 años me permitió formar parte de ella aceptándome en el programa de *Iniciación Universitaria* formándome como estudiante, como individuo y como ciudadano. A la *Facultad de Ciencias* a la cual le debo mi formación profesional, como físico y futuro científico. Espero poder estar a la altura de las expectativas que la Universidad tiene para con sus estudiantes y ser un digno representante de la misma ante la sociedad mexicana y el mundo.

Agradezco el apoyo del Dr. Carlos Gershenson García, quien me ha dado apoyo invaluable para la realización de este trabajo y me ha dado acceso a una perspectiva más amplia en el mundo de la investigación. También agradezco a Dr. Germinal Cocho, Dr. Jorge Flores, Dr. Rosalío Rodríguez, Dr. Carlos Pineda, a Dr. Gerardo Iniguez y Sergio Sánchez quienes desde 2015 me acogieron dentro de su grupo de investigación y que me han permitido participar activamente en trabajos diversos y gracias a éstos pude realizar esta tesis que es sólo una prueba de las inmensas conclusiones a las que se han llegado en dicho grupo, *grupo de lenguajes*.

A mis padres, que siempre estuvieron allí para mí, moral, económica, emocional y físicamente cuando los necesité. Sin su amor ni su apoyo jamás habría logrado llegar a donde estoy. Espero que con estos logros pueda corresponder a ese apoyo y amor que me han otorgado incondicionalmente toda mi vida y hacerlos sentir orgullosos. A mi hermana, que siempre me ha apoyado y divertido desde que estamos juntos y que en los peores momentos nos hacíamos reír mutuamente para hacer más llevadero todo; sin duda, eres mi mejor amiga. Ustedes son las personas más importantes en mi vida. También quiero agradecer a mis tíos Carmen Morales y Arturo Morales, quienes siempre están atentos de mí y de mi hermana y quienes siempre se emocionan por cada uno de nuestros logros.

A Ximena Pérez Baena, que en este último año de mi vida ha sido la persona más importante, todo tu amor y apoyo han sido fundamentales para que yo siga adelante. Gracias por tu sonrisa, por tus chistes, por tus palabras y por estar siempre conmigo; eres uno de los elementos más importantes que me impulsó a terminar este trabajo.

El camino hasta aquí habría sido muy complicado si no fuera por los amigos, los que hacen más divertido el día a día. A los mejores amigos que he tenido en la vida, Jonathan López Ruiz y Marisol Villegas Jiménez, desde que los conozco me divierto

mucho cada vez que nos encontramos y su apoyo en los momentos más difíciles ha sido invaluable para mí. También amigos que me fui encontrando en el camino y que conocí en la preparatoria como Geraldine, Miguel Vilchis, Abril Manzanarez, Silvana, Rafael Díaz, Ricardo González y Brenda Cobos y a todos los amigos que olvidé mencionar, les pido una disculpa pero siempre los tendré presente en el desarrollo de mi vida como estudiante y como persona.

A uno de mis mejores amigos, John Progatsky, con quien siempre he podido contar en cualquier momento y con su amistad incondicional hemos salido adelante juntos. A Nancy Sánchez, cuya amistad y sentido del humor siempre hace las cosas más llevaderas. A Diego Alcalá a quien desde que conozco me ha apoyado con todo y escuchado mis problemas.

Gracias a los amigos que encontré durante la carrera; a Luis Enrique, Juan Sánchez, Fernando Jefe con quienes siempre trabajé en equipo para salir adelante durante la carrera. A Pablo Reséndiz, Marcela Ordorica y Daniel Martínez, los mejores amigos en la práctica, en la diversión y en la informalidad, muchas gracias por hacer muy ameno este camino. También a amigos que tendré siempre presente y con quienes he pasado muy buenos momentos, Charlie Olmos, Luis Hernández, Miguel Hernández, Omar Figueroa, Uriel Luviano, Abraham Sánchez, David Dávalos, Álvaro Díaz, Adolfo Hernández, Laura Torres, Sergio Sánchez y muchos más a los que les pido disculpa si olvidé mencionarlos.

Y sin duda los profesores son quienes forjan el qué camino específico se seguirá tomando en cuenta su ejemplo, inspiración y conocimientos. Al profesor Juan Pablo Balderas quien es el primer profesor en mi vida que consolida mi gusto por las matemáticas. A la profesora Judith que se convirtió en una guía importante. A la profesora de la primaria Bernarda que me dio la primera perspectiva formal del mundo. Al profesor Héctor Méndez Lango quien con su pasión por la enseñanza me inspiró a querer aprender de la manera más formal todo lo que estudie siempre teniendo presente un poco de humor y de manera colorida, usted hizo amena la primera mitad de la carrera. Y a Saúl Ramos Sánchez que ha sido mi guía más importante en el área de la Física y con quien espero trabajar muy bien en el futuro.

Finalmente a la DGAPA-UNAM, dentro del programa de apoyo a proyectos de investigación e innovación tecnológica por medio del proyecto PAPIIT IG100518. Y al CONACYT que por medio del apoyo a ayudantes de investigadores SNI III, pude enfocarme en los proyectos que dieron nacimiento a este trabajo.

Resumen

Los deportes y juegos son sistemas complejos jerárquicos (sistemas donde sus elementos interaccionan para dar lugar a un ordenamiento, ‘ranqueo’, de los mismos) muy estudiados debido a su influencia en muchos ámbitos sociales; en este trabajo se estudió el caso de 12 deportes y juegos. El ordenamiento de los elementos de estos sistemas se lleva a cabo a partir de sus puntajes, que es un número asignado a los elementos por una organización reguladora de la disciplina en cuestión.

En esta tesis se estudió la manera en que estos puntajes se distribuyen, lo que llamaremos distribución de rango, y un primer acercamiento al estudio de la evolución del sistema partiendo de estas distribuciones. Se proponen 5 modelos teóricos para aproximar de manera funcional dichas distribuciones y las ponemos a prueba con dos bondades de ajuste: el coeficiente de determinación R^2 y el índice p de Kolmogorov-Smirnov.

Se definen medidas dinámicas: diversidad de rango, probabilidad de cambio, entropía de rango y complejidad de rango; medidas que nos permiten hacer un análisis cuantitativo de la evolución de los “rankings” u ordenamientos para los sistemas en cuestión. Estas medidas estudian la forma en que los rangos (lugares en el ordenamiento, como el primer o segundo lugar) son ocupados a lo largo del tiempo desde diferentes perspectivas.

Al final se presentan dos modelos que intentan reproducir esta evolución de los sistemas basándose en los resultados obtenidos con las medidas dinámicas anteriormente mencionadas: modelo del caminante aleatorio y el Modelo Nulo.

Índice general

Índice de figuras	xv
Índice de tablas	xxiii
1. Introducción	1
1.1. Sistemas complejos y su importancia.	2
1.2. Los deportes y juegos desde el punto de vista de los sistemas complejos.	5
1.3. Objetivos y estructura general.	6
2. Bases de datos.	9
2.1. Las bases de datos de las disciplinas.	9
2.2. Características de ordenamiento.	12
2.2.1. Jugadores de ajedrez, masculino y femenino (FIDE)	12
2.2.2. Clubes de Fútbol, (FCWR-C) y Goleadores de Fútbol en equipos de clubes (FCWR-G)	14
2.2.3. Equipos nacionales de Fútbol, (FIFA)	15
2.2.4. Jugadores de Golf, (OWGR)	16
2.2.5. Corredores de NASCAR, para Busch Grand Nation (NASCAR-B) y Winston Cup Grand National (NASCAR-W)	17
2.2.6. Jugadores de Póquer, (GPI)	17
2.2.7. Jugadores de tabla sobre nieve, (WSD)	18
2.2.8. Jugadores de tenis, (ATP)	19
2.2.9. Ganancias de jugadores de E-Sports, (ESE)	20
2.3. Algunas consideraciones adicionales	20
3. Distribución de rango. Un análisis estadístico.	23
3.1. Definición y motivación.	24
3.2. Modelos para la distribución de rango.	25
3.2.1. Distribución de rango vs. distribución de frecuencias relativas . .	28
3.2.2. Los modelos a utilizar	29
3.2.3. Ley de Zipf. Distribuciones β y γ	30
3.3. Equivalencia entre distribución de rango (DRe) y distribución acumulativa de los puntajes (DCe).	32

3.4.	Métodos y bondad de ajuste.	34
3.4.1.	Coefficiente de determinación R^2	35
3.4.2.	El índice p de Kolmogorov-Smirnov.	35
3.5.	Comparación entre modelos y distribución de rango.	37
3.6.	Bondades de ajuste para los modelos de distribución de rango a lo largo del tiempo.	42
4.	Dinámica de rango para deportes y juegos: diversidad, probabilidad de cambio, entropía y complejidad.	49
4.1.	Los Espaguetis.	50
4.1.1.	¿Podemos encontrar regularidades en los espaguetis?	51
4.2.	Diversidad de rango.	53
4.2.1.	Definición.	54
4.2.2.	Un primer vistazo a la diversidad de rango.	55
4.2.3.	Modelo para la diversidad de rango.	58
4.2.4.	Posible origen analítico de la diversidad de rango.	61
4.2.5.	Diversidad de rango versus distribución de rango.	63
4.2.6.	Universalidad.	64
4.3.	Probabilidad de cambio.	66
4.3.1.	Definición.	67
4.3.2.	Probabilidad de cambio para deportes y juegos.	67
4.4.	Entropía y complejidad de rango.	70
4.4.1.	Motivación.	70
4.4.2.	Definiciones.	71
4.4.3.	Entropía y complejidad de rango para deportes y juegos.	72
4.5.	Discusión final para la dinámica de rango.	76
5.	Modelos.	77
5.1.	Modelo del caminante aleatorio	78
5.1.1.	Descripción de la implementación del modelo.	79
5.1.2.	Comparativa entre los sistemas generados con el modelo del caminante aleatorio y los sistemas reales.	81
5.2.	Modelo Nulo.	85
5.2.1.	Descripción de la implementación del modelo.	87
5.2.2.	Comparativa entre los sistemas generados con el modelo nulo y los sistemas reales.	88
5.3.	Modelo Nulo versus Modelo del caminante aleatorio	92
6.	Conclusiones.	93
A.	Cálculo del índice p de Kolmogorov-Smirnov para distribuciones de rango.	97
B.	Espaguetis ejemplos.	113

C. Figuras Adicionales.	119
D. Artículo original publicado.	127
Bibliografía	145

Índice de figuras

1.1. Imagen que ilustra redes de distinta índole. a. Representa un anillo de 10 nodos, los cuales están conectados con los vecinos más cercanos. b. Ahora los 10 nodos se encuentran conectados todos contra todos. c. Gráfica construida aleatoriamente. d. Gráfica invariante de escala. Imagen y descripción obtenidas de [1]	4
3.1. Distribuciones de rango para las 12 bases de datos. En todos los casos se observa que las distribuciones de rango son decrecientes, como uno esperaría de acuerdo a la definición que anteriormente se dio. Las fechas correspondientes a las distribuciones de rango aquí graficadas son, para cada deporte las siguiente: FIDE-F(Abril 2016), FIDE-M(Abril 2016), FCWR-C(Semana 53 del 2014), FIFA(Junio 2017), FCWR-G(Semana 33 de 2017), OWGR(21/05/2017), NASCAR-B(2015), NASCAR-W(2013), GPI(31/05/2017), WSD(26/03/2018), ATP(27/12/2010), ESE(2016). Estas fechas corresponden a la última disponible en las bases de datos.	26
3.2. Proceso para transformar la distribución de rango empírica (DRe) a la distribución cumulativa empírica (DCe) Vemos que el proceso de transformación va desde a) a d), mostrando la equivalencia entre la distribución de rango empírica y la distribución cumulativa empírica.	33
3.3. Comparación de los datos de ranking con los modelos m_1 y m_2. Gráfica que muestra la distribución de rango (puntaje contra rango k) para todos los deportes considerados aquí, en una rodaja temporal, (las mismas rodajas mencionadas en la Figura 3.1) así como los ajustes a los modelos correspondientes a las ecuaciones 3.9 y 3.10 . En general, la ley de Zipf (m_1) no reproduce de manera satisfactoria los datos para ningún deporte o juego estudiado aquí. La distribución Gamma (m_2) parece ser más apropiadas en algunos casos. Sólo se presentan las gráficas de ajustes para una rodaja temporal en cada base de datos para ilustrar la forma funcional de los modelos y evidenciar algunas diferencias en sus comportamientos.	38

<p>3.4. Comparación de los datos de ranking con los modelos m_3 y m_4. Gráfica que muestra la distribución de rango (puntaje contra rango k) para todos los deportes aquí considerados, en una rodaja temporal, (las mismas rodajas mencionadas en la Figura 3.1) así como los ajustes a los modelos correspondientes a las ecuaciones 3.11 y 3.12. Observamos que el modelo m_3 sufre de caídas abruptas cuando se adquieren valores grandes en k, las cuales tal vez dificultan el parecido con la distribución empírica de rango, sin embargo, esto sólo se ve para las rodajas temporales aquí graficadas. El modelo m_4 parece ser el más adecuado en la mayoría de los casos como podemos observar, y esto se puede deber a los grados de libertad adquiridos por mayor número de parámetros.</p>	39
<p>3.5. Comparación de los datos de ranking con el modelo m_5. Gráfica que muestra la distribución de rango (puntaje contra rango k) para todos los deportes aquí considerados en una rodaja temporal, (las mismas rodajas mencionadas en la Figura 3.1) así como los ajustes al modelo correspondiente a la Ecuación 3.13. Podemos apreciar que en muchos casos el modelo parece reproducir bien los datos. Los casos de NASCAR-B, NASCAR-W y ESE son interesantes, pues parece ser que hay cambios bruscos en el comportamiento funcional de los puntajes a partir de cierto rango, pudiéndose dar el caso de que la ley doble Zipf pueda ser la adecuada para diferenciar dos regímenes en el ranqueo para ciertos deportes o juegos.</p>	40
<p>4.1. Evolución temporal en el ranking para los 8 equipos punteros entre la Semana 9 del 2013 y la semana 14 del 2013. Vemos que 9 diferentes equipos ocupan los primeros ocho rangos en estas fechas. Los equipos que en algún momento ocupan el primer lugar son: Barcelona, Real Madrid y Atlético Madrid. Algo interesante es que en estas fechas, el Manchester United se mantiene siempre en el rango $k = 6$.</p>	50
<p>4.2. Evolución temporal de los rangos para jugadores y equipos de las primeras 6 disciplinas. Gráfica que muestra el cambio en el rango k a lo largo del tiempo t para todos los jugadores/equipos en cada deporte y juego considerado en este estudio: FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G y OWGR. Aquí incluyo los espaguetis de todos los equipos/jugadores que alguna vez ocuparon uno de los primeros 30 lugares a lo largo de todo el tiempo que se tiene disponible. Nótese que los jugadores/equipos en los rangos bajos tienden a cambiar menos que los de rango más altos, incluso cuando los rankeos en todas las actividades varían a diferentes tasas y la resolución temporal correspondiente varía entre semanas a meses (ver Tabla 2.1).</p>	52

4.3.	Evolución temporal de los rangos para jugadores y equipos de las últimas 6 disciplinas. Gráfica que muestra el cambio en el rango k a lo largo del tiempo t para todos los jugadores/equipos en cada deporte y juego considerado en este estudio: NASCAR-B, NASCAR-W, GPI, WSD, ATP, ESE. Aquí incluyo los espaguetis de todos los equipos/jugadores que alguna vez ocuparon uno de los primeros 30 lugares a lo largo de todo el tiempo que se tiene disponible. Para este último conjunto de datos el comportamiento, de los elementos den los primeros lugares no es tan estable como vimos en la Figura 4.2 , para ATP sí se puede apreciar la presencia de la regularidad antes mencionada. Sin embargo, en ESE apreciamos que la permanencia de los elementos en el primer lugar es básicamente sólo de un tiempo, recordemos que aquí se deben considerar muchos factores, tales como la resolución temporal correspondiente varía entre semanas a meses (ver Tabla 2.1).	53
4.4.	Diversidad de rango de deportes y juegos. Gráfica que muestra la diversidad de rango $d(k)$ para todos los conjuntos de datos (puntos azules), también se presenta el índice de apertura para cada una de las disciplinas Ω , como definimos en la Ecuación 4.2	57
4.5.	Diversidad de rango de deportes y juegos en escala semilogarítmica. Gráfica que muestra la diversidad de rango $d(k)$ para todos los conjuntos de datos (puntos azules), así como los ajustes a Φ (líneas rojas). Incluimos los valores de μ , σ , y el parámetro de bondad de ajuste R^2 también.	60
4.6.	Similaridad en la diversidad de rango normalizada entre los deportes y juegos. Gráfica que muestra una comparación de la diversidad de rango $d(k)$ para todas las actividades consideradas. Con los valores de μ y σ obtenemos el ajuste a Φ , hemos reescalado la abscisa. Como referencia incluimos la forma básica de la Ecuación 4.3 (delgada línea roja), con $\mu = 0$, y $\sigma = 1$. Estos resultados indican que todas las actividades tienen la misma forma funcional para la diversidad de rango.	65
4.7.	Probabilidad de cambio para deportes y juegos en escala semilogarítmica. Gráfica que muestra la probabilidad de cambio $d(k)$ para todos los conjuntos de datos (puntos verdes), así como los ajustes a Φ (líneas rojas). Incluimos los valores de μ y σ así como el parámetro de bondad de ajuste R^2	69
4.8.	Entropía de rango para deportes y juegos. Esta medida dinámica es calculada de acuerdo a lo que definimos en la Ecuación 4.15 , observamos la entropía de rango es muy grande (con valores cercanos a 1) para la mayoría de los rangos en todos los deportes y juegos aquí considerados.	74

4.9.	Complejidad de rango para deportes y juegos. Esta medida dinámica es calculada de acuerdo a lo definido en la Ecuación 4.17 . Observamos que la complejidad es alta para los rangos pequeños en la mayoría de los sistemas, las clara excepciones son las de siempre NASCAR-B, NASCAR-W y ESE.	75
5.1.	Comparación entre las diversidades de rango empíricas y simuladas con el Modelo del caminante aleatorio. Gráfica que muestra la diversidad de rango $d(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ, σ para los datos simulados. El modelo del caminante aleatorio parece reproducir cualitativamente las diversidades de rango observadas en todos los deportes y juegos considerados aquí, a pesar de las claras diferencias que revelan que este modelo es ineficiente para obtener resultados satisfactorios cuantitativamente.	83
5.2.	Comparación entre las probabilidades de cambio empíricas y simuladas con el Modelo del caminante aleatorio. Gráfica que muestra la probabilidad de cambio $p(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ, σ para los datos simulados.	84
5.3.	Descripción de la implementación del modelo nulo para una rodaja temporal. Diagrama que ilustra de manera detallada cada uno de los pasos que se llevan a cabo en el proceso de reordenamiento para el modelo nulo.	86
5.4.	Comparación entre las diversidades de rango empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la diversidad de rango $d(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ, σ para los datos simulados. El modelo nulo parece reproducir cualitativamente las diversidades de rango observadas en todos los deportes y juegos considerados aquí, pero en menor medida que con el caso del modelo del caminante aleatorio. Se registran caídas más abruptas a rangos altos y $d(k)$ de los datos simulados quedan por encima de la $d(k)$ de los datos empíricos para los rangos bajos.	90

<p>5.5. Comparación entre las probabilidades de cambio empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la probabilidad de cambio $p(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ, σ para los datos simulados.</p>	91
<p>A.1. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIDE-F Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	100
<p>A.2. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIDE-M Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	101
<p>A.3. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FCWR-C Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	102
<p>A.4. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIFA Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	103
<p>A.5. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FCWR-G Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	104
<p>A.6. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de OWGR Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	105
<p>A.7. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de NACAR-B Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.</p>	106

A.8. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de NACAR-W Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.	107
A.9. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de GPI Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.	108
A.10. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de WSD Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.	109
A.11. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de ATP Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.	110
A.12. Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de ESE Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.	111
B.1. Espaguetis para FIDE-F para los rangos $k = 12681, 6340, 1$	113
B.2. Espaguetis para FIDE-M para los rangos $k = 13500, 6750, 1$	113
B.3. Espaguetis para FCWR-C para los rangos $k = 850, 425, 1$	114
B.4. Espaguetis para FIFA para los rangos $k = 200, 100, 1$	114
B.5. Espaguetis para FCWR-G para los rangos $k = 400, 200, 1$	115
B.6. Espaguetis para OWGR para los rangos $k = 1150, 575, 1$	115
B.7. Espaguetis para NASCAR-B para los rangos $k = 76, 38, 1$	116
B.8. Espaguetis para NASCAR-W para los rangos $k = 50, 25, 1$	116
B.9. Espaguetis para GPI para los rangos $k = 1795, 897, 1$	117
B.10. Espaguetis para WSD para los rangos $k = 1413, 706, 1$	117
B.11. Espaguetis para ATP para los rangos $k = 1600, 800, 1$	118
B.12. Espaguetis para ESE para los rangos $k = 400, 200, 1$	118

- C.1. **Similaridad en la probabilidad de cambio normalizada para deportes y juegos.** Gráfica que muestra una comparación de la probabilidad de cambio $p(k)$ para todas las actividades consideradas. Con los valores de μ y σ obtenemos el ajuste de Φ , hemos reescalado la abscisa como hicimos en el caso de la diversidad de rango. Como referencia concluimos que forma básica de la [Ecuación 4.3](#) (línea roja en la gráfica), con $\mu = 0$ y $\sigma = 1$. Estos resultados indican que todas las actividades tienen la misma forma funcional para la probabilidad de cambio. 119
- C.2. **Diferencias entre la probabilidad de cambio y la diversidad de rango $p(k) - d(k)$ en escala semilogarítmica.** En la figura podemos apreciar que todos los valores de la diferencia entre estas dos medidas es siempre mayor o igual a cero, lo que comprueba que $p(k) \geq d(k)$ 120
- C.3. **Distribución de los tamaños relativos del cambio de frecuencias $[k_{t+1} - k_t]/k_t$, para el caso de las palabras en inglés. Las palabras correspondientes a las cabezas de la diversidad (en dorado), las que están en el cuerpo de la diversidad (azul) y las que están en la cola (en verde).** La distribución gaussiana con una desviación estándar $\sigma = 0.0575$ está graficada en rojo para hacer una comparativa. En verde se grafica una curva estrecha que representa una distribución Lorentziana que mejor se ajusta con el promedio de las tres distribuciones empíricas que se muestran aquí (las de los tres grupos de palabras). Notemos que las palabras que están en la cabeza, los saltos relativos son no se asemejan a la distribución Gaussiana. Para el caso de las distribuciones de las palabras en la cola y cabeza tenemos mucha similitud mutua. Notemos que, en promedio, los saltos relativos parecen muy independientes de los valores de k , por lo que parece que reproducir los rangos requiere de un modelo invariante de escala. Las curvas de la distribución gaussiana y lorentziana centradas en cero son las que mejor se ajustaron a los datos presentados aquí. Aunque la distribución lorentziana parece ajustar mejor a los datos que la gaussiana, usamos la distribución gaussiana en el modelo del caminante aleatorio, pues las largas colas de la lorentziana implicaría grandes saltos de los rangos de las palabras (algo que no se observa en las bases de datos, pues $d(k)$ indica lo contrario). Tal vez una lorentziana truncada en las colas funcionaría mejor, pero complicaría bastante el modelo; optaremos por utilizar la distribución gaussiana. Imagen y descripción tomados de [\[2\]](#). 121
- C.4. **Comparación entre las entropías de rango empíricas y simuladas con el Modelo del Caminante Aleatorio.** Gráfica que muestra la entropía de rango $E(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados). 122

C.5. **Comparación entre las complejidades de rango empíricas y simuladas con el Modelo del Caminante Aleatorio.** Gráfica que muestra la complejidad de rango $C(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados). 123

C.6. **Comparación entre las entropías de rango empíricas y simuladas con el Modelo Nulo.** Gráfica que muestra la entropía de rango $E(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados). 124

C.7. **Comparación entre las complejidades de rango empíricas y simuladas con el Modelo Nulo.** Gráfica que muestra la complejidad de rango $C(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados). 125

Índice de tablas

2.1. Resumen de los datos de ordenamiento para cada deporte y juego considerado en este estudio. La tabla enlista las propiedades más importantes utilizadas aquí (incluyendo la fuente de los datos, el periodo de tiempo, la resolución del ordenamiento, y el número de jugadores/equipos). Para poder tener una distribución homogénea de jugadores/equipos en cada rodaja de tiempo para cada actividad, se ignoraron algunos datos que conformaban las bases completas. Cada una de las bases de datos, o en este caso los deportes y juegos, estarán representados por unas siglas características. Por ejemplo, las siglas correspondientes a los equipos nacionales de fútbol son FIFA, y siempre que usemos dichas siglas nos estamos al sistema en cuestión y todo lo que implica.	11
3.1. Parámetros de ajuste de los 5 modelos m_i a las DRe de los 12 deportes y juegos correspondientes a las últimas fechas disponibles en las bases de datos. FIDE-F(Abril 2016), FIDE-M(Abril 2016), FCWR-C(Semana 53 del 2014), FIFA(Junio 2017), FCWR-G(Semana 33 de 2017), OWGR(21/05/2017), NASCAR-B(2015), NASCAR-W(2013), GPI(31/05/2017), WSD(26/03/2018), ATP(27/12/2010), ESE(2016)	41
3.2. Promedios y desviaciones estándar las bondades de ajuste R^2 , p y la estadística de Kolmogorov D para los ajustes realizados a todas las fechas con las que cuentan las bases de datos de los deportes y juegos: FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G, OWGR	47
3.3. Promedios y desviaciones estándar las bondades de ajuste R^2 , p y la estadística de Kolmogorov D para los ajustes realizados a todas las fechas con las que cuentan las bases de datos de los deportes y juegos: NASCAR-B, NASCAR-W, GPI, WSD, ATP, ESE	48

Introducción

El presente trabajo estudiará una clase de sistemas complejos: los sistemas complejos jerárquicos, en los cuales se presentan interacciones entre los elementos que los conforman para que haya formaciones jerárquicas, es decir, compiten entre sí para resultar ganadores o vencedores. Una clase muy evidente de sistemas complejos jerárquicos son justamente los deportes y juegos y pueden ser descritos como tales debido a la gran cantidad de factores que influyen la dinámica competitiva y el rendimiento en ellos, como interacciones de red, heterogeneidades humanas y ambientales y otras tendencias a nivel individual y grupal. En este tipo de sistemas se tiene una noción entonces de cuál elemento es el mejor, cuál es el peor, cuál es el segundo mejor, dando lugar a un ordenamiento de los elementos de acuerdo a su rendimiento siguiendo ciertos criterios; a este proceso de ordenamiento le llamaremos *ranqueo*, y consiste en asignarle a cada elemento un rango, por ejemplo, al mejor elemento se le asigna el rango 1, al segundo mejor elemento se le asigna el rango 2 y así sucesivamente, siempre tomando como criterio de ordenamiento al rendimiento de los elementos cuando interactúan entre ellos. En particular, el rendimiento de los jugadores y equipos está influenciado por una gran variedad de condiciones de aspecto: económico, político o geográfico que determinan sus rangos y se pueden utilizar para predecir el rendimiento de los mismos.

Más aún, las reglas de competición relativamente simples y las medidas de rendimiento asociadas con deportes y juegos permiten explorar mecanismos básicos de interacción conduciendo a una formación jerárquica, que es común a muchos sistemas dominados por la competición, no sólo para actividades deportivas sino para otros sistemas sociales, biológicos y económicos. Con estos objetivos en mente, la disponibilidad de una gran estructura de bases de datos relacionadas con deportes, equipos, y jugadores permite a los investigadores desarrollar múltiples análisis estadísticos, en particular, con respecto a la estructura y dinámica del desempeño en diversos rankeos. Para estudiar este tipo de sistemas, debemos definir qué aspecto queremos abarcar. Por ejemplo, la forma en que los elementos son ordenados, que es asignándoles un puntaje bajo ciertos criterios o estudiar la evolución de estas formaciones jerárquicas o rankeos, es decir, analizar cómo cambia el ordenamiento de los elementos conforme avanza el tiempo.

En este trabajo estudiaremos ambos aspectos desde una perspectiva específica. La disponibilidad de datos ha hecho posible no sólo el estudio de la distribución de puntajes, los cuales determinan el ranqueo de los elementos del sistema, sino que también ha

sido posible estudiar la evolución de los rankeos mismos. Por lo que en este trabajo seguiremos esas dos líneas: el estudio de la distribución de los puntajes asignados a los elementos (este aspecto es estático en el tiempo, pues analizaremos los puntajes correspondientes a una fotografía/rodaja en el tiempo) y la evolución de ranqueo del sistema, que como veremos más adelante, lo más conveniente es ocuparse de la evolución en la forma que se ocupan los rangos y no en cómo evolucionan sus elementos.

1.1. Sistemas complejos y su importancia.

Dado que consideramos a los deportes y juegos como sistemas complejos, vale la pena entender qué es un sistema complejo y cuáles son las formas posibles en que se pueden analizar y estudiar. Un sistema complejo es un sistema compuesto por componentes que pueden interactuar entre sí. En muchos casos es conveniente representar a dichos sistemas como redes donde los nodos representan las componentes y las líneas que unen a esos nodos son las interacciones que tienen entre sí. [3] En la [Figura 1.1](#) se ilustran varios ejemplos de redes.

De hecho, las redes que representa a los sistemas complejos (redes complejas) son el principal objeto de estudio en el campo de dichos sistemas. Las redes complejas las encontramos en todas partes: sistemas relacionados con la obra humana, en materia orgánica e inorgánica, estructuras naturales y antropogénicas. Ejemplos: estructuras moleculares, redes climáticas, redes de comunicación e infraestructura, redes sociales y económicas. Es importante estudiar la topología de estas redes (la interrelación en la estructura) y su dinámica de los sistemas complejos. Se ha visto recientemente que en sistemas reales, existe una estrecha relación entre los microestados y macro estados del mismo. La teoría de redes permite estudiar esta relación claramente, en donde los nodos y los vértices representan los microestados, mientras que la red en sí, su topología y dinámica representa el macroestado del sistema. Los nuevos retos de la teoría de redes es el acoplamiento entre redes, la dinámica de las redes, interrelación entre la estructura, interdependencia en una red dada, co-evolución de las redes, y propiedades espaciales de las mismas. [4]

En el artículo [5] se justifica de cierta manera que la caracterización de redes complejas reales dan lugar de manera natural a la aparición de una invariancia de escala y una estructura jerárquica. Donde la invariancia de escala se evidencia del hecho de que la distribución de grado de los vértices de la red se comporta como una ley de potencias. Dado que los modelos propuestos para generar redes que cumplan tanto la propiedad de invariancia de escala como la de una estructura jerárquica, también se presenta un modelo que intenta capturar ambas características. Muchos resultados recientes en la topología de redes reales indican que la aparente aleatoriedad de los sistemas complejos que representan esconden mecanismos genéricos y cierto orden que son cruciales para el estudio de muchos fenómenos en la naturaleza.

En [6] se mencionan cuáles son los retos para poder hacer predicciones en sistemas tecno-sociales. Siendo de gran importancia lograr hacer predicciones de éstos, pues estamos en constante interacción o formamos parte de estos sistemas. Algunos de estos

sistemas tecno-sociales son sistemas como la "World Wide Web", la tecnología de comunicación por Wifi, o infraestructuras de movilidad y transporte. Como se menciona, es importante lograr hacer predicciones en estos sistemas pues son de vital importancia, pues darían información útil tal como anticipar tendencias, evaluar riesgos y el control de futuros eventos. Un ejemplo importante en el que la predictibilidad juega un papel importante es el caso de la meteorología, pues gracias a la colección de datos históricos que se tienen y a varios modelos que ayudan a describir la dinámica de fluidos, se puede predecir la formación de algunos huracanes. Sin embargo, en el caso de sistemas tecno-sociales, llegar a predecir es más complejo, pues no se tiene un conocimiento tan profundo del comportamiento social y humano como en el caso de las leyes de la Física. Pero muchos esfuerzos se han hecho para el estudio de sistemas tecno-sociales, y se han hecho algunas caracterizaciones con ayuda de la teoría de redes.

En [1] se enfatiza la aparición de las redes en todas las áreas de la ciencia. El estudio de una red se enfoca en el estudio de su estructura, ya que el entenderla, da información sobre el sistema que se esté describiendo. Pero sin duda, un comentario muy importante que aquí se menciona es la importancia de caracterizar la anatomía(estructura) de las redes y es porque, en efecto, la estructura de la red afecta la función total del sistema que se desea describir. Las redes con complicadas de describir y algunas de las dificultades que se han presentado son: complejidad estructural, la evolución de la red, diversidad de conexión (está relacionada con los distintos pesos que pueden llegar a tener los vértices que conectan a diversos nodos), complejidad dinámica (puede que la dinámica del sistema sea no lineal), diversidad de los nodos (donde los nodos en la red pueden llegar a tener distinta naturaleza), y la metacomplejidad (la cual es la descripción conjunta con todas las sutilezas posibles, incluyendo las anteriormente mencionadas). **Para eliminar muchas de estas complicaciones, dependiendo del campo en el que se esté trabajando, se realizan ciertas suposiciones haciendo más simples los problemas, tales como considerar al sistema estático, y por tanto estudiar la estructura de la red en el caso estacionario.** En este artículo también se estudian las propiedades de redes construidas con ciertas características estructurales, para así poder estudiar el comportamiento a nivel de sistema completo y posteriormente realizar una analogía con las redes correspondientes a sistemas reales.

Lo interesante de todas las investigaciones que se acaban de citar es que se enfocan mucho en la construcción de modelos con redes para caracterizar a los sistemas complejos que son objeto de estudio. En los modelos se hacen suposiciones que simplifican el problema de caracterizarlos. Incluso mencionamos una investigación [5] donde la red correspondiente a sistemas evidencia una estructura jerárquica, es decir, redes donde los algunos nodos tienen mayor importancia o interaccionan más que el resto de los nodos. En los temas que competen a esta tesis, la jerarquía presente en los sistemas se refiere a que ciertos elementos obtiene mayor peso que el resto debido a que destacan más gracias a su desempeño al interactuar con el resto de los elemento. ¿La jerarquía vista en redes por [5] estará íntimamente relacionada con la jerarquía que nosotros trabajaremos? ¿La teoría de redes es necesaria para nuestra investigación?

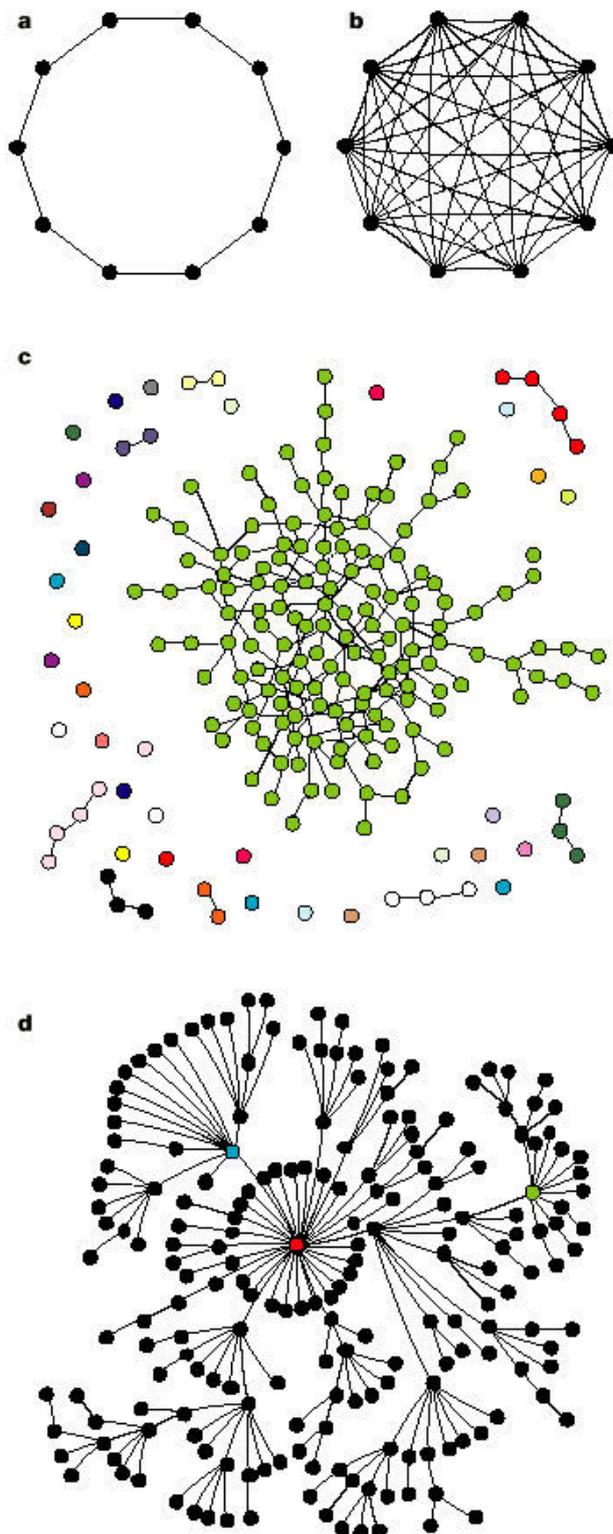


Figura 1.1: Imagen que ilustra redes de distinta índole. **a.** Representa un anillo de 10 nodos, los cuales están conectados con los vecinos más cercanos. **b.** Ahora los 10 nodos se encuentran conectados todos contra todos. **c.** Gráfica construida aleatoriamente. **d.** Gráfica invariante de escala. Imagen y descripción obtenidas de [1]

1.2. Los deportes y juegos desde el punto de vista de los sistemas complejos.

Claramente los deportes y juegos son sistemas complejos, pues están constuidos por jugadores/equipos que interaccionan entre sí. Y claramente han sido estudiados desde el punto de vista de la redes. Por ejemplo, en [7] se utiliza la teoría de redes para estudiar deportes. Afirma que la *centralidad* de las redes, que es una cantidad que mide la importancia relativa de los nodos en la red, es equivalente a un sistema de clasificación para jugadores o equipos en deportes, por lo que un enlace dirigido representa el resultado de un enfrentamiento. Proponen una variable dinámica basada en teoría de redes y la aplican a la disciplina de jugadores profesionales de Tenis masculino. Su sistema de clasificación predice el resultado de futuros juegos.

Sin embargo, se pueden realizar análisis e intentar predicciones de resultados en deportes y juegos sin utilizar la teoría de redes. Simplemente, analizando los puntajes obtenidos por los integrantes del sistema y los resultados en cuanto a ganar-perder de los mismos a lo largo del tiempo, justo como lo hacen en [8].

En [9] se presenta un análisis estadístico para 12 deportes y se reporta una escala universal en los rankeos, a pesar del hecho de que los deportes considerados tienen distintos sistemas de ranqueo entre sí. Aquí se estudia la distribución acumulativa empírica de los puntajes y se comparan con distribuciones teóricas como la distribución Gamma o Beta, pero sólo estudia eso, los puntajes y el otro factor determinante en este tipo de sistemas es la forma en cómo son ocupados los rangos. En principio, los rangos se ven afectados en el tiempo por eventos aparentemente insignificantes como un mal desayuno antes de un evento importante, o el clima durante la competencia como se vio en [10]. Ya que estos factores son inherentes para todas las actividades, se esperaría que la evolución del ranqueo tenga comportamientos genéricos entre los deportes y juegos.

Esta tesis está motivada por ejemplo en lo que hicieron para analizar la distribuciones de puntajes y que se estudió en [9], incluso introducen una bondad de ajuste llamada índice p de Kolmogorov-Smirnov para corroborar si cierto conjunto de datos aleatorios se distribuyen, en efecto, de acuerdo a un modelo teórico propuesto.

En esta tesis contaré con bases de datos que conforman múltiples deportes/juegos con información histórica en cierto rango de tiempo sobre los rankeos de los jugadores/equipos que los conforman y los puntajes respectivos que éstos tienen en cada uno de los tiempos disponibles en la evolución temporal del mismo. Este trabajo se enfocará en las trayectorias temporales del desempeño concerniente a jugadores y equipos, es decir, la evolución del rango, ésto con el objetivo de encontrar regularidades estadísticas que indiquen tendencias jerárquicas de los jugadores y equipos. Como ya se verá, conviene estudiar la evolución temporal en la forma que los rangos son ocupados y propongo utilizar una medida recientemente introducida llamada *diversidad de rango*. Con la ayuda de la base de datos de Google *N-Gram* [11], la diversidad de rango ya ha sido utilizada con anterioridad para estudiar cómo cambia el vocabulario con el tiempo [2]. Ese trabajo muestra que la diversidad de rango tienen la misma forma funcional para todos los idiomas estudiados, y es capaz de discriminar el tamaño del núcleo de cada idioma.

1.3. Objetivos y estructura general.

El trabajo de tesis presente consistirá en analizar los dos aspectos ya mencionados que describen la estructura jerárquica de sistemas complejos (distribución de puntajes y evolución concerniente a la ocupación de los rangos a lo largo del tiempo) para distintos conjuntos de datos correspondientes a deportes y juegos, es decir, bases de datos que recopilen los ordenamientos de los elementos que participen en disciplinas correspondientes por un periodo de tiempo y por saltos de tiempo regulares, como cada semana o cada mes, éstos saltos los identificaremos como rodajas temporales. Los ordenamientos de los elementos se llevan a cabo por medio de la asignación de puntajes de acuerdo a criterios establecidos por federaciones u organismos reguladores de los eventos competitivos entre los competidores. Las disciplinas aquí consideradas están relacionadas con: juegos de ajedrez, fútbol, golf, póquer, corredores de autos, patinadores sobre nieve y jugadores de videojuegos. Las características específicas de las bases se proporcionan a detalle en el [Capítulo 2](#).

El objetivo de este trabajo consiste en analizar la estructura jerárquica de estos sistemas, que básicamente consiste en:

- Enfocarnos en la distribución de los puntajes asignados a los elementos y proponer modelos teóricos que pretendan reproducir esas distribuciones.
- Estudiar la dinámica que rige la ocupación de los rangos a lo largo del tiempo con el manejo de medidas bien definidas como la diversidad de rango que ya se mencionó anteriormente y a partir de ella buscar posibles regularidades asociadas al conjunto de sistemas considerados y que podrían ser de interés para entender la naturaleza de los mismos.
- Introducir medidas dinámicas adicionales a la diversidad de rango que describan aspectos distintos sobre la evolución de la ocupación de los rangos.
- Utilizar la diversidad de rango como una herramienta para entender la dinámica de rango en deportes, juegos, y otros sistemas complejos jerárquicos, para así identificar la dependencia del rango en un cambio en la jerarquía del sistema. Usando este análisis, se considera la posibilidad de estimar qué tan bien puede ser predecido el cambio en un rango, ignorando las particularidades de un fenómeno de estudio.
- Diseñar modelos matemáticos que reproduzcan, al menos cualitativamente, la evolución de los sistemas aquí descritos y que presenten el mismo comportamiento dinámico observado. Ésto utilizando la intuición obtenida del comportamiento de las dinámicas observadas.

La estructura de la tesis llevará un orden lógico que nos permita ir desarrollando las ideas necesarias para construir nuevas. Este trabajo cuenta con 6 capítulos y 4 apéndices que tienen la siguiente estructura: (evidentemente, el primer capítulo es el presente)

- El [Capítulo 2](#) tendrá la descripción detallada de las bases de datos utilizadas en este trabajo y que representan el comportamiento de los elementos dentro de las disciplinas ya mencionadas. En total, son 12 bases de datos con características propias y distintas entre sí. También se describirán a detalles los criterios utilizados en cada disciplina para la asignación de puntajes a los competidores.
- El [Capítulo 3](#) cubre el aspecto mencionado sobre la forma en que se distribuyen los puntajes, se definirá una función llamada *distribución de rango* que captura lo propio. Se propondrán 5 modelos teóricos para modelar el comportamiento de dichas distribuciones: la ley de Zipf, la distribución Gamma γ , la distribución Beta β , la combinación de ambas anteriores $\gamma\beta$ y la ley doble Zipf. Utilizando dos bondades de ajuste, el coeficiente de determinación R^2 y el índice p de Kolmogorov-Smirnov, se pondrán a prueba estos modelos con las distribuciones de rango correspondientes de todas las rodajas temporales disponibles para todos los sistemas para determinar cuál de los modelos es el mejor en todo el intervalo de tiempo correspondiente.
- El [Capítulo 4](#) trata a detalle la dinámica de rango de los sistemas. Primero se define lo que son los *espaguetis* que son trayectorias en el espacio de rangos versus tiempo que llevan la evolución temporal de la ocupación de rango para elementos específicos, ésto en busca de alguna regularidad. Después se introduce el concepto de *diversidad de rango* y se propone un modelo teórico que aproxime esta función. Se buscan regularidades a partir de lo observado en la diversidad de rango. Además, se introducen cantidades dinámicas adicionales: *probabilidad de cambio*, *entropía de rango* y *complejidad de rango*, viendo que éstas últimas capturan distintos aspectos de la evolución del sistema.
- El [Capítulo 5](#) se introducen dos modelos matemáticos que intentan reproducir, de manera artificial, la evolución de ocupación de los rangos a lo largo del tiempo partiendo de lo observado en el [Capítulo 4](#). Los dos modelos que aquí describimos son: el modelo del caminante aleatorio y el modelo nulo. Se comparan los resultados obtenidos en dichos modelos respecto a lo obtenido en los sistemas reales.
- El [Capítulo 6](#) consiste en la redacción de los comentarios finales de la tesis, conclusiones sobre lo realizado y proyecciones a futuro para continuar con esta investigación.
- El [Apéndice A](#) describe en detalle el proceso para calcular la bondad de ajuste: índice p de Kolmogorov-Smirnov. El [Apéndice B](#) contiene ejemplos de espaguetis que pasan por ciertos rangos para el caso de todos los sistemas. El [Apéndice C](#) contiene algunas gráficas que complementan la explicación de algunos capítulos. El [Apéndice D](#) anexa el artículo publicado del cual se basó esta tesis y del que soy uno de los autores.

Cabe recalcar que esta tesis es un análisis extenso a lo publicado en [\[12\]](#) y del cual soy uno de los autores. Cabe recalcar que la construcción de las medidas dinámicas, la

1. INTRODUCCIÓN

construcción de los modelos nulo y del caminante aleatorio, la derivación analítica de la diversidad de rango, la derivación analítica de los modelos teóricos para la distribución de rango y la proposición de los mismos se realizaron como trabajo del grupo de investigación en el que participé y que son parte del resto de los autores en la misma publicación. Mi participación más específica e individual fue el de obtener las bases de datos, realizar las figuras que se incluirían y encontrar una adaptación al cálculo del índice p de Kolmogorov-Smirnov para lo que aquí nos compete.

El artículo completo se encuentra también en el [Apéndice D](#) tal y como se publicó y se enlistan algunas de las diferencias entre la publicación y el presente trabajo. También es necesario informar que a lo largo de la tesis se utilizará la palabra *rankeo* para referirnos a ordenamiento, en algunas ocasiones también utilizamos la palabra *ranking* como el compilado de ordenamiento de los elementos para los sistemas; adicionalmente, la palabra *rankeado* se utiliza para referirnos a que un elemento rankeado es un elemento ordenado dentro del ranking.

Bases de datos.

Los deportes y juegos que analizaré en este trabajo presentan diversos criterios de ordenamiento para sus elementos (jugadores o equipos). Estos criterios están adaptados para las condiciones en que se llevan a cabo dichas disciplinas, tales como: la periodicidad en que se realizan los eventos de competición, si los elementos del sistema consisten en entidades individuales o de equipos, el rango que tienen los jugadores/equipos al momento que se enfrentan para competir, y muchos otros criterios específicos que platicaré para cada una de las disciplinas aquí reportadas. Lo interesante a destacar aquí es que las reglas con las cuales se llevan a cabo los rankeos son bastante variadas y encontrar una universalidad de algún tipo sería de mucha utilidad y gran interés. Cabe destacar también que no todas las disciplinas consisten de competición con esfuerzo físico, algunas, como veremos, requieren de habilidades con aparatos (automóviles o consolas de videojuegos) que se adquieren con mucha práctica. El hecho de que las disciplinas estudiadas también tengan características variadas es de gran importancia para poder concluir que se está identificando algún comportamiento genérico para sistemas complejos jerárquicos y que tal vez se podría generalizar a sistemas con índoles más variadas y diferentes a los de este trabajo.

En este capítulo corto proporcionaré las características que tienen las bases de datos que consisten en la evolución de los rankeos para 12 deportes/juegos, éstas bases deben consistir en un historial de listas de ranqueo para que se pueda estudiar una dinámica. Una vez descritas las características de las bases de datos, proporcionaré una descripción de las reglas con las cuales los elementos que conforman a la disciplina son ordenados o rankeados, estableciendo la forma en que éstos interactúan y los criterios para determinar quiénes son mejores o peores al ejercer estas actividades, dando lugar a la estructura jerárquica de estos sistemas complejos.

2.1. Las bases de datos de las disciplinas.

Proporcionaré, a continuación, las características generales de las bases de datos de los deportes y juegos aquí estudiados, indicando el nombre de la federación u organización encargada de hacer el ranqueo, la resolución temporal, el periodo de tiempo en el que se están usando los datos y la cantidad de jugadores o equipos considerados. La cantidad de elementos en cada base es la misma en todas las rodajas temporales, pues las medidas

2. BASES DE DATOS.

que utilizaré en este trabajo requieren de esta característica.

Se utilizaron datos de ordenamiento para jugadores y equipos en 12 deportes y juegos diferentes: (a) Jugadores de Ajedrez, femenino, ordenado por "Fédération Internationale des Échecs"(FIDE-F) [13]; (b) Jugadores de Ajedrez, masculino, ordenado por "Fédération Internationale des Échecs"(FIDE-M) [13]; (c) Equipos (clubes) de fútbol, ordenados por "Football Club World Ranking"(FCWR-C) [14]; (d) Equipos nacionales de fútbol, ordenados por "Fédération Internationale de Football Association"(FIFA) [15]; Goleadores de Fútbol en equipos de clubes (FCWR-G), ordenados por "Football Club World Ranking"[14]; (e) Jugadores de Golf, ordenados por "Official World Golf Ranking"(OWGR) [16]; (f) Corredores de NASCAR de la Busch Grand Nation, ordenados por "National Association for Stock Car Auto Racing"(NASCAR) [17], (g) Corredores de NASCAR de la Winston Cup Grand National, ordenados por "National Association for Stock Car Auto Racing"(NASCAR) [17]; (h) Jugadores de Póquer, ordenados por "Global Poker Index"(GPI) [18]; (i) Jugadores de tabla sobre nieve, ordenados por "World Snowboarding Points Lists"[19]; (j) Jugadores de Tenis, masculino, ordenado por la "Association of Tennis Professionals"(ATP) [20]; (k) Ganancias en videojuegos, ordenados por E-Sports Game (ESE) [21].

La [Tabla 2.1](#) tiene los detalles de todos estos sistemas especificados: el nombre del deporte o juego en cuestión, la fuente de los datos que es la página web de dónde se obtuvo la información y la organización o federación encargada de realizar los ordenamientos; el periodo de tiempo que comprende la base de datos, es decir, la desde la primera fecha disponible y la última; la resolución temporal que es el periodo de tiempo que separa cada una de la lista de rankeos a lo largo del tiempo, por ejemplo, si el ranqueo se hace semanalmente, mensualmente, etcétera; y también se incluye el número de jugadores/equipos que se tiene en cada lista de ranqueo correspondiente a una rodaja temporal.

Para obtener una distribución homogénea del número de jugadores/equipos en cada rodaja temporal para una determinada actividad, ignoramos algunos datos de ATP, FIDE, OWGR, GPU y FIFA. En todos los casos, el tiempo entre las publicaciones de los rankings varía mucho (desde menos de una semana hasta más de un mes), y el número de jugadores/equipos a lo largo de las rodajas en los rankeos puede cambiar también. Siendo así, para cada base de datos escogimos una constante de resolución temporal en los rankeos (semanas o meses) que maximice y mantenga constante el número de jugadores/equipos rankeados a lo largo del tiempo. Como ya mencioné anteriormente, los criterios para el ranqueo de los elementos en cada disciplina varía considerablemente; a continuación explicaré a detalle el proceso para cada una de las bases de datos.

Tabla 2.1: Resumen de los datos de ordenamiento para cada deporte y juego considerado en este estudio. La tabla enlista las propiedades más importantes utilizadas aquí (incluyendo la fuente de los datos, el periodo de tiempo, la resolución del ordenamiento, y el número de jugadores/equipos). Para poder tener una distribución homogénea de jugadores/equipos en cada rodaja de tiempo para cada actividad, se ignoraron algunos datos que conformaban las bases completas. Cada una de las bases de datos, o en este caso los deportes y juegos, estarán representados por unas siglas características. Por ejemplo, las siglas correspondientes a los equipos nacionales de fútbol son FIFA, y siempre que usemos dichas siglas nos estamos al sistema en cuestión y todo lo que implica.

Deporte/juego	Fuente de los datos	Periodo de tiempo	Resolución temporal	# jugadores/equipos
Jugadores de ajedrez (femenino)	Fédération Internationale des Échecs (FIDE-F) [13]	Jul 2012 – Abr 2016	Mensual	12681
Jugadores de ajedrez (masculino)	Fédération Internationale des Échecs (FIDE-M) [13]	Jul 2012 – Abr 2016	Mensual	13500
Clubes de Fútbol	Football Club World Ranking (FCWR-C) [14]	Feb 1 2012 – Dec 29 2014	Semanal	850
Equipos nacionales de Fútbol	Fédération Internationale de Football Association (FIFA) [15]	Jul 2010 – Dec 2015	Mensual	150
Goleadores de Fútbol en equipos de clubes	Football Club World Ranking (FCWR-G) [14]	Semana 33 2016 – Semana 33 2017	Semanal	400
Jugadores de Golf	Official World Golf Ranking (OWGR) [16]	Sept 10 2000 – Abr 19 2015	Semanal	1000
NASCAR Busch Grand Nation	National Association for Stock Car Auto Racing (NASCAR-B) [17]	1982 – 2015	Anual	76
NASCAR Winston Cup Grand National	National Association for Stock Car Auto Racing (NASCAR-W) [17]	1979 – 2013	Anual	50
Jugadores de Póquer	Global Poker Index (GPI) [18]	Jul 25 2012 – Jun 10 2015	Semanal	1799
Jugadores de Tabla sobre nieve	World Snowboarding (WSD) [19]	Enero 5 2015 – Marzo 26 2018	Semanal	1413
Jugadores de tenis (masculino)	Association of Tennis Professionals (ATP) [20]	May 5 2003 – Dic 27 2010	Semanal	1600
Ganancias en videojuegos	E-Sports Earnings (ESE) [21]	2003 – 2016	Anual	400

2.2. Características de ordenamiento.

A continuación trataré de describir lo más detalladamente posible el procedimiento en que se realizan los ordenamientos (rankeos) de los elementos de cada uno de los sistemas. Teniendo en cuenta que siempre se tiene un criterio cuantitativo, el cual, dependiendo de su valor, es el que establece el orden de cada uno de los elementos. A esta cantidad numérica asignada a cada elemento del sistema lo llamaremos "puntaje". Es importante que se compactarán las reglas de cada disciplina lo más que se pueda, pues el único objetivo de presentarlas aquí, es para apreciar las diferencias tan marcadas en las formas que las federaciones cuantifican el rendimiento de los jugadores/equipos de los sistemas.

2.2.1. Jugadores de ajedrez, masculino y femenino (FIDE)

La FIDE (Fédération Internationale des Échecs) es la organización internacional oficial encargada de conectar a las federaciones nacionales de ajedrez de distintos países alrededor del mundo. Además de organizar el Campeonato del mundo de ajedrez, la FIDE calcula el rango Elo de los jugadores, redacta las reglas del ajedrez, publica libros y nombra a Maestros Internacionales, Grandes Maestros y árbitros. Los rankings de la FIDE se hacen en base a su puntaje Elo, que es el puntaje que la FIDE asigna a los jugadores. [22]

La FIDE usa el sistema Elo para rankear a los jugadores. Este método calcula las habilidades relativas de los jugadores. El creador del método fue Arpad Elo, un físico húngaro-americano. El sistema Elo varía de disciplina en disciplina. Ahora bien, para los jugadores con rangos más altos el ranqueo de FIDE es el más importante. Desde julio de 2012, la FIDE actualiza las listas de rankings mensualmente. [23]

También ha establecido una serie de títulos para los jugadores según su puntuación Elo a partir de 2003 (aunque recientemente agregó el título de Candidato a Maestro a partir de 2002). Además de estos títulos, la puntuación Elo permite ubicar a un jugador en ciertas categorías no oficiales. Según datos de septiembre de 2016, la FIDE cuenta en su escalafón con 147,008 jugadores activos (aquellos que han disputado al menos una partida oficial de ritmo clásico en los últimos dos años), clasificados por rangos así (en la modalidad clásica y a partir de una puntuación Elo de 1000) [24]

Lo interesante del sistema Elo es que determina su puntaje en función de: qué tan fuerte es el oponente (es decir, cuál es su rango al momento del encuentro), el número de encuentros y los resultados de los mismos. El sistema Elo que emplea la FIDE es bastante complicado, así que pasaré a explicarlo paso por paso. La puntuación Elo se establece a partir de los resultados contra otros jugadores. Por ejemplo, dos jugadores que se enfrentarán en una partida tienen determinados puntajes Elo, con ellos se pueden calcular una cantidad llamada probabilidad estimada o puntuación estimada. Ahora bien, supongamos que tenemos dos jugadores, A y B. Si el jugador A tiene una puntuación Elo R_A y el jugador B una puntuación Elo R_B , la puntuación esperada del jugador A se calcula como: [24]

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \quad (2.1)$$

Por otro lado, la puntuación esperada del jugador B es:

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}} \quad (2.2)$$

o equivalentemente, estas cantidades pueden ser expresadas como:

$$E_A = \frac{Q_A}{Q_A + Q_B}, \quad E_B = \frac{Q_B}{Q_A + Q_B}$$

donde declaramos las nuevas variables $Q_A = 10^{R_A/400}$ y $Q_B = 10^{R_B/400}$. Si calculamos la razón entre E_A y E_B obtenemos que:

$$\frac{E_A}{E_B} = \frac{Q_A}{Q_B} = 10^{(R_A - R_B)/400}$$

Eso quiere decir que la proporcionalidad entre la probabilidades de ganar de ambos jugadores depende del número $(R_A - R_B)/400$, por ende, si el jugador A tiene al menos 400 puntos Elo de ventaja sobre el jugador B, entonces la probabilidad de que el jugador A gane el encuentro es 10 veces mayor, y por cada 400 puntos de ventaja, la probabilidad de victoria de A es 10 veces mayor. Ésto es muy interesante porque conociendo los puntos Elo previos al encuentro se pueden hacer estimaciones. Es claro que $E_A + E_B = 1$ lo cual es consistente probabilísticamente. Por otro lado, ésto sólo es una estimación de quién sería el ganador del encuentro entre dos jugadores; después del encuentro se debe asignar un nuevo puntaje Elo a ambos jugadores dependiendo del resultado. El sistema Elo incrementará o decrementará la puntuación previa al encuentro dependiendo de si la puntuación obtenida es menor o mayor a la puntuación esperada. La puntuación obtenida sólo puede valer tres cantidades: 0 si se perdió la partida, 0.5 si se presentó un empate y 1 si se ganó el encuentro. Entonces lo que se hará es comparar ese puntaje obtenido y el esperado que se calcula a partir de los puntajes Elo expresado en las ecuaciones 2.1 y 2.2. Para calcular el nuevo puntaje Elo se utiliza un ajuste lineal proporcional a la diferencia entre la puntuación esperada y la obtenida por el jugador en cuestión. El factor de proporcionalidad (K) dependerá del título asignado al jugador que depende de su puntuación, de los cuales ya hablamos anteriormente. Por ejemplo, para los que tengan puntaje Elo entre 2300 y 2400 la constante K es $K = 20$, mientras que los que tengan un puntaje Elo superior a 2400 les corresponderá una constante $K = 10$. Entonces el ajuste de su puntuación después del encuentro se calcula como: [24]

$$R'_A = R_A + K(S_A - E_A) \quad (2.3)$$

La constante K tiene menor valor para los que tengan mayor puntaje Elo, por lo que si pierden, implicaría que $S_A - E_A$ es negativo, pero como K es pequeña con respecto a los de menores puntajes, les afectará menos la derrota, caso contrario con los que tenían

menor puntaje Elo antes de la partida. El ajuste Elo que se menciona en la ecuación 2.3 se puede realizar luego de cada partida, al finalizar un torneo o contando las partidas en un periodo computable. [24]

Estos puntajes Elo son los que determinan el rango de un jugador en la lista que publica la FIDE. Lo interesante aquí es la importancia que se le da al nivel del oponente, pues sí importa si se está jugando contra un experto o contra un principiante para saber cómo evolucionará en el ranking.

2.2.2. Clubes de Fútbol, (FCWR-C) y Goleadores de Fútbol en equipos de clubes (FCWR-G)

Todos los Rankings (Clubs, Máximos Goleadores y entrenadores) se calculan de forma idéntica y se emiten semanalmente. El cálculo se establece igual que, por ejemplo, el ATP Ranking Mundial de Tenis (el cual explicaremos más adelante), lo que significa que sólo los resultados de partidos de las últimas 52 semanas se tienen en cuenta. Cada lunes a las 12.00 h CET todos los Rankings se actualizan con los resultados de la semana anterior, mientras que los resultados de la semana correspondiente hace un año pierden su valor.

Los puntos del Ranking Mundial se ganan en 40 ligas superiores seleccionadas y 10 torneos de clubes internacionales en todo el mundo.

Los resultados del partido se ponderan cuando se convierten en Puntos de Clasificación Mundial, porque:

- No todos los partidos son igual de importantes.
- No todas las competiciones nacionales o torneos internacionales de copas son igual de fuertes.
- No todas las competiciones nacionales tienen un número igual de partidos para jugar.
- No todas las Confederaciones de la FIFA son igual de fuertes

Los puntos del Ranking Mundial se calculan de forma diferente para partidos nacionales e internacionales. El acumulado de los puntos nacionales e internacionales determina el total de puntos del Ranking Mundial. (Texto completo tomado de [14])

Ahora bien, los puntos en el ranqueo de los equipos o goleadores se agregan cuando juegan un partido específico, pero como ya mencionamos, no todos los encuentros tienen el mismo peso. Si por ejemplo, el partido es doméstico (dentro del país correspondiente al club, o club del goleador) los puntos para el ranqueo mundial se calculan como:

$$S = 100 \times RE \times IE \times FCD \times EE \times FCF \quad (2.4)$$

donde RE es un factor relacionado con el resultado del encuentro, IE está relacionado con la importancia del encuentro, FCD está relacionada con la fuerza que tiene el jugador o equipo dentro de la competición doméstica, EE es un factor que tiene que ver

con ecualizar el encuentro dependiendo del número de participantes del torneo en cuestión y FCF es la fuerza que tiene el equipo o jugador en la confederación establecida por la FIFA. S es justamente el puntaje que obtiene el equipo o jugador por el encuentro dentro del ranqueo mundial. Por otro lado, si el encuentro tiene índole internacional, entonces el puntaje a agregar se calcula como:

$$S = 100 \times RE \times IE \times FO \times FCF \quad (2.5)$$

donde el nuevo factor desconocido es FO que es un factor relacionado con la fuerza del oponente. Estos factores tendrán valores bien establecidos según sean las condiciones de los encuentros, pero no incluiremos aquí qué valores adquieren, pues se sale de nuestros objetivos que es mostrar a grandes rasgos el proceso de ranqueo para esta disciplina.

2.2.3. Equipos nacionales de Fútbol, (FIFA)

Lo que necesitamos es calcular un puntaje P para los equipos de fútbol, dependiendo de qué puntaje obtenga serán rankeados, siendo los mejores equipos los que tienen el puntaje más alto. El método de cálculo es sencillo: cualquier equipo que consiga buenos resultados en el fútbol internacional obtendrá puntos que le permitirán ascender en la clasificación mundial.

El total de puntos acumulado por un equipo en un cuatrienio se obtiene sumando:

- el número de puntos ganados en un partido;
- la media de puntos ganados en partidos durante los últimos 12 meses; y
- la media de puntos ganados en partidos anteriores a los últimos 12 meses (depreciación anual).

El número de puntos por partido que pueden obtenerse en un partido depende de los siguientes factores:

- ¿Victoria o empate? ($M =$ encuentro)
- ¿Fue un partido importante (desde un amistoso hasta un partido de la Copa Mundial de la FIFA)? ($I =$ importancia)
- ¿Cuál era la fuerza de los contendientes con respecto a su puesto en la clasificación y la confederación a la que pertenecen? ($T =$ tabla y $C =$ confederación)

Estos factores se sintetizan en una fórmula para determinar el número total de puntos ($P =$ puntos).

$$P = M \times I \times T \times C$$

Para el cálculo de puntos, se aplicarán los siguientes criterios: M : Puntos por victoria

2. BASES DE DATOS.

- Los equipos ganan 3 puntos por victoria, 1 punto por empate y 0 puntos por derrota. En una tanda de tiros penales, el ganador obtiene 2 puntos y el perdedor 1 punto.

I: Importancia del partido

- Amistoso (incluidos los torneos menores): $I = 1.0$
- Eliminatoria mundialista o en el ámbito de la confederación: $I = 2.5$
- Competición final de confederación o Copa FIFA Confederaciones: $I = 3.0$
- Competición final de la Copa Mundial de la FIFA: $I = 4.0$

T: Fuerza de los contendientes

- La fuerza de los contendientes se basa en la siguiente fórmula: $200 - \text{el puesto en la clasificación de los contendientes}$. Como excepción de esta fórmula, se asigna siempre al equipo a la cabeza de la clasificación el valor 200 y a los equipos clasificados en el puesto 150.º y subsiguientes se les asigna un valor mínimo de 50. El puesto en la tabla se obtiene de la última Clasificación Mundial FIFA/Coca-Cola publicada.

C: Fuerza de la confederación

- Al calcular partidos entre equipos de distintas confederaciones, se emplea el valor medio de las confederaciones a las que pertenecen los equipos que compiten. La fuerza de una confederación se calcula de acuerdo con el número de victorias que ha obtenido en las últimas tres ediciones de la Copa Mundial de la FIFA. Los valores son los siguientes:

CONMEBOL 1.00

UEFA 0.99

CONCACAF/AFC/CAF/OFC 0.85

La descripción del método de puntaje para el caso de FIFA fue íntegramente tomado de [15], y es el método vigente a la fecha que se revisó esta página por última vez.

2.2.4. Jugadores de Golf, (OWGR)

Los torneos oficiales que se pueden considerar para considerar una suma de puntajes en el ranking mundial a los jugadores son los siguientes: El Campeonato mundial de Golf, Los juegos Olímpicos y la Copa mundial de Golf, sólo en sus modalidades individuales y no en equipos. Cualquier jugador que juegue en cualquiera de estas competencias recibirá puntos para el ranking mundial tomando en cuenta su posición final que será ganada de acuerdo a la importancia de los torneos.

Los torneos que actualmente se consideran para sumar puntos en el sistema (Original Worl Golf Ranking, OWGR) son: Alps Tour Golf, Asian Development Tour, Asian Tour,

Big Easy Tour, China Tour, EuroPro Tour, European Challenge Tour, Japan Golf Tour, KPGA Korean Tour, MENA Golf Tour, Nordic Golf League, PGA European Tour, PGA Tour, PGA Tour Canada, PGA Tour China Series, PGA Tour Latinoamérica, PGA Tour of Australasia, ProGolf Tour, Sunshine Tour and Web.com Tour.

Los puntajes para ser considerados en el ranking mundial son acumulados de los puntos obtenidos a lo largo de dos años en torneos que ser desarrollados en periodos de al menos 13 semanas. Los puntajes se reducen en el mismo decremento por el resto de las 91 semanas del rango de dos años. Cada jugador es rankeado de acuerdo al promedio de los puntos por torneo, que se determina dividiendo el número total de puntos por el número de torneos que haya jugado durante el periodo de dos años. Hay un mínimo divisor de 40 torneos durante el periodo de dos años y un máximo divisor de los 52 últimos torneos de los jugadores.

También la importancia de los torneos son considerados. Al final, lo que ordena a los jugadores en la tabla de rankeos es justamente ese promedio de puntajes obtenidos durante el periodo de 2 años ya mencionado. Información tomada de [16].

2.2.5. Corredores de NASCAR, para Busch Grand Nation (NASCAR-B) y Winston Cup Grand National (NASCAR-W)

Para el caso de este sistema tenemos una problemática importante, y es que el sistema que asigna puntos a los corredores ha cambiado a lo largo del tiempo por lo que no podemos proporcionar un sistema específico que se haya utilizado aquí. El campeonato *Busch Grand Nation* y *Winston Cup Grand National* son dos campeonatos de distintas categorías dentro de la NASCAR, sin embargo ambas se rigen por el mismo sistema de puntos. El sistema de lista de puntos que rige la NASCAR se usa desde 1949. El campeonato se le otorga cada año al conductor que acumule la mayor cantidad de puntos en el campeonato correspondiente durante la temporada que ocupa el mismo. Los puntajes han cambiado, como ya se mencionó, pero éstos dependen por ejemplo del dinero ganado, contando las vueltas recorridas en las carreras dándoles cierto peso, la posición resultante al final del evento, etc.[25]

2.2.6. Jugadores de Póquer, (GPI)

El GPI, por sus siglas en inglés, (Índice Global de Póquer) es una clasificación de los jugadores de torneos en vivo en el mundo desde el día en que se publica. Los jugadores se rankean semanalmente en función de su rendimiento finalizando en posiciones debido al efectivo ganado en torneos clasificatorios que ocurren durante el periodo anterior de 36 meses. Una posición debido al efectivo ganado es cualquier posición en la que el jugador recibe una porción del premio total por su desempeño en el evento. Normalmente, el 10 % al 20 % de los participantes en un evento finalizan en una posición debido al efectivo. Los jugadores sólo pueden recibir una puntuación por torneo. Si un jugador cobra varias veces en el mismo torneo, su puntaje se basará en su posición final más alta. Los torneos clasificatorios son eventos con 32 o más jugadores en un "buy-in" de 1 dólar (u otro equivalente en moneda) o superior que están abiertos al público y no son

2. BASES DE DATOS.

eventos especiales o seleccionados del público como caridad, personas mayores, dobles, satélites, mujeres, equipos, empleados y ejecutivos.

Los jugadores se clasifican de acuerdo con sus puntajes de finalización en los torneos clasificatorios. El puntaje GPI individual de cada jugador es un agregado de puntajes en eventos durante el periodo anterior de 36 meses, medido desde el día en que se calcula el GPI. El puntaje para un evento dado se deriva de una combinación de su porcentaje de lugar de llegada, el "buy-in" factor de envejecimiento. Porcentaje de finalización se refiere al porcentaje del campo de inicio que un jugador supera en su final. Buy-in se refiere a la cantidad relativa del buy-in del evento al buy-in inicial de 1000 USD (los eventos con buy-in por debajo de 1000 USD aún se comparan en relación con una línea de base de 1000 USD). Factor de envejecimiento se refiere a la ponderación de los resultados por su actualidad, donde los resultados más recientes se ponderan más que los resultados anteriores. El GPI limita los resultados a cinco (5) resultados por período de medio año para los últimos 18 meses y cuatro (4) resultados por período de medio año durante los 18 meses anteriores para un total máximo de 27 puntajes por período de agregación de 36 meses. Estos factores a considerar tienen varias peculiaridades y escenarios considerados, sin embargo, aquí no se describirá a detalle eso.

El puntaje compuesto de GPI es la suma de todos los puntajes de eventos individuales. Para ayudar a garantizar que el puntaje general compuesto de GPI no esté sesgado por instancias de jugadores que se desempeñan en un número extremadamente grande de eventos, el GPI limita el número de puntajes de eventos individuales para cada período de medio año. La cantidad de eventos limitados para un período de medio año dado se determina tomando la cantidad promedio de finales en torneos en vivo para los jugadores que están en el GPI. La administración para determinar la cantidad media de acabados se realiza anualmente. Una vez que se determina el número medio de finales, ese es el límite para el número de puntajes individuales que se cuentan para cada período de medio año. Para cualquier jugador con puntajes más individuales que el límite para el período de medio año, los puntajes más altos de eventos individuales se calculan como parte de la puntuación compuesta, mientras que los puntajes de eventos individuales más bajos se descartan. Estos eventos descartados pueden incluirse más adelante en la puntuación compuesta a medida que se mueven de un período de medio año a otro. (Texto tomado íntegramente de [18]).

2.2.7. Jugadores de tabla sobre nieve, (WSD)

Las Listas de puntos World Snowboarding (WSPL) son el resultado de un esfuerzo global de colaboración para crear un sistema de clasificación universal, transparente y justo para el snowboard competitivo. En cualquier momento dado, los WSPL apuntan a proporcionar la representación más precisa y transparente de los mejores patinadores del mundo y sus correspondientes resultados. El WSPL evalúa los resultados de eventos de todas las competiciones de snowboard a nivel mundial. La posición de un patinador en la Lista de Puntos se determina calculando un promedio de puntos de WSPL de un patinador de sus mejores tres resultados en la disciplina respectiva de estilo libre dentro de un período de 52 semanas. Las listas de puntos World Snowboarding se calculan

y publican todos los lunes y jueves a las 6:00 a.m. PST (3:00 p.m. CET) para las disciplinas Halfpipe, Slopestyle y Big Air, tanto para hombres como para mujeres.

El sistema de puntos utilizado para el cálculo de WSPL se basa en un Sistema de puntos dinámico de 10 escalas con diez niveles de puntos que van de 100 a 1000 puntos. Tres factores determinan el nivel de puntos de un evento y los puntos de clasificación que obtiene un usuario para su ubicación individual en un evento: Categoría del evento", Calidad del campo "Tamaño del campo". Cada evento se clasificará en una de las siguientes tres categorías de eventos: regional/nacional, internacional y élite. Para cada categoría, se define un nivel mínimo de puntos (Regional / Nacional: 100 puntos, Internacional: 300 puntos, Elite: 600 puntos) mientras que el nivel máximo de puntos para un evento no está limitado. Independientemente de la categoría del evento, cualquier evento puede alcanzar un nivel de puntos de 1000 puntos. La determinación del nivel de puntos de un evento se basa en la calidad del campo. (Texto tomado íntegramente de [19])

2.2.8. Jugadores de tenis, (ATP)

El ranqueo de la ATP es reconocido como el sistema oficial en el cual se ordenan a los jugadores de Tenis al rededor del mundo. Los criterios de ranqueo han cambiado a lo largo de los años desde que en 1972 se hizo por primera vez un ranqueo oficial. El ranqueo de la ATP es un método basado en el rendimiento del jugador en todos los torneos simples y dobles (aquellos en los cuales se permite perder una y dos veces, respectivamente), excepto las finales de Nitto ATP. El periodo de ranqueo de la ATP comprende las 52 semanas anteriores inmediatas, excepto para algunos torneos como la final de Nitto ATP y torneos a los cuales se registren en ciertas fechas fuera del rango considerado para la clasificación en este método. [20]

Actualmente, 2092 jugadores están enlistados en el método de 52 semanas. Durante cada encuentro, el supervisor de la computadora de la ATP está actualizando en vivo la información de los puntajes. Al final de cada semana, el supervisor de la ATP proporciona al coordinador de los rankings de la ATP con la información de sus respectivos torneos. La información es compilada en la base de datos el domingo en la noche donde se hacen las estadísticas correspondientes y se hace pública para que todos los jugadores y los medios de comunicación tengan acceso a ella vía *ATPWorldTour.com* [26]

El ranqueo de un jugador de la ATP está basado en los puntos que acumule en cualquiera de los 19 torneos siguientes (excepto si no califica para las finales de la ATP, en ese caso sólo 18) :[27]

- Los cuatro torneos *Grand Slam*
- Los 8 torneos correspondientes a *ATP World Tour Masters 1000*.
- Las finales de la ATP previas hasta antes del lunes siguiente al evento final de la temporada regular de la ATP del siguiente año.
- Los 6 mejores resultados de los torneos *ATP World Tour 1000*, *ATP World Tour*

2. BASES DE DATOS.

500, ATP World Tour 250, ATP Challenger Tour, Future Series, y la Davis Cup jugados en el calendario correspondiente.

Ahora bien, la participación en los torneos antes mencionados está condicionada por muchos factores; algunos son opcionales, otros son obligatorios, algunos tienen más peso, otros tienen menos peso. Mencionar estos detalles no tiene sentido aquí; sólo haré hincapié en el hecho de que todos estos torneos están en consideración dentro de los cuales se hará el registro de los puntajes obtenidos. Este puntaje asignado a cada jugador de la ATP es el que determina su posición en el ranking y se les asigna dependiendo de su desempeño en los torneos antes mencionados. [27]

2.2.9. Ganancias de jugadores de E-Sports, (ESE)

Los deportes electrónicos o e-sports son competiciones de videojuegos que se han convertido en eventos de gran popularidad. Por lo general los deportes electrónicos son competiciones de videojuegos multijugador, particularmente entre jugadores profesionales aunque no de manera exclusiva. Los géneros más comunes en los videojuegos asociados a los esports son: estrategia en tiempo real, disparos en primera persona y arenas de batalla multijugador online (mejor conocido por sus siglas en inglés MOBA, multiplayer online battle arena). Torneos como The International (el torneo anual del videojuego Dota), el “League of Legends World Championship” (torneo mundial del videojuego League of Legends), la Battle.net World Championship Series (una serie de torneos de los diversos títulos de la compañía Blizzard Entertainment), el Evolution Championship Series (evento anual que se centra exclusivamente en los juegos de lucha), la Intel Extreme Masters (serie de torneos internacionales de deportes electrónicos celebrados en diferentes países alrededor del mundo por la compañía Intel) y el Smite World Championship (campeonato mundial del videojuego Smite) entre otros, proveen al público de transmisiones en vivo de sus competiciones así como premios monetarios y salarios a los competidores. (Texto tomado íntegramente de [28]).

Básicamente, lo que nosotros estamos estudiando es el ranqueo que hace [21] de acuerdo a las ganancias que tienen los participantes en los torneos ya mencionados por año. Generando así una base de datos que representa un sistema complejo jerárquico, justo como hace el resto de los sistemas incluidos.

2.3. Algunas consideraciones adicionales

Para realizar este trabajo se encontraron numerosas bases de datos para deportes y juegos de diferente índole a la trabajada aquí, sin embargo, no pudimos considerarlas porque no tenían realmente una evolución registrada de esos sistemas, entonces no tenía caso incluirlas.

Evidentemente, pudimos observar que los puntajes asignados por las organizaciones o federaciones afines a cada disciplina son los que automáticamente hacen un ordenamiento de los elementos del sistema, por lo que es obligatorio que estudiemos este aspecto; es decir, observar cómo es que estos puntajes se comportan a lo largo del tiempo

o ver cómo es que se comportan en una rodaja temporal específica, que es justamente lo que haremos en el [Capítulo 3](#).

Más aún, parece importante el hecho de estudiar la evolución del sistema en el aspecto de cómo son ocupados los rangos a lo largo del tiempo, ver si hay diferencias entre los sistemas una vez realizado ese análisis y si depende completamente de la forma en que se comporten los puntajes. Un hecho interesante es que todas las organizaciones tiene formas muy peculiares de asignar puntajes a los competidores, la variedad de criterios es evidente, pero también nos preguntamos si esa variedad tendrá alguna influencia para diferenciar los deportes/juegos entre sí en cuanto a su evoluciones.

A partir de ahora y por el resto del trabajo, usaremos exclusivamente las siglas que asignamos a las disciplinas (ver [Tabla 2.1](#)) para referirnos a ellas. En vez de referirnos al sistema conformado por los equipos nacionales de fútbol, diremos únicamente FIFA; o para referirnos al sistema conformado por los jugadores de tabla sobre nieve, simplemente diremos WSD. **Las gráficas y tablas que manejaremos en el resto de esta tesis únicamente harán alusión a esas siglas y debe sobreentenderse a lo que nos referimos.**

Ahora en el siguiente capítulo analizaremos el comportamiento de los puntajes que tienen los elementos de los sistemas y que acabamos de describir de manera detallada la forma en que éstos son asignados.

Distribución de rango. Un análisis estadístico.

Los deportes y juegos aquí estudiados tienen una característica en común: Los elementos (jugadores o equipos) que participan en cierta disciplina son rankeados de acuerdo a un puntaje "score" que se les asigna a criterio de cierta organización encargada de regular las actividades del deporte o juego en cuestión. Ésto quiere decir que el desempeño de un jugador o equipo usualmente se mide por este puntaje que varía con el tiempo. A cada tiempo en que el ranqueo es publicado, los elementos están ordenados de mayor a menor respecto a este puntaje asignado, es decir, el primer lugar tiene el puntaje más alto y el segundo lugar tiene el segundo puntaje más alto; éste va disminuyendo conforme vamos observando los rangos más altos. Lo que haré en este capítulo es estudiar el comportamiento funcional de la distribución de puntajes contra rangos (ésta es la que llamaremos "distribución de rango") y posiblemente estudiar su evolución a lo largo del tiempo o tiempos disponibles para cada disciplina. Propongo 5 modelos que sospecho se parecen a nuestros resultados y justificaré por qué esos. Las distribuciones de rango provenientes directamente de nuestros datos serán comparadas con 5 distribuciones muy bien conocidas de la literatura: La ley de Zipf, distribución β , distribución γ , la distribución $\beta\gamma$ y la distribución doble Zipf.

Abordaré de manera detallada la descripción de distintos métodos estadísticos que proporcionan un parámetro para cuantificar la bondad de ajuste de ciertos modelos a un conjunto determinado de datos experimentales. Existen bondades de ajuste que simplemente nos indican qué tanto se parece la gráfica de una función a un conjunto de puntos provenientes de ciertas mediciones; pero también existen bondades que cuantifican el parecido de una cierta distribución de probabilidad modelo con la distribución empírica de datos provenientes de un sistema real. Lo que haré aquí será estudiar algunas de estas medidas pues serán de vital importancia para concluir si los modelos propuestos son adecuados para los sistemas que aquí trataré. Sobre todo, estas bondades de ajuste serán utilizadas para las "distribuciones de rango", también la utilizaremos para el caso de "la diversidad de rango" que es el punto central de mi trabajo en el siguiente capítulo.

En este capítulo, primero definiremos "la distribución de rango", que es una distribución fija en el tiempo que describe la manera en que se distribuyen los puntajes o "scores" para un determinado deporte/juego en una fecha específica, asimismo proporcio-

naré y justificaré los 5 modelos teóricos para la distribución de rango que compararemos con los valores empíricos de la misma. Posteriormente, introduciré criterios cuantitativos que nos permitirán concluir objetivamente si un modelo teórico para la distribución de rango es el mejor respecto a otros, los llamados criterios de bondades de ajuste. Finalmente, compararé los modelos teóricos con los datos empíricos de una rodaja temporal, a un tiempo, ésto para observar la forma funcional de cada uno de los modelos y darse una idea del comportamiento cualitativo de las mismas. Al final del capítulo, calcularé las bondades de ajuste de los modelos con las distribuciones de rango empíricas para cada una de las rodajas temporales para cada uno de los deportes/juegos para que al final proporcione un valor promedio y sus desviaciones estándar; así podremos concluir objetivamente cuál de los modelos es el más adecuado.

3.1. Definición y motivación.

Se han realizado muchos estudios sobre las frecuencias de aparición de las palabras en diversos textos. George Zipf [29] encontró que cuando tenemos un cuerpo extenso de palabras, podemos clasificarlas de acuerdo a su frecuencia de aparición, es decir, las palabras que tengan frecuencias más altas ocuparán los rangos o posiciones más bajos y las palabras que tengan los rangos más altos corresponderán con las que tienen menores frecuencias de aparición. Zipf propuso que las frecuencias están relacionadas con su rango de la forma $f \sim 1/k$, donde k representa el rango de la palabra y f es la frecuencia relativa de la misma. En honor a Zipf, esta regularidad se le conoce como ley de Zipf. Esta idea, es un buen comienzo para comenzar a dar una distribución de los rangos de la palabras, lo que se podría llama "distribución de rango".

Hace poco se realizó un estudio de estas ideas para el caso de un conjunto de datos muy extenso proveniente de "Google Books Ngram Viewer"[11], el cual consiste en un compilado de masivo proveniente del aproximado 4% de los libros jamás impresos alrededor del mundo hasta el año 2009. De este compilado se pueden extraer las frecuencias por año de todas las palabras que se hayan usado en cualquiera de estos libros. En [2], procedieron a estudiar esa distribución de rango que Zipf propuso anteriormente (la ley de Zipf se había visto sólo para el caso de un texto específico a la vez), sólo que aquí se está tomando en cuenta una cantidad mucho más extensa de libros en 6 idiomas diferentes: ruso, alemán, inglés, español, italiano y francés. Además se puede estudiar una evolución temporal de la distribución de rango en cuestión, pues la base de datos consiste en una clasificación por año de la frecuencia de aparición de las palabras. Lo que encontraron en [2] es que la ley de Zipf no necesariamente es adecuada cuando se toma en cuenta una gran cantidad de libros, además la diferencia aquí es que no se calcula la distribución de rango de un texto solamente, sino que la distribución de rango por año del compilado de libros que tiene disponible la base de datos de "Google". Y esta conclusión no parece ser una gran sorpresa pues en [30] se intenta demostrar que textos generados de manera artificial concantenando caracteres que incluyen espacios en blanco para marcar la separación de palabras, observándose que la correspondiente distribución de rango de dichos textos también muestran un comportamiento tipo Zipf,

lo que implicaría que esta ley no es relevante para el caso de los idiomas.

Motivado por esta idea, en el contexto de deportes y juegos, definiré una distribución de rango similar a la utilizada en el caso de las palabras. En este caso, los jugadores o equipos jugarán el papel de las palabras, mientras que los puntajes jugarán el papel de las frecuencias (pues ambas cantidades son las encargadas de asignarle un rango específico al elemento del sistema en cuestión). Por lo tanto la definición de rango en nuestro trabajo será: *Es la relación entre los puntajes o "scores" los rangos de los jugadores o equipos que conforman al sistema complejo jerárquico en cuestión. Si graficamos esta relación, tendríamos que hacer una gráfica de puntajes vs rangos.* En la figura 3.1 observamos las gráficas de las distribuciones de rango para las últimas fechas disponibles en las bases de datos utilizadas aquí.

3.2. Modelos para la distribución de rango.

En [31] se hace un estudio profundo de la distribución de rango y su posible aproximación a un modelo funcional bien definido que la reproduzca. En este artículo se discute que pueden llegar a presentarse discrepancias entre modelos de un sólo parámetro (como la ley de Zipf), y los datos empíricos. Por eso parece conveniente considerar modelos con más de un parámetro. Igual en [31], comienzan a estudiar algunas leyes con uno y dos parámetros, tales como la distribución Beta ($\sim (n+1-r)^b/r^a$), la distribución exponencial ($\sim e^{-ar}$), la distribución de Yule ($\sim b^r/r^a$) o la distribución de Mandelbrot ($\sim 1/(r+b)^a$). Menciona que, por ejemplo, la ley de Menzerath-Altman [32] relaciona la longitud de dos unidades lingüísticas, las de las oraciones y las de las palabras; si y es la longitud de las oraciones respecto a la unidad lingüística (las palabras) y x es la longitud de las palabras respecto al número de letras, entonces la ley dice que la relación entre estas longitudes es $y \sim x^b e^{-c/x}$ y donde tenemos dos parámetros libres, b y c . Basados en estas ideas, suena natural comenzar a considerar modelos para la distribución de rango con más parámetros libres, la cuestión es cuáles utilizar.

En el artículo de estudio de idiomas que motivó inicialmente este trabajo [2] y al cual ya hemos hecho referencia en numerosas ocasiones, se tiene material suplementario donde se discute sobre la distribución de rango ya bien definida anteriormente para el estudio de los idiomas. En ese trabajo se menciona que de manera analítica se puede dar origen a variados modelos para la distribución de rango considerando procesos de nacimiento o muerte. Con esto queremos decir que si se tiene $N(k, t)$ como el número de veces que cierta palabra (en este caso jugador o equipo) aparece con el rango k en el tiempo t . Ahora bien, si $B(k)$ y $D(k)$ denotan la probabilidad por unidad de tiempo que una palabra entra o deja el rango k , se tiene la relación: (desarrollado también en [33])

$$\begin{aligned} \frac{\partial}{\partial t} N(k, t) = & \{\xi(k) - F[\Sigma(t)]\} N(k, t) + \\ & \{D(k+1)N(k+1, t) + B(k-1)N(k-1, t) - [D(k) + B(k)]N(k, t)\} \quad (3.1) \end{aligned}$$

3. DISTRIBUCIÓN DE RANGO. UN ANÁLISIS ESTADÍSTICO.

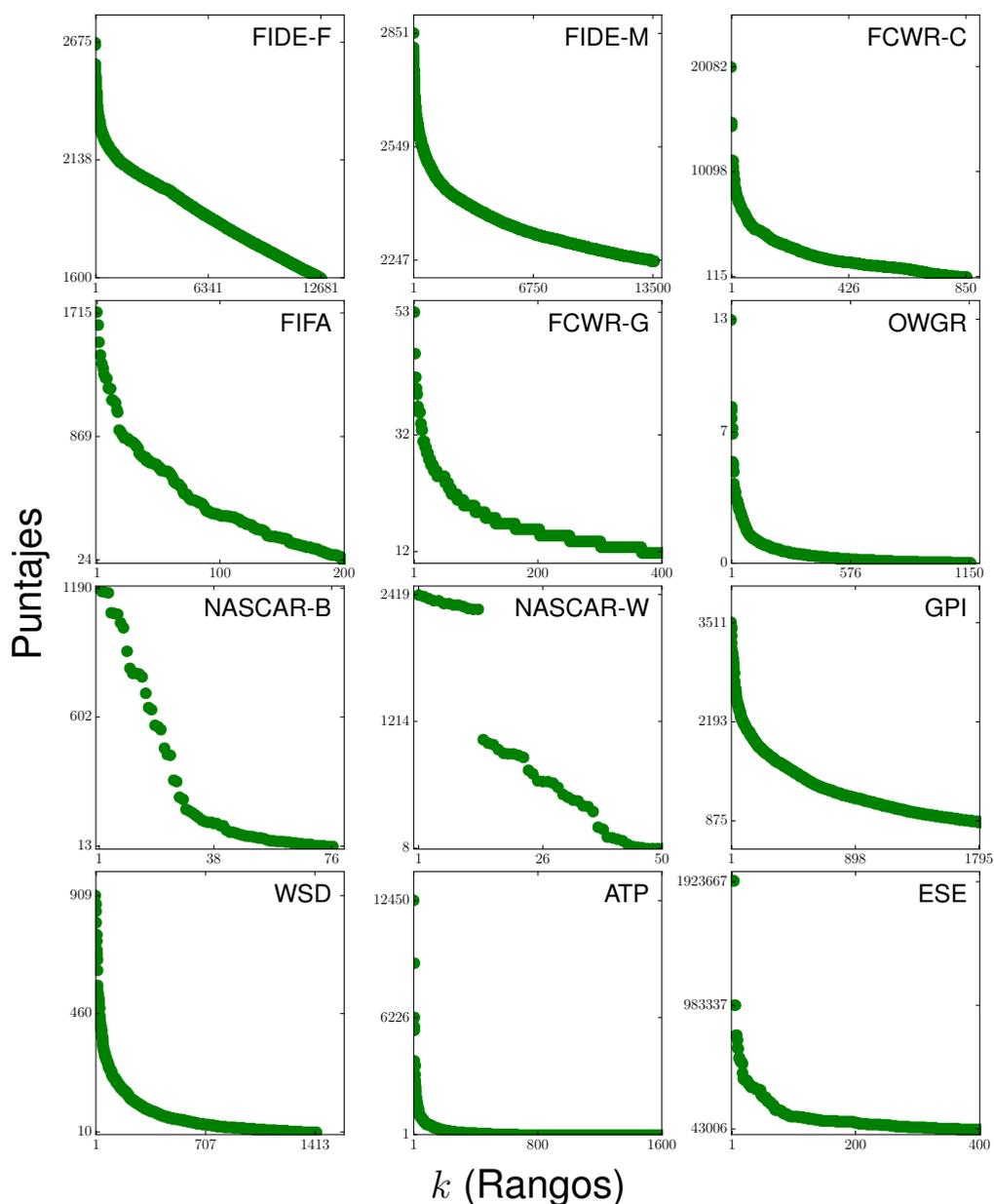


Figura 3.1: Distribuciones de rango para las 12 bases de datos. En todos los casos se observa que las distribuciones de rango son decrecientes, como uno esperaría de acuerdo a la definición que anteriormente se dio. Las fechas correspondientes a las distribuciones de rango aquí graficadas son, para cada deporte las siguiente: FIDE-F(Abril 2016), FIDE-M(Abril 2016), FCWR-C(Semana 53 del 2014), FIFA(Junio 2017), FCWR-G(Semana 33 de 2017), OWGR(21/05/2017), NASCAR-B(2015), NASCAR-W(2013), GPI(31/05/2017), WSD(26/03/2018), ATP(27/12/2010), ESE(2016). Estas fechas corresponden a la última disponible en las bases de datos.

donde podemos observar que el primer coeficiente de $N(k, t)$ tiene dos términos, el primero describe la razón de crecimiento local y el segundo la razón de crecimiento global actuando sobre $N(k, t)$. La cantidad total de palabras en el tiempo t está dada por $\Sigma(t)$ y F es una función que determina las constricciones globales al comportamiento de las palabras. Los términos que se encuentran en las segundas llaves se refiere a un equilibrio que proviene de las contribuciones de nacimiento y muerte $B(k)$ y $D(k)$ a primer vecino, es decir con respecto a los rangos inmediatos superiores o inferiores, es decir $k \pm 1$ en el tiempo t . Ahora bien, siguiendo la suposición de que

$$\Sigma(t) = \sum_k N(k, t)$$

se define la densidad de probabilidad de encontrar una palabra en el rango k , o también se le puede llamar la distribución de las frecuencias relativas como:

$$n(k, t) = \frac{N(k, t)}{\Sigma(t)} \quad (3.2)$$

Ahora bien, esta es la indentificación clave que necesitaba. La distribución de frecuencias relativas es la análoga a la distribución de rango que describí anteriormente para deportes y juegos. Entonces me tengo que fijar en el modelo teórico obtenido para 3.2 y hacer una analogía con la distribución de rango que utilizaré para este trabajo. Continuando con el proceso seguido en [2], se llega a una ecuación maestra para las frecuencias relativas y que está dada por:

$$\dot{n}(k, t) = D(k+1)n(k+1, t) + B(k-1)n(k-1, t) - [D(k) + B(k)]n(k, t) \quad (3.3)$$

Recordemos que una ecuación maestra es aquella utilizada para describir la evolución de un sistema que puede ser modelado por la combinación probabilística de estados intermedios del sistema a cualquier tiempo. [34] Esta ecuación determina de manera completa la evolución de dicho sistema. En este caso, se obtuvo la ecuación maestra para la distribución de frecuencias relativas (describiendo al sistema en el que las palabras, jugadores o equipos se van acomodando en distintos rangos a lo largo del tiempo). De la misma forma, en [2] mencionan que haciendo la suposición de que k es pequeña, se puede tratar como una variable continua, convirtiendo a la ecuación maestra 3.3 de una forma discreta a una continua, obteniendo la conocida ecuación de Fokker-Planck:

$$\frac{\partial n(k, t)}{\partial t} = -\frac{\partial}{\partial k}[g(k)n(k, t)] + \frac{1}{2}\frac{\partial^2}{\partial k^2}[f(k)n(k, t)] \quad (3.4)$$

donde $f(k) = B(k) + D(k)$ y $g(k) = B(k) - D(k)$. Ahora bien, el caso estacionario de esta ecuación y su respectiva solución $m(k)$ se escribiría de la forma:

$$g(k)m(k) = \frac{1}{2}\frac{d}{dk}[f(k)m(k)] \quad (3.5)$$

Es decir, esta sería la ecuación cuya solución nos proporcionaría la distribución de rango a un cierto tiempo. En el caso de las bases de datos que estamos trabajando, lo que

obtuvimos es una aproximación funcional al comportamiento de los puntajes vs. rangos en una rodaja temporal específica de los rankings para cada disciplina, donde $m(k)$ es la distribución de rango; sin embargo, debemos tomar en cuenta que hay claras diferencias entre la distribución de frecuencias relativas de los idiomas con la distribución de rango para deportes y juegos, mientras que en [2], las aproximaciones que acabo de mencionar se utilizan pensando en palabras y sus frecuencias. El caso de deportes y juegos es completamente equivalente, y más adelante aclararé que la $m(k)$ también se puede utilizar sin preocupación para los casos que aquí nos competen. Continuando con el desarrollo para la solución estacionaria de la ecuación de Fokker-Planck, si aproximamos $g(k)/m(k)$ de acuerdo a la aproximación de Padé [35] se tiene que $g_n(k)/f_n(k) = A_0 + \sum_{k=1}^n \frac{A_k}{(k+c_k)}$, se tiene que la solución estacionaria de 3.5 es:

$$m(k) = N \exp(A_0 k) \prod_{k=1}^n (k + c_k)^{-A_k} \quad (3.6)$$

Y se asume que $D(k)$ y $B(k)$, las transiciones de probabilidad, tienen las formas cuadráticas más simples:

$$D(k) = \lambda_1 (c_1 + k)(N_1 - k)$$

$$B(k) = \lambda_2 (c_2 + k)(N_2 - k)$$

entonces la solución más general estacionaria suponiendo que las transiciones de probabilidad tienen esta forma es:

$$m(k) = N \exp(-bk) \frac{(\bar{N} - k)^q}{(\bar{c} + k)^a} \quad (3.7)$$

donde $\bar{N} = \frac{1}{2}(N_1 + N_2)$, $\bar{c} = \frac{1}{2}(c_1 + c_2)$, $a = c_1 - c_2 + 1$ y $q = N_1 - N_2 - 1$; a la constante A_0 la renombramos como $-b$. Es de esta ecuación 3.7 que sacaré los modelos que utilizaré para compararlos con los datos de las bases de datos con las que cuento.

3.2.1. Distribución de rango vs. distribución de frecuencias relativas

Hasta este momento, he utilizado resultados obtenidos pensando en las distribuciones de frecuencias relativas. A partir de procesos de nacimiento y muerte (nacimiento: incremento de rango, muerte: descenso de rango; ambos casos con respecto al rango consecutivo, es decir, decaimiento o aumento a primeros vecinos) se obtuvo una ecuación maestra para la distribución de frecuencias relativas (en realidad, es la densidad de probabilidad de frecuencias, formalmente) a lo largo del tiempo. Posteriormente, haciendo varias simplificaciones de acuerdo a [2], se llegó a una solución estacionaria de esta distribución de frecuencias. Matemáticamente, esta distribución debe estar normalizada, para el caso discreto esto quiere decir que:

$$\sum_n p_n = 1$$

donde, si la variable aleatoria es tal que puede tomar los valores $X = \{x_1, x_2, \dots, x_n\}$, en este caso toma una cantidad finita de valores, entonces $p_n = P(x_n) \geq 0$. Es decir, si $m(k)$ es el valor en la distribución de frecuencias, y la palabra *pepe* es la palabra correspondiente al rango k , entonces $m(k) = p(k)$, en nuestra definición, donde $p(k) < 1$ es el número de veces que aparece *pepe* en un texto dividido entre el número total de palabras en el texto; de esta forma la suma de los $p(k)$ para todos los k es igual a 1 como debe ser. Sin embargo, en la definición de distribución de rango que di para el contexto en el que estoy trabajando, el valor de $m(k)$ para el jugador Messi, por ejemplo, que ocupa el rango 1 en cierta fecha, es el puntaje o "score" que tiene asignado por el organismo encargado de hacer el ranking, y este puntaje no es necesariamente menor que 1, de hecho, esa llamada "distribución de rango" no está normalizada necesariamente. **En otras palabras, la distribución de rango aquí definida en el contexto de deportes y juegos no es una distribución de probabilidad (tampoco una función de densidad de probabilidad), pues no cumple las condiciones necesarias para serlo.** Sin embargo, como veré con detalle más adelante, sí está relacionada con una distribución de probabilidad en el sentido estricto matemático, pues su equivalencia a una es evidente después de cierta transformación, así que optaré por seguirla llamando distribución de rango.

La ecuación 3.7 nos proporcionará los modelos que aquí se utilizarán, sin embargo, ésta fue obtenida en el contexto de la distribución de frecuencias, y se interpreta a la constante \mathcal{N} como una constante de renormalización. En nuestro caso, seguiré utilizando a la constante en cuestión pero no representará una constante de renormalización, sino simplemente otro parámetro de ajuste del modelo a los datos empíricos.

3.2.2. Los modelos a utilizar

Como en el caso de [2] utilizaré 5 modelos para la distribución de rango, los cuales tienen su origen en la ecuación general 3.7. Ahora bien, consideraremos siempre $\bar{c} = 0$; la constante \bar{N} se relaciona directamente con el número de elementos en el sistema (pues en la expresión en 3.7 se está operando con el rango), entonces nombraremos a esta constante como $\bar{N} = N + 1$ y el sumamos un 1 para evitar divergencias no deseadas en el caso de que q obtenga valores negativos, por lo tanto, 3.7 se convierte en:

$$f(k) = \mathcal{N} \frac{(N + 1 - k)^q \exp(-bk)}{k^a} \quad (3.8)$$

entonces por la definición de distribución de rango se tiene f es el puntaje asociado al rango k , a es un exponente que domina casi toda la curva, b un exponente que controla la caída exponencial y q un decaimiento algebraico que regula la caída pronunciada de la curva para k muy grande. Finalmente, N es el número total de elementos (es decir, el número total de jugadores o equipos) en el sistema, y \mathcal{N} es una constante que controlará el dominio numérico en el cual se encuentran los puntajes.

En las ecuaciones 3.9-3.13 enlisto estos 5 modelos. Vemos que en el caso que si de la ecuación general 3.8 hacemos $b = q = 0$, se recupera el modelo m_1 de la Ecuación 3.9; por otro lado si $b \neq 0$ y $q = 0$ se obtiene la Ecuación 3.10, que le llamaremos distribución

γ por ser regularmente así conocida en la literatura; por otro lado, si $b = 0$ y $q \neq 0$ se llega a la [Ecuación 3.11](#) que se le conoce como distribución β ; finalmente el modelo de la [Ecuación 3.12](#) se simplemente el producto de $\beta\gamma$ y justamente coincide con la [Ecuación 3.8](#).

Los primeros cuatro modelos son:

$$m_1(k) = \mathcal{N} \frac{1}{k^a} \quad (3.9)$$

$$m_2 = \mathcal{N} \frac{\exp(-bk)}{k^a} \quad (3.10)$$

$$m_3(k) = \mathcal{N} \frac{(N+1-k)^q}{k^a} \quad (3.11)$$

$$m_4(k) = f(k) \quad (3.12)$$

mientras que el quinto modelo está dado por la ley doble Zipf [\[36\]](#)

$$m_5(k) = \mathcal{N} \begin{cases} \frac{1}{k^a}, & k \leq k_c \\ \frac{k_c^{a'-a}}{k^{a'}} & k > k_c \end{cases} \quad (3.13)$$

donde a' es un exponente alternativo que regula el comportamiento de la curva después de un rango crítico k_c , este modelo ha sido utilizado exitosamente en otros contextos, lo cual me motiva a utilizarlo también para el caso de deportes y juegos. Antes de continuar con la comparación de estos modelos con los datos empíricos que poseemos, contaré brevemente la importancia de estos modelos en el estudio de sistemas complejos, algunas de sus aplicaciones y sus implicaciones.

3.2.3. Ley de Zipf. Distribuciones β y γ .

Este es tal vez el modelo más famoso que trataremos. La ley de Zipf es un comportamiento genérico que siguen varios sistemas físicos y sociales. De hecho, esta ley tomó fuerza por primera vez cuando el lingüista George Kingsley Zipf encontró que la distribución de frecuencia de las palabras en textos escritos se comportan como una ley de potencias, justamente este comportamiento es lo que empezó a llamarse Ley de Zipf. [\[29\]](#) Su forma funcional se define, en general de la forma, $f \sim 1/k^a$, donde el exponente es generalmente mayor que 1. Como ya mencioné anteriormente, la ley de Zipf describe de manera muy adecuada a las distribuciones de frecuencias de palabras en textos para muchos idiomas, se ha hecho el estudio en una gran cantidad de ellos, y la ley de Zipf parece siempre ser la descripción adecuada. Una de las implicaciones más interesantes de esta ley es que deja al descubierto el hecho de que si $f(1)$ es la frecuencia correspondiente a la palabra en el rango 1 entonces la palabra con rango 2, tiene una frecuencia $f(2) = f(1)/2$, es decir que se repite la mitad de las veces en el texto respecto a la de

rango 1, similarmente, la palabra con el tercer rango se repite en el texto una tercera parte de las veces que lo hace la de primer rango.

Existen algunas generalizaciones de la ley, de Zipf, como la ley de Zipf-Mandelbrot [37], la cual consiste en una modificación de la ley de Zipf como $f(k) \sim 1/(h+k)^a$, donde a y h son parámetros de ajuste. De hecho, nosotros obtuvimos esta ley en la derivación de la forma más general de nuestro modelos que llevamos a cabo en la [Sección 3.2](#), también obtuvimos esta ley de Mandelbrot. Se ha observado que esta ley funciona mejor con cuerpos gigantescos de texto de palabras.

En [38] se intenta dar una explicación del por qué no sólo en el campo de los idiomas se presenta esta ley de Zipf, también ocurriendo en sistemas que aparentemente no tienen nada que ver entre sí. Allí mismo se intenta argumentar que la razón por la que se aprecia ese comportamiento se debe a que los sistemas que se estudian se pueden dividir en grupos aleatorios, como los ciudadanos de una ciudad. Se observa, de igual forma en [38] que haciendo el estudio para diferentes sistemas, se deben tomar a consideración tres factores: el número total de elementos del sistema, el número de grupos en los que se divide el sistema, y el número de elementos del sistema más grande, obteniendo que las distribuciones para cada sistema tienen una forma muy parecida a $P(k) \sim \exp(-bk)/k^\gamma$, donde γ es una función de los parámetros anteriormente mencionados. Curiosamente, en este trabajo también estudiaré ese modelo, aquí lo llamamos distribución γ y corresponde al modelo de la [Ecuación 3.10](#), lo cual da una motivación más a tratarlo aquí.

Las distribución β corresponde al modelo presentado en la [Ecuación 3.11](#) y junto con la distribución γ tiene numerosas aplicaciones. Por ejemplo, la distribución β se utiliza en estadísticas de orden (que se encarga de obtener valores especiales de un conjunto de datos, tales como el mínimo o el máximo; en inferencia Bayesiana, análisis de ondas o lógica subjetiva.[39] Por otro lado, la distribución γ ha sido utilizada, por ejemplo, para predecir la cantidad de lluvia en un lugar de reserva, en la comunicación inalámbrica pues modela el desvanecimiento de la potencia de una señal, en neurociencia modela la distribución de los intervalos entre los picos de una señal anatómica del cerebro, y demaás aplicaciones en distintas áreas. [40]

Naturalmente, la razón por la que incluyo el modelo m_4 descrito en la [Ecuación 3.12](#) es porque consiste en una combinación de las distribuciones γ y β , agregando más parámetros a la forma funcional de la distribución. Al tener más grados de libertad es posible que permita modelar mejor el comportamiento empírico de los datos. El modelo m_5 que se encuentra representado en la [Ecuación 3.13](#) consiste en la concatenación de dos leyes de potencia. La razón por la que se consideran dos leyes de potencia es debido a que en [36] encuentran que existen dos regímenes de palabras, uno que consiste en palabras con altas frecuencias que no afectan la aparición de otras, y el otro régimen consiste en palabras menos frecuentes pero que pueden incluir la aparición de nuevas. En [29] hacen una derivación de esta nueva ley doble Zipf y que ha sido utilizada en muchos contextos como el que acabo de mencionar [36]. Por estas razones, aunque m_5 no fue obtenido analíticamente en la [Sección 3.2](#) como un caso de la [Ecuación 3.8](#) lo incluiré.

3.3. Equivalencia entre distribución de rango (DRe) y distribución acumulativa de los puntajes (DCe).

Como mencioné en la [Subsección 3.2.1](#), la distribución de rango, tal y como la definí al inicio de este capítulo, no es formalmente una distribución de probabilidad, pues no cumple con las condiciones necesarias para serlo. Ahora bien, en [\[31\]](#) se hace un análisis muy interesante al respecto. Demuestran que la distribución de rango empírica (DRe)(que consiste graficar los puntajes contra los rangos) es equivalente a la distribución acumulativa de los puntajes. Para entender ésto, primero recordemos la definición de la distribución empírica de un conjunto de puntos. Sea $\{s_1, \dots, s_n\}$ una muestra finita de variables aleatorias distribuidas idénticamente con la distribución acumulativa común $F(s)$. Entonces la función de distribución empírica de este conjunto se define como: [\[41\]](#)

$$M(s) = \frac{\text{número elementos en la muestra} \leq s}{N} = \frac{1}{N} \sum_{i=1}^N \theta(s - s_i) \quad (3.14)$$

donde la función θ es la función escalera. Claramente tenemos que $\lim_{s \rightarrow \infty} M(s) = 1$, justo como se desea pues es una distribución de probabilidad en el sentido estricto matemático. Una vez que tenemos esta definición, procederé a justificar la equivalencia entre la distribución empírica como se definió en la [Ecuación 3.14](#); para ésto seguiré la idea dada en [\[31\]](#). Observemos la [Figura 3.2](#), en la cual se cuenta el cómo transformar la distribución de rango empírica en 4 pasos. Primero que nada [3.2a](#)) tenemos la distribución de rango empírica, lo que se hace es transformar a) en b) reflejando a) con respecto al eje y , más una transformación en el eje k de la forma $(N + 1 - k)/(N + 1)$, donde N es el número total de elementos del ranqueo, de esa forma se está invirtiendo el antiguo eje k y normalizando entre los valores $(0, 1)$. El objetivo de esta operación es empezar la distribución empírica acumulada desde el punto de menor rango.

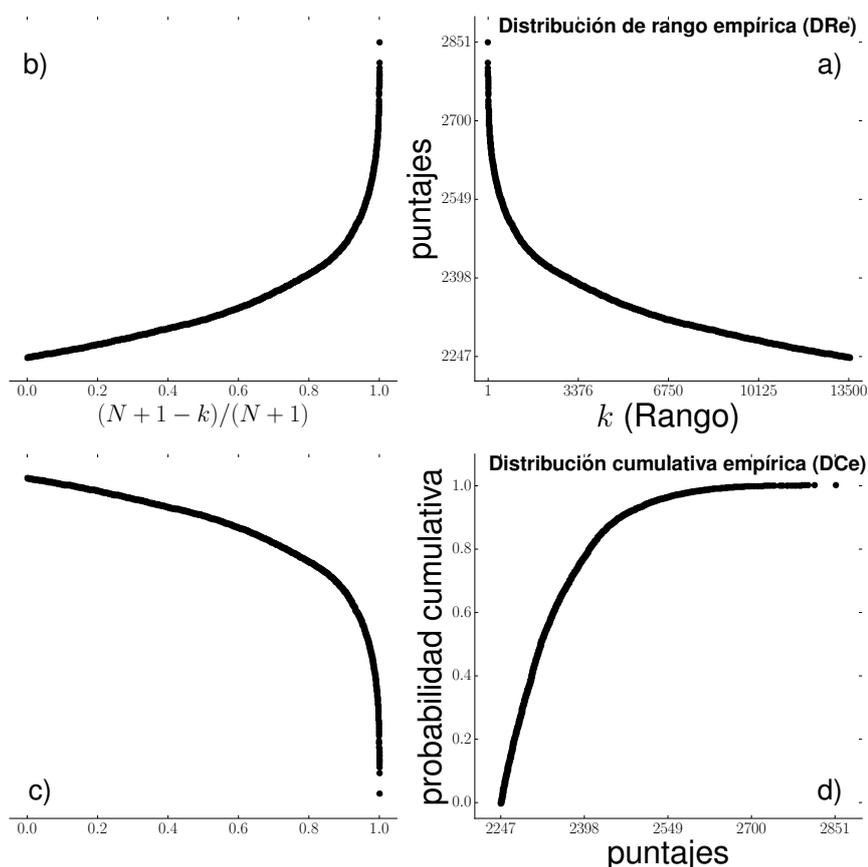
Posteriormente, para pasar de [3.2b](#)) a [3.2c](#)) se hace una reflexión respecto al eje k y para llegar de [3.2c](#)) a [3.2d](#)) se hace una rotación de 90° en sentido contrario a las manecillas del reloj. La [figura 3.2d](#)) representa la distribución acumulativa empírica del conjunto de datos (en este caso los datos consisten en los puntajes que se tienen para cada uno de los elementos–jugadores o equipos).

Formalmente, al llegar a d) en la [figura 3.2](#), es solamente un conjunto de puntos y no una función continua como se definió en la [Ecuación 3.14](#). Para obtener formalmente la distribución acumulativa empírica sólo se hace una extensión continua de d), de la siguiente forma: Si $\{s_1, s_2, \dots, s_N\}$ es el conjunto de puntajes ordenado que poseen los elementos del sistema y $M(s_i)$ es el valor del punto s_i en la [figura 3.2d](#)), entonces la distribución acumulativa empírica de los mismos es haciendo que $\forall s$ tal que $s_i < s < s_{i+1}$, entonces $M(s) = M(s_i)$, de tal manera que también se obtiene una función escalera como lo sugiere la [Ecuación 3.14](#).

Aquí entra una dificultad de lenguaje, y proviene del hecho de que siempre se nombra a las funciones de densidad de probabilidad con las funciones de distribución de probabilidad. Recordemos que la distribución de rango se

3.3 Equivalencia entre distribución de rango (DRe) y distribución cumulativa de los puntajes (DCe).

Figura 3.2: Proceso para transformar la distribución de rango empírica (DRe) a la distribución cumulativa empírica (DCe) Vemos que el proceso de transformación va desde a) a d), mostrando la equivalencia entre la distribución de rango empírica y la distribución cumulativa empírica.



definió motivados de las distribuciones de frecuencias relativas, que como ya se mencionó anteriormente, en realidad es la densidad de frecuencias para las palabras de un texto. Por ello, la distribución de rango empírica, pues proviene de los datos, (que es una función discreta) no es una distribución de probabilidad. El proceso que acabo de describir hace una relación entre la mal llamada distribución de rango empírica para un ranqueo que usa los puntajes de jugadores o equipos y que es una función discreta con la distribución cumulativa del conjunto de puntajes (ésta sí es una distribución de probabilidad) y que es una función continua de acuerdo a la definición dada en la [Ecuación 3.14](#).

De esto se puede obtener un resultado aún más interesante. Supongamos que queremos aproximar la distribución de rango empírica con algún modelo, llamémosle convenientemente $m_i(k)$, donde su variable k se encuentra en el dominio de los rangos. Esto

de aproximar el modelo m_i quiere decir que ajustaremos la función m_i al conjunto de puntos que conforma la distribución de rango empírica obteniendo el conjunto de parámetros adecuados en la definición de m_i . Entonces, este modelo se puede interpretar como la versión continua de la distribución de rango empírica del sistema. Para obtener su respectiva distribución acumulativa, que llamaremos $M_i(m_i)$ siguiendo el mismo procedimiento que en la [Figura 3.2](#) se puede deducir que el transformar m_i sería equivalente a:

$$M_i(m_i) = \frac{N + 1 - k(m_i)}{N + 1} \quad (3.15)$$

donde m_i ya vimos que es la versión continua de la distribución de rango empírica; $k(m_i)$ es claramente la función inversa de $m_i(k)$, notando que el valor $m_i(k)$ está en el dominio de los puntajes, entonces la [Ecuación 3.15](#) está recibiendo valores de puntajes. Más adelante daré un ejemplo de cómo hacer esta transformación para el caso de modelos teóricos continuos.

En conclusión, esta equivalencia que acabamos de delucidar me será muy útil más adelante por la siguiente razón: obtener un modelo teórico $m_i(k)$ adecuado para la distribución de rango empírica es equivalente a obtener un modelo teórico $M_i(m_i)$ adecuado para el caso de la distribución acumulativa empírica de los datos. Todo esto implica que trabajaré indistintamente los modelos M_i y m_i a lo largo de este trabajo. Pero ¿por qué pasar por tantos problemas para transformar a la distribución de rango empírica a una distribución de probabilidad empírica? Bueno, la respuesta radica en el hecho de que usaremos técnicas más sofisticadas de probabilidad y estadística para ver si un modelo es el adecuado para nuestros datos empíricos y el lenguaje de la distribución de rango no es suficiente pues uno de los métodos que usaremos para ver qué tan bien se ajusta un modelo requiere del uso estricto del lenguaje de distribuciones de probabilidad.

3.4. Métodos y bondad de ajuste.

Para poder determinar si un modelo teórico es el adecuado para describir el comportamiento empírico de cierto conjunto de datos, es necesario utilizar técnicas cuantitativas bien determinadas. En este momento ya puedo mencionar que justamente los modelos presentados en las ecuaciones [3.9-3.13](#) serán los utilizados para la distribución de rango empírica que se analizará directamente de los conjuntos de datos con que contamos y que describí a detalle en el [Capítulo 2](#). Ahora bien, es de suma importancia poder concluir cuál de esos modelos es el más adecuado, y además poder concluirlo de manera cuantitativa y objetiva. En este trabajo usaré dos medidas de bondad de ajuste cuyos valores permiten concluir si un modelo teórico es el adecuado, y no sólo eso, ver si funciona mejor que cualquiera de los otros modelos teóricos que se están poniendo a prueba. Las medidas de bondad de ajuste que utilizaré serán: el coeficiente de determinación R^2 y el índice p de Kolmogorov-Smirnov, los cuales explico en detalle a continuación.

3.4.1. Coeficiente de determinación R^2

En el trabajo [31] que ya he referencia anteriormente, también se hizo una comparativa de distintos modelos con datos empíricos para modelar teóricamente las distribuciones de rangos de variados sistemas. Para ver objetivamente si los modelos propuestos son adecuados utilizan justamente el coeficiente de determinación R^2 . Platicaré la utilización de este coeficiente respecto a los datos que aquí analizaré. Por ejemplo, si observamos una rodaja temporal del ranqueo para una disciplina, es decir, el ranqueo de una fecha, tendremos un conjunto de datos de la siguiente forma $\{(k_1, s_1), (k_2, s_2), \dots, (k_n, s_n)\}$; es decir, tendremos un conjunto de n puntos, donde los k_j son los rangos y s_j es el puntaje correspondiente al rango $k_j = j$. Por lo tanto, si tenemos que al ajustar el modelo $m_i(k)$, donde $i = 1, \dots, 5$, a este conjunto de datos (es decir, encontrar los parámetros tales que $m_i(k)$ se ajuste a los puntos antes mencionados) el coeficiente R^2 se define como:[31]

$$R^2 = 1 - \frac{\sum_{j=1}^n |m_i(k_j) - s_j|^2}{\sum_{j=1}^n |s_j - \langle s \rangle|^2} \quad (3.16)$$

donde $\langle s \rangle$ es el promedio de los puntajes en esa rodaja temporal específica. Entre más cercano esté R^2 al valor 1, el conjunto de puntos se ajustan mejor al modelo. Ésta definición tiene muchas variaciones equivalentes, pero el que acabamos de dar en la Ecuación 3.16 es la que aquí utilizaré. Observemos que el cociente en el segundo término del lado derecho de la misma ecuación $\sum_{j=1}^n |s_j - \langle s \rangle|^2$ es justamente la varianza de los puntajes en esa rodaja temporal, mientras que el numerador $\sum_{j=1}^n |m_i(k_j) - s_j|^2$ es una medida de la desviación de los puntos con el modelo, el hecho de que ese cociente se aproxime mucho a cero, quiere decir que el numerador es mucho más pequeño que el denominador, es decir, que la desviación de los datos al modelo es mucho menor que la desviación estándar de los puntajes. En [31] se argumenta formalmente el por qué ésto permite concluir que los datos se comportan de acuerdo al modelo conforme $R^2 \rightarrow 1$.

3.4.2. El índice p de Kolmogorov-Smirnov.

Esta medida de bondad de ajuste es complicada de explicar, pues tiene mucha teoría matemática por detrás. En este trabajo me limitaré a explicar el proceso con el cual se calcula, la razón por la que la usaré, y algunas de las interpretaciones que se le dan a sus posibles valores. La razón que me motivó el uso de esta medida, fue a que en el trabajo [9] se estudiaron posibles modelos teóricos para la distribución acumulativa de los puntajes que determinan el ranqueo de jugadores y equipos en 40 disciplinas deportivas. Algo muy parecido a lo que estoy haciendo aquí. Usando modelos motivados en las mismas ideas (también utilizan las distribuciones γ y β pero en el contexto de las distribuciones acumulativas) buscan de manera objetiva el modelo más adecuado para dichas distribuciones pero utilizando una bondad de ajuste más fina que el coeficiente de determinación R^2 y se trata del parámetro que aquí explicaremos, el llamado índice p de Kolmogorov-Smirnov. La virtud de este parámetro radica es que sólo se puede utilizar para determinar si una distribución de probabilidad teórica corresponde a la distribución empírica de cierto conjunto de datos.

Lo último mencionado es el punto clave: sólo se puede aplicar esta bondad de ajuste cuando estamos comparando una distribución de probabilidad teórica a la distribución empírica de un conjunto de datos. Por ello, fue fundamental lo desarrollado en la [Sección 3.3](#), donde argumentamos que la distribución de rango empírica (DRe) es equivalente a una distribución de probabilidad (la distribución acumulativa empírica, DCe). Por lo tanto podremos usar de manera indirecta esta bondad de ajuste.

Un artículo muy famoso [\[42\]](#) sobre leyes de potencias, métodos y bondades de ajuste a distribuciones de probabilidad describe el proceso con el cual se calcula el índice p de Kolmogorov-Smirnov y sus implicaciones. Como el proceso es bastante largo, el proceso lo detallaré en el [Apéndice A](#). Lo que haré será una adaptación del proceso descrito en [\[42\]](#) al caso de la distribución de rango. Es decir, obtendré el parámetro de bondad de ajuste p para un modelo m_i ajustado a la distribución de rango empírica. Recordando que la p sólo se puede calcular para comparar una distribución teórica con la distribución acumulativa empírica de los puntajes. **La distribución de probabilidad teórica es justamente la transformación de m_i a M_i como se describió en la Ecuación 3.15, ésta será comparada con la distribución acumulativa empírica de los puntajes y se obtendrá el índice de Kolmogorov-Smirnov correspondiente a esta distribución de probabilidad teórica, diremos que gracias a la equivalencia del modelo m_i con M_i y a la equivalencia de la DRe y DCe, la p obtenida es una bondad de ajuste del modelo m_i con la DRe.**

En el proceso se define una cantidad importante llamada *estadística de Kolmogorov* D , que es el valor máximo de las distancias entre la DCe de los puntajes y la distribución de probabilidad teórica M_i , y que se define como:

$$D = \sup_s |M_i(s) - M(s)| \quad (3.17)$$

donde $M(s)$ es la DCe como la definimos en la [Ecuación 3.14](#) de los puntajes. Es una norma de funciones, si lo vemos desde el punto de vista matemático. Siempre que aquí proporcionemos el valor D del modelo m_i , en realidad estamos calculando la D como en la [Ecuación 3.17](#) de M_i con la DCe. Vemos que D es la diferencia vertical máxima entre dos curvas.

Ahora bien, resulta que si un modelo (una distribución de probabilidad teórica, en este caso nos referimos a M_i) es consistente con la DCe de los puntajes, se debe tener una p larga o al menos $p > 0.1$. La gran diferencia con el coeficiente R^2 es que, mientras este coeficiente sólo nos indica si un conjunto de datos se parece a una función determinada en el mismo dominio, la p nos indica si efectivamente un conjunto de datos está distribuido de cierta forma (M_i). Es decir, la p no nos indica si los datos se asemejan mucho a una curva, va mucho más allá, nos indica si el conjunto de datos, en efecto, se distribuye de la forma que indica el modelo propuesto. Otra gran ventaja del índice p es que permite considerar que un pequeño conjunto de datos tendrá algo de ruido debido a una pobre estadística. Siendo así, si el modelo en efecto es el correcto, pero tenemos una estadística pobre, podremos aún tener una buena (larga) p .

3.5. Comparación entre modelos y distribución de rango.

Nuestro deseo es poder comparar los modelos de las ecuaciones 3.9-3.13 con las distribuciones empíricas que poseemos para estudiar en este trabajo. En el [Capítulo 2](#) proporcioné de manera detallada las características de las bases de datos con las que cuento. Para cada base de datos, contamos con una distribución empírica generada por los puntajes por cada fecha que forma parte de los datos de dichas disciplinas. Es decir, tenemos una gran cantidad de distribuciones empíricas que se compararán con todos los modelos que aquí me interesan. A modo de ilustración, los resultados de los ajustes entre los modelos 3.9-3.13 y la distribución empírica de puntajes para una fecha representativa de cada deporte o juego se muestran en las figuras 3.3, 3.4 y 3.5 mientras que la [Tabla 3.1](#) resume los parámetros obtenidos en dichos ajustes. Las rodajas temporales (o fechas) presentados en las figuras son: FIDE-F(Abril 2016), FIDE-M(Abril 2016), FCWR-C(Semana 53 del 2014), FIFA(Junio 2017), FCWR-G(Semana 33 de 2017), OWGR(21/05/2017), NASCAR-B(2015), NASCAR-W(2013), GPI(31/05/2017), WSD(26/03/2018), ATP(27/12/2010), ESE(2016) que coinciden con los ejemplos presentados en la [Figura 3.1](#). Para que se pueda apreciar mejor el comportamiento de cada distribución de rango, las gráficas son presentadas en escala log – log.

De las figura 3.3 es claro que la ley de Zipf (m_1), líneas en rojo, no es adecuada, pues en escala log – log estamos hablando de una recta por tratarse de una ley de potencias; en todos los casos los puntajes tienen cierto cambio en la pendiente o caídas de algún tipo, por lo que la ley de Zipf parece no asemejarse lo suficiente a las curvas. Por otro lado, la distribución Gamma (m_2) ajusta mejor para ciertas bases de datos, particularmente aquellas que no muestran una caída abrupta de los puntajes como función del rango, el caso de FIFA y NASCAR-W sí presentan una caída abrupta y vemos que el la distribución Gamma no parece ser buen modelo para esos casos. Las bases de datos con una caída abrupta en los puntajes (FIFA y NASCAR-W) tampoco parecen ajustar bien a la distribución Beta (m_3) [Figura 3.4](#), lo que parece indicar que esta distribución tiene una caída demasiado pronunciada a rangos más altos poco controlable y que no reproduce lo visto en la distribuciones de rango empíricas. Sin embargo, la mayoría de los deportes y juegos parecen tener un caso intermedio donde ambas funciones representan mejor al comportamiento global de los puntajes, y consecuentemente el ajuste es considerablemente mejor para la combinación de ambas distribuciones, refiriéndonos a m_4 [Figura 3.4](#); de hecho, justamente observamos que el modelo m_4 parece acercarse más al comportamiento de FIFA y NASCAR-W, dándonos cuenta de que las caídas del modelo son más controlables en m_4 debido a que en su contenido, la función Gamma interviene. También podemos apreciar que la función doble Zipf (m_5) tiene buenos ajustes para algunos casos [Figura 3.5](#). Algo curioso que se puede observar de las distribuciones de rango de los casos de NASCAR-W, NASCAR-B y ESE es un cambio muy marcado en los valores de los puntajes en ciertos valores del rango, como si en efecto, la forma funcional de los mismos no fuera la misma en todo el dominio, es por ello que posiblemente el modelo m_5 pueda ser el adecuado para estos casos. Eso es interesante, pues recordemos que la ley doble Zipf implica la existencia de dos regímenes de elementos

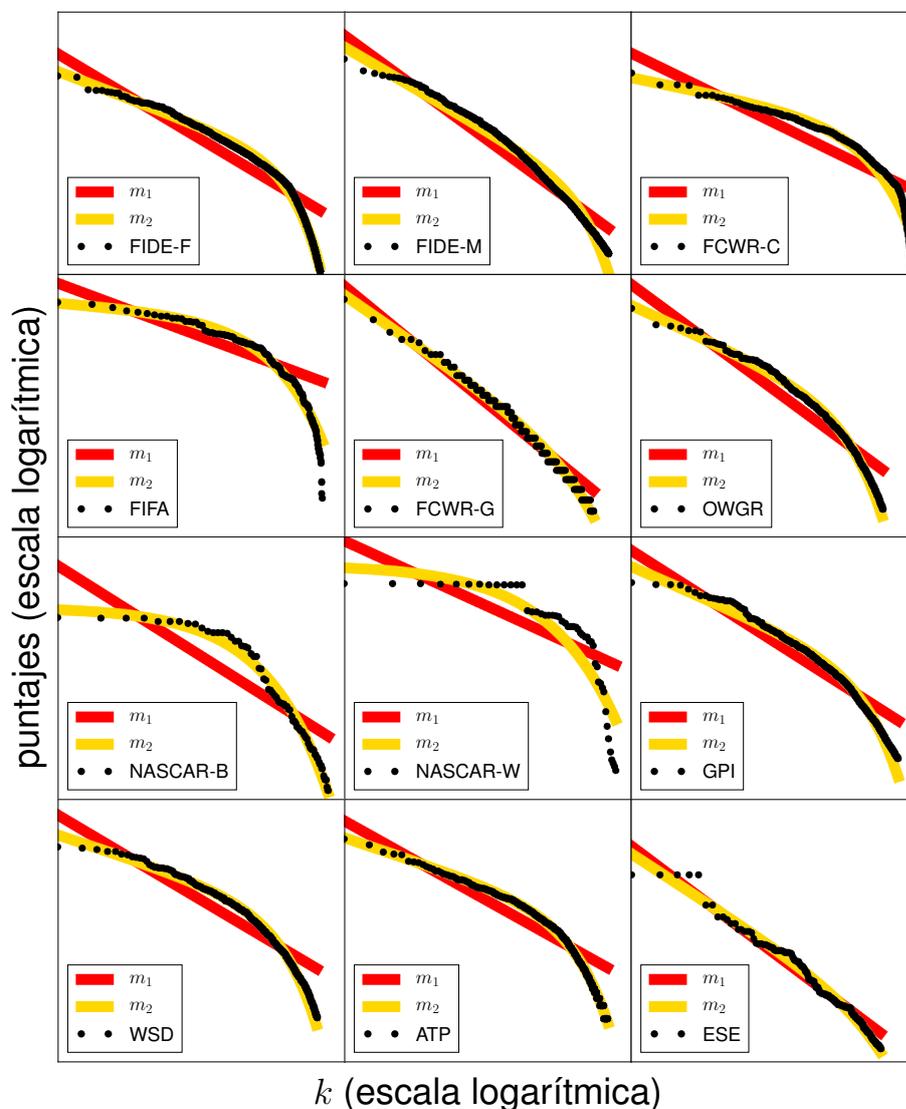


Figura 3.3: Comparación de los datos de ranking con los modelos m_1 y m_2 . Gráfica que muestra la distribución de rango (puntaje contra rango k) para todos los deportes considerados aquí, en una rodaja temporal, (las mismas rodajas mencionadas en la [Figura 3.1](#)) así como los ajustes a los modelos correspondientes a las ecuaciones 3.9 y 3.10. En general, la ley de Zipf (m_1) no reproduce de manera satisfactoria los datos para ningún deporte o juego estudiado aquí. La distribución Gamma (m_2) parece ser más apropiadas en algunos casos. Sólo se presentan las gráficas de ajustes para una rodaja temporal en cada base de datos para ilustrar la forma funcional de los modelos y evidenciar algunas diferencias en sus comportamientos.

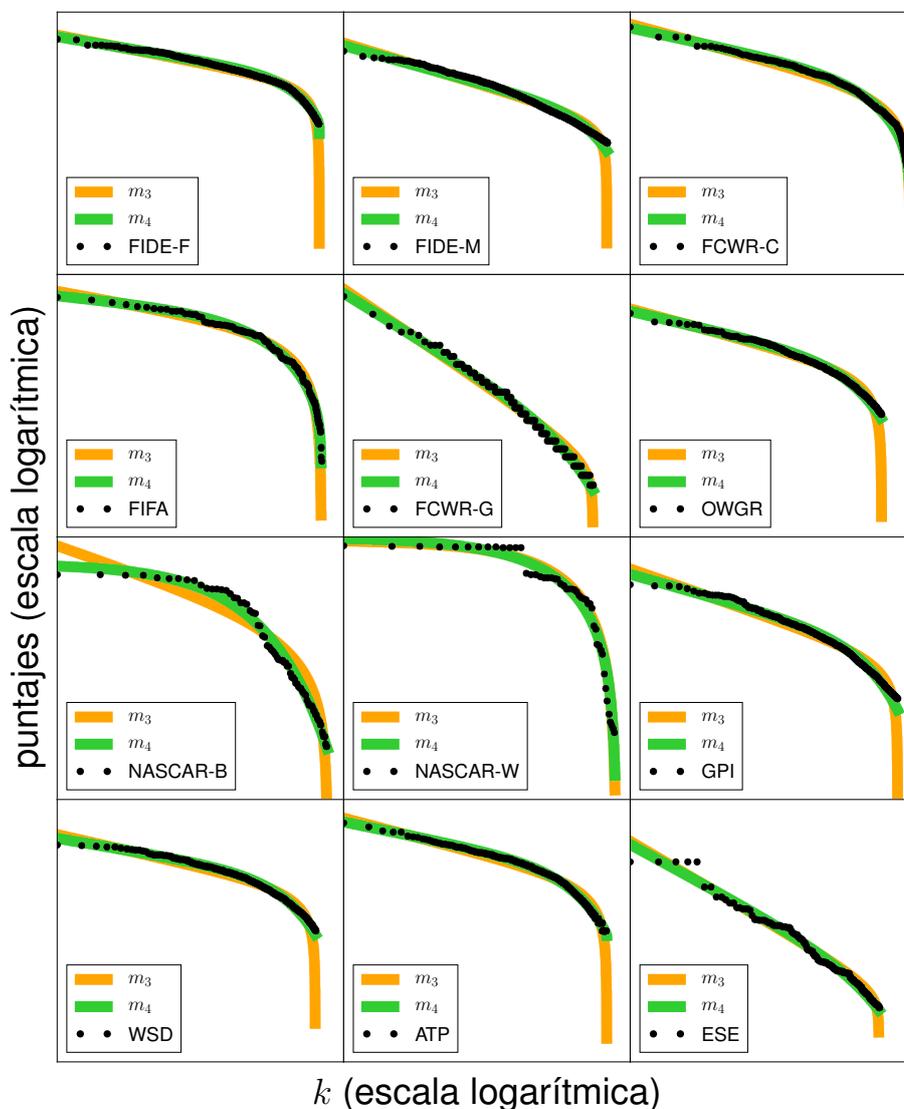


Figura 3.4: Comparación de los datos de ranking con los modelos m_3 y m_4 . Gráfica que muestra la distribución de rango (puntaje contra rango k) para todos los deportes aquí considerados, en una rodaja temporal, (las mismas rodajas mencionadas en la Figura 3.1) así como los ajustes a los modelos correspondientes a las ecuaciones 3.11 y 3.12. Observamos que el modelo m_3 sufre de caídas abruptas cuando se adquieren valores grandes en k , las cuales tal vez dificultan el parecido con la distribución empírica de rango, sin embargo, esto sólo se ve para las rodajas temporales aquí graficadas. El modelo m_4 parece ser el más adecuado en la mayoría de los casos como podemos observar, y esto se puede deber a los grados de libertad adquiridos por mayor número de parámetros.

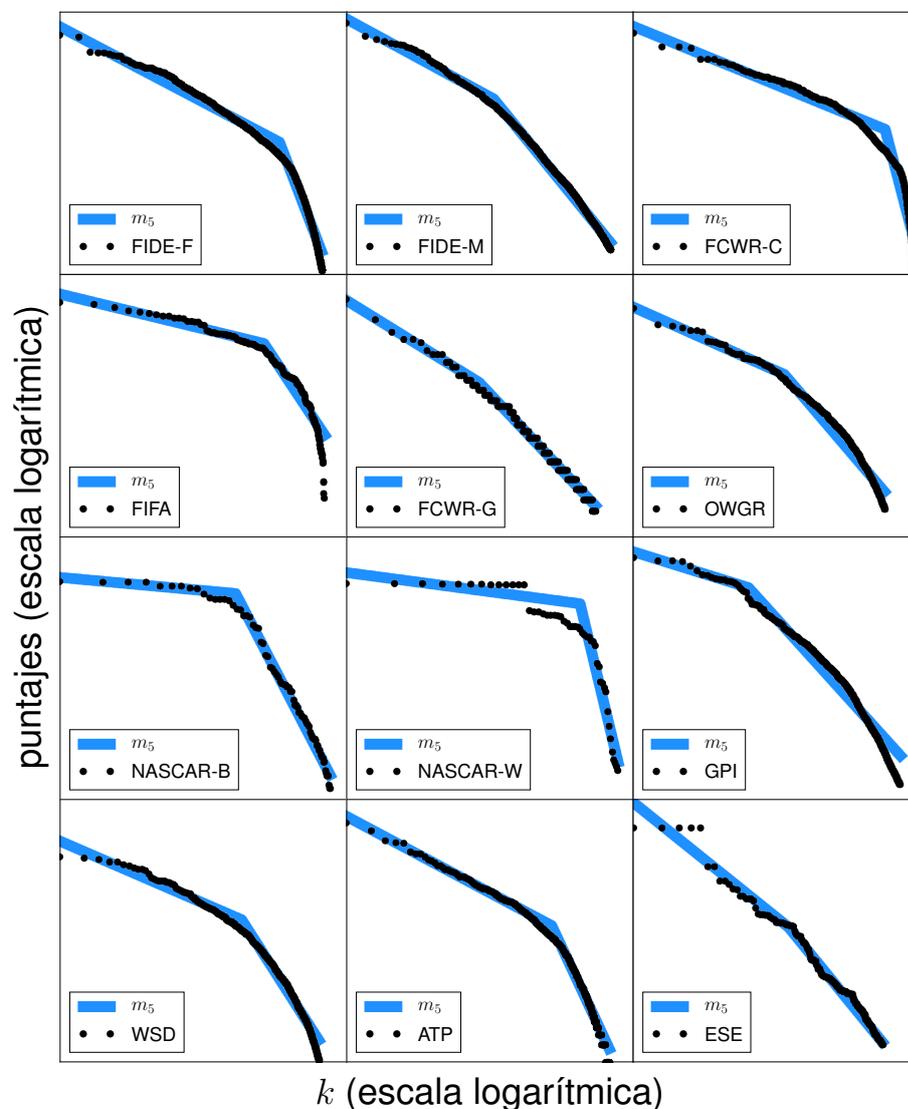


Figura 3.5: Comparación de los datos de ranking con el modelo m_5 . Gráfica que muestra la distribución de rango (puntuaje contra rango k) para todos los deportes aquí considerados en una rodaja temporal, (las mismas rodajas mencionadas en la [Figura 3.1](#)) así como los ajustes al modelo correspondiente a la [Ecuación 3.13](#). Podemos apreciar que en muchos casos el modelo parece reproducir bien los datos. Los casos de NASCAR-B, NASCAR-W y ESE son interesantes, pues parece ser que hay cambios bruscos en el comportamiento funcional de los puntajes a partir de cierto rango, pudiéndose dar el caso de que la ley doble Zipf pueda ser la adecuada para diferenciar dos regímenes en el ranqueo para ciertos deportes o juegos.

3.5 Comparación entre modelos y distribución de rango.

	Modelo m_1		Modelo m_2			Modelo m_3		
	log N	a	log N	a	b	log N	a	q
FIDE-F	3.452	4.33×10^{-2}	3.432	2.78×10^{-2}	1.94×10^{-5}	3.021	3.3×10^{-2}	0.102
FIDE-M	3.468	2.49×10^{-2}	3.461	1.99×10^{-2}	6.27×10^{-6}	3.327	2.19×10^{-2}	3.3×10^{-2}
FCWR-C	4.522	0.529	4.242	0.219	3.06×10^{-3}	2.191	0.341	0.733
FIFA	3.418	0.407	3.229	8.37×10^{-2}	1.29×10^{-2}	1.29	0.235	0.875
FCWR-G	1.764	0.239	1.736	0.2	8.77×10^{-4}	1.527	0.221	8.54×10^{-2}
OWGR	1.388	0.691	1.145	0.429	2.19×10^{-3}	-1.958	0.511	1.035
NASCAR-B	3.673	1.031	3.185	6.1×10^{-10}	6.37×10^{-2}	1.577	0.56	0.965
NASCAR-W	3.943	0.964	3.627	1.58×10^{-10}	9.48×10^{-2}	0.137	5.54×10^{-8}	1.943
GPI	3.652	0.178	3.598	0.122	4.13×10^{-4}	2.902	0.152	0.222
WSD	3.323	0.554	3.094	0.31	1.94×10^{-3}	0.228	0.423	0.942
ATP	4.511	1.042	4.117	0.626	3.18×10^{-3}	-1.467	0.817	1.795
ESE	6.551	0.679	6.485	0.589	2.02×10^{-3}	6.001	0.636	0.198

	Modelo m_4				Modelo m_5			
	log N	a	b	q	log N	a	a'	log k_c
FIDE-F	3.392	2.81×10^{-2}	9.83×10^{-3}	1.79×10^{-5}	3.436	3.19×10^{-2}	0.158	2.8×10^3
FIDE-M	3.461	1.99×10^{-2}	6.66×10^{-13}	6.27×10^{-6}	3.457	1.58×10^{-2}	3.59×10^{-2}	2.03×10^2
FCWR-C	2.937	0.27	0.458	1.4×10^{-3}	4.357	0.371	3.472	4.27×10^2
FIFA	2.338	0.132	0.397	7.83×10^{-3}	3.308	0.263	1.64	59.585
FCWR-G	1.736	0.2	3.88×10^{-10}	8.77×10^{-4}	1.721	0.18	0.309	24.494
OWGR	1.145	0.429	1.19×10^{-9}	2.19×10^{-3}	1.125	0.416	1.094	69.565
NASCAR-B	3.185	3.5×10^{-11}	5.86×10^{-8}	6.37×10^{-2}	3.112	0.125	2.626	16.795
NASCAR-W	0.685	2.06×10^{-9}	1.644	1.65×10^{-2}	3.524	0.284	8.886	28.962
GPI	3.598	0.122	6.68×10^{-12}	4.13×10^{-4}	3.561	6.8×10^{-2}	0.242	25.385
WSD	3.094	0.31	2.29×10^{-8}	1.94×10^{-3}	3.104	0.338	1.221	1.64×10^2
ATP	4.018	0.628	3.12×10^{-2}	3.14×10^{-3}	4.2	0.747	3.004	3.19×10^2
ESE	6.485	0.589	3.7×10^{-9}	2.02×10^{-3}	6.474	0.583	0.89	42.08

Tabla 3.1: Parámetros de ajuste de los 5 modelos m_i a las DRe de los 12 deportes y juegos correspondientes a las últimas fechas disponibles en las bases de datos. FIDE-F(Abril 2016), FIDE-M(Abril 2016), FCWR-C(Semana 53 del 2014), FIFA(Junio 2017), FCWR-G(Semana 33 de 2017), OWGR(21/05/2017), NASCAR-B(2015), NASCAR-W(2013), GPI(31/05/2017), WSD(26/03/2018), ATP(27/12/2010), ESE(2016)

que interactúan de manera distinta. Los parámetros de ajuste de cada uno de los modelos a los datos se presentan de manera detallada en la [Tabla 3.1](#).

Sin embargo, estos comentarios son simplemente un análisis cualitativo de lo observado para los ajustes de los modelos a una sola rodaja temporal, que no necesariamente es representativa del resto con las que se cuentan. Por ello, he dedicado gran parte del capítulo a describir los detalles de la definición de la distribución de rango para el caso de deportes y juegos, así como en introducir algunas cantidades estadísticas que proporcionan un criterio adecuado para determinar si un modelo es el adecuado al momento de describir un conjunto de datos. Me gustaría poder analizar todas las rodajas temporales para todas las bases de datos, comparando en cada una de ellas los modelos que aquí propongo para trabajar. Eso es justamente lo que intentaré hacer a continuación.

3.6. Bondades de ajuste para los modelos de distribución de rango a lo largo del tiempo.

Justo como mencioné, me gustaría poder hacer un análisis objetivo (cuantitativo) de qué tan buenos son los modelos reproduciendo las distribuciones de rango empíricas de los datos y no sólo en una rodaja temporal para cada uno de los deportes, si no en todos los tiempos con los cuales contamos para cada deporte o juego. En la [Sección 3.4](#) introduje 3 cantidades que son de mucha ayuda para concluir si cierta función modelo es adecuado o no: la norma R^2 , la estadística de Kolmogorov D y el índice p de Kolmogorov-Smirnov; de igual forma, la definición, significado y valores óptimos están descritos a detalle en la [Sección 3.4](#).

La [Tabla 3.2](#) y la [Tabla 3.3](#) muestran los valores esperados $\langle R^2 \rangle$, $\langle D \rangle$ y $\langle p \rangle$ (y sus respectivas desviaciones estándar σ_D , σ_{R^2} y σ_p), promediadas a lo largo de todas las rodajas temporales disponibles, para el proceso de ajuste entre las seis bases de datos y los cinco modelos m_i utilizados aquí. Recordemos que si $\langle R^2 \rangle$ es lo más grande posible, $\langle D \rangle$ es lo más pequeño posible implican un mejor ajuste, mientras que si $\langle p \rangle \geq 0.1$ implica que los datos sí se distribuyen, en promedio, de acuerdo al modelo m_i en cuestión. Ahora, para tener un análisis más ordenado, haremos comentarios específicos de los resultados en cada una de las disciplinas en cuestión.

- **FIDE-F:** Para el caso de los rankings de las ajedrecistas tenemos que con la $\langle R^2 \rangle$ se concluye que el mejor de los modelos es m_4 , mientras que para el caso de la $\langle D \rangle$, de igual manera el mejor modelo parece ser m_4 , adicionalmente sus respectivas desviaciones estándar tienen valores considerablemente pequeños respecto al resto de los modelos. Sin embargo, la $\langle p \rangle$ no permite concluir que los puntajes se distribuyan conforme a un modelo, pues su valor nunca fue ≥ 0.1 , por ende concluimos que el modelo más adecuado (sólo en forma funcional) es el modelo m_4 .
- **FIDE-M:** En el caso de los rankings de los ajedrecistas observamos que curiosamente el modelo m_5 es el más adecuado pues para ese modelo $\langle R^2 \rangle$ tiene el valor más alto respecto al resto de los modelos y lo mismo para $\langle D \rangle$ tiene el valor más bajo. Adicionalmente, las desviaciones estándar para estas bondades de ajuste son más pequeñas respecto al resto. Sin embargo, $\langle p \rangle$ no permite que concluir que los puntajes, en promedio, se distribuyan conforme a uno de los modelos aquí trabajados, pues en todos los casos $\langle p \rangle = 0$.
- **FCWR-C:** Los equipos clubes de fútbol tienen un comportamiento similar a FIDE-F, pues $\langle R^2 \rangle$ y $\langle D \rangle$ indican que el modelo m_4 es el que mejor reproduce teóricamente los valores empíricos de las distribuciones de rango. Nuevamente $\langle p \rangle$ indica que los puntajes, en promedio, no se distribuyen respecto a alguno de los modelos aquí propuestos.
- **FIFA:** En este caso ocurre algo interesante. Las bondades de ajuste $\langle R^2 \rangle$ y $\langle D \rangle$ indican que el modelo m_4 es el más adecuado. La bondad de ajuste $\langle p \rangle$ arroja

3.6 Bondades de ajuste para los modelos de distribución de rango a lo largo del tiempo.

conclusiones más interesantes, pues tanto el modelo m_3 como el modelo el modelo m_4 son adecuados para describir la forma en que se distribuyen los puntajes para este deporte a lo largo del tiempo. Ahora bien, es importante notar que $\langle p \rangle$ es mayor para m_4 que para m_3 . Esta observación coincide con la intuición generada en la sección anterior, pues vimos que los puntajes de FIFA (al menos en la rodaja temporal mostrada) tienen una caída que no puede ser modelada por el modelo m_2 , el modelo m_3 tiene inherente una caída en su forma teórica, sin embargo, los datos empíricos no tienen una caída tan pronunciada como se aprecia en la [Figura 3.4](#), entonces un modelo intermedio (una caída controlada) debería describir mejor a los datos, y es justamente lo que apreciamos, el modelo m_4 tiene incluido m_2 , el cual controla la caída abrupta en los rangos más altos y justamente m_4 el que tiene la $\langle p \rangle$ más alta de entre los modelos, como ya habíamos observado.

- **FCWR-G:** Para los goleadores de fútbol ocurre algo aún más interesante, pues al parecer 3 de los 5 modelos parecen ser adecuados. La bondad de ajuste $\langle R^2 \rangle$ arroja el mismo valor para m_2 , m_4 y m_5 , pero para el caso de $\langle D \rangle$ se concluye que los mejores modelos son sólo m_2 y m_4 de los cuales se obtiene el mismo valor. Sin embargo, lo más interesante es que la bonda de ajuste $\langle p \rangle$ indica que, en promedio, los puntajes sí se distribuyen de acuerdo a m_2 , m_3 y m_5 . Debido a que $\langle p \rangle$ es mucho mayor en los casos de m_2 y m_4 que para m_5 , se concluye que esos dos modelos son mejores que la ley doble Zipf. Sin embargo, la prueba de Kolmogorov-Smirnov nos indica que los datos sí parecen distribuirse de acuerdo a la ley doble Zipf a lo largo del tiempo. Recordemos que anteriormente describí que la interpretación de esta ley es que existen dos regímenes de los elementos, en este caso los goleadores, uno que son los goleadores con los mayores puntajes que no se ven afectados por la aparición de nuevos goleadores y un segundo régimen con menores puntajes que sí incluyen la aparición de nuevos jugadores o goleadores; ésto es muy interesante porque tenemos una primera descripción analítica de un fenómeno de este tipo para un sistema complejo. Este resultado parece coincidir con una pequeña intuición de la forma en que el sistema se comporta, pues los mejores goleadores siempre parecen ser los mismos y casi no se ven afectados por la incursión de novatos en los distintos clubes de fútbol.
- **OWGR:** Este es el caso de los golfistas. Vemos que con la bondad de ajuste $\langle R^2 \rangle$ se concluye que los modelos adecuados son tanto m_2 y m_4 pues comparten valores para dicha bondad de ajuste, el caso de $\langle D \rangle$ tenemos que sólo un modelo es adecuado y corresponde a m_4 . Sin embargo, con el criterio de Komogorov-Smirnov, tenemos que el promedio del índice $\langle p \rangle$ se tienen dos modelos adecuados, tanto m_2 como m_4 . Coincide con nuestras observaciones cualitativas de la sección anterior, donde notamos que la caída de los puntajes a rangos altos no es tan pronunciada y predijimos que m_3 no sería un modelo adecuado en este caso.
- **NASCAR-B:** Aquí tenemos el primer caso donde se nota que las bondades $\langle R^2 \rangle$ y $\langle p \rangle$ son completamente exclusivas con interpretaciones independientes. Mientras que con $\langle R^2 \rangle$ se concluye que el modelo m_5 es el mejor a lo largo del tiempo, con

$\langle D \rangle$ el mejor modelo es m_4 . Ahora bien, $\langle p \rangle$ indica que, en promedio, los modelos m_2 , m_4 y m_5 describen mejor la forma en que se distribuyen los puntajes, viendo que los valores de $\langle p \rangle$ para m_2 y m_4 son mayores que para m_5 . Ahora bien, es importante notar que las desviaciones estándar del índice de Kolmogorov son pequeñas, por lo que lo observado en las figuras 3.3-3.5 sí es representativo, y se observa de los valores empíricos de la distribución de rango que a cierto valor de k hay un cambio de la forma funcional de los puntajes, cualitativamente se puede predecir que la ley doble Zipf podría llegar a ser importante en la descripción teórica para este sistema, y de hecho, observamos que el índice de Kolmogorov promediado nos permite concluir que sí, observando nuevamente la existencia de dos regímenes en los corredores de NASCAR-B. A pesar de que la $\langle R^2 \rangle$ indica que m_5 es el mejor modelo, el coeficiente de Kolmogorov indica que en realidad los datos se distribuyen de acuerdo a lo que indica m_2 o m_4 sobre lo indicado por m_5 , reafirmando que ambas bondades de ajuste tienen distintas interpretaciones.

- **NASCAR-W**: Éste debería ser un sistema muy similar al anterior, pues tienen criterios similares al momento de hacer el ordenamiento de los corredores; se ve que no es el caso. La bondad de ajuste $\langle R^2 \rangle$ indica que m_3 es el que mejor describe al sistema, mientras que para la bondad $\langle D \rangle$ el modelo m_4 es el más adecuado. Ahora bien, $\langle p \rangle$ nos permite concluir que, en promedio, los modelos m_3 , m_4 y m_5 son los mejores para reproducir la forma en que los puntajes se distribuyen, contrario a lo obtenido en NASCAR-B, donde el modelo m_2 era mejor. Esto quiere decir que los puntajes tienen una caída más abrupta para rangos altos, tal y como lo indica el comportamiento funcional de m_3 . Pero en este sistema se tiene un resultado sorprendente por no haberse visto en el resto de los sistemas anteriores: el modelo m_5 funciona mejor sobre m_3 y m_4 , evidenciando la existencia de dos regímenes en los elementos (corredores), y de hecho, es consistente con el análisis cualitativo que hicimos en la sección anterior. La desviación estándar para p es la más pequeña en m_5 respecto a los otros modelos, mostrando que la muestra observada en la Figura 3.5 es representativa; ya habíamos concluido que el comportamiento empírico de los datos allí graficados mostraba un cambio abrupto en la forma funcional de los puntajes desde cierto rango, lo cual nos permitía predecir que la ley doble Zipf dominaría en este caso.
- **GPI**: Los resultados para este sistema, los jugadores de póquer, no son tan significativos. La bondad $\langle R^2 \rangle$ indica que el modelo m_5 es el que mejor describe la forma de las curvas inducidas por los datos empíricos de puntajes, mientras que la bondad $\langle D \rangle$ nos dice que el modelo m_3 es el más adecuado. Desafortunadamente, el criterio de Kolmogorov-Smirnov indica que ninguno de los modelos aquí trabajados describe satisfactoriamente la forma en que los puntajes se distribuyen a lo largo del tiempo. No podemos hacer mayores conclusiones más lejos de esto.
- **WSD**: En el caso de los jugadores de tabla sobre nieve se obtiene el valor más alto de $\langle R^2 \rangle$ tanto para el modelo m_2 y el modelo m_4 , mientras que con $\langle D \rangle$ observamos que el modelo que menos distancia tiene hacia los datos es m_4 . De

3.6 Bondades de ajuste para los modelos de distribución de rango a lo largo del tiempo.

igual manera, el criterio de Kolmogorov-Smirnov indica que tanto el modelo m_2 como el modelo m_4 son adecuados. Ésto de nuevo va en concordancia a lo visto en la [Figura 3.3](#) y la [Figura 3.4](#), pues se nota que la caída de los puntajes conforme los rangos aumentan es suave y no abrupta como lo describiría el modelo m_3 .

- **ATP:** Del mismo modo, para los jugadores de tenis, tenemos que $\langle R^2 \rangle$ indica que tanto m_2 y m_4 son los mejores modelos que describen la forma funcional de la distribución de rango empírica. Por otro lado, $\langle D \rangle$ nos hace ver que el modelo m_4 es el que en promedio tiene menor distancia con los datos empíricos. Finalmente, $\langle p \rangle$ permite concluir que tanto m_2 , m_4 y m_5 describen adecuadamente la forma en que los puntajes se distribuyen, en promedio, a lo largo del tiempo, de estos tres, m_4 es el mejor por tener el valor más grande de $\langle p \rangle$, y coincide con el hecho de que $\langle D \rangle$ dice que es el modelo con menor distancia a los datos empíricos. Hay que notar que el modelo m_5 es también buen modelo de acuerdo al criterio de Kolmogorov-Smirnov y evidencia la existencia, como en algunos sistemas anteriores, de dos regímenes de jugadores.
- **ESE:** Para el caso de los jugadores de videojuegos tenemos que $\langle R^2 \rangle$ indica que el modelo m_5 es el mejor al momento de describir la forma funcional de la distribución de rango empírica, mientras que $\langle D \rangle$ dice que el modelo m_2 es el que menos dista de los datos, pudiendo indicar que la caída de los puntajes conforme aumenta k es suave. Pero el promedio de p , $\langle p \rangle$ permite concluir que el modelo m_5 es el que mejor describe la manera en que los puntajes son distribuidos a lo largo del tiempo, de nueva cuenta, se tiene que hay dos regímenes en los jugadores que conforman al sistema.

Ahora que analizamos en detalle cada uno de los sistemas, se pueden concluir cosas muy interesantes. Se observa de manera contundente que la ley de Zipfm, m_1 nunca fue el mejor modelo, de hecho, se observa que siempre obtiene los peores valores en las bondades de ajuste. Es curioso que la ley de Zipf que es utilizada y aceptada en muchos ámbitos para la distribución de frecuencias de palabras en idiomas no funcione para los sistemas analizados en este trabajo. Ésto debe indicar que la naturaleza jerárquica de estos sistemas tiene diferencias con la forma en que las palabras son rankeadas a lo largo del tiempo.

Otro comportamiento interesante, y esperado diría yo, es que siguiendo el criterio de Kolmogorov-Smirnov, siempre que m_2 o m_3 eran modelos adecuados, también lo era el modelo m_4 , siendo consistente con la naturaleza de m_4 que es una combinación de ambos modelos m_2 y m_3 . Sin embargo, nunca ocurrió que los modelos m_2 y m_3 funcionaran al mismo tiempo, marcando otra consistencia pues ambas tienen distinta naturaleza, ya habíamos discutido que el modelo m_3 se caracterizaba por tener una caída abrupta mientras el valor de k se incrementa, m_2 no presenta esta característica, entonces ambas eran necesarias para describir posibles diferencias entre deportes/juegos en cuanto este comportamiento. El modelo m_4 representó siempre un punto medio entre ambos modelos, suaviza la caída abrupta en los casos que m_3 funcione.

3. DISTRIBUCIÓN DE RANGO. UN ANÁLISIS ESTADÍSTICO.

Algo inesperado en muchos de los casos fue que la ley doble Zipf funcionara como en el caso de FCWR-G, NASCAR-B, NASCAR-W, ATP y ESE, casi la mitad de las bases de datos. El caso de NASCAR-B y NASCAR-W no fue tan sorprendente, pues como vimos en la sección anterior y en las figuras 3.3-3.5, sí se apreciaba un cambio evidente en la forma funcional de los datos empíricos para la distribución de rango. Y como mencioné en el análisis para cada uno de los sistemas, la implicación más importante del funcionamiento de este modelo es que existen al menos dos regímenes de los jugadores/equipos que son independientes entre sí, en el sentido de que los elementos con los rangos más bajos y puntajes más altos no se ven afectados por la influencia de nuevos elementos que puedan estar ingresando a los rankings, mientras que el segundo régimen o los jugadores/equipos con los rangos más altos o puntajes más bajos sí se ven influenciados por la posible entrada de nuevos elementos.

Un aspecto más a considerar es el hecho de que las bondades de ajuste R^2 y p arrojan distintas conclusiones respecto a qué modelo es mejor. Por ejemplo, tuvimos los casos en que para algunas disciplinas, $\langle R^2 \rangle$ indicaba que ciertos modelos eran los mejores, mientras que $\langle p \rangle$ permitía concluir que en realidad los puntajes encargados de hacer el ordenamiento de los elementos no se distribuyen de acuerdo a esos modelos; pudimos apreciar que ninguno de los modelos funcionaba en los casos de FIDE-F, FIDE-M, FCWR-C y GPI. Recordemos que estas dos cantidades conllevan a criterios que no provienen del mismo concepto. El caso de la norma R^2 únicamente describe el parecido entre los puntos empíricos de la distribución de rango y los modelos ajustados a dichos puntos; sin embargo, el índice p es un poco más sensible, pues esta cantidad indica, en nuestro caso, si los puntajes se distribuyen de cierta forma, es decir, si se distribuyen conforme indica cierta distribución de probabilidad modelo. (ver Apéndice A) Por lo tanto, estas dos bondades de ajuste conducen a distintas conclusiones en cuanto a ver si un modelo es "bueno." "maloz se debe a dos factores: la cantidad de datos disponibles y el número de parámetros en el modelo. Entre más grande sea la cantidad de datos, más fácil es distinguir el mejor modelo en una buena aproximación. Por otro lado, entre más parámetros tenga el modelo (como m_4), será más fácil ajustarlo a cualquier conjunto de datos. Estos dos aspectos están considerados en la definición de p , pero no de R^2 .

La conclusión más importante de este capítulo es que un modelo (de los que aquí presentamos) no es capaz de describir de manera genérica a todos los sistemas aquí estudiados, sin embargo, esa no universalidad nos permitió identificar claras diferencias entre todos los sistemas y de qué naturaleza son. Pero notemos que existe una debilidad muy clara en la distribución de rango: ésta no logra estudiar describir la evolución del sistema respecto al tiempo, de hecho, es ciega respecto al tiempo pues es una función que se construye con los puntajes para una sola rodaja temporal.

Como el principal objetivo de este trabajo es estudiar la dinámica de cómo los rankings van cambiando o evolucionando respecto al tiempo, en el siguiente capítulo daremos ese paso a encontrar formas de ver al sistema con tiempo variable.

3.6 Bondades de ajuste para los modelos de distribución de rango a lo largo del tiempo.

		m_1	m_2	m_3	m_4	m_5
FIDE-F	$\langle R^2 \rangle$	0.445	0.981	0.89	0.995	0.965
	$\langle D \rangle$	0.495	0.086	0.081	0.036	0.133
	$\langle p \rangle$	0.0	0.0	0.0	0.0	0.03
	σ_{R^2}	0.047	0.016	0.031	0.002	0.011
	σ_D	0.003	0.027	0.019	0.002	0.004
	σ_p	0.0	0.0	0.0	0.0	0.004
	FIDE-M	$\langle R^2 \rangle$	0.778	0.936	0.657	0.936
$\langle D \rangle$		0.477	0.201	0.189	0.201	0.141
$\langle p \rangle$		0.0	0.0	0.0	0.0	0.0
σ_{R^2}		0.007	0.005	0.017	0.005	0.001
σ_D		0.007	0.005	0.003	0.005	0.004
σ_p		0.0	0.0	0.0	0.0	0.001
FCWR-C		$\langle R^2 \rangle$	0.727	0.987	0.982	0.997
	$\langle D \rangle$	0.295	0.115	0.057	0.056	0.172
	$\langle p \rangle$	0.0	0.0	0.01	0.0	0.0
	σ_{R^2}	0.028	0.005	0.005	0.001	0.01
	σ_D	0.019	0.018	0.01	0.011	0.027
	σ_p	0.002	0.0	0.017	0.006	0.001
	FIFA	$\langle R^2 \rangle$	0.763	0.987	0.993	0.997
$\langle D \rangle$		0.44	0.107	0.046	0.042	0.114
$\langle p \rangle$		0.0	0.01	0.41	0.48	0.07
σ_{R^2}		0.021	0.004	0.004	0.001	0.005
σ_D		0.018	0.014	0.013	0.015	0.011
σ_p		0.0	0.012	0.32	0.351	0.038
FCWR-G		$\langle R^2 \rangle$	0.973	0.992	0.985	0.992
	$\langle D \rangle$	0.265	0.101	0.137	0.101	0.152
	$\langle p \rangle$	0.04	0.75	0.03	0.74	0.31
	σ_{R^2}	0.013	0.003	0.006	0.003	0.002
	σ_D	0.059	0.021	0.015	0.021	0.033
	σ_p	0.071	0.234	0.052	0.229	0.206
	OWGR	$\langle R^2 \rangle$	0.65	0.985	0.971	0.985
$\langle D \rangle$		0.377	0.033	0.1	0.031	0.093
$\langle p \rangle$		0.0	0.61	0.0	0.52	0.09
σ_{R^2}		0.132	0.014	0.011	0.012	0.009
σ_D		0.019	0.007	0.02	0.008	0.027
σ_p		0.0	0.242	0.0	0.337	0.157

Tabla 3.2: Promedios y desviaciones estándar las bondades de ajuste R^2 , p y la estadística de Kolmogorov D para los ajustes realizados a todas las fechas con las que cuentan las bases de datos de los deportes y juegos: FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G, OWGR

3. DISTRIBUCIÓN DE RANGO. UN ANÁLISIS ESTADÍSTICO.

		m_1	m_2	m_3	m_4	m_5
NASCAR-B	$\langle R^2 \rangle$	0.392	0.971	0.853	0.974	0.976
	$\langle D \rangle$	0.323	0.082	0.154	0.078	0.089
	$\langle p \rangle$	0.0	0.33	0.0	0.29	0.25
	σ_{R^2}	0.218	0.017	0.068	0.015	0.013
	σ_D	0.03	0.024	0.022	0.021	0.024
	σ_p	0.002	0.293	0.025	0.293	0.226
NASCAR-W	$\langle R^2 \rangle$	0.469	0.845	0.952	0.95	0.932
	$\langle D \rangle$	0.322	0.213	0.136	0.133	0.208
	$\langle p \rangle$	0.02	0.0	0.15	0.12	0.18
	σ_{R^2}	0.162	0.068	0.022	0.025	0.066
	σ_D	0.043	0.047	0.034	0.035	0.08
	σ_p	0.024	0.018	0.24	0.192	0.16
GPI	$\langle R^2 \rangle$	0.796	0.977	0.938	0.977	0.982
	$\langle D \rangle$	0.521	0.194	0.139	0.194	0.224
	$\langle p \rangle$	0.0	0.0	0.0	0.0	0.0
	σ_{R^2}	0.014	0.005	0.005	0.005	0.007
	σ_D	0.011	0.014	0.005	0.014	0.049
	σ_p	0.0	0.0	0.0	0.0	0.001
WSD	$\langle R^2 \rangle$	0.799	0.989	0.954	0.989	0.972
	$\langle D \rangle$	0.474	0.058	0.115	0.058	0.172
	$\langle p \rangle$	0.0	0.11	0.0	0.11	0.0
	σ_{R^2}	0.016	0.003	0.006	0.003	0.004
	σ_D	0.009	0.019	0.008	0.019	0.04
	σ_p	0.0	0.148	0.0	0.154	0.0
ATP	$\langle R^2 \rangle$	0.272	0.982	0.88	0.982	0.965
	$\langle D \rangle$	0.434	0.044	0.08	0.039	0.077
	$\langle p \rangle$	0.0	0.43	0.0	0.59	0.1
	σ_{R^2}	0.211	0.013	0.067	0.013	0.029
	σ_D	0.097	0.017	0.009	0.012	0.029
	σ_p	0.003	0.232	0.0	0.411	0.11
ESE	$\langle R^2 \rangle$	0.826	0.94	0.903	0.919	0.965
	$\langle D \rangle$	0.36	0.082	0.135	0.097	0.097
	$\langle p \rangle$	0.01	0.07	0.02	0.06	0.16
	σ_{R^2}	0.104	0.043	0.06	0.105	0.037
	σ_D	0.082	0.024	0.043	0.061	0.054
	σ_p	0.026	0.11	0.056	0.112	0.243

Tabla 3.3: Promedios y desviaciones estándar las bondades de ajuste R^2 , p y la estadística de Kolmogorov D para los ajustes realizados a todas las fechas con las que cuentan las bases de datos de los deportes y juegos: NASCAR-B, NASCAR-W, GPI, WSD, ATP, ESE

Dinámica de rango para deportes y juegos: diversidad, probabilidad de cambio, entropía y complejidad.

Ya vimos en el capítulo anterior que la distribución de rango intenta hacer una descripción matemática de cómo los puntajes asignados a cada elemento dentro del sistema en cuestión se comportan en una distribución, dando de manera directa origen a un ordenamiento de los elementos. Aunque puede llegar a ser útil esta descripción, y no lo es tanto pues vimos que no hay un comportamiento universal entre las distintas disciplinas; ésta no proporciona un punto de vista dinámico, pues sólo se puede obtener la distribución de rango para una rodaja temporal por separado, es ciega ante la evolución del sistema a lo largo del tiempo. Por ende, el objetivo de este trabajo es buscar una manera de describir la dinámica de cómo los elementos de una disciplina en particular ocupan distintos rangos, que podamos estudiar tal vez, la evolución de ciertos elementos o cómo es que un rango determinado es ocupado a lo largo del tiempo. De igual manera, nos basaremos en el estudio que se ha realizado para el caso de idiomas y que ya he utilizado anteriormente [2]. En ese estudio se da una forma de estudiar la dinámica de rango para las palabras a lo largo de varios años, donde también se pudo definir un ranqueo instantáneo en cada uno de los años en que se tenía disponibilidad. Aquí seguiré algunas de esas ideas.

A lo largo de este capítulo desenmascaremos la forma en que podemos estudiar la dinámica de los rankeos. Primero definiremos el concepto de "espagueti" que son trayectorias en una gráfica de rango contra tiempo y que traza la evolución en el ranking de un jugador/equipo determinado. Después definiremos una medida que cuantifica la variación en la que un rango específico es ocupado por distintos elementos, esta medida se llamará *diversidad de rango*, obtendremos algunas propiedades interesantes de ellas, sus consecuencias, y la posibilidad de encontrar cierta regularidad entre todas las disciplinas; también haré un pequeño análisis comparativo entre esta medida dinámica y la distribución de rango que definimos en el [Capítulo 3](#). Posteriormente, definiremos otras tres medidas dinámicas de los rangos: *probabilidad de cambio*, *entropía de rango* y *complejidad de rango*, de las cuales obtendremos conclusiones adicionales respecto a la dinámica de rango.

4.1. Los Espaguetis.

Un primer intento para estudiar la evolución de los rankings a lo largo del tiempo es tomando un elemento (jugador o equipo) dentro de una disciplina, y desde la primer fecha en la que aparece, ver qué rango va adquiriendo conforme avanzamos en el tiempo y hacer eso para todos los elementos del sistema. Lo que estamos haciendo es ir trazando un camino que conforme avanzamos en el tiempo adquiere distintos rangos. Para ser más precisos con esta idea presentaré un ejemplo muy específico. Por ejemplo, en el caso de los clubes de fútbol, si vemos qué equipos ocupan los primeros 8 lugares entre la semana 9 del 2013 y la semana 14 del 2013, obtendremos que los equipos son: Grêmio, Barcelona, Real Madrid, Ulsan, Bayern München, Corinthians, Atlético Madrid, Chelsea y Manchester United. Entonces, para cada uno de esos equipos graficamos puntos de rango contra fechas y unimos esos puntos para el caso de cada uno de estos equipos, obtenemos lo que se aprecia en la [Figura 4.1](#). Esta gráfica describe básicamente una trayectoria en los rangos que ha asquirido el equipo a lo largo del tiempo.

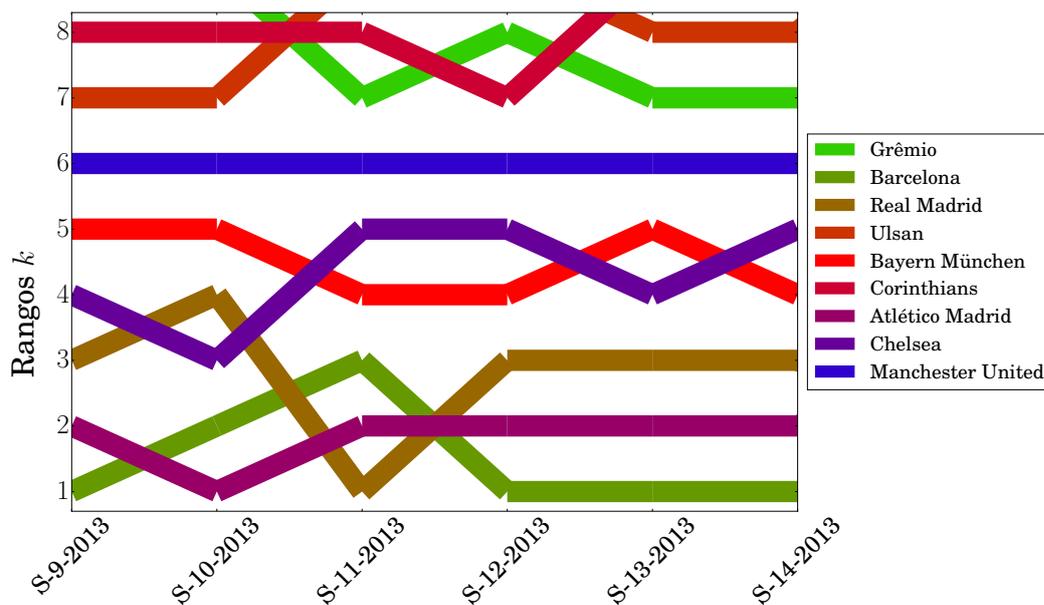


Figura 4.1: Evolución temporal en el ranking para los 8 equipos punteros entre la Semana 9 del 2013 y la semana 14 del 2013. Vemos que 9 diferentes equipos ocupan los primeros ocho rangos en estas fechas. Los equipos que en algún momento ocupan el primer lugar son: Barcelona, Real Madrid y Atlético Madrid. Algo interesante es que en estas fechas, el Manchester United se mantiene siempre en el rango $k = 6$.

Lo que se aprecia en la [Figura 4.1](#) es que los equipos de Barcelona, Real Madrid y Atlético Madrid se disputaron el primer lugar en los ranking publicados estas semanas. El caso curioso es el de Manchester United que se mantuvo en el sexto lugar para las

mismas fechas. Las líneas que representan la evolución temporal para cada uno de los equipos se asemejan mucho a la forma del platillo italiano, el espagueti. Es por ello, que a partir de ahora, llamaré a este tipo de gráficas como los espaguetis para los elementos que conforman un deporte o juego. Con esta idea ya tenemos una primer descripción de la evolución del sistema, pero ¿existe una manera de caracterizarlo o alguna regularidad apreciable a considerar? Además, el ejemplo sólo describe el caso de los clubes de fútbol y en este trabajo tenemos 12 disciplinas para analizar. Intentemos cumplir nuestro objetivo aplicando esta idea a todas nuestras bases de datos.

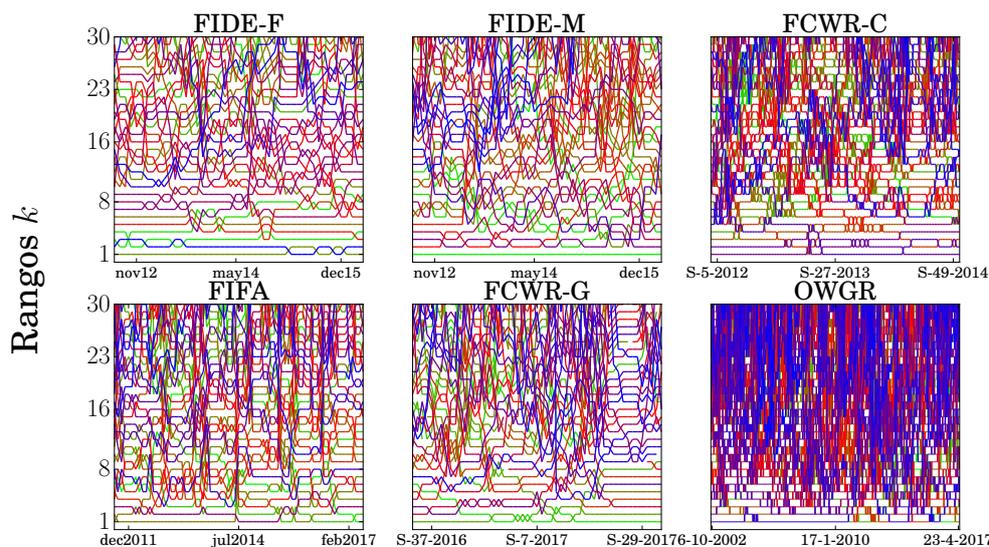
4.1.1. ¿Podemos encontrar regularidades en los espaguetis?

Las figuras 4.2 y 4.3 proporcionan los espaguetis para todos los equipos que alguna vez ocuparon uno de los primeros 30 rangos en todo el dominio temporal que cuenta cada una de las bases de datos. Primero que nada, en la Figura 4.2 tenemos estos espaguetis para el caso de FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G y OWGR. En todos los casos podemos apreciar que los elementos que ocupan los primeros lugares se mantienen estables a lo largo del tiempo. Es decir, no hay mucha variación de lo espaguetis para esos equipos/jugadores, y conforme vemos los casos de los rangos más altos, podemos apreciar mayor variabilidad, esos elementos ocupan rangos distintos en los periodos de tiempo disponibles. Sin embargo, en el caso de los espaguetis presentados en la Figura 4.3 podemos notar que esa regularidad no se aprecia, pues el caso de ESE (el más extremo) se observa que los elementos que ocupan el primer lugar en algún momento sólo se mantienen allí una rodaja temporal. En los casos de WSD y ATP (que son deportes que exigen rendimiento físico) sí se nota más estabilidad para los primeros lugares. Sin embargo, los dos casos de NASCAR se ve prácticamente la misma variabilidad en todas las escalas de rangos, pero notemos que la resolución temporal de los rankings es anual, desde los años 80 hasta la década de 2010, y es evidente que a lo largo de todos esos años, los corredores punteros se retirarán o entrarán nuevos competidores en algún momento.

También de las figuras 4.2 y 4.3 podemos observar que los jugadores y equipos con rangos bajos cambian muy lentamente o nada, mientras que aquellos con rangos más altos tienen una mayor variación de su rango a lo largo del tiempo. Esta intuición es clara de experiencias recientes en el caso de deportes como Tenis y Fútbol: De acuerdo con los conjuntos de datos analizados, Hewitt, Nadal, Roddick, Ferrero, Agassi y Federer han sido los únicos jugadores número uno desde Mayo de 2003 hasta Diciembre de 2010. Lo mismo ocurre en el caso de clubes de Fútbol: Real Madrid, Atlético Madrid, Barcelona, y el Bayern München han sido los equipos mejor rankeados desde Enero de 2012 hasta Diciembre de 2014. En otras palabras, los jugadores y equipos con k pequeñas tienden a tener una diversidad de rango pequeña, salvo las claras excepciones que pudimos apreciar. De estas observaciones comenzamos a darnos cuenta de algo; parece inútil tratar de describir la evolución de un jugador/equipo a lo largo del tiempo. En conjunto, la variación del posicionamiento de un elemento determinado parece ser un completo desaste, impredecible, sobre todo por los rangos más altos donde parece que los elementos pueden moverse hacia rangos que no tienen nada que ver con el que tenían en un inicio. Pero lo que acabamos de mencionar parece ser una regularidad

4. DINÁMICA DE RANGO PARA DEPORTES Y JUEGOS: DIVERSIDAD, PROBABILIDAD DE CAMBIO, ENTROPÍA Y COMPLEJIDAD.

o una característica de la mayoría de los sistemas aquí presentados; tenemos que los elementos con rangos más bajos tienen mayor estabilidad, mientras que los elementos con rangos altos parecen evolucionar adquiriendo rangos muy variados.

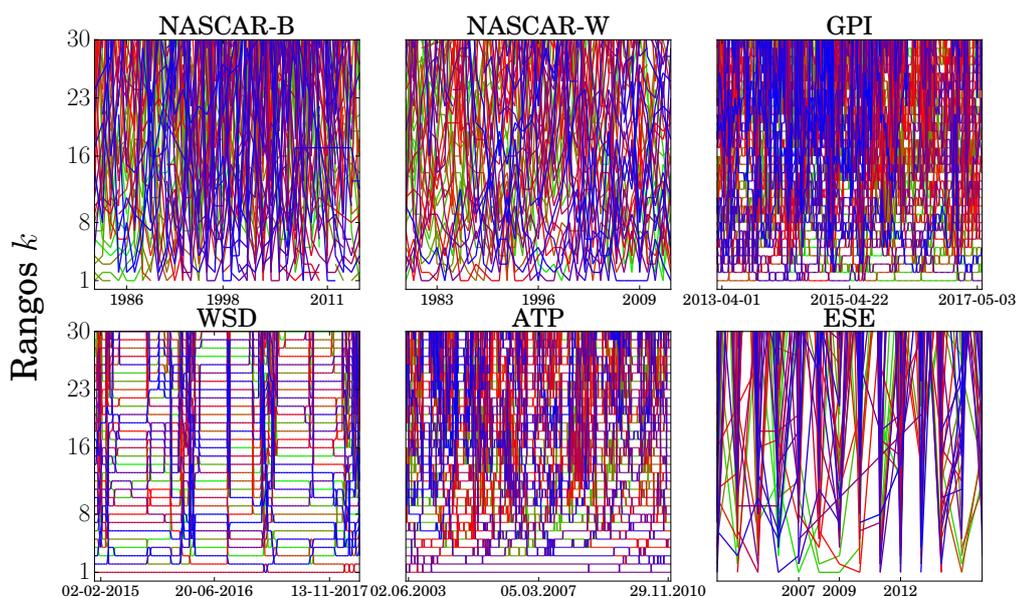


Rodajas temporales

Figura 4.2: Evolución temporal de los rangos para jugadores y equipos de las primeras 6 disciplinas. Gráfica que muestra el cambio en el rango k a lo largo del tiempo t para todos los jugadores/equipos en cada deporte y juego considerado en este estudio: FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G y OWGR. Aquí incluyo los espaguetis de todos los equipos/jugadores que alguna vez ocuparon uno de los primeros 30 lugares a lo largo de todo el tiempo que se tiene disponible. Nótese que los jugadores/equipos en los rangos bajos tienden a cambiar menos que los de rango más altos, incluso cuando los rankeos en todas las actividades varían a diferentes tasas y la resolución temporal correspondiente varía entre semanas a meses (ver [Tabla 2.1](#)).

Intentemos hacer un análisis más exhaustivo de las ideas ya mencionadas. En el [Apéndice B](#) incluyo una serie de gráficas que contienen los espaguetis de todos los equipos que adquieren los rangos $k = 1, N/2, N$, donde N es el número total de jugadores/equipos en el rankeo realizado en cada una de las fechas, o también es el rango máximo que se tiene en los rankings para cada base de datos. Ésto con el objetivo de entender qué pasa en los rangos de 3 diferentes regímenes: el primer lugar, la mitad del rankeo, el final del rankeo o rango máximo. Las figuras [B.1-B.12](#). Es necesario hacer un análisis exhaustivo gráfica por gráfica, el cual realizo en el [Apéndice B](#). Pero de ese análisis exhaustivo podemos comenzar a notar que en vez de estudiar la evolución de elementos, parece ser más conveniente estudiar la variabilidad de ocupación de un

rango, uno por uno. Y la siguiente definición es la de una medida dinámica que captura esta idea.



Rodajas temporales

Figura 4.3: Evolución temporal de los rangos para jugadores y equipos de las últimas 6 disciplinas. Gráfica que muestra el cambio en el rango k a lo largo del tiempo t para todos los jugadores/equipos en cada deporte y juego considerado en este estudio: NASCAR-B, NASCAR-W, GPI, WSD, ATP, ESE. Aquí incluyo los espaguetis de todos los equipos/jugadores que alguna vez ocuparon uno de los primeros 30 lugares a lo largo de todo el tiempo que se tiene disponible. Para este último conjunto de datos el comportamiento, de los elementos den los primeros lugares no es tan estable como vimos en la [Figura 4.2](#), para ATP sí se puede apreciar la presencia de la regularidad antes mencionada. Sin embargo, en ESE apreciamos que la permanencia de los elementos en el primer lugar es básicamente sólo de un tiempo, recordemos que aquí se deben considerar muchos factores, tales como la resolución temporal correspondiente varía entre semanas a meses (ver [Tabla 2.1](#)).

4.2. Diversidad de rango.

Como mencioné desde el inicio del presente trabajo, el desarrollo de esta Tesis está basado en el artículo de investigación que publiqué en colaboración con investigadores de la UNAM y las ideas que a continuación mencionaré se desarrollaron ya en [\[12\]](#), lo que haré es seguir exactamente esas ideas para el caso de las bases de datos con las que contamos aquí. Y el primer paso importante es definir *diversidad de rango*, medida que es el objeto de estudio central en este trabajo.

El análisis previo de la forma funcional de la distribución de rango en varios deportes (incluso aunque la bondad de ajuste ha sido promediada en el tiempo) está restringida por el hecho de que la distribución de rango es inherente a una medida instantánea (sólo estudia una rodaja temporal en el tiempo), en el sentido de que captura el ranqueo en un punto del tiempo determinado y no toma en cuenta la dinámica de los jugadores y equipos cambiando de rango conforme el tiempo avanza. Incluso en la [Sección 4.1](#) anterior, intentamos aproximarnos a la idea de analizar la evolución de los rankings siguiendo cada uno de los jugadores/equipos a lo largo del tiempo, pero concluimos que no se puede obtener una regularidad evidente con este punto de vista, observamos que es más fácil estudiar cómo son ocupados los rangos, por ejemplo, ver cómo es que el primer lugar es obtenido por los jugadores/equipos conforme se avanza en el tiempo; observamos que los rangos bajos tienden a ser ocupados por menor cantidad de elementos respecto a los rangos altos, lo que nos lleva a la definición de la nueva medida.

Con el fin de superar el problema de no tener una descripción dinámica de los rangos, aquí contribuiré con el análisis de los rankeos en los deportes y juegos calculando la diversidad de rango, una medida del número de elementos ocupando un rango específico a lo largo del tiempo. De trabajos previos [2] y del actual [12], parece ser que la diversidad de rango tiene la misma forma funcional, no sólo para deportes sino para otros sistemas complejos, como países clasificados por su complejidad económica, las 500 empresas cabeza rankeadas por la revista *Fortune*, o por el conjunto de millones de palabras en seis idiomas indo-europeos, lo cual es muy interesante; pero por el momento sólo nos concentraremos en los sistemas que nos competen.

4.2.1. Definición.

Antes de dar la definición formal, haremos unas definiciones notacionales y que serán de gran ayuda para las definiciones formales de las medidas dinámicas que se estudiarán en este trabajo. Primero que nada denotaremos $X(k, t)$ como el elemento que ocupa el rango k al tiempo t . El tiempo t es por ejemplo la semana 24 del 2017 para el caso de FCWR-C, refiriéndose entonces a la rodaja temporal correspondiente a la semana 33 del 2017. Por tanto, $X(1, S - 33 - 2017)$, donde $S - 33 - 2017$ denota la semana 33 del 2017, denota por ejemplo al Barcelona que ocupó el primer lugar en la semana 33 de 2017. Ahora bien, un sistema tendrá un total de T rodajas temporales, por lo que cada paso en el tiempo estará denotado por t_1, t_2, \dots, t_T . Por otro lado, denotaremos $X(k) = \{X(k, t_1), X(k, t_2), \dots, X(k, t_T)\}$ como el conjunto de elementos que ocuparon el rango k en todas las rodajas temporales, como estamos hablando de un conjunto, no hay repeticiones.

Tomando en cuenta la idea de sólo estudiar cómo son ocupados los rangos mientras avanza el tiempo haremos la siguiente definición. La diversidad de rango $d(k)$ se define como el número de distintos elementos en un sistema complejo (en este caso deporte o juego) que ocupan el rango k considerando un periodo de tiempo dado. En términos matemáticos tenemos entonces que la definición está dada por:

$$d(k) = \frac{|X(k)|}{T} \quad (4.1)$$

donde, $|X(k)|$ es la cardinalidad del conjunto $X(k)$. En otras palabras, escogemos enfocarnos en la dependencia temporal de los rangos, en vez del enfoque estático (es decir, la distribución de rango $f(k)$ en una rodaja temporal). De los espaguetis, la diversidad de rango tienen una pequeña interpretación geométrica. La diversidad de rango $d(k)$ es simplemente el número normalizado de los diferentes elementos (curvas o espaguetis) que ocupan al menos un tiempo (o intervalo de tiempo) el rango k . La diversidad de rango para varios deportes y juegos se muestra en la [Figura 4.5](#) como los puntos azules. La gráfica se muestra en escala semilogarítmica, poniendo en el eje de las abscisas $\log_{10}(k)$.

4.2.2. Un primer vistazo a la diversidad de rango.

De la [Figura 4.4](#) podemos ver que los valores empíricos de la diversidad de rango son (en esencia) monótonas crecientes. Analicemos con más detalle la [Figura 4.4](#). Vemos que para la mayoría de los casos (FIDE-F, FIDE-M, FCWR-C, FIFA, FCWR-G, OWGR, GPI, WSD y ATP) los valores de diversidad de rango para valores pequeños de k también son pequeños, conforme k crece, también lo hace la diversidad de rango. Esta idea coincide con las gráficas que presentamos en el [Apéndice B](#), donde vimos que para rangos pequeños había una menor cantidad de elementos que ocupaban dichos valores de rangos y para valores grandes aumentaba el número de espaguetis. De hecho, este fenómeno es consistente a lo observado en las figuras [4.2](#) y [4.3](#), donde se ve mayor estabilidad de los espaguetis en valores bajos de k y un mayor desorden en los espaguetis para rangos altos. Las diversidades un poco más particulares son las de NASCAR-B y NASCAR-W. Para NASCAR-W vemos que la diversidad de rango de los rangos bajos no empieza tan abajo, pero sí se aprecia que aparece una tendencia monótona creciente. El caso de NASCAR-B es interesante porque desde el rango $k = 1$, $d(k)$ alcanza valores muy grandes. Más aún, para ESE, tenemos que existen prácticamente 3 valores para la diversidad de rango, en definitiva no tiene la forma empírica que sugiere el resto de las disciplinas y de ella no podríamos obtener grandes conclusiones; este comportamiento proviene del hecho de que se tienen muy pocas rodajas temporales, sólo 13, pues la resolución temporal es anual desde el 2004 al 2016.

Hay una par de bases de datos con resultados curiosos. Vemos que para el caso de FIFA, OWGR y FCWR-C los valores de $d(k)$ empiezan a decrecer para los valores grandes de k . Las consecuencias fenomenológicas son interesantes de considerar, pues quiere decir que al igual que los rangos más bajos, los rangos altos empiezan a tener menos variabilidad en su ocupación. Sin embargo, en la mayoría de los sistemas no se aprecia esa caída. ¿Qué podría estar generando ese comportamiento en los que sí? El caso más marcado de este fenómeno es el de FIFA, pues los valores de diversidad no son monótonos crecientes ya que a partir de cierto rango éstos comienzan a ser monótonos decrecientes. Para tratar de entender por qué pasa esto, definiremos una medida interesante. Es natural pensar que los jugadores, por ejemplo, tienen un tiempo

4. DINÁMICA DE RANGO PARA DEPORTES Y JUEGOS: DIVERSIDAD, PROBABILIDAD DE CAMBIO, ENTROPÍA Y COMPLEJIDAD.

de vida dentro del sistema. No es posible que Lionel Messi juegue para siempre, éste se debe retirar en algún momento, lo mismo si pensamos el caso de Roger Federer, uno de los mejores tenistas del mundo. Es natural pensar que nuevos jugadores entren a interactuar en los rankings, pues están iniciando su participación en alguna disciplina deportiva. En nuestros sistemas seguro se debe ver reflejada esa dinámica, en cada uno de las listas de ranqueo no es posible que se encuentren los mismos jugadores a lo largo del tiempo, pues algunos se retiran y otros debutan en algún momento.

Para cuantificar la idea anterior de entrada y salida de elementos en los rankeos, definiremos una cantidad nueva: *índice de cerradura*. Sabemos que N es el número total de elementos rankeados en cada una de las rodajas temporales que conforman al sistema, como ya habíamos definido anteriormente. Ahora, definimos Γ como el número total de elementos que alguna vez apareció en las listas de ranqueo para todas las rodajas temporales, es decir, estamos contando a los elementos que alguna vez salieron o entraron en las listas aunque sea una sola vez. Por lo tanto, el índice de cerradura, Ω , lo definimos como:

$$\Omega = \frac{N}{\Gamma} \quad (4.2)$$

Por la definición de este *índice de cerradura*, sabemos que $0 < \Omega \leq 1$. Si pasara que siempre los mismos elementos aparecen en todas las listas de los rankeos, entonces $\Gamma = N \Rightarrow \Omega = 1$, por lo que no habrá entrado ni salido ningún elemento de las listas nunca, entonces el sistema es muy cerrado, de allí el nombre de este índice. Si en cambio, entran y salen elementos conforme avanzamos en el tiempo, naturalmente Γ irá incrementando su valor (el mínimo valor que puede adquirir es $\Gamma = N$), por lo que conforme éste aumenta, el valor del índice de cerradura decrementará. El que entren y salgan elementos es parecido a la idea de un sistema abierto, se hace menos cerrado, por lo que si el sistema es más abierto el valor de Ω decrementa. En las gráficas de la [Figura 4.4](#) pongo los valores del índice de cerradura para cada uno de los sistemas estudiados.

Apreciamos que el sistema más cerrado de ellos es el de FIFA, pues es el que tiene un valor más cercano a 1. El sistema más abierto, o menos cerrado aunque en matemáticas estos conceptos no son contrarios, es NASCAR-B. Tienen sentido estos resultados, pues los equipos nacionales de fútbol nunca cambian, son los mismos países con sus respectivos equipos, aunque jugadores se retiren o debuten, las selecciones nacionales seguirán allí. Algo así debería pasar con el caso de FCWR, pues los clubes de fútbol tampoco deberían de cambiar, sin embargo, como mencioné en el [Capítulo 2](#), las bases no están completas como se obtendrían sin modificarlas desde la fuente de donde se extrajeron, ya que realicé un corte a las listas para que todas tuvieran la misma longitud y como existen más rangos disponibles pero yo no los considero, entran y salen equipos de la lista que se encuentran más allá del rango máximo de las bases y con rangos por abajo cercanos al mismo.

Los sistemas con menor cerradura como NASCAR-B, NASCAR-W y ESE son justamente los que obtuvieron las diversidades de rango más peculiares y debe haber una clara correlación con que sean los menos cerrados. Es interesante notar que el sistema

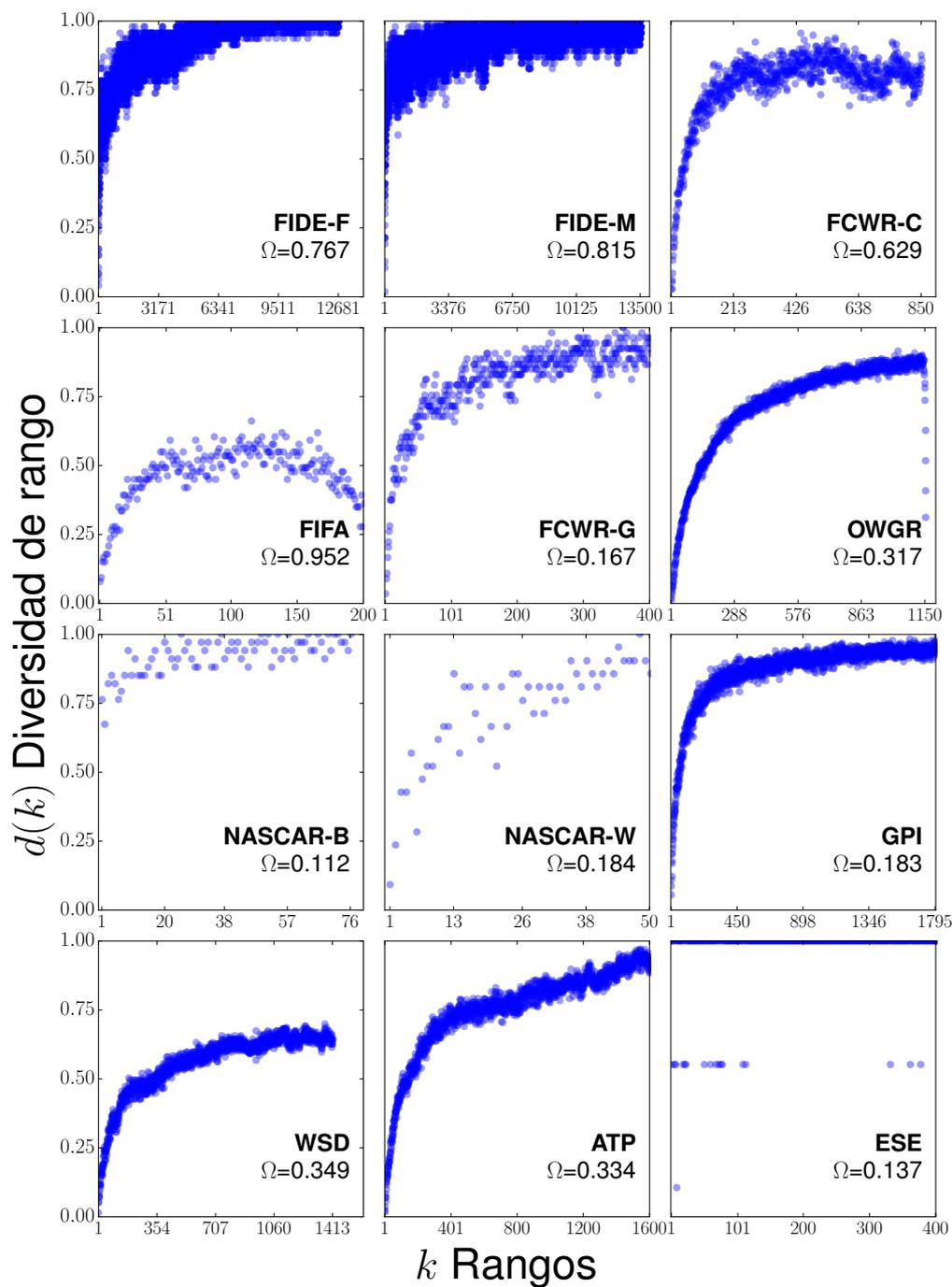


Figura 4.4: Diversidad de rango de deportes y juegos. Gráfica que muestra la diversidad de rango $d(k)$ para todos los conjuntos de datos (puntos azules), también se presenta el índice de apertura para cada una de las disciplinas Ω , como definimos en la Ecuación 4.2

más cerrado, FIFA, es justamente el que tiene la forma no monótona en la forma empírica de la diversidad de rango, es decir, el hecho de ser cerrada puede llegar a implicar que los elementos de los últimos rangos no cambien mucho su rango a lo largo del tiempo, sean más estables. Dicho de otra forma, los últimos rangos son ocupados prácticamente por los mismos elementos siempre. Sin embargo, aquí sólo tenemos un ejemplo de ese comportamiento tan marcado, por lo que no podemos concluir que el hecho de que los sistemas sean cerrados implica que se presente ese comportamiento. A pesar de ello, podemos encontrar otra regularidad a partir de presentar estas gráficas de otra forma.

4.2.3. Modelo para la diversidad de rango.

Ya tenemos los valores de la diversidad de rango graficados para las 12 disciplinas tal cual dimos su definición y es lo que observamos en la [Figura 4.4](#). Pero de ellas aún no podemos apreciar inmediatamente una regularidad que pueda modelar la dinámica de rango para estos sistemas. Si ahora graficamos $d(k)$ en escala semilogarítmica (valores logarítmicos de k en las abscisas y escala normal en el eje de las ordenadas) obtenemos los puntos azules (ignoremos un momento las curvas rojas porque eso es algo del futuro) pintados en las gráficas que configuran la [Figura 4.5](#). Nuevamente, vemos que los valores de la diversidad de rango no parecen tener una regularidad clara para el caso de NASCAR-B, NASCAR-W y ESE, ésto se puede deber por motivos ya expresados anteriormente. Los espaguetis evidencian trayectorias anómalas para ESE, además son sistemas para los cuales se tienen menor cantidad de rodajas temporales respecto al resto de los trabajados. Otro de los factores que influyen es el que nos indica el índice de cerradura, pues son justamente los sistemas menos cerrados. Sin duda $d(k)$ es el más anómalo para ESE, pero en los dos casos de NASCAR se nota claramente la tendencia monótona que ya había mencionado con anterioridad.

El resto de los sistemas aquí muestran una clara regularidad. La diversidad de rango sigue mostrando un comportamiento monotónico creciente (en esencia) y más aún tienen una forma que asemeja a la de una sigmoide. Una sigmoide corresponde a la gráfica de funciones muy particulares, un ejemplo es el de la función logística, y se asemeja a la forma de una S. [\[43\]](#)

La gráfica de la distribución acumulativa de una función cuadrado integrable con un sólo punto crítico tendría estas mismas propiedades (la forma sigmoide) y una Gaussiana (o distribución normal) es la elección más simple, pues conocemos muchas de sus propiedades y es matemáticamente manipulable con cierta facilidad. [\[44\]](#) Y éste es el punto clave, tal vez la acumulativa de una distribución normal sea el modelo teórico que aproxime a la distribución de rango de acuerdo a los sistemas que hemos trabajado a lo largo de esta tesis. Más aún, basaré esta suposición del hecho de que ya se probó anteriormente este modelo en el estudio similar a este pero en el caso de palabras y su frecuencia de aparición en textos para seis idiomas indo-europeos y en el trabajo que tanto he citado aquí [\[2\]](#). En este mismo trabajo se introduce el concepto de diversidad de rango $d(k)$ y modelan a la misma con la hipótesis de que se pueda aproximar a la acumulativa de una distribución normal. Confiando en el juicio de ese artículo y con la similitud de nuestros resultados, en la mayoría de los sistemas, con los que allí se

obtuvieron, procederemos a usar ese mismo modelo. El modelo es:

$$\Phi_{\mu,\sigma}(\log k) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log k} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy. \quad (4.3)$$

Tengamos en cuenta que el dominio no es precisamente el de los rangos k , ya que la [Ecuación 4.3](#) modela lo observado en las gráficas de la [Figura 4.5](#), las cuales se encuentran en escala semilogarítmica. Por lo tanto la variable independiente en la sigmoide es $\log k$ y los parámetros libres o de ajuste de este modelo son μ y σ pues dependiendo de sus valores la forma sigmoide de la gráfica de esta función varía. El valor medio μ es el parámetro encargado de centrar a la sigmoide muy cerca a $N/2$ una vez hecho el ajuste, mientras que el ancho σ se ajusta y proporciona la escala tal que $\Phi(\log k)$ se acerque a los valores extremos de $d(\log k)$. Si k_{\pm} están dados por $\log_{10} k_{\pm} = \mu \pm 2\sigma$, la parte más empinada de la diversidad se encuentra entre k_- y k_+ , ésto lo sabemos pues k_{\pm} representa los dos puntos de inflexión en la distribución normal [\[44\]](#). En la [Figura 4.5](#) mostramos los ajustes de Φ para todos los deportes y juegos aquí considerados (Los valores de R^2 para las curvas Φ también se muestran en la [Figura 4.5](#)), la misma R^2 explicada en la [Sección 3.4](#) y que es aplicable de igual modo para este caso pues es sólo una medición de qué tanto se acerca la tendencia de un conjunto de datos a un modelo teórico. En este caso no consideramos ni D ni p , ya que esas medidas de bondad de ajuste sólo son significativas para distribuciones cumulativas de probabilidad, mientras que $d(k)$ no lo es pues su definición no coincide con algo parecido. Claro, el modelo de la [Ecuación 4.3](#) para la diversidad de rango indica que existe una variable aleatoria que proviene de la mutiplicación de otras variables aleatorias y que se distribuyen de acuerdo a la distribución log-normal, que es justamente lo que indica la presencia de un proceso multiplicativo. Sin embargo, ¿qué es lo que distribuye de acuerdo a la distribución log-normal?, la respuesta más inmediata sería que los rangos como variable aleatoria son los que se distribuyen de esa forma, pero no está claro que sea así. [\[45\]](#) La diversidad de rango es sólo una función de los rangos, la distribución empírica cumulativa de los rangos no la hemos presentado aquí y no coincidirá con $d(k)$ pues ésta es estrictamente creciente y como veremos en la [Figura 4.5](#), es claro que $d(k)$ no lo es. Es por ello que no aplicaremos las bondades de ajuste p o D para este caso.

De la [Figura 4.5](#) podemos apreciar que, al menos visualmente, los ajuste de Φ asemejan bien a la diversidad de rango para la mayoría de los deportes a excepción, claro está, de los tres casos anómalos ya discutidos (NASCAR y ESE), además debido a que algunos sistemas como OWGR, FCWR-C y FIFA presentan caídas en los valores de $d(k)$ para k grande, Φ no puede modelar esas caídas, pues es una función monótona, sin embargo vemos que cualitativamente y visualmente logra capturar la esencia de la tendencia de la diversidad de rango para la mayoría de los deportes/juegos.

4. DINÁMICA DE RANGO PARA DEPORTES Y JUEGOS: DIVERSIDAD, PROBABILIDAD DE CAMBIO, ENTROPÍA Y COMPLEJIDAD.

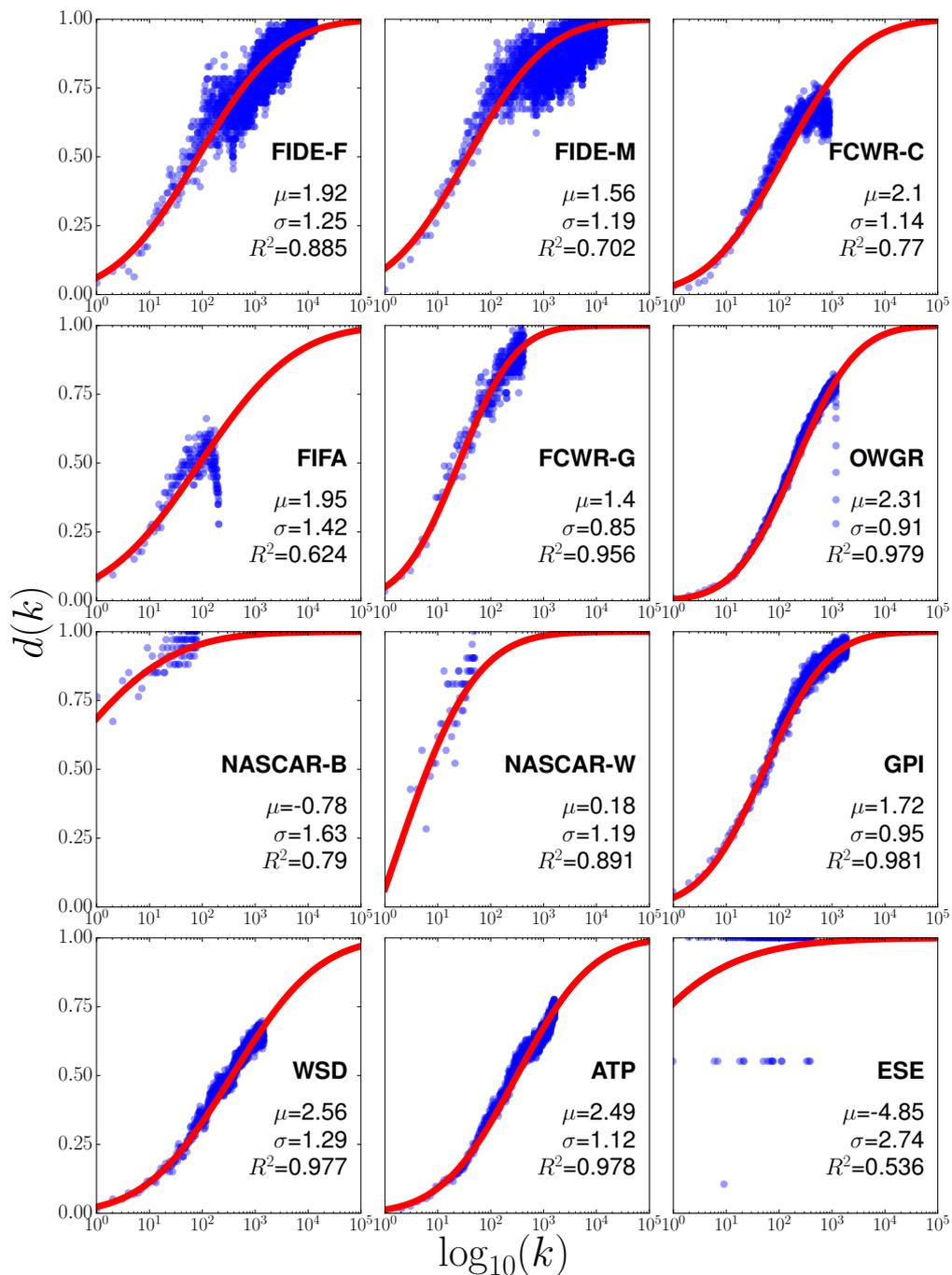


Figura 4.5: Diversidad de rango de deportes y juegos en escala semilogarítmica. Gráfica que muestra la diversidad de rango $d(k)$ para todos los conjuntos de datos (puntos azules), así como los ajustes a Φ (líneas rojas). Incluimos los valores de μ , σ , y el parámetro de bondad de ajuste R^2 también.

La bondad de ajuste R^2 también proporciona algunas observaciones importantes; por ejemplo, para OWGR, GPI, WSD y ATP tenemos valores tales que $R^2 > 0.9$ indicando que son los casos mejor modelados con la [Ecuación 4.3](#), y en efecto, la diversidad de rango sí parece aproximarse visualmente a la forma sigmoideal y la bondad de ajuste lo confirma. Lo interesante es que a pesar de tener caída para algunos rangos en OWGR, la sigmoide trabaja muy bien para el resto de los puntos, la caída es claramente una excepción significativa. En el caso de NASCAR-W tenemos una R^2 muy próxima a 0.9, los puntos tienen esa tendencia sigmoideal en cierta aproximación pero no se aprecia en primera instancia pues es poca la cantidad de rangos con las que cuenta este sistema (ver [Tabla 2.1](#)). El caso de la caída tan abrupta para FIFA se ve marcada en la bondad de ajuste, pues es una de las más bajas, la sigmoide alcanza a describir ese régimen fenomenológico. Para el caso de ESE concluimos claramente que un sistema con pocos elementos y escasas rodajas temporales no permiten hacer una descripción satisfactoria de su dinámica de rango, esto se comprueba por su clara diferencia con los resultados del resto de las disciplinas, y era de esperarse pues no se tiene suficiente estadística.

Dejando de lado el caso de ESE, del cual no podemos obtener conclusiones, es muy interesante que esa forma sigmoideal de la diversidad de rango en la mayoría de las disciplinas sea evidente, al menos visualmente, y en algunos otros casos la bondad de ajuste R^2 también nos acercan a la conclusión de que este modelo es adecuado. Incluso en [\[2\]](#) se concluyó que la diversidad de rango para la evolución del vocabulario escrito en 6 idiomas es sigmoideal con el mismo modelo presentado en la [Ecuación 4.3](#), por lo que estos resultados parecen sugerir que la formación jerárquica de los sistemas complejos, medida con $d(k)$, puede tener similitudes a lo largo de diversos sistemas complejos, es una característica en común. Pero tal vez el origen de este fenómeno es más profundo y proviene de características más propias de sistemas teóricos bien conocidos, a continuación intento ahondar en el tema.

4.2.4. Posible origen analítico de la diversidad de rango.

En la Física existen diversos sistemas compuestos por elementos diversos con interacciones similares entre ellos, y a pesar de que estas interacciones pueden no tener una regularidad, el comportamiento macroscópico del sistema está determinado usualmente por leyes generales, tal como la ecuación de estado (como en Termodinámica que es una relación matemática bien definida entre variables de estado y/o potenciales termodinámicos).

Una argumentación analítica realizada en el artículo original publicado que se anexa en el [Apéndice D](#) sugeriría que éste podría ser un *ansatz* apropiado ([Ecuación 4.3](#)) bajo condiciones generales, al menos cualitativas. Sin embargo, para sistemas dinámicos con dinámica competitiva, existen comportamientos genéricos descritos por las distribuciones Gamma m_2 , Beta m_3 y m_4 , [\[33\]](#), [\[46\]](#), [\[2\]](#) y pueden haber también diferencias entre las realizaciones, siguiendo una dinámica multiplicativa, que como ya comenté, quiere decir que se tiene el involucramiento de variables aleatorias que se distribuyen de acuerdo a la distribución log-normal. [\[45\]](#). Esta fenomenología multiplicativa también se presenta en el caso para varios idiomas indo-europeos [\[2\]](#) como ya habíamos mencionado

y ahora lo haremos más evidente para los conjuntos de datos de deportes y juegos aquí considerados.

En el artículo original publicado (ver [Apéndice D](#)) se retoma la idea de que la distribución de rango es la solución estacionaria de la ecuación de Fokker-Planck que se presentó en el [Capítulo 3](#), es decir, la [Ecuación 3.4](#) que aquí escribimos de nuevo:

$$\frac{\partial m(k, t)}{\partial t} = -\frac{\partial}{\partial k}[g(k)m(k, t)] + \frac{1}{2} \frac{\partial^2}{\partial k^2}[f(k)m(k, t)]$$

Como vemos que la presencia de la distribución log-normal evidencia que existe un proceso multiplicativo, entonces a $m(k, t)$ ahora lo escribimos en términos de la nueva variable $x = \log k$, en el mismo artículo se demuestra que entonces la ecuación de Fokker-Planck se puede transformar con el cambio de variable $\nu(x, y) = f(x)m(x, t)$ y $\tau = f(x)t$ en:

$$\frac{\partial}{\partial \tau} \nu(x, \tau) = -\Lambda \frac{\partial \nu}{\partial x} + \frac{\partial^2 \nu}{\partial x^2} \quad (4.4)$$

con Λ una constante. El carácter multiplicativo ya mencionado se introduce con un cambio de variable $u(x, \tau)$ de la forma:

$$\log \frac{\nu(x, \tau)}{u(x, \tau)} = \Lambda x - \frac{\Lambda^2}{4} \tau \quad (4.5)$$

y como resultado, la [Ecuación 4.4](#) se transforma en:

$$\frac{\partial}{\partial \tau} u(x, \tau) = \frac{\partial^2 u}{\partial x^2} \quad (4.6)$$

que es justamente la ecuación de difusión y cuya solución está dada por una Gaussiana de la forma:

$$u(x, \tau) = \frac{1}{\sqrt{4\pi\tau}} \int_{-\infty}^{\infty} e^{-(x-x')^2/4\tau} u(x', 0) dx' \quad (4.7)$$

Ahora bien, se menciona en el artículo que considerando la frontera absorbente $u(x_c, t) = 0$ y donde x_c representa el punto de absorción, ahora denotemos $u(x, t; x_0, x_c)$ a la densidad de probabilidad que satisface esta condición de frontera para $x < x_c$ (x_0 denota una condición inicial). La probabilidad de sobrevivencia $S(t, x_c)$ de que la partícula se mantenga en la posición $x < x_c$ para todos los tiempos hasta t estaría dada por:

$$S(t, x_c) = \int_{-\infty}^{x_c} u(x, t; x_0, x_c) dx \quad (4.8)$$

que de hecho es también la cumulativa de la distribución de x al tiempo t . Ahora, la probabilidad de que una partícula alcance el punto de absorción entre los tiempos t y $t + dt$ sería $h(t)$ tal que:

$$h(t)dt = S(t) - S(t + dt)$$

por lo que de manera infinitesimal esto equivale a:

$$h(t) = -\frac{\partial S(t)}{\partial t} \quad (4.9)$$

entonces entre dos tiempos arbitrarios t_1 y t_2 tenemos que:

$$S(t_1) - S(t_2) = \int_{t_1}^{t_2} h(t') dt' \quad (4.10)$$

entonces la [Ecuación 4.10](#) se interpreta como la probabilidad de que la partícula (en nuestro caso sería un jugador/equipo) haya alcanzado el punto de absorción x_c (recordemos que $x_c = \log k_c$ son los rangos en escala logarítmica). Y el artículo identifica el lado derecho de la [Ecuación 4.10](#) con $d(k)$, pues es la que cuenta el número de eventos que hayan alcanzado k_c en un periodo de tiempo. Ahora bien, la [Ecuación 4.7](#), es decir $u(x, t)$ está relacionada con caminatas aleatorias por su forma funcional y de eso además se puede ver que los pasos de esas caminatas están distribuidos conforme a la distribución normal. Más aún, de estas ecuaciones se evidencia una relación entre $d(k)$ y ese modelo de caminante aleatorio y la distribución normal, que es justamente lo que queríamos justificar. Un resultado también muy interesante es que la distribución de rango $m(k)$ está íntimamente relacionada con $d(k)$ y no son conceptos separados como creíamos.

4.2.5. Diversidad de rango versus distribución de rango.

El planteamiento de la diversidad de rango $d(k)$ es muy interesante, pues en el caso de las bases de datos que tengan contenido adecuado para el estudio de una dinámica de rango, podemos observar la presencia de algo que se acerca mucho a un comportamiento genérico, y de hecho, no exclusivo de los deportes/juegos, si no que también se ha observado en el estudio de la evolución del vocabulario en 6 idiomas indo-europeos [2] y en un estudio reciente en el que participé que analiza la evolución de los idiomas yendo más allá que ver la dinámica de rango para palabras, sino que se estudia la frecuencia de aparición en textos de estructuras más amplias, los N-gramas (1-grama es una palabra, 2-grama el conjunto de dos palabras y así sucesivamente) y que arroja resultados interesantes pero que mantiene la tendencia en una forma funcional similar a la sigmoide aquí propuesta [47]. Por lo que estamos frente a una aparente regularidad genérica a lo largo de una variedad de sistemas complejos que no tienen relación alguna entre sí en la forma que se da lugar a la estructura jerárquica propia.

En este trabajo se han desarrollado dos conceptos fundamentalmente: la diversidad de rango $d(k)$ y la distribución de rango $m(k)$ y concluyo que estos dos conceptos son complementarios y ajenos pues analizan distintos aspectos que conforman a los sistemas complejos, en general, no sólo los aquí vistos. Por lo que $d(k)$ y $m(k)$ miden diferentes aspectos de la estructura jerárquica de un sistema complejo:

- La diversidad de rango incluye información sobre cómo cambia la ocupación de los rangos a través del tiempo por medio de una sola función, mientras que la distribución de rango captura la jerarquía del sistema para un sólo intervalo de tiempo

(rodaja temporal) y es ciega ante la evolución del sistema; pretender describir la dinámica del sistema a partir de la distribución de rango implicaría comparar las funciones obtenidas (para $m(k)$) por cada rodaja temporal, lo cual no suena práctico.

- La diversidad de rango ignora cualquier información sobre los puntajes de los elementos más allá de su orden, y por ende la misma $d(k)$ puede ser obtenida para distintos comportamientos de $m(k)$ (ley de potencias, Gamma, Beta, etc.). Como ejemplo, consideremos cualquier transformación en el tiempo de los puntajes de los elementos en el sistema, tal que el orden de los rangos se mantenga invariante; entonces $m(k)$ podría variar entre distintas formas funcionales mientras transcurre el tiempo, mientras que $d(k)$ se mantiene invariante. El caso inverso también es posible, y cualquier distribución de rango podría reproducir una gran variedad de diversidades de rango. Por ejemplo, podríamos construir muchas dinámicas distintas para los puntajes que mantengan constante el número de elementos con un determinado puntaje, pero eso cambia la cantidad de tiempo que un elemento conserva cierto puntaje, así manteniendo $m(k)$ invariante y $d(k)$ cambiando.
- Tanto $d(k)$ como $m(k)$ miden algunos aspectos de la estructura y dinámica de la jerarquía en un sistema complejo, pero sólo la diversidad de rango captura la forma en que los elementos cambian su posición en la jerarquía, más allá de cambios menores en los puntajes que pueden ser atribuidos, por ejemplo, a diferentes formas de medir el desempeño y este aspecto es fundamental para ser una descripción común a todos los sistemas, pues es ciega ante la forma en que la estructura jerárquica nace en cada una de las rodajas temporales.
- Una diferencia menos relevante pero importante de mencionar es que $m(k)$ no necesita tener una muestra estadística representativa, pues con pocos elementos como el caso de FIFA ya se puede hacer un análisis con ayuda de bondades de ajuste finas como p el índice de Kolmogorov-Smirnov. Mientras que para tener una dinámica apreciable y descriptible se necesita una cantidad "aceptable" de rodajas temporales, justo como lo vimos en el caso de ESE.

En pocas palabras, para tener una descripción completa de la estructura jerárquica de los sistemas necesitamos ambas medidas pues ven distintos aspectos. Sin embargo, parece más apropiado usar $d(k)$ para tener conclusiones sobre la dinámica de rango de los sistemas y de hecho, como veremos más adelante, podemos incluso considerar un par adicional de medidas que describen la dinámica los mismos.

4.2.6. Universalidad.

El coeficiente R^2 no nos permite concluir que la sigmoide propuesta (Ecuación 4.3) sea adecuada para modelar a la diversidad de rango para todos los sistemas aunque en muchos casos parece ser así. Todas las curvas tienen distintos dominios y evidentemente para cada k tenemos que $d(k)$ es distinta a lo largo de todos los sistemas puestos aquí.

Con una perspectiva distinta podemos ver la regularidad a través de todos los sistemas, es decir, la haremos ver más evidente y es sólo un ejercicio matemático para tratar de justificar mejor cualitativamente.

Para comparar las diferentes curvas de diversidad de rango de manera más objetiva, los rangos se pueden normalizar a $\frac{\log(k)-\mu}{\sigma}$, ésto para que cualquier sigmoide dada por la Ecuación 4.3 sea vista desde el punto de vista de la distribución cumulativa normal unitaria (donde $\mu = 0$ y $\sigma = 1$). Si realizamos dicha transformación en el eje de las abscisas entonces todas las sigmoides ajustadas se transformarán en la sigmoide normal unitaria. Por otro lado, del mismo modo, si para los datos empíricos de la diversidad de rango para las 12 disciplinas realizamos la misma transformación, los puntos aparentemente se acumularán en torno a la sigmoide unitaria. Para hacer más evidente este fenómeno vamos a promediar los datos de diversidad por ventanas (en subintervalos del dominio de $\log_{10}(k)$) y graficaremos dichos promedios para el valor medio de cada uno de estos subintervalos. Lo que obtenemos es lo que se aprecia en la Figura 4.6.

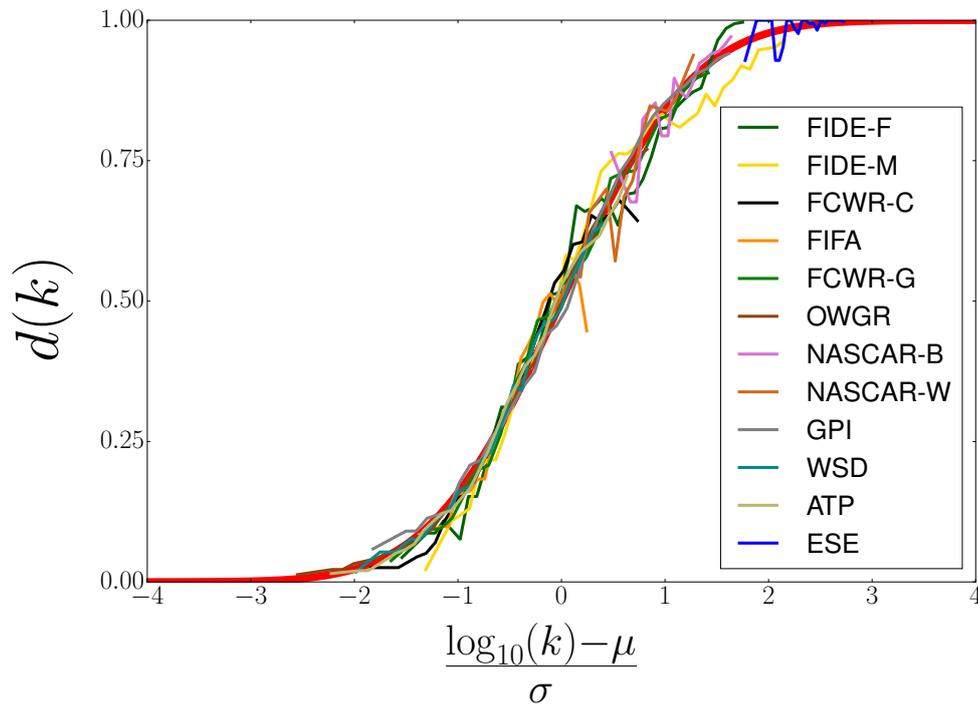


Figura 4.6: Similaridad en la diversidad de rango normalizada entre los deportes y juegos. Gráfica que muestra una comparación de la diversidad de rango $d(k)$ para todas las actividades consideradas. Con los valores de μ y σ obtenemos el ajuste a Φ , hemos reescalado la abscisa. Como referencia incluimos la forma básica de la Ecuación 4.3 (delgada línea roja), con $\mu = 0$, y $\sigma = 1$. Estos resultados indican que todas las actividades tienen la misma forma funcional para la diversidad de rango.

El reescalamiento de las abscisas y el promedio de la diversidad por ventanas muestra de manera más evidente la tendencia sigmoïdal de los datos. De hecho, aquí sí estamos comparando las 12 disciplinas en un solo punto de vista. Es claro, al menos cualitativamente, que esta gráfica evidencia más satisfactoriamente un comportamiento genérico. Hay que notar que incluso el caso más anómalo, el de ESE, que está representado por las líneas azules en el gráfico y se agrupan sólo para valores grandes del dominio (el dominio reescalado) también siguen esa tendencia.

Además, la universalidad que aquí presentamos también puede ser vista desde la perspectiva de lo que hicimos en la [Subsección 4.2.4](#), donde vimos que la forma de la diversidad de rango de la [Ecuación 4.3](#) emerge naturalmente de modelar a la interacción entre los jugadores/equipos en términos de procesos aleatorios gaussianos markovianos de un paso, y el hecho de que las $d(k)$ para todas las disciplinas puedan ser aproximadas por la acumulativa de una distribución normal, indica que, en efecto, las interacciones de los elementos de las disciplinas son procesos aleatorios y que parecen ignorar las diferencias en que los sistemas jerárquicos se forman por la peculiaridades entre sí.

4.3. Probabilidad de cambio.

La diversidad de rango fue la primera medida que intenta proporcionarnos una descripción de la dinámica de rango del sistema, vimos que es bastante exitosa pues hallamos una regularidad en la forma funcional de la misma. La diversidad de rango mide la cantidad de elementos distintos que ocupan un rango en una ventana de tiempo determinada, sin embargo, no dice nada respecto a cómo va cambiando la ocupación de cierto rango.

Supongamos un sistema hipotético en el que el Barcelona y el Real Madrid (equipos de fútbol) son los únicos que se disputan el primer lugar a lo largo de 52 semanas. Entonces, la diversidad de rango del rango 1 está definida como $d(1) = 2/52$ pues sólo dos equipos diferentes ocuparon ese rango en este ventaneo temporal. Ahora bien, si mantenemos esa característica pero resulta que semana a semana van supliendo al otro el ocupar el primer lugar, es decir, la primera semana el primer lugar lo ocupa el Real Madrid, la siguiente el Barcelona, la siguiente el Real Madrid y así sucesivamente. La diversidad de rango en esta situación es $d(1) = 2/52$. Ahora supongamos otra situación, que la primer semana el Barcelona ocupe el primer lugar y el resto de las semanas sea ocupado por el Real Madrid. En este caso seguimos manteniendo que $d(1) = 2/52$. Otra situación hipotética es que las primeras 26 semanas el Barcelona ocupe el primer lugar y las 26 últimas semanas sea ocupado por el Real Madrid. En este caso la diversidad de rango $d(1) = 2/52$ se mantiene.

En las situaciones hipotéticas mencionadas tenemos que la diversidad de rango es la misma para dinámicas completamente diferentes. La diversidad de rango no captura la forma en que los lugares se van ocupando, sólo se fija en cuántos equipos distintos lo ocuparon. Para poder cuantificar esa dinámica no descrita aún, definiremos una nueva cantidad llamada *probabilidad de cambio*. Esta cantidad dinámica también ha sido utilizada en el trabajo de N-gramas que participé y que ya mencioné con anterioridad.

[47]

4.3.1. Definición.

Proponemos una medida diferente a la diversidad de rango, que llamaremos probabilidad de cambio $p(k)$. La probabilidad de cambio es una función del rango k y se interpreta como la probabilidad de que una palabra que se encuentre en el rango k cambie el valor de ese rango. Para calcularla, nuevamente debemos tener un sistema complejo jerárquico con una lista de elementos rankeados para una secuencia de tiempos. Si el ventaneo temporal consiste en T rodajas temporales, entonces $p(k)$ se calcula dividiendo el número de veces que un elemento ocupando el rango k adquiere un rango distinto por el número de transiciones temporales $T - 1$. Formalmente, esta cantidad se puede calcular como: [47]

$$p(k) = \frac{\sum_{t=0}^{T-1} 1 - \delta(X(k, t), X(k, t + 1))}{T - 1} \quad (4.11)$$

donde $X(k, t)$ denota al elemento que ocupa el rango k al tiempo t y $\delta(X(k, t), X(k, t + 1))$ es la delta de Kronecker, que sería uno si $X(k, t)$ y $X(k, t + 1)$ coinciden, lo que implica que el mismo elemento ocupa el rango k en dos tiempos sucesivos y $\delta(X(k, t), X(k, t + 1)) = 0$ en caso de que $X(k, t)$ y $X(k, t + 1)$ fueron distintos, entonces dos elementos diferentes ocuparon el rango k en tiempos sucesivos, es decir, hubo un cambio en la ocupación de ese rango.

4.3.2. Probabilidad de cambio para deportes y juegos.

Siguiendo la definición de arriba, presento en la [Figura 4.7](#) los valores de la probabilidad de cambio para los 12 sistemas complejos aquí estudiados (puntos en verde) en escala semilogarítmica motivados en los resultado que obtuvimos en el caso de la diversidad de rango. Nuevamente apreciamos una tendencia sigmoïdal en la forma que los valores de $p(k)$, claro está, a excepción de los sistemas que desde la diversidad de rango parecían anómalos. Vemos de nuevo una caída ahora más pronunciada para el caso de FIFA, también seguimos viendo la caída en algunos valores de k para OWGR. El caso de ESE sigue siendo inconcluyente, no podemos decir nada sobre ese sistema, al parecer los valores de $p(k)$ no adquieren una tendencia clara. Mientras que NASCAR-B vemos que siempre tiene valores muy altos, algo similar en el caso de NASCAR-W, salvo que un par de valores pequeños de k sí tienen un $d(k)$ pequeño.

Ahora bien, como en el caso de la diversidad de rango, dado que también se observa la tendencia sigmoïdal, propondremos el mismo modelo teórico para aproximar esta nueva medida, es decir, el modelo dado por la [Ecuación 4.3](#). En la [Figura 4.7](#) también presento los ajustes correspondientes de este modelo a los valores empíricos de la probabilidad de cambio (líneas rojas), así como los valores de los parámetros de ajuste μ y σ y el parámetro de ajuste R^2 .

Al igual que en el caso de la diversidad de rango tenemos que en esencia, la probabilidad de cambio es creciente. Es más probable que un rango cambie su ocupación

mientras k es más grande. Esta medida describe el aspecto anteriormente discutido y que la diversidad de rango no logra ver, que es cómo va cambiando la ocupación del rango a lo largo del tiempo. Aunque $p(k)$ tiene una interpretación distinta de $d(k)$ se mantiene la idea de que los rangos más grandes tienen menor estabilidad, mientras que los k pequeños tienen una mayor estabilidad. Por ejemplo, en el caso de FCWR-C vemos que en los rangos pequeños $p(k)$ tiene un valor pequeño en rangos pequeños, lo que implica que no puede pasar la situación hipotética anteriormente mencionada de que el Barcelona y Real Madrid van ocupando el primer lugar alternadamente tiempo a tiempo, los valores pequeños de $p(k)$ implican que por ejemplo el Barcelona se mantiene en el primer lugar un periodo de tiempo largo, y puede llegar a darse una alternancia con otro equipo pero pasa pocas veces y cuando pasa el nuevo equipo se mantiene ahora estable en el primer lugar por varios pasos en el tiempo. Es una dinámica que no se podía describir con $d(k)$ y es por ello que $p(k)$ es tan relevante.

Algo que a primera vista se nota es que la probabilidad de cambio es siempre más grande que la diversidad de rango, es decir, $p(k) \geq d(k)$, una de las gráficas donde más se nota es el caso de FCWR-C, donde se aprecia que la sigmoide de $p(k)$ se acerca más a 1 que $d(k)$. Para comprobar si esta observación resulta útil graficar la diferencia entre ambas medidas, es decir, $p(k) - d(k)$ y es justo lo que hago y muestro en la gráfica de la [Figura C.2](#) en el [Apéndice C](#). Vemos que en todos los sistemas pasa que $p(k) - d(k) \geq 0$ que comprueba nuestra aseveración. Además un efecto, tal vez importante, es que en los rangos más altos esa diferencia va decreciendo, mientras que en la mitad del dominio de rangos la diferencia se vuelve máxima, indicando que en los rangos intermedios la ocupación de los rangos es más variable conforme se avanza en el tiempo que el número de equipos que ocupan esos mismos rangos en la ventana temporal. Otra observación importante es que en los casos de OWGR, WSD y ATP las diferencias comienzan creciendo desde rangos pequeños y a partir de cierto valor comienzan a decrecer.

Acabamos de darnos cuenta que el modelo de la [Ecuación 4.3](#), el igual que la diversidad de rango, funciona relativamente bien para la probabilidad de cambio. En la [figura C.1](#) del [Apéndice C](#) se hace el mismo ejercicio que para la diversidad en donde reescalamos el eje de las abscisas de la forma $\frac{\log_{10}(k) - \mu}{\sigma}$ para poder comparar todos los sistemas en una sola gráfica. Con ese reescalamiento vemos que de nuevo, salvo el caso de ESE que se omite en la figura, se aglutinan alrededor de la sigmoide unitaria, es decir, al modelo de la [Ecuación 4.3](#) con $\mu = 0$ y $\sigma = 1$, dejando en evidencia un comportamiento genérico nuevamente pero para el caso de la probabilidad de cambio. El caso de ESE se omitió pues se vio que los valores promedios de $p(k)$ y con los rangos reescalados no seguían la misma tendencia, pero eso es evidente desde los valores de los parámetros de ajuste presentados en la [Figura 4.7](#) pues el valor de $\sigma = 0.09$ es muy pequeño comparado con la obtenida en el resto de los sistemas; adicionalmente se ve que el conjunto de puntos que grafican los valores empíricos de $p(k)$ no se pueden caracterizar de ninguna manera por las características de la base de datos y que ya se había discutido con anterioridad.

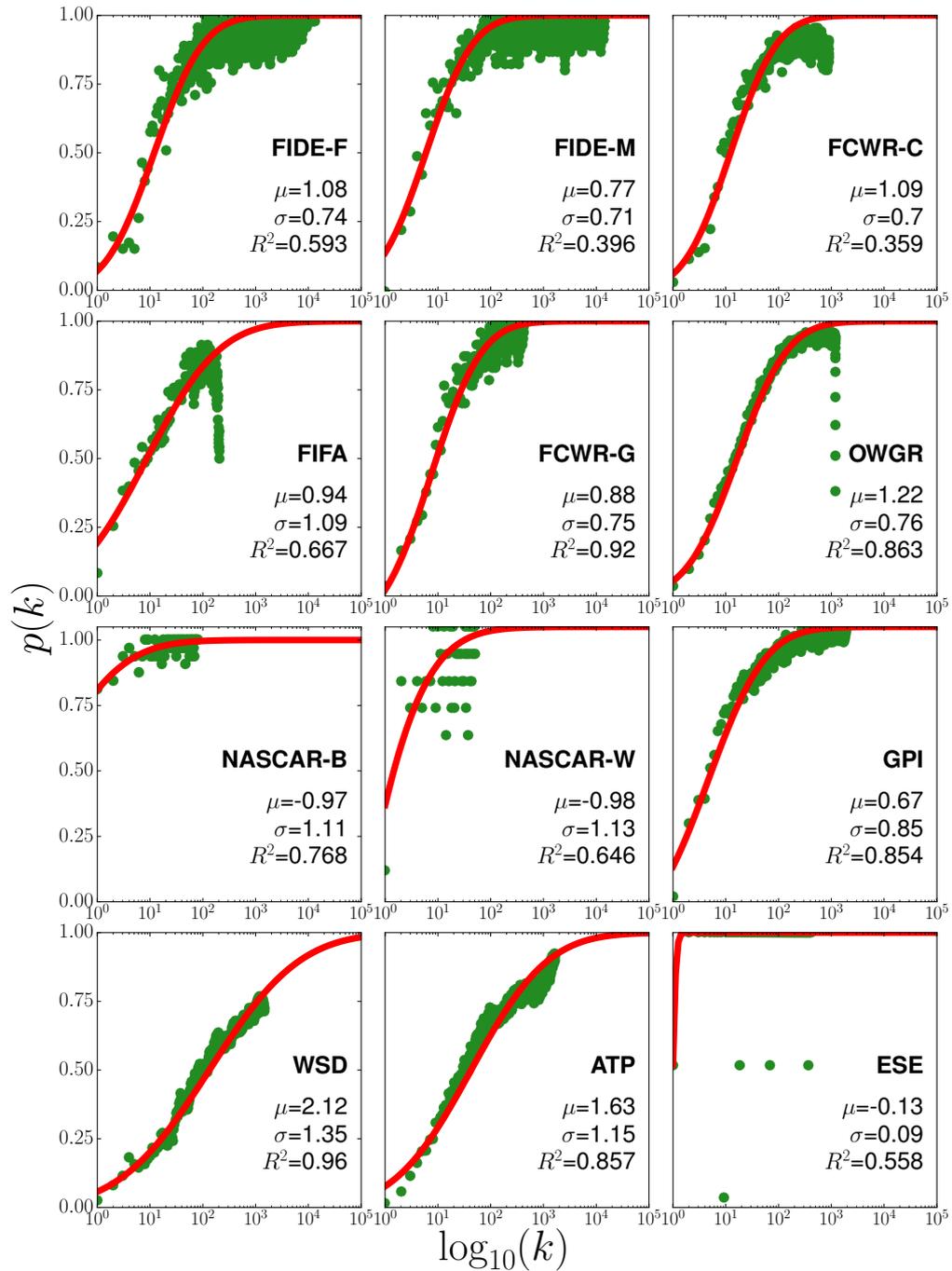


Figura 4.7: Probabilidad de cambio para deportes y juegos en escala semilogarítmica. Gráfica que muestra la probabilidad de cambio $d(k)$ para todos los conjuntos de datos (puntos verdes), así como los ajustes a Φ (líneas rojas). Incluimos los valores de μ y σ así como el parámetro de bondad de ajuste R^2 .

4.4. Entropía y complejidad de rango.

En el trabajo que colaboré sobre la evolución de los idiomas a diferentes escalas y que ya he citado anteriormente [47] se introducen dos conceptos nuevos llamados *entropía de rango* y *complejidad de rango*, los cuales se utilizaron para estudiar la dinámica de rango en el contexto de los idiomas a distintas escalas. Aquí también aplicaremos esas ideas y definiciones, pero antes hagamos una pequeña motivación para que quede clara la interpretación que haremos más adelante de estas dos nuevas medidas dinámicas.

4.4.1. Motivación.

La teoría de la información ha crecido mucho en los últimos años. Fue creada por Claude Shannon en 1948 para ser aplicada a las telecomunicaciones. Por ejemplo, si un mensaje era transmitido como una cadena de símbolos que formaban parte de un alfabeto de la forma $Y = x_0x_1x_2\dots$ donde cada uno de los x_i forman parte del alfabeto ya mencionado; cada uno de los símbolos tendrá una probabilidad $P(x)$ de aparecer en la cadena, en otras palabras, los símbolos más frecuentes en las transmisiones de información tendrán entonces las probabilidades más altas, mientras que los símbolos menos frecuentes tendrán, evidentemente, las probabilidades más bajas. El interés de Shannon radicaba en que quería encontrar una función capaz de cuantificar cuánta información es producida en un proceso. [48]

Platicaré de manera rápida cuál es la idea de Shannon para construir dicha función. Supongamos que tenemos un conjunto de posibles eventos cuyas probabilidades de ocurrencia son p_1, p_2, \dots, p_n . Estas probabilidades son conocidas y cada una de ellas están asociadas a los caracteres del alfabeto en cuestión. Shannon demostró que la función que a continuación pongo mide qué tan seguros estamos de cuál será el siguiente caracter que se obtendrá en la transmisión de cierta información con ciertos caracteres formando un alfabeto, otra manera de decirlo es qué tanta información se está produciendo: [48]

$$E = -K \sum_{i=1}^n p_i \log p_i \quad (4.12)$$

donde k es una constante positiva. Un ejemplo muy ilustrativo que se menciona en [48] es que si tenemos la cadena "0001000100010001...", se puede estimar la probabilidad de los caracteres 0 y 1 donde el alfabeto está conformado sólo por esos dos. Entonces podemos estimar las probabilidades de esos dos caracteres con esa cadena, siendo $P(0) = 0.75$ y $P(1) = 0.25$, por lo tanto si usamos $K = 1$ obtenemos que $E \approx 0.811$ y vamos a interpretar este resultado.

En [48] se habla sobre el concepto de *emergencia* que se refiere a las propiedades de un fenómeno que está presente ahora pero no antes. Es decir, que si estas propiedades están presentes se podría decir que es más difícil reproducir dicho fenómeno. Entonces, podemos considerar que la *emergencia* se puede considerar como el tránsito de un proceso que requiere poca información para ser descrito a un procesos que requiere más información para ser descrito. Es decir, si hay *emergencia* en un fenómeno que

está produciendo información y es justamente lo que la ecuación de Shannon cuantifica [Ecuación 4.12](#). Ahora bien, la función E tiene valores en el intervalo $[0, 1]$, si $E = 1$ quiere decir que nueva información emergerá y cuando $E = 0$ no emergerá nueva información. Tenemos que la *emergencia* implica un incremento en la información que es análogo a la entropía y desorden. Para que E se encuentre en el rango correcto de valores, la constante K debe ser adecuada para esos fines. En el caso de que el alfabeto tenga h elementos, entonces el valor adecuado de la constante de renormalización debe ser:

$$K = \frac{1}{\log h} \quad (4.13)$$

Otro concepto importante a considerar es el de complejidad C que se interpreta como el balance entre cambio (caos) y estabilidad (orden). Como la entropía E es una medida del desorden, entonces la medida de estabilidad debe ser lo contrario, es decir $S = 1 - E$. Por lo que la complejidad debe tener las siguientes características: [\[48\]](#)

- El rango de sus valores debe estar en el intervalo $[0, 1]$.
- $C = 1$ si y sólo si $S = E$.
- $C = 0$ si y sólo si $S = 0$ o $E = 0$.

Para que satisfaga esas condiciones se propone:

$$C = 4 \cdot S \cdot E = 4 \cdot E \cdot (1 - E) \quad (4.14)$$

El factor 4 para normalizar a la función. Tenemos entonces que mientras E mide el desorden, C mide el equilibrio entre desorden y estabilidad. Mi propósito es llevar estas ideas al contexto de la dinámica de rango que compete a este trabajo.

4.4.2. Definiciones.

Si tenemos un sistema complejo jerárquico que cuenta con información sobre el ranqueo de sus elementos para T rodajas temporales o pasos en el tiempo. Tenemos que para cada k hay un conjunto de elementos del sistema (en nuestro caso jugadores/equipos) que ocuparon ese rango en algún momento dentro de la ventana de tiempo disponible. Este conjunto de elementos será el diccionario (del que se habló hace un momento) el diccionario será entonces $X(k) = \{X(k, t_1), X(k, t_2), \dots, X(k, T)\}$ donde ya vimos anteriormente que $X(k, t)$ simboliza al elemento que ocupa el rango k en el tiempo t . La probabilidad de aparición de un elemento X del diccionario en el rango k se calcula como el número de apariciones del ese elemento en el rango k entre el número total de rodajas temporales, ese cálculo resulta en la probabilidad p_X . Por lo tanto, siguiendo la definición de entropía anterior, definiremos a la entropía de rango como:

$$E(k) = -K \sum_{X \in X(k)} p_X \log p_X \quad (4.15)$$

donde K es la constante de renormalización dada en la [Ecuación 4.13](#) y en este caso su valor sería:

$$K = \frac{1}{\log |X(k)|} \quad (4.16)$$

pues en este caso el número de elementos en el diccionario es justamente la cardinalidad del conjunto $X(k)$. De manera análoga, podemos definir a la complejidad de rango motivados por la definición de entropía de rango y estaría dada por:

$$C(k) = 4 \cdot E(k) \cdot (1 - E(k)) \quad (4.17)$$

Y son justamente las definiciones que utilizaremos para nuestros sistemas aquí descritos. La entropía de rango $E(k)$ cuantifica el desorden de información que tenemos respecto a la ocupación del rango k , es decir, nos dice si tenemos mucha información o no sobre la ocupación del sistema (con esto quiero decir si tenemos muchos datos para poder determinar cuáles elementos ocuparán el mismo rango en una rodaja temporal extrapolado). La complejidad de rango $C(k)$ cuantifica qué tanto equilibrio hay entre el desorden y la estabilidad en la ocupación de dicho rango.

4.4.3. Entropía y complejidad de rango para deportes y juegos.

Con las definiciones para entropía de rango dada en la [Ecuación 4.15](#) y para complejidad de rango dada en la [Ecuación 4.17](#) calculamos dichas cantidades para las 12 bases de datos que representan las disciplinas aquí estudiadas. Ahora tenemos 2 medidas dinámicas adicionales que enriquecerán nuestro entendimiento de la dinámica de rango.

En la [Figura 4.8](#) se presentan las gráficas correspondientes a la entropía de rango para las 12 disciplinas en cuestión. Ahora las gráficas se presentan en escala lineal para ambos ejes. Observamos que en la mayoría de los deportes/juegos la entropía de rango tiene valores muy cercanos a 1, lo que indica que este tipo de sistemas tiene poca predictibilidad, pues la ocupación de los rangos no es estable. Esto es una sorpresa pues el concepto de estabilidad que adquirimos con la diversidad de rango y la probabilidad de cambio, pues vimos que eran más variables los rangos más altos en cuanto a su ocupación en distintos criterios: la diversidad de rango sólo ve cuántos elementos distintos ocupan un rango k a lo largo de todos los pasos temporales; la probabilidad de cambio cuantifica las veces que cambia el elemento que estaba ocupando el rango k a lo largo de las rodajas de tiempo. La entropía de rango, como ya discutimos, nos dice cuánta información tenemos sobre la ocupación del rango para poder hacer una predicción de qué jugador/equipo ocupará ese rango k en un tiempo posterior. Recordemos que si $E(k)$ es muy próximo a 1, quiere decir que la forma en que los jugadores/equipos ocupan el rango k es muy caótica. A diferencia de $d(k)$ y $p(k)$, la entropía de rango nos dice que no sólo los rangos altos son poco caracterizables, sino que también los rangos medios tienen una entropía considerable, no hay forma, al menos a primera vista, de predecir fácilmente qué elementos adquirirán dichos rangos.

En algunos sistemas como los FIDE, WSD y ATP se aprecia que los rangos pequeños tienen una entropía baja, lo que implica que hay mayor estabilidad en esos rangos, pero

en el resto de los sistemas la entropía de los rangos pequeños incluso tienen una entropía por encima de 0.5. Ésta es una medida más fina que el resto porque, en efecto, vimoa con la diversidad y probabilidad de cambio que pocos elementos ocupan los rangos bajos y que son pocas las ocasiones en que se presentan cambios en la ocupación de dichos rangos y a pesar de ello la información que tenemos es considerable para no poder predecir de manera sencilla qué elemento ocupará un rango bajo en tiempos posteriores.

En la [Figura 4.9](#) se aprecian las gráficas de la complejidad de rango en escala lineal para ambos ejes. Recordemos que $C(k)$ no es lo contrario a $E(k)$, sino que cuantifica qué tanto equilibrio hay entre el desorden y estabilidad. $C(k) \approx 1$, es máxima cuando $E(k) \approx 0.5$. Es bastante curioso que en los resultados que obtuvimos la complejidad de rango es muy grande en rangos bajos para los FIDE, OWGR, FIFA, FCWR-G, WSD y ATP. No hay un desorden completo, hay un balance entre ese desorden medio y estabilidad. Lo que nos lleva a pensar que la estabilidad de la que hablábamos en $d(k)$ y $p(k)$ para rangos bajos es en realidad un punto medio entre estabilidad y aleatoriedad. En el resto de los sistemas la complejidad es pequeña, y como vimos en la figura [Figura 4.8](#) no se debe a que va adquiriendo mayor estabilidad, sino que la entropía se incrementa poco a poco para rangos más altos. Vemos que el caso de ESE tiene los valores mínimos de $C(k)$ y máximos para $E(k)$ en todo el dominio de rangos, que junto con los resultados obtenidos para la diversidad y la probabilidad de cambio nos indican que este sistema es matemáticamente aleatorio, no hay forma de caracterizarlo y es por ello que obtuvimos los resultados anómalos para este sistema; su dinámica no es analizable debido a la poca estadística que se tiene para dicha disciplina.

4. DINÁMICA DE RANGO PARA DEPORTES Y JUEGOS: DIVERSIDAD, PROBABILIDAD DE CAMBIO, ENTROPÍA Y COMPLEJIDAD.

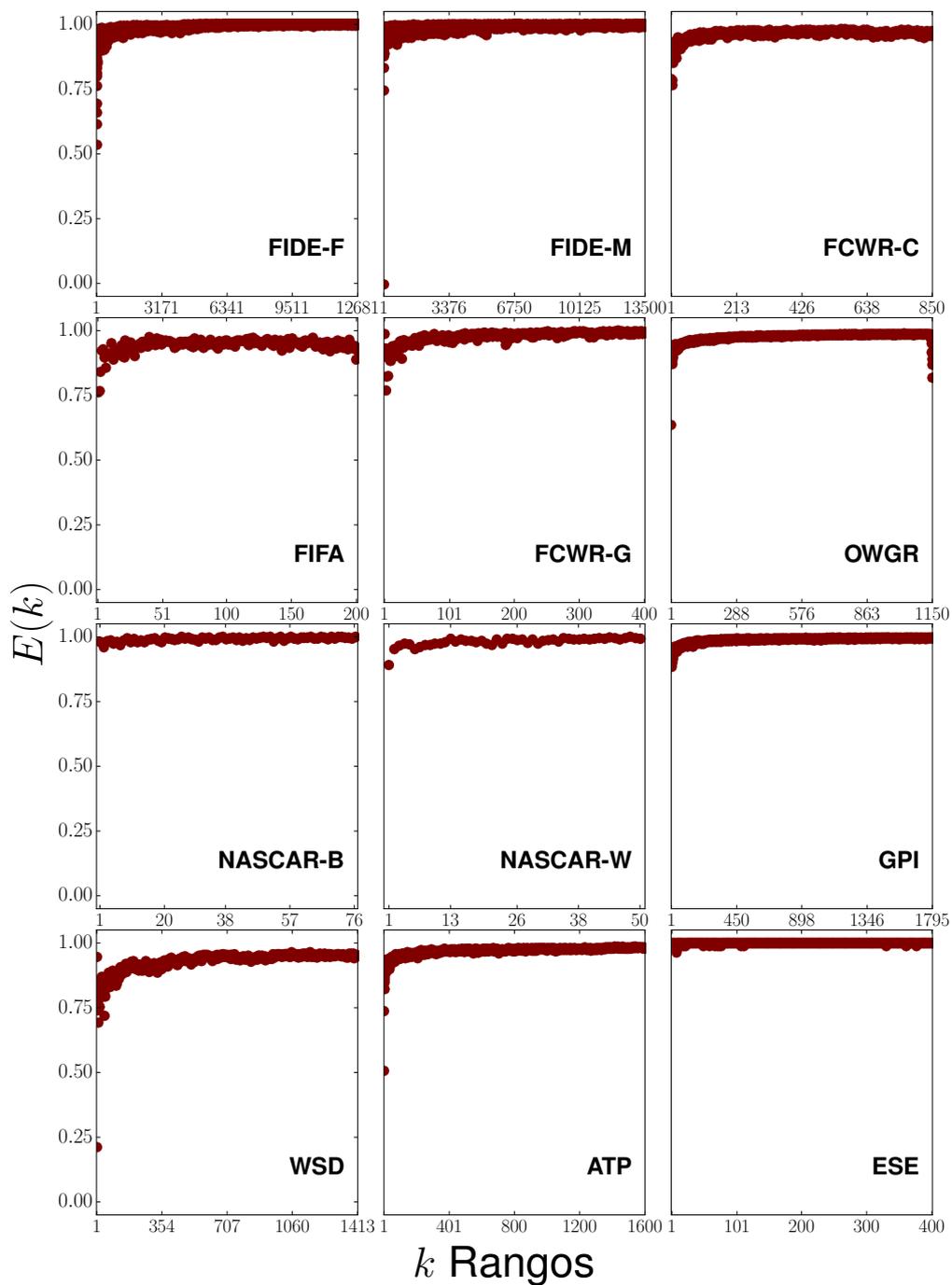


Figura 4.8: Entropía de rango para deportes y juegos. Esta medida dinámica es calculada de acuerdo a lo que definimos en la [Ecuación 4.15](#), observamos la entropía de rango es muy grande (con valores cercanos a 1) para la mayoría de los rangos en todos los deportes y juegos aquí considerados.

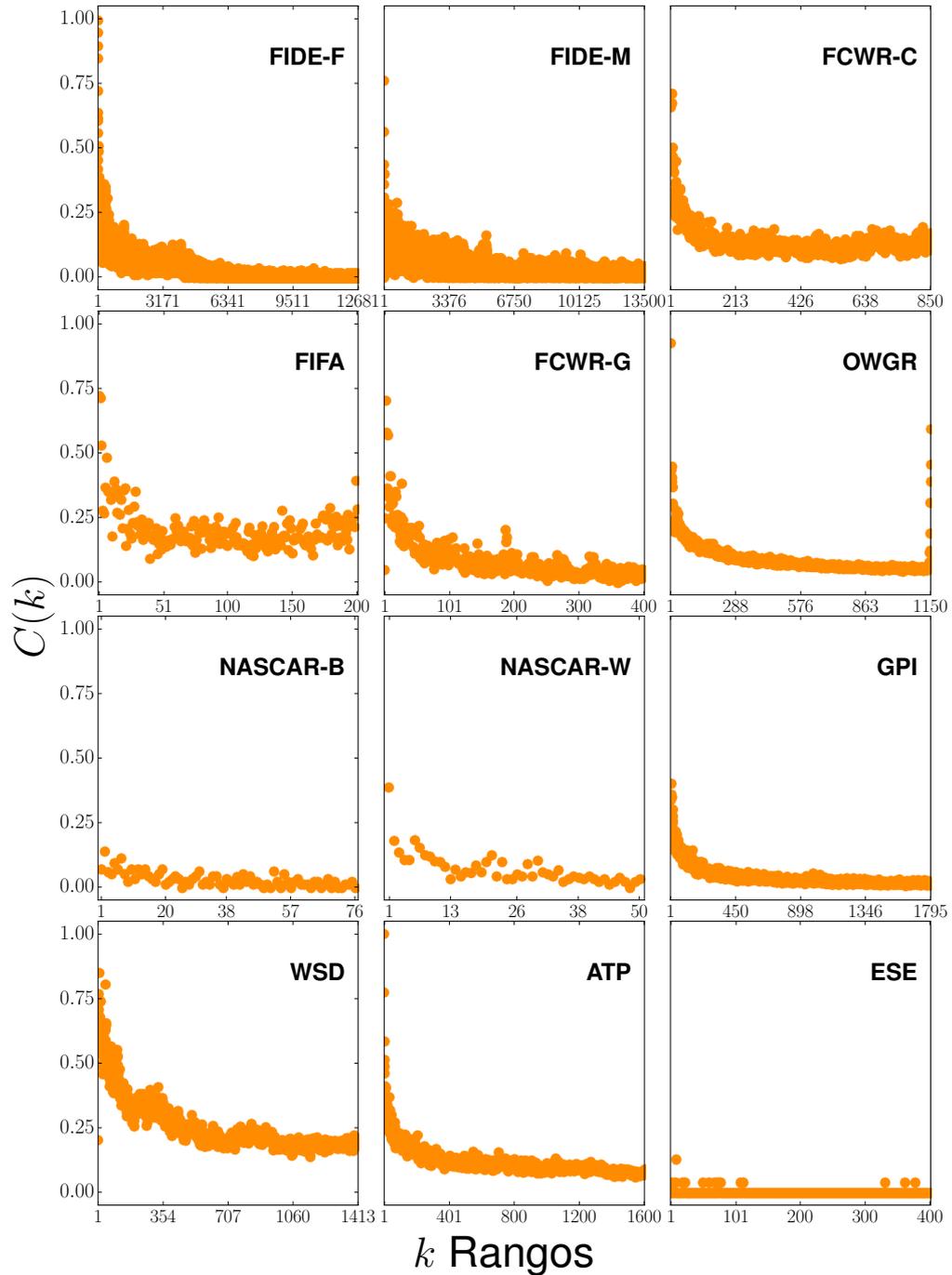


Figura 4.9: Complejidad de rango para deportes y juegos. Esta medida dinámica es calculada de acuerdo a lo definido en la [Ecuación 4.17](#). Observamos que la complejidad es alta para los rangos pequeños en la mayoría de los sistemas, las clara excepciones son las de siempre NASCAR-B, NASCAR-W y ESE.

4.5. Discusión final para la dinámica de rango.

Las medidas dinámicas trabajadas en este capítulo captan distintos aspectos de la evolución de los sistemas jerárquicos aquí estudiados. Para ejemplificar un poco más las diferencias y complementariedad entre las medidas dinámicas, presento a continuación dos ejemplos que ya había utilizado anteriormente pero que ponen en evidencia estas afirmaciones.

1. El Barcelona y Real Madrid se turnan el primer lugar fecha a fecha, es decir, la primera fecha es de Barcelona, la siguiente del Real Madrid, la siguiente del Barcelona y así sucesivamente. La base cuenta con un total de $T = 850$ semanas, es decir, rodajas temporales. Tenemos entonces:
 - $d(1) = 2/850$, pues sólo esos dos equipos ocupan el primer lugar.
 - $p(1) = 1$, pues siempre va cambiando la ocupación del primer lugar, es turnada.
 - Como el Barcelona aparece la mitad de las ocasiones, $p_{Barcelona} = 1/2$, y de la misma forma $p_{Madrid} = 1/2$, por lo tanto, $E(k) = 1$, lo que indica que tenemos mucha información, no hay estabilidad y un indicador de que no se puede predecir qué elemento ocupará el primer lugar en un tiempo posterior.
2. El Barcelona ocupa el primer lugar la primera mitad de las rodajas temporales y el Real Madrid ocupa la segunda mitad de las rodajas temporales. La base cuenta con un total de $T = 850$ semanas, es decir, rodajas temporales. Tenemos entonces:
 - $d(1) = 2/850$, pues sólo esos dos equipos ocupan el primer lugar.
 - $p(1) = 1/425$, pues sólo en una ocasión hay un cambio del elemento que ocupa el lugar.
 - Como el Barcelona aparece la mitad de las ocasiones, $p_{Barcelona} = 1/2$, y de la misma forma $p_{Madrid} = 1/2$, por lo tanto, $E(k) = 1$, igual que en el caso anterior. Es similar, porque ambos elementos tienen la misma probabilidad de aparecer.
3. Un equipo diferente ocupa cada una de las rodajas temporales, es decir, el primer lugar siempre es ocupado por elementos distintos. La base cuenta con un total de $T = 850$ semanas, es decir, rodajas temporales. Tenemos entonces:
 - $d(1) = 1$, pues 850 elementos diferentes ocupan el primer lugar.
 - $p(1) = 1$, pues siempre hay cambio en la ocupación del primer lugar.
 - Como todos los elementos tuvieron la misma probabilidad de aparecer, es decir $p_{X(1,t)} = 1/850$, y esto $\forall t = t_1, t_2, \dots, t_T$, tenemos que $E(1) = 0$, es el caso contrario a los ejemplos anteriores.

Como desde el inicio de este trabajo se mencionó, nuestro deseo es entender la formación jerárquica de sistemas complejos y su evolución a través del tiempo. Tal y como se describió en el [Capítulo 2](#), en este trabajo nos enfocamos en 12 sistemas complejos y que son justamente sistemas asociados a deportes y juegos. Es claro que los elementos de dichos sistemas interactúan de diferente manera sistema a sistema, pues cada deporte o juego tiene reglas distintas y compiten por medio de eventos de distinta naturaleza, como torneos o encuentros aislados. Lo importante es que compiten obteniendo más puntos y los valores de éstos dan lugar a una formación jerárquica. En el [Capítulo 4](#) vimos que parecía muy complicado caracterizar los espaguetis de los jugadores/equipos, por lo que pasamos a estudiar la dinámica de ocupación de los rangos a lo largo del tiempo construyendo tres medidas: diversidad de rango, probabilidad de cambio y entropía de rango.

Aunque pudimos estudiar de manera satisfactoria la dinámica de ocupación de los rangos en el [Capítulo 4](#), durante un desarrollo analítico para justificar el origen de la diversidad de rango hecho en la [Subsección 4.2.4](#). Aunque la derivación analítica mencionada de la diversidad de rango fue sólo para justificar cualitativamente el comportamiento de dicha medida, una consecuencia bastante interesante es que la forma en que los elementos de un sistema complejos jerárquico van ocupando los rangos parece estar ligada a caminatas aleatorias. Con esta idea podríamos modelar la forma en que los elementos se mueven en la tabla de los rankeos y si podemos reproducir los resultados obtenidos en el [Capítulo 4](#) tendríamos una primera aproximación a lo que ocurre en la realidad respecto a cómo se van ocupando los rangos, más allá de ver únicamente su dinámica. Claramente, para crear modelos de ese tipo, nos tenemos que basar en los observado para las medidas dinámicas de los sistemas reales.

En este capítulo presentaremos dos modelos que justamente pretenden reproducir, al menos cualitativamente, los resultados del [Capítulo 4](#) que son justamente las medidas dinámicas de los sistemas reales, creando sistemas sintéticos (no existentes con un comportamiento matemático especificado) cuyas evoluciones se aproximen a lo visto en los sistemas reales. Trataremos dos modelos: el modelo del caminante aleatorio, que está motivado en el resultado tan interesante proveniente de la derivación analítica de la diversidad de rango, y el modelo nulo que es un modelo bastante interesante pues ignorará las características específicas de los sistemas para reproducir los resultados ya

vistos, por eso el título de *nulo*. Ambos modelos podrán darnos una visión más amplia sobre la forma en que los sistemas jerárquicos se configuran y que pueden dar lugar a nuevas regularidades.

5.1. Modelo del caminante aleatorio

El modelo del caminante aleatorio para aproximar la formación jerárquica de un sistema complejo ya ha sido utilizado anteriormente en el contexto de la evolución de los idiomas, en el artículo que hemos citado en numerosas ocasiones anteriormente. [2] Presentaré primero que nada el modelo y posteriormente justificaré las razones por las que parece adecuado, justificándolo de la manera más formal posible.

Como discutimos en [Capítulo 4](#), detrás de la formación jerárquica de los sistemas, tenemos la posible presencia de procesos aleatorios (caminatas aleatorias) que se distribuyen de acuerdo a la distribución normal. Por lo que conjeturamos que el cambio en el rango de los elementos cambian paso a paso temporal con una perturbación aleatoria de los rangos de los elementos que obedece a una distribución gaussiana. Ahora bien, inicialmente podríamos proponer los rangos, al tiempo t , fueran perturbados por un número aleatorio generado por una distribución gaussiana, si un rango al tiempo t fuera denotado por k_t entonces la perturbación estaría dada por:

$$k_t \rightarrow k_t + G(0, \sigma)$$

donde $G(k, \sigma)$ representa un número aleatorio generado de acuerdo a una distribución gaussiana con media μ y desviación estándar σ . En nuestro caso consideraremos $\mu = 0$, pues es sólo una perturbación al rango original y la desviación estándar del ruido se tendría que determinar y allí está la clave del modelo. Vamos a determinar la naturaleza de esa desviación estándar a partir de lo que ya conocemos. Vimos en el [Capítulo 4](#) y en las figuras del [Apéndice B](#) que los elementos que ocupan rangos bajos tienen menor variabilidad a lo largo del tiempo respecto a los elementos que ocupan rangos altos. Entonces, todo indica que la desviación estándar del ruido gaussiano tendría que ser proporcional al rango al tiempo discreto t , es decir, proporcional a k_t . De esa forma, el número aleatorio generado será más grande, en general, para rangos altos y éste se verá perturbado en mayor medida, mientras que para los rangos bajos, la perturbación gaussiana tendrá menor valor; ésto capura la idea de lo que se observa respecto a la evolución de los jugadores/equipos en las tablas de ranqueo. Por tanto proponemos que:

$$\sigma = k_t \hat{\sigma}$$

donde $\hat{\sigma}$ es una constante por determinar. Más adelante explicaré cómo determinar dicha constante. Por el momento, estamos capurando la idea de menor variabilidad a rangos pequeños y mayor variabilidad a rangos grandes. Para comenzar a implementar el modelo comenzamos con un conjunto de datos con N elementos (este número debe coincidir con la N del sistema real que se intenta reproducir. Los elementos están ordenados de acuerdo a su correspondiente rango k_1 , que es el rango al primer tiempo.

Tenemos T rodajas temporales, y a cada rodaja la nombramos como t_1, t_2, \dots, t_T . Cada rango k_{t_i} corresponde al tiempo t_i .

Lo que aquí tratamos es un modelo del caminante aleatorio Gaussiano invariante de escala, ya que un miembro con rango k_t , en un tiempo discreto t , se convierte en el rango k_{t+1} de acuerdo al siguiente procedimiento que llamamos el *procedimiento de reordenamiento* y detallo a continuación:

1. Sea $t \in \{t_1, t_2, \dots, t_T\}$ el tiempo correspondiente a una rodaja temporal arbitraria en el conjunto de tiempos que se tienen. A ese tiempo, los elementos tienen sus respectivos k_t . Definimos una variable auxiliar l_{t+1} , que llamaremos pre-rango, en el tiempo $t + 1$ con la relación:

$$l_{t+1} = k_t + G(0, k_t \hat{\sigma}), \quad (5.1)$$

donde $G(0, k_t \hat{\sigma})$ es un número aleatorio generado con una distribución Gaussiana con desviación estándar $k_t \hat{\sigma}$ y valor medio 0; el valor adecuado de $\hat{\sigma}$ ya se debió haber determinado anteriormente. Esto significa que la variable aleatoria l_{t+1} tiene una distribución con un ancho proporcional a k_t , y por ende tendrá, para k_t pequeña, pequeños cambios también. Entonces para el tiempo $t + 1$ se calculan los pre-rangos de cada uno de los elementos.

2. Una vez que los valores de los pre-rangos l_{t+1} han sido establecidos, los ordenamos de acuerdo a su magnitud. Este nuevo orden, da nuevos rangos, es decir, el valor de k en el tiempo $t + 1$, es decir, k_{t+1} .

Este procedimiento está descrito y se realiza para una de las rodajas temporales t . El proceso de reordenamiento está definido para tiempos arbitrarios. Y como tal, el modelo completo va más allá, pues dejamos un parámetro libre en esta descripción que es $\hat{\sigma}$. El modelo del caminante aleatorio es aplicar el proceso de reordenamiento de los elementos tiempo a tiempo y en orden. Estamos simulando una dinámica y automáticamente generando una base de datos sintética. Una vez implementado el modelo, podremos calcular todas las medidas dinámicas que hemos definido en este trabajo y compararlas con los sistemas reales. Lo difícil del modelo es encontrar el parámetro $\hat{\sigma}$ que genere el sistema sintético que se asemeje más al real. A continuación describimos el procedimiento por el cual se busca este parámetro.

5.1.1. Descripción de la implementación del modelo.

El implementar el modelo del caminante aleatorio que acabamos de describir es, como ya se explicó, usar el proceso de reordenamiento tiempo a tiempo. Lo importante es encontrar el parámetro $\hat{\sigma}$ que asemeje al sistema sintético al sistema real lo mejor posible. Para encontrar el valor de $\hat{\sigma}$ adecuado simplemente se aplicará el modelo del caminante para varios valores de esta variable, el conjunto de datos generado correspondiente a un valor específico de $\hat{\sigma}$ que más se asemeje a los datos reales será el valor del parámetro que estamos buscando. La implementación requiere de mucho procesamiento de

datos de manera computacional y a continuación lo describimos a detalle. Se queremos simular uno de los sistemas complejos que estamos considerando en este trabajo (como WSD, ATP o FIFA) entonces contamos el número de rodajas temporales T con las que cuenta su respectiva base de datos. El conjunto de tiempos que tiene la base está dada, nuevamente, por el conjunto $\{t_1, t_2, \dots, t_T\}$, en cada rodaja temporal se tendrá N . Con estas cantidades que definen al sistema tenemos entonces, fijamos el valor de $\hat{\sigma} = 0.001$:

1. Generamos una lista de N elementos diferentes ordenados (pueden ser lo que sea, pero que sean N). Este orden induce un rango al tiempo t_1 , es decir, k_1 para cada uno de los elementos.
2. Aplicamos el procedimiento de reordenamiento T veces, el cual ya se explicó con anterioridad. Los números aleatorios sumados a los rangos ahora estarán generados por $G(0, k_i, \hat{\sigma})$ con el valor de $\hat{\sigma}$ que hemos fijado.
3. Calculamos la diversidad de rango correspondiente a este sistema que se ha generado, lo llamamos sistema sintético.
4. Calculamos la R^2 de la diversidad de rango sintética obtenida para este sistema artificial con la $\Phi_{\mu, \sigma}(\log k)$ (Ecuación 4.3) obtenida del ajuste de este modelo con la diversidad de rango calculada del sistema real.
5. Guardamos esta R^2 obtenida correspondiente a el valor de $\hat{\sigma}$ fijado inicialmente a este proceso.
6. Al valor de $\hat{\sigma}$ le sumamos 0.001

El proceso anterior lo repetimos 1500 veces. De modo que estamos generando sistemas artificiales con el modelo del caminante aleatorio para valores de $\hat{\sigma}$ desde 0.001 hasta 1.5 con una resolución de 0.001, este rango es el más conveniente pues realicé pruebas con distintos valores para todos los sistemas y encontré que éste es el más adecuado para hacer nuestro ajuste. Entonces, tenemos un conjunto de valores de R^2 , cada una de ellas pertenece a valores distintos de $\hat{\sigma}$. De ese conjunto vemos cuál tiene el valor más grande (la más cercana a 1 de todas) y vemos a qué $\hat{\sigma}$ corresponde, éste es el parámetro buscado. Lo que hicimos básicamente es aplicar el modelo del caminante aleatorio para diferente valores de $\hat{\sigma}$, de hacer ésto se crea una base de datos correspondiente a un sistema ficticio para cada $\hat{\sigma}$, vimos cuál de esos sistemas generados tiene la diversidad de rango más parecida a la del sistema real. Así es como con el modelo intentamos reproducir los sistemas reales.

Claramente, escogimos a la diversidad de rango como la medida dinámica a comparar entre los sistemas reales y los sintéticos para obtener la mayor similitud posible entre ellos. Iba a ser muy complicado hacer la comparativa con la entropía o complejidad de rango, pues no tenemos un modelo teórico funcional que se asemeje al comportamiento de los valores empíricos. Y escogimos a la diversidad de rango sobre la probabilidad de cambio pues en el [Capítulo 4](#) hicimos una posible derivación teórica de la misma, dando a lugar a una correlación entre esta medida y las caminatas aleatorias.

También existe una justificación adicional por la cual se utiliza la distribución gaussiana para reproducir los cambios en el rango de los elementos para reproducir una dinámica similar a lo observado en los sistemas reales. En el artículo [2], que estudia la evolución del uso de palabras en distintos idiomas, los rangos no se trabajan como valores discretos, sino $k = 1/f$, el inverso de la frecuencia de aparición de las palabras cada cierto tiempo (resolución temporal). En el mismo trabajo se estudió la distribución de los saltos relativos de los rangos de una rodaja temporal a otra, es decir, de $[k_{t+1} - k_t]/k_t$, que es justamente tratar de ver cómo es que se distribuyen las perturbaciones a los rango tiempo a tiempo en el modelo del caminante. Se encontró que la distribución Lorentziana parece modelar mejor esas distribuciones que una Gaussiana, sin embargo se ve una no concordancia de los valores en las colas de la Lorentziana respecto de los valores empíricos, por lo que se optó trabajar con las distribuciones gaussianas para generar las perturbaciones a los rangos. En el Apéndice C se proporciona la Figura C.3 que ejemplifica lo aquí explicado para el grupo de palabras en inglés. Motivados por esta idea optamos por seguir el mismo modelo propuesto en [2] perturbando a los rangos por medio de números aleatorios generados con distribuciones gaussianas.

5.1.2. Comparativa entre los sistemas generados con el modelo del caminante aleatorio y los sistemas reales.

En la Figura 5.1 mostramos la diversidad de rango para sistemas sintéticos con el mismo número de elementos observados en la Figura 4.5, pero generados con el modelo del caminante aleatorio. La primera impresión que se tiene de los resultados obtenidos es que que este par de conjuntos de datos son similares cualitativamente, aunque claras diferencias revelan que el modelo es insuficiente para justificarlo completamente de manera cuantitativa. En todos los casos salvo ESE y los dos de NASCAR, vemos que las simulaciones reproducen de manera aceptable el comportamiento que ya habíamos notado: tenemos que la diversidad de rango es monótona creciente y más aún, que la forma sigmoideal sí reproduce el comportamiento de la sigmoide obtenida con los sistemas reales correspondientes. En la Figura 5.1 se proporciona el valor de $\hat{\sigma}$ utilizado en el modelo del caminante aleatorio para generar los sistemas sintéticos; así como los valores de los parámetros de ajuste del modelo sigmoideal (Ecuación 4.3) a la diversidad a estos sistemas artificiales, la bondad de ajuste R^2 de dichos ajustes también se incluyen.

Vemos que la diversidad de rango de los datos simulados en el caso de OWGR se parece mucho a la de los datos reales en casi todo el dominio, algo similar pero en menor medida ocurre para el caso de FCWR-G. Como era de esperarse, no se pueden sacar conclusiones del caso de ESE, pero incluso con una base de datos simulada con las mismas características de la original, no induce una diversidad de rango que nos indique algo interesante. Para todos los casos podemos observar que para rangos bajos, la diversidad de rango de los datos sintéticos siempre se mantienen por debajo a sus correspondientes bases reales. El hecho de que tanto la diversidad de rango simulada como la empírica tienen un comportamiento en forma de sigmoide sugiere que los cambios en el rango en sistemas reales debe ser resultado de una enorme número de procesos multiplicativos, tal y como justificamos en el Capítulo 4. Sin embargo, la no concordancia entre el modelo

y los datos como se observa en la [Figura 5.1](#) muestra que no todo el comportamiento característico del proceso empírico es capturado por nuestro modelo, y un estudio más profundo es necesario, sobre todo porque la mayor discrepancia de los datos entre sí se observa para los rangos bajos.

Las bases de datos sintéticas son sistemas generados con el modelo del caminante aleatorio, y que en esencia tiene las mismas características que los sistemas reales: el número de elementos por lista de ranqueo N y el número de rodajas temporales. Por lo que al igual que sus correspondientes empíricas, se les puede calcular las otras tres medidas dinámicas que introdujimos en el [Capítulo 4](#). En la [Figura 5.2](#). Evidentemente la concordancia entre los sistemas reales y los sintéticos es mucho menor si lo vemos desde el punto de vista de la probabilidad de cambio. Nuevamente, se ve que los valores de $p(k)$ de los datos simulados es menor que para los reales en el dominio de los rangos bajos, esto es consistente también por lo observado en $d(k)$ pues los cambios de ocupación en los rangos bajos es menor en el caso sintético respecto al real. En los casos de WSD y ATP notamos que las diferencias son evidentes.

En el [Apéndice C](#) adjunto también en las figuras [C.4](#) y [C.5](#) que tienen las gráficas de la entropía y complejidad de rango, respectivamente, para los datos sintéticos generados con este modelo del caminante aleatorio y comparándolos con los observados en los datos reales. Notamos que en estas dos medidas se encuentran mayores similitudes entre sí. Lo que implica que el modelo reproduce bastante bien las condiciones de qué tanta información se tiene sobre la ocupación de ciertos rangos.

Algo que no considera este modelo, y que puede ser la fuente de la diferencias entre los datos reales y los simulados, es que no se considera el factor de cerradura (Ω) del sistema que discutimos en el [Capítulo 4](#) y cuyo cálculo se realiza por lo explicado en la [Ecuación 4.2](#). Recordemos que Ω cuantifica qué tantos elementos van entrando o saliendo de las listas de ranqueo a lo largo de las rodajas temporales. Evidentemente, los sistemas simulados tienen todos $\Omega = 1$, pues la lista de ranqueo nunca cambia, los mismos elementos tienen perturbación en sus rangos y reordenados de acuerdo a lo que expliqué del modelo anteriormente. Éste factor no considerado en el modelo puede ser una de las razones por las que dicho modelo es incompleto y no reproduce fielmente los datos empíricos.

Este modelo es muy interesante, pues captura la idea de poca variabilidad a rangos bajos, contrario a lo que se ve en rangos altos; y justamente, esta idea reproduce cualitativamente lo observado en los datos empíricos. A continuación explicaré un modelo mucho más sencillo y que aún así nos arroja resultados interesantes.

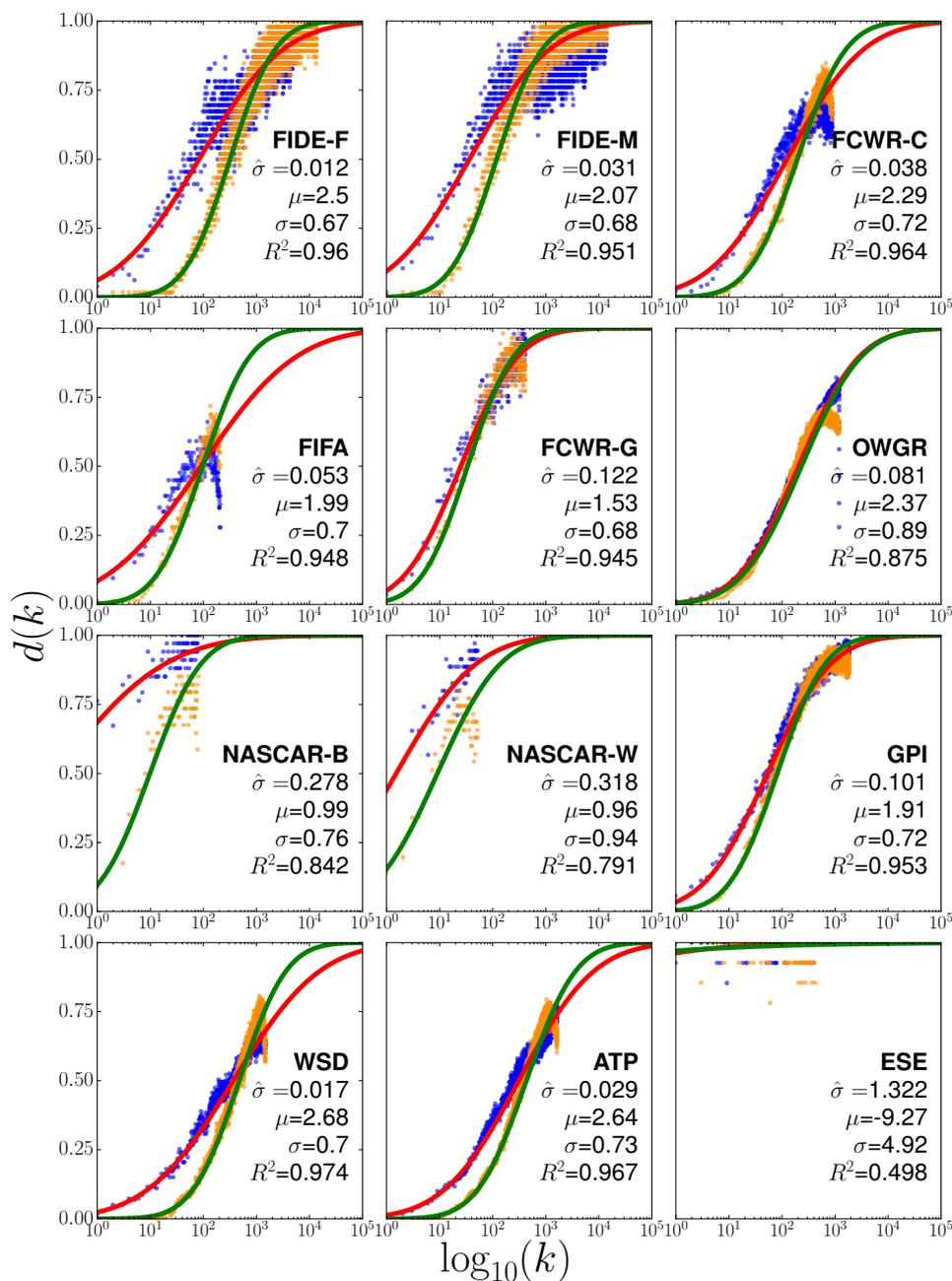


Figura 5.1: Comparación entre las diversidades de rango empíricas y simuladas con el Modelo del caminante aleatorio. Gráfica que muestra la diversidad de rango $d(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ , σ para los datos simulados. El modelo del caminante aleatorio parece reproducir cualitativamente las diversidades de rango observadas en todos los deportes y juegos considerados aquí, a pesar de las claras diferencias que revelan que este modelo es ineficiente para obtener resultados satisfactorios cuantitativamente.

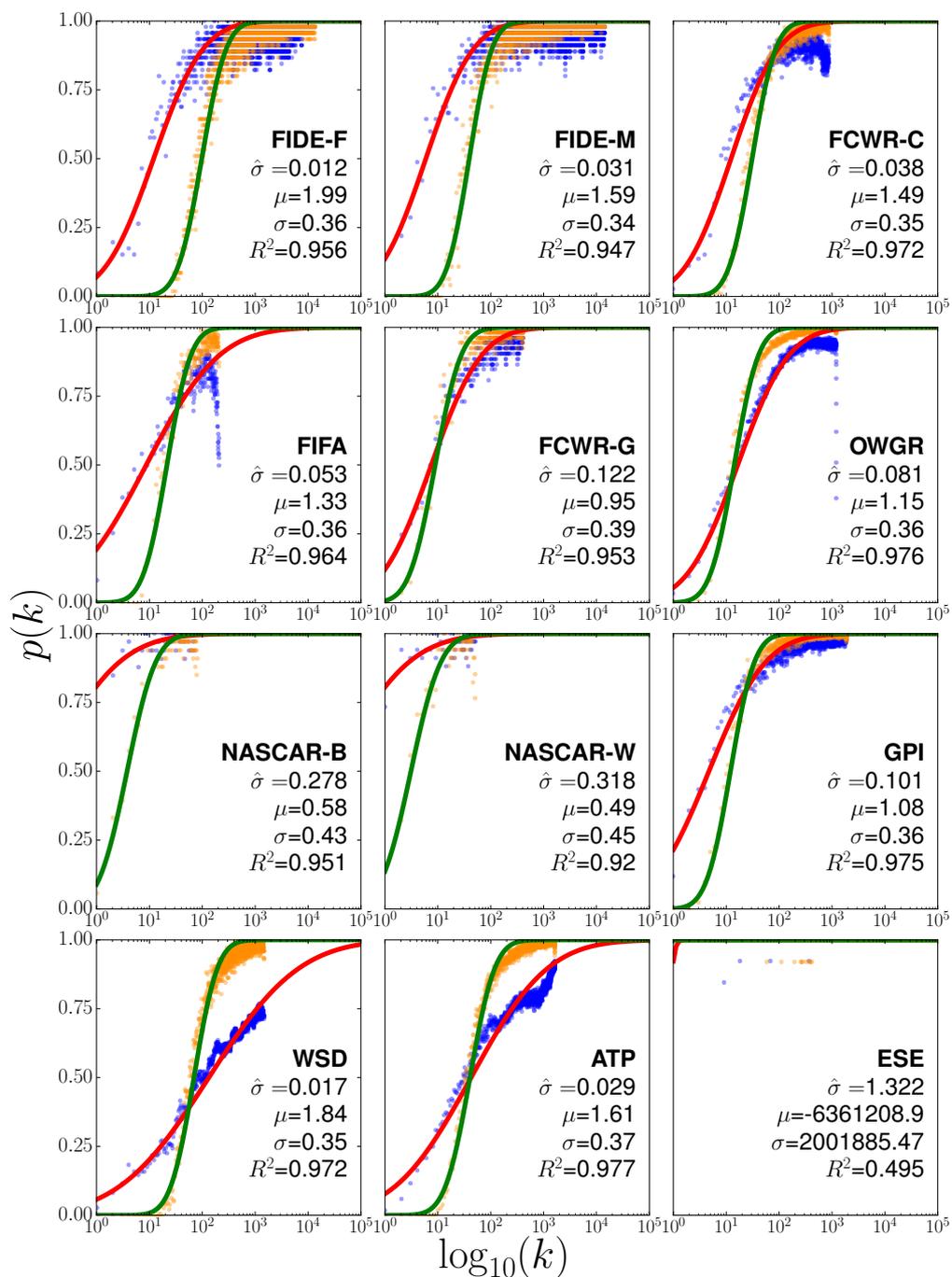


Figura 5.2: Comparación entre las probabilidades de cambio empíricas y simuladas con el Modelo del caminante aleatorio. Gráfica que muestra la probabilidad de cambio $p(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ , σ para los datos simulados.

5.2. Modelo Nulo.

Antes de comenzar a describir este nuevo modelo quiero hacer algunas advertencias. El modelo no se justificará formalmente, pero sí se motivará un poco; el modelo principal de este trabajo de tesis es el del caminante aleatorio, mientras que el modelo nulo se introduce como una posibilidad adicional para explicar la interacción entre elementos de un sistema complejo jerárquico. Por la forma en que se planteará el modelo, también se pueden hacer aproximaciones matemáticas que darían origen a mayor entendimiento de la naturaleza aleatoria de estos sistemas. El modelo nulo presentado aquí es sólo una primera aproximación al mismo, pues se puede mejorar de muchas maneras como veremos más adelante. Con todo esto en mente, ahora podemos pasar a explicar en qué consiste.

En analogía al modelo del caminante aleatorio, primero definiré un proceso de reordenamiento de la lista de ranqueo paso a paso. Como siempre, como simularemos un sistema con N elementos, entonces comenzamos generando una lista de N elementos ordenados, ese ordenamiento induce un rango k a cada uno de los elementos de dicho sistema sintético. Paso a paso del proceso de reordenamiento se encuentra ilustrado en la [Figura 5.3](#):

1. Imaginamos que cada uno de los N elementos se encuentran dentro de N casillas ordenadas de acuerdo al rango que tienen, denotaremos a cada uno de los elementos ordenados como a_k con $k \in \{1, 2, \dots, N\}$.
2. Escogemos aleatoriamente un rango $R \in \{1, 2, \dots, N\}$ y el elemento a_R correspondiente a ese rango y es removido de su correspondiente casilla. Los elementos que se encuentren a la derecha de esa casilla se recorren una posición hacia la derecha para cubrir el espacio vacío. Evidentemente, la última casilla quedará desocupada. El rango a remover se escoge aleatoriamente de tal manera que todos los rangos tengan la misma probabilidad de ser removidos, es decir, cualquier rango tiene $1/N$ de probabilidad de ser removido. Si el elemento elegido fue el a_N entonces no se recorre nada.
3. Ahora bien, insertamos de manera aleatoria al elemento removido a_R de alguna de las formas siguientes:
 - Entre los espacios de dos de algunas de las $N - 1$ casillas ocupadas consecutivas. Digamos que en el espacio entre la casilla del elemento a_{M-1} y el elemento a_M , entonces los elementos a la izquierda de a_M y el mismo a_M se recorren una casilla a la derecha y a_R ocupa la casilla que le correspondía a a_M
 - Puede ser insertado justo antes de la primer casilla, entonces todos los elementos desde el primero al último se recorren a la derecha y a_R ahora ocupará la primer casilla.
 - También puede ser introducido en la última casilla, la cual había sido desocupada. En este caso, nada más se hace.

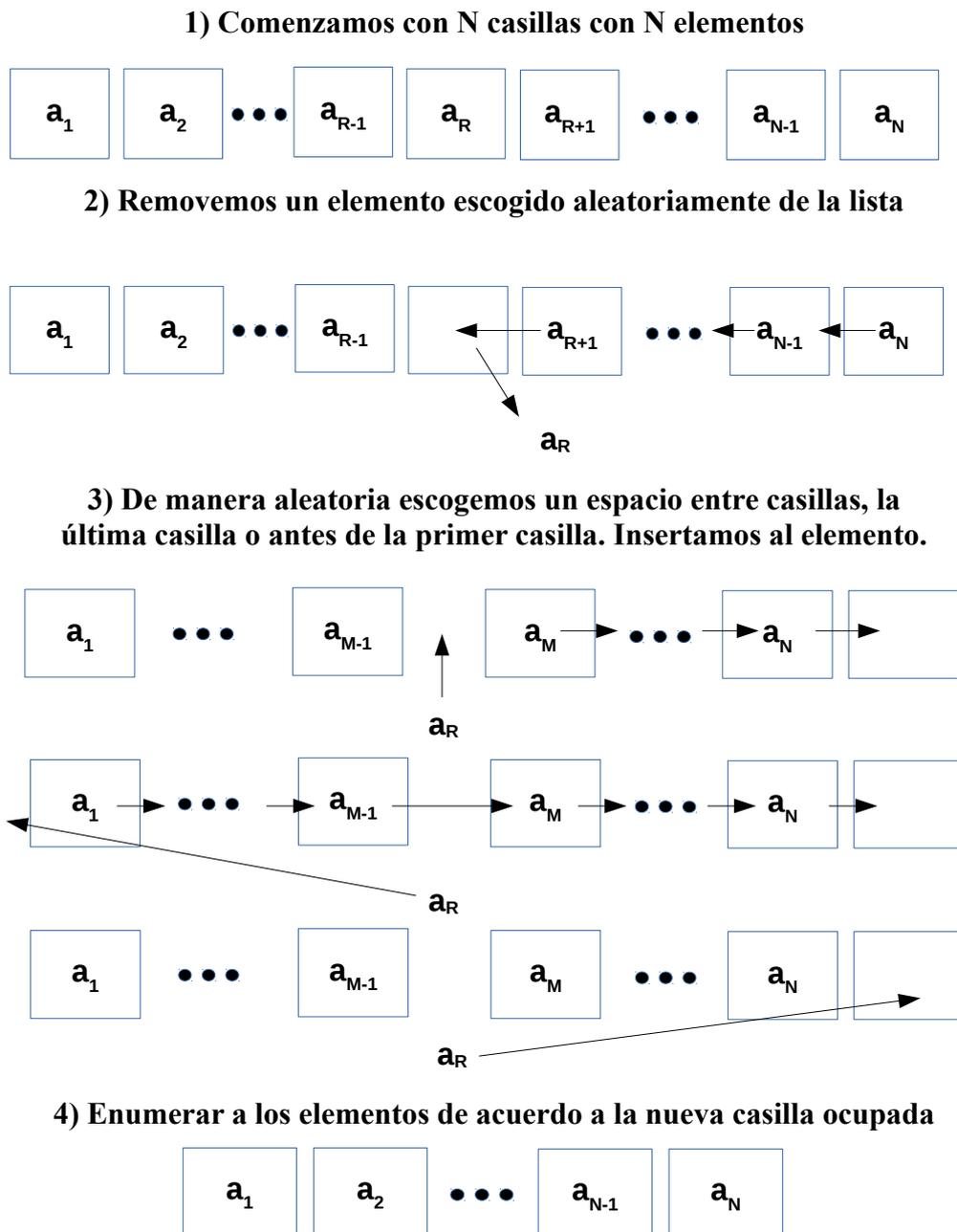


Figura 5.3: Descripción de la implementación del modelo nulo para una rodaja temporal. Diagrama que ilustra de manera detallada cada uno de los pasos que se llevan a cabo en el proceso de reordenamiento para el modelo nulo.

Alguno de estos casos, que son N , también tienen la misma probabilidad de ser escogidos, es decir, con probabilidad $1/N$.

4. Una vez que el elemento a_R fue insertado de nuevo a la lista de las casillas, todos los elementos se reenumeran de acuerdo a la casilla ocupada. Este nuevo índice es su nuevo rango y termina el proceso de reordenamiento.

El proceso de reordenamiento consiste entonces de sacar a un elemento escogido al azar, y reingresarlo a la lista en un espacio al azar, al principio o al final. El proceso es demasiado sencillo, incluso más que el correspondiente al modelo del caminante aleatorio. Ahora bien, el modelo nulo consistirá en repetir este proceso muchas veces, pero ¿qué tantas veces?.

5.2.1. Descripción de la implementación del modelo.

Ahora bien, como en el caso del modelo del caminante aleatorio, este proceso se realizará varias veces, pero repito la pregunta ¿cuántas veces y cómo generar el sistema que más se parezca al sistema real que se quiere simular?. Entonces, como en el modelo anterior, queremos simular los sistemas que hemos presentado en este trabajo de tesis, tales como WSD, ATP o FIFA. Digamos que el sistema que queremos simular tiene un total de T rodajas temporales. El conjunto de tiempos que tiene la base dada estará dada entonces, como ya habíamos descrito, por el conjunto $\{t_1, t_2, \dots, t_T\}$ y en cada rodaja temporal tendremos N jugadores/equipos.

El modelo consistirá básicamente en repetir el proceso de reordenamiento muchas veces, demasiadas, nombraremos a esas demasiadas veces como $T_\tau \gg T$. Entonces, lo que haremos es ir generando bases de datos con T rodajas temporales a partir de saltos de distinta longitud a lo largo de las T_τ , es decir, con saltos de 1 en 1, T veces, con saltos de 2 en 2, T veces y así sucesivamente; entonces haremos saltos de $\Delta\tau$ en $\Delta\tau$ por T veces para generar una nueva base de datos. Lo que haremos es comparar las bases de datos generadas, cada una proveniente de los saltos de longitud $\Delta\tau$ para escoger la base que mejor se asemeje con la base real que se deseaba simular. A continuación explico los detalles de este proceso.

Fijaremos $\Delta\tau_{max} = N$ como la longitud máxima de saltos en la megabase de datos generada por hacer T_τ veces el proceso de reordenamiento; esta elección está planteada así para considerar que tal vez en la iteración de N procesos de reordenamiento haya ocurrido que el primer lugar fue bajando de lugar en lugar hasta alcanzar el último o viceversa, es un caso límite y motiva esta elección de $\Delta\tau_{max}$. Siendo así que tendremos que fijar a $T_\tau = \Delta\tau \cdot T$, esto para que se puedan generar bases de datos de T rodajas temporales a partir de los saltos de $\Delta\tau = 1, 2, \dots, \Delta\tau_{max} = N$. Una vez hecho el proceso de reordenamiento T_τ veces entonces procedemos de la siguiente forma:

1. Generamos una base de datos similar a la que se quiere simular, T rodajas temporales y con N elementos por ranqueo, distinta haciendo saltos de valor en valor de $\Delta\tau = 1, 2, \dots, \Delta\tau_{max}$.

2. Calculamos la diversidad de rango correspondiente a este nuevo sistema sintético generado.
3. Calculamos la R^2 de la diversidad de rango sintética obtenida para este sistema artificial con la $\Phi_{\mu,\sigma}(\log k)$ (Ecuación 4.3) obtenida del ajuste de este modelo con la diversidad de rango calculada del sistema real.
4. Guardamos esta R^2 obtenida correspondiente al valor $\sigma\tau$ que indujo la respectiva base de datos.

Esto se hace iteradamente para el valor de $\Delta\tau = 1, 2, \dots, \Delta\tau_{max}$, y de las R^2 calculadas, se calcula cuál es la máxima (la más cercana a 1) y de ella se obtiene el sistema que más se asemeja al real. Es así con este procedimiento que obtenemos la mejor simulación para nuestro sistema empírico. A continuación se muestran los resultados obtenidos.

5.2.2. Comparativa entre los sistemas generados con el modelo nulo y los sistemas reales.

Una vez que encontramos el sistema sintético más similar al sistema empírico, podemos calcular la diversidad de rango del sistema artificial y compararlo con $d(k)$ del real. En la Figura 5.4 presento los valores de la diversidad de rango para el sistema simulado (puntos naranja) y los valores empíricos de la diversidad de rango para los sistemas reales (puntos azules). También se muestran los respectivos ajustes a la función sigmoide Ecuación 4.3, se agregan los valores de los parámetros para la sigmoide ajustada a la $d(k)$ del sistema sintético, es decir, μ y σ . Se aprecia también la bondad de ajuste R^2 entre los datos sintéticos y el ajuste del modelo sigmoide a ellos. En las mismas gráficas se aprecia el valor $\Delta\tau$ que viene del modelo nulo y que nos indica qué sistema sintético de todos los generados fue el que mejor simuló los datos reales.

Ahora analicemos los resultados obtenidos. Justo como observamos a lo largo del Capítulo 4, los sistemas de NASCAR y ESE no son tan adecuados para poder caracterizar una dinámica. Notamos de manera interesante que, contrario a lo observado en el modelo del caminante aleatorio, $d(k)$ para valores de k pequeños de los sistemas sintéticos quedan por arriba de los valores de la diversidad para los datos reales. Ahora la variabilidad de los rangos pequeños es superior que para los datos reales y evidentemente para los generados con el caminante aleatorio. Afortunadamente, se mantiene la característica monótona desde los rangos pequeños simulando de manera aceptable $d(k)$ para ciertos valores en el dominio. Sin embargo, en casi todos los casos se aprecia que hay una caída abrupta de los valores de $d(k)$ para valores grandes de k , algo muy parecido a lo que apreciamos en la Figura 4.5 en el caso de FIFA y donde conjeturamos que se debía a la cerradura Ω del sistema. Claramente, en los sistemas generados por el modelo nulo tenemos que $\Omega = 1$ en todos los casos, pues siempre realizamos el mismo proceso de reordenamiento sobre los mismos elementos, no permitimos la entrada de nuevos elementos o la salida de los que ya se encuentran en consideración.

Las caídas tan pronunciadas no nos permiten afirmar que logramos reproducir cualitativamente los resultados empíricos observados en los sistemas reales. Aunque no

lo justificaremos formalmente, es importante notar que el proceso de reordenamiento implica que al mover un elemento, toda la cola o los elementos posteriores al removido cambian su rango, mientras que al reingresarlo, los elementos a posteriores al lugar de reingreso regresan a su rango actual y los anteriores cambiaron, ésto en caso de que ingrese en una posición posterior de la que fue removido. Por otro lado si es reinsertado en un rango anterior del que fue removido, todos los que se movieron inicialmente regresan a su posición pero algunos otros anteriores cambian de rango. Lo que está pasando es que estamos registrando intuitivamente **mayor variabilidad en los rangos intermedios**, y es justo lo que se registra en los resultados obtenidos y que se aprecian en la [Figura 5.4](#).

También podemos calcular el resto de las medidas dinámicas vistas en el [Capítulo 4](#) de los sistemas sintéticos. En la [Figura 5.5](#) se aprecian los valores de la probabilidad de cambio del sistema sintético (puntos naranjas) y los valores de la misma medida para los sistemas reales (puntos en azul), así como las sigmoides ajustadas a cada uno de ellos y los valores de los parámetros de ajuste de Φ para la $d(k)$ de los datos sintéticos, el valor de R^2 para ese ajuste y la $\Delta\tau$ que obtuvimos de la utilización del modelo nulo. Nuevamente se aprecia que en la mayoría de los casos hay una caída abrupta en los valores de $d(k)$ para k grandes, lo que provoca que la semejanza entre los sistemas disminuya, al menos cualitativamente.

Finalmente, también podemos calcular los valores de la entropía de rango y complejidad de rango para los sistemas sintéticos generados con el modelo nulo. En las figuras [C.6](#) y [C.7](#) que se encuentran en el [Apéndice C](#) se tienen los valores de la entropía de rango y complejidad de rango, respectivamente, de los datos simulados (naranja) y los datos provenientes de los sistemas reales (puntos azules). Vemos de nuevo que no se aprecia una diferencia significativa. Lo cual, al igual que en el caso del modelo del caminante aleatorio, parece no haber diferencia en cuanto a la información que se tiene respecto a la ocupación de los rangos. El caso de discrepancia más significativa se aprecia para WSD.

El modelo nulo es bastante interesante pues no tomamos consideración alguna de lo observado en el [Capítulo 4](#) respecto a que a rangos bajos, menor variabilidad y a rangos altos, mayor variabilidad. Simplemente quitamos y ponemos elementos, es un proceso aleatorio, y al menos para los rangos bajos se presenta una similitud no ignorable con los datos reales. De hecho, no podemos escapar ante la sorpresa de que FIFA es bastante bien reproducida por el modelo nulo, y la alta cerradura del sistema FIFA debe estar jugando un papel importante, logramos reproducir la caída para los rangos altos en $d(k)$ y $p(k)$.

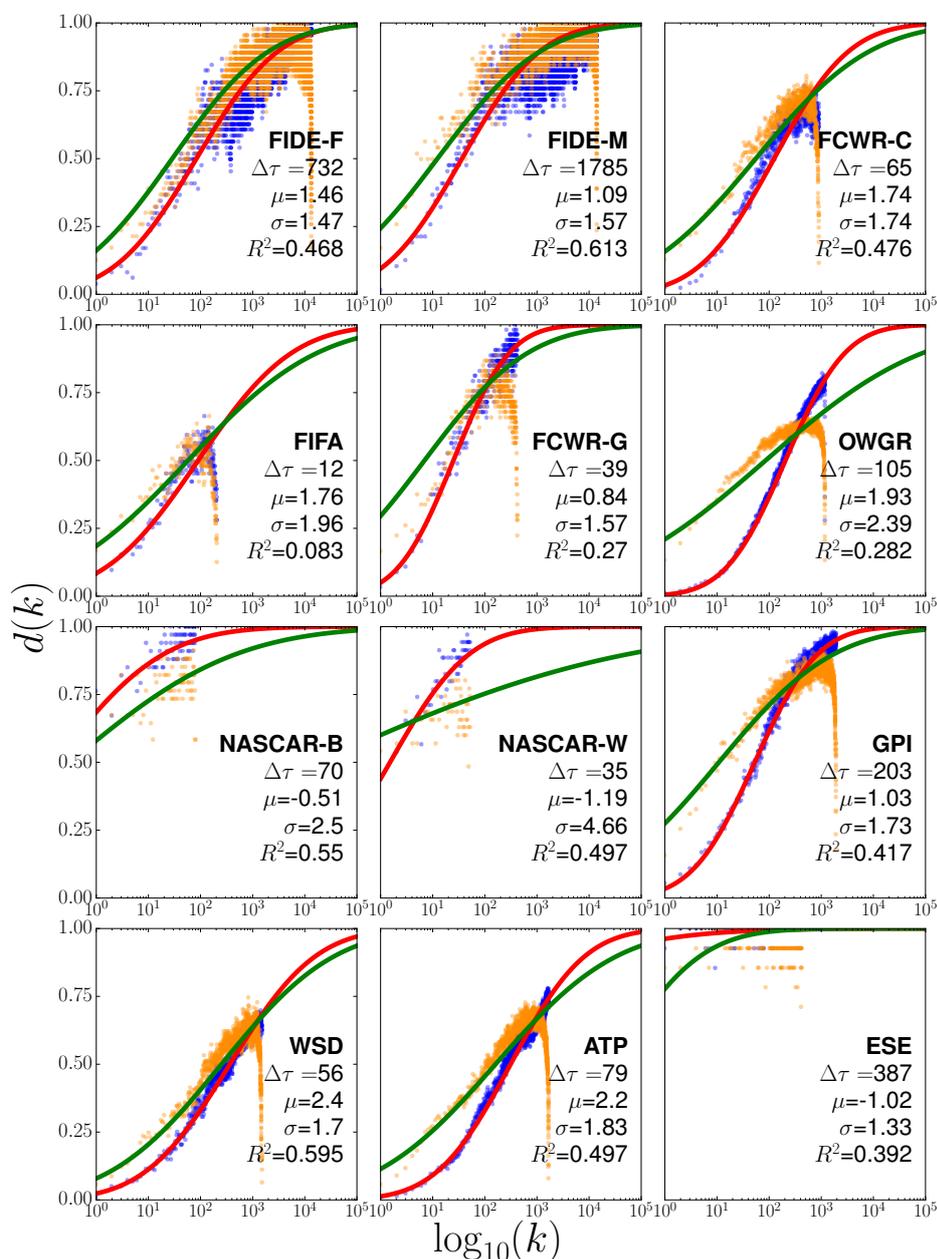


Figura 5.4: Comparación entre las diversidades de rango empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la diversidad de rango $d(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ , σ para los datos simulados. El modelo nulo parece reproducir cualitativamente las diversidades de rango observadas en todos los deportes y juegos considerados aquí, pero en menor medida que con el caso del modelo del caminante aleatorio. Se registran caídas más abruptas a rangos altos y $d(k)$ de los datos simulados quedan por encima de la $d(k)$ de los datos empíricos para los rangos bajos.

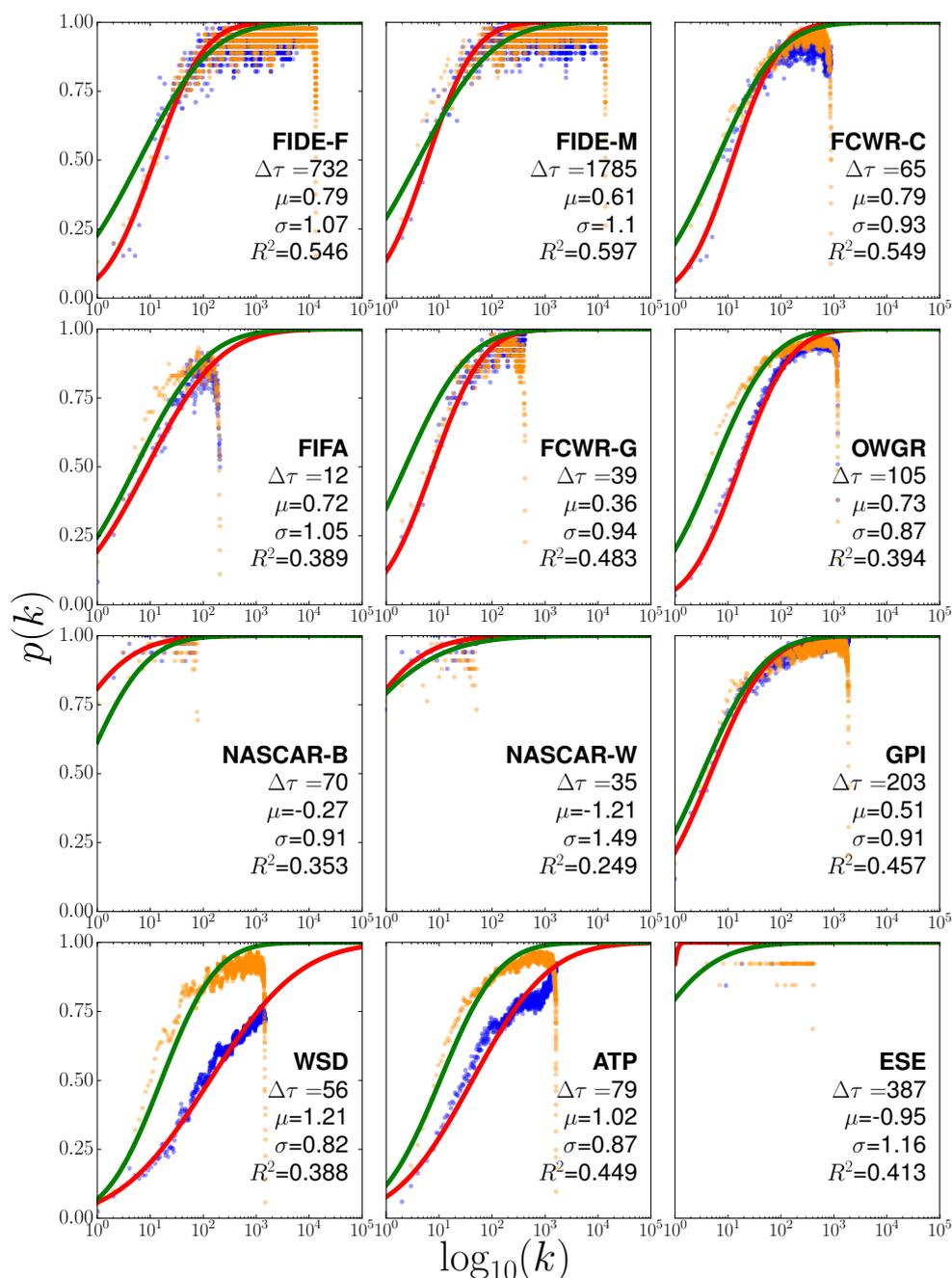


Figura 5.5: Comparación entre las probabilidades de cambio empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la probabilidad de cambio $p(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados), así como el ajuste a Φ (línea roja/verde para los datos empíricos/simulados, respectivamente). También incluimos los valores de μ , σ para los datos simulados.

5.3. Modelo Nulo versus Modelo del caminante aleatorio

Lo natural es preguntarse ahora ¿cuál de los dos modelos es el mejor?. Pienso que no hay respuesta correcta a esa pregunta. Queda claro que ambos modelos no logran capturar a la perfección la esencia de los sistemas aquí estudiados. Por un lado, vimos que en el caso del modelo del caminante aleatorio provoca que la diversidad de rango y la probabilidad de cambio sea menor a la observada en los datos reales. Esta discrepancia evidencia la incompletez del modelo, más allá de hacer que los sistemas se parezcan cualitativamente, necesitamos que cuantitativamente haya una mayor similitud.

El caso del modelo nulo también fue interesante, pues los sistemas simulados siempre arrojaban menor variabilidad del sistema para los rangos altos, y eso es algo que no observamos en los sistemas reales, salvo el caso de FIFA, el cual fue bastante bien reproducido por el modelo nulo, pues para para FIFA desde el capítulo pasado hicimos hincapié al hecho de que hay una caída abrupta en los valores de $d(k)$ en rangos altos, algo que hicimos más evidente cuando graficábamos a la diversidad de rango en escala lineal, [Figura 4.4](#). Una posible modificación al modelo nulo podría llegar a producir simulaciones más fieles a lo observado, si además tomamos en cuenta el parámetro de cerradura de los sistemas. Por ejemplo, si tomamos en cuenta a todos los elementos que entraron o salieron de las listas de los rankings, y que el modelo considere la interacción de ese total, y al momento de compararlo con el sistema real, hacer el corte de la base de datos proveniente de la simulación a sólo N elementos por rodaja temporal. Es posible que registremos una subida de $d(k)$ a rangos altos (dentro del dominio original $k \in [1, N]$), es por ello que el modelo nulo aquí presentado parece estar incompleto y requiere del uso de un parámetro más que conjeturamos está relacionado con la cerradura Ω de los sistemas.

La razón por la que el modelo del caminante aleatorio no presenta caídas abruptas en los valores de $d(k)$ para rangos altos es porque nosotros obligamos a que no fuera así. La perturbación de los rangos $G(0, k_t \hat{\sigma})$ está relacionada a la magnitud del rango que se quiere perturbar, entonces mientras más alto sea el rango mayor será la perturbación, y como FIFA nos demostró, eso no se presenta necesariamente en toda clase de sistemas. Sería interesante recopilar más bases de datos que representen sistemas con una cerradura $\Omega \approx 1$ para comprobar esta hipótesis. De ser así, el modelo del caminante aleatorio podría no ser útil en todos los casos y el modelo nulo podría representar una mejor alternativa, claro está, incluyendo ese factor de cerradura. Incluso el modelo del caminante aleatorio se podría modificar de la misma forma que se mencionó para el modelo nulo.

Aunque ninguno de los modelos reproduzca satisfactoriamente el comportamiento de los sistemas, en el aspecto cuantitativo, revelan que la estructura jerárquica de muchos sistemas pueden estar ligados a procesos aleatorios, y genéricos a lo largo de los sistemas, lo cual pone en evidencia una regularidad. A pesar de que los sistemas empíricos aquí estudiados hacen interactuar a sus miembros de maneras distintas, a gran escala se encuentra que esas diferencias desaparecen cuando se estudia su evolución temporal.

Conclusiones.

La competición y desempeño heterogéneo son característicos en elementos de muchos sistemas en el ámbito biológico, social y económico. A pesar del hecho de que estos sistemas muestran grandes variaciones en las definiciones de sus constituyentes y la relevancia de la interacción entre ellos, aún se debe analizar cuándo el surgimiento de una estructura jerárquica está determinada mayormente por las peculiaridades de cada fenómeno, o si hay mecanismos de estratificación comunes a la evolución temporal de muchos sistemas. En esta tesis hemos explorado esta noción considerando un conjunto de datos relativamente simplificado y controlado dominados por la competición: deportes y juegos humanos, donde las reglas de dedicación y medidas de rendimiento están bien definidas, en contraste con, digamos, el ranqueo de físicos (la pregunta de quién es el mejor físico debería tener una respuesta ambigua, por decir). Esto nos permite caracterizar el surgimiento de una jerarquía heterogénea comparando comportamiento temporal de los rangos de individuos y equipos en actividades bien determinadas. Explícitamente, nosotros analizamos las propiedades estadísticas de las distribuciones de rango en 12 bases de datos asociadas a deportes y juegos, cada una de ellas con diferente número de miembros y reglas para calcular los puntajes (y, por ende, rangos).

Primero analizamos la forma en que se distribuyen los puntajes asignados a los elementos de los sistemas, definiendo la distribución de rango. Comparamos las distribuciones de rango (de todas las rodajas temporales para todas las bases) con 5 modelos, encontramos que la ley de Zipf (modelo $m_1(k)$) no provee un ajuste satisfactorio para los datos empíricos. Incluso el modelo más genérico m_4 (una combinación de las distribuciones Gamma y Beta) que tiende a ofrecer mejores ajustes, no siempre es el mejor. Pero la consistencia de los resultados es satisfactoria porque a cuando $m_2(k)$ o $m_3(k)$ funcionaban, esto ocurriendo siempre de manera exclusiva, entonces $m_4(k)$ también lo hacía, pues las distribuciones Beta y Gamma son casos particulares de $m_4(k)$. Lo sorprende fue que en algunos casos, el modelo $m_5(k)$ también era adecuado, teniendo la interesante implicación de que, en los sistemas para los cuales funcionó, existen dos regímenes entre los elementos que no influyen entre sí para la evolución del sistema al cual pertenecen. Para concluir objetivamente se utilizaron dos bondades de ajuste: el índice p de Kolmogorov-Smirnov y el coeficiente de determinación R^2 , siendo el índice p el más robusto, pues no sólo indica si de manera funcional un conjunto de puntos son bien aproximados por un modelo, sino que nos indica si realmente un conjunto de datos

6. CONCLUSIONES.

se distribuye de acuerdo a una distribución de probabilidad teórica específica. Por lo que para concluir lo ya mencionado, se utilizó dicha bondad de ajuste.

Sin embargo, no pudimos notar una regularidad común a todos los sistemas, en algunos, incluso en algunos casos, el índice p nos hizo concluir que ninguno de los modelos era el apropiado. De esto se puede deducir que necesitamos o un modelo más general o modelos adicionales que capten el comportamiento de los sistemas para los cuales no se encontró la distribución adecuada. Y aunque estudiamos la distribución de rango a lo largo de las rodajas temporales para todos los sistemas, no pudimos encontrar una manera general de describirlos. Pero sí se obtuvieron resultados que hacen notar particularidades de la forma en que se asignan puntajes: como que para los de la ley doble Zipf hay dos regímenes de elementos, que en algunos se asignan puntajes que empiezan a decaer muy rápido para los que tienen los rangos más altos, como se constata con los sistemas adecuadamente descritos por el modelo $m_3(k)$. Pero al final de cuentas, la distribución de rango sólo nos proporciona un aspecto estático, es ciega ante la evolución en el tiempo.

También estudiamos el comportamiento temporal de los rangos calculando explícitamente la diversidad de rango $d(k)$, una medida del número de individuos o equipos ocupando un rango determinado a lo largo del tiempo. Encontramos que $d(k)$ tiene la misma tendencia funcional a una forma sigmoide, incluso para sistemas relativamente pequeños como FIFA (con sólo 150 elementos por cada rodaja temporal). Adicionalmente al hecho de que el comportamiento en forma de sigmoide de la diversidad de rango también se encuentra en la forma que cambia el vocabulario a lo largo del tiempo [2], nuestros resultados sugieren el emergimiento de una complejidad jerárquica, medida con $d(k)$, que puede ser común a muchos sistemas. Esta afirmación se resalta por el hecho de que un modelo simple (el modelo del caminante aleatorio Gaussiano invariante de escala) puede reproducir la diversidad de los deportes y juegos estudiados aquí, y también el de lenguajes [2]. Uno podría sospechar inicialmente que el rango cambia dependiendo en la fortaleza intrínseca o cualidades de los jugadores y equipos. Sin embargo, dado el hecho de que nuestro modelo del caminante aleatorio reproduce relativamente bien la dinámica de rango de muchos deportes y juegos, parece ser que el cambio de rango puede ser caracterizado mejor por un proceso aleatorio. Esto no implica que el cambio de rango es aleatorio, sino que los mecanismos específicos asociados con cada actividad y sistema de ranqueo son irrelevantes para el cálculo de la diversidad de rango.

Pero debemos ser cuidadosos al afirmar que la diversidad de rango representa una generalidad sigmoideal entre los sistemas complejos jerárquicos, pues tenemos pocos ejemplos aquí presentados y el caso de FIFA representó un contraejemplo a lo afirmado en cierta medida, ya que $d(k)$ comienza a decaer para rangos grandes. Y gracias al concepto de cerradura de un sistema Ω que cuantifica qué tan cerrado es un sistema en el sentido si entran y salen elementos conforme avanza el tiempo a la lista de los rankeos; FIFA resultó ser el sistema más cerrado, por lo que puede haber una correlación entre la caída de $d(k)$ y este concepto, implicando que la monotonía del modelo sigmoideal Φ no necesariamente es correcto, tendríamos que analizar más sistemas cerrados para corroborarlo, y de ser así, limitar el alcance de nuestro modelo sigmoideal.

Siempre vimos que los casos de ESE y NASCAR tendían a dar los resultados más anómalos en cuestión de dinámica, y ésto puede estar relacionado con la pobre estadística del sistema, pues contábamos con pocos elementos o pocas rodajas temporales. (Ver [Tabla 2.1](#)). A lo largo del [Capítulo 4](#) nos dimos cuenta de que la diversidad de rango era ciega ante otros posibles aspectos de la evolución del sistema. Dado ese problema introdujimos también a la *probabilidad de cambio* $p(k)$, que nos dice, por rango k cuál es la probabilidad de que el elemento que ocupa el rango al siguiente paso temporal cambie. Fue notable el hecho de que esta cantidad también tendía a una forma sigmoideal. También se definieron la *entropía de rango* y la *complejidad de rango* que son medidas que cuantifican qué tan estables o caóticas son las ocupaciones de los rangos a lo largo del tiempo, ésto en el contexto de la teoría de la información. Notamos que, en general, el desorden o la inestabilidad de ocupación es muy grande para casi todo el dominio de rangos. La introducción de estas tres medidas son muy afortunadas porque tenemos diferentes perspectivas de la dinámica de rango para los sistemas aquí estudiados.

Además del modelo del caminante aleatorio que ya mencionamos, también se introdujo la idea de otro modelo, el modelo nulo. Este modelo es bastante sencillo, pues sólo consiste en quitar y poner elementos en la lista de ranqueo aleatoriamente. Los resultados obtenidos por el modelo no fueron satisfactorios, salvo el caso de FIFA, pues siempre presentaban una caída abrupta de $d(k)$ para rangos altos. Lo cual puede estar relacionado con las cerraduras de los sistemas, pues también se registra una caída de $d(k)$ para rangos altos en el caso de FIFA. Podría ser interesante realizar una modificación del modelo nulo tomando en cuenta este aspecto de la cerradura y ver si bajando el valor de este parámetro se puede aproximar más a lo observado en los sistemas reales. El modelo del caminante aleatorio, evidentemente, no iba a tener esas caídas a rangos altos pues obligamos a que las variaciones de rango en ese dominio fueran mayores, cosa que no está establecido formalmente en el modelo nulo. Por lo que parece posible que el modelo nulo funcione mejor para sistemas cerrados e incluso modelos abiertos si se toma en consideración ese aspecto también.

El siguiente paso a seguir en un futuro cercano es estudiar el comportamiento de la diversidad de rango en otros fenómenos competitivos más allá de actividades deportivas o lenguajes, tales como sistemas físicos o procesos de estratificación sociales y económicos. Si, en efecto, cierta universalidad se presenta en el comportamiento temporal en otros sistemas complejos, podría indicarse que los fenómenos jerárquicos pueden estar manejados por los mismos mecanismos inherentes de la formación ordenada, ignorando la naturaleza de sus componentes. Potencialmente, podríamos explorar estas regularidades para predecir el tiempo de vida en la ocupación de cierto rango, incrementando nuestra habilidad de formar estrategias en presencia de competición.

También se pueden proponer más modelos para la distribución de rango en el futuro, pues como vimos, no en todos los sistemas funcionó alguno de los modelos propuestos. Adicionalmente, se puede modificar el modelo del caminante aleatorio tomando en cuenta el parámetro Ω de los sistemas, para que en muchos casos podamos eliminar esas caídas abruptas de $d(k)$ para rangos altos.

Cálculo del índice p de Kolmogorov-Smirnov para distribuciones de rango.

El objetivo de este apéndice es describir de manera detallada la forma en que el índice p de Kolmogorov es calculado, concepto introducido en el [Capítulo 3](#), para los modelos que se intenten aproximar a las distribuciones de rango. Ésto para que con su valor se pueda utilizar el criterio de Kolmogorov-Smirnov y que nos indica que una distribución de probabilidad teórica efectivamente es la adecuada para modelar los datos empíricos. Primero explicaremos introduciremos los conceptos matemáticos requeridos y después se explicará en detalle cómo calcular p para que al final podamos utilizar el criterio mencionado que nos permitirá concluir adecuadamente sin un modelo de distribución de probabilidad es el adecuado.

En el [Capítulo 3](#) se definió el concepto de distribución de rango y es simplemente una relación funcional entre puntajes y rangos, lo que llamamos distribución de rango empírica (DRe). Y aunque en su nombre tiene la palabra *distribución*, vimos que no es formalmente una distribución de probabilidad. Existe una equivalencia entre la distribución de rango empírica y la distribución acumulativa empírica de los puntajes (DCE), conceptos también definidos en el [Capítulo 3](#), la equivalencia se prueba por medio del proceso descrito en la [Figura 3.2](#). La DCE de un grupo de puntajes $\{s_1, s_2, \dots, s_N\}$ se calcula de acuerdo a:

$$M(s) = \frac{\text{número elementos en la muestra } \leq s}{N} = \frac{1}{N} \sum_{i=1}^N \theta(s - s_i) \quad (\text{A.1})$$

que es una función continua y sí es una distribución de probabilidad en el sentido matemático estricto. Del mismo modo, los modelos que proponemos en este trabajo de tesis para la distribución de rango empírica están expresados en las ecuaciones [3.9–3.13](#), que genéricamente denotaremos por $m_i(k)$ donde $i = 1, 2, 3, 4, 5$. Ahora bien, por una derivación hecha en [\[31\]](#), se demuestra que mientras que $m_i(k)$ modela a la DRe, $M_i(m_i)$ modela a la DCE, siendo que $M_i(m_i)$ es equivalente a $m_i(k)$ por medio de la transformación:

$$M_i(m_i) = \frac{N + 1 - k(m_i)}{N + 1} \quad (\text{A.2})$$

donde $k(m_i)$ es la función inversa de $m_i(k)$. Por lo que gracias a lo hecho en [31] estamos justificando que el modelar a la DRe con los modelos $m_i(k)$ es equivalente a modelar la DCe con $M_i(m_i)$. Estamos hablando del mismo fenómeno pero en dos idiomas diferentes, **si probamos que $M_i(m_i)$ es un modelo adecuado para la DCe entonces el modelo $m_i(k)$ es el modelo adecuado para la DRe correspondiente, y viceversa.**

Acabamos de traducir el concepto de distribución de rango al lenguaje de las distribuciones de probabilidad en el sentido matemático formal de la palabra y esto es importante porque el criterio de Kolmogorov-Smirnov es aplicable sólo a distribuciones de rango.

Ahora definimos a la estadística de Kolmogorov D como el supremo de las distancias entre una distribución acumulativa empírica (en este caso hablamos de los puntajes) y la distribución de probabilidad teórica con la que queremos modelarla (en este caso hablamos de $M_i(m_i)$). Por tanto, si tenemos un conjunto de puntajes $\{s_1, s_2, \dots, s_N\}$, podemos calcular la DCe de los datos y si queremos aproximarla al modelo $M_i(m_i)$ tenemos que la estadística de Kolmogorov se calcula como:

$$D = \sup_s |M_i(m_i) - M(s)| \quad (\text{A.3})$$

Notemos que s es el nombre de la variable que representa a los puntajes. También recordemos que por la definición de m_i , ésta corresponde a puntajes también, por lo que está bien definida esta cantidad. [42] Lo que haremos ahora es justamente compara estas dos distribuciones de probabilidad $M_i(m_i)$ y $M(s)$. El procedimiento que a continuación describo es el que se utiliza para calcular el llamado índice p de Kolmogorov-Smirnov y que se obtuvo de [42]. Si comenzamos entonces con un grupo de puntajes $\{s_1, s_2, \dots, s_N\}$ que corresponden a los elementos rankeados en una rodaja temporal, cada puntajes corresponde a un rango y de allí obtenemos la DRe.

1. Calculamos los parámetros de ajuste a $m_i(k)$ para la distribución de rango empírica DRe que induce este grupo de puntajes.
2. Obtenemos la distribución acumulativa de los puntajes DCe $M(s)$ y la respectiva distribución acumulativa $M_i(m_i)$ usando la [Ecuación A.2](#)
3. Calculamos la estadística de Kolmogorov D entre $M_i(m_i)$ y $M(s)$ como la definimos en la [Ecuación A.3](#).
4. Generamos 2500 conjuntos de datos artificiales de puntajes, generados aleatoriamente de acuerdo a la distribución de probabilidad $M_i(m_i)$ inducida por la $m_i(k)$ ajustada a la DRe. Para cada uno de estos conjuntos de datos, ajustamos una $M_{i,art}(s)$ artificial y que proviene de una $m_i(s)$ como definimos. Para cada uno de esos conjuntos de datos obtenemos la D_{art} , como se definió en la [Ecuación A.3](#), entre la DCe ($M(s)$) de los datos con la $M_{i,art}$.

-
5. Contamos cuántos de los 2500 D_{art} son más grandes que D de los datos empíricos y dividimos por 2500. El resultado es el valor p , el índice de Kolmogorov-Smirnov.

Hay que notar algunas cosas muy importantes del proceso, en el paso 4 tuvimos que generar un conjunto de datos por medio de una distribución de probabilidad $M_i(m_i)$, si no se tratara estrictamente de una distribución de probabilidad, no tendría sentido ese paso. En [42] se hace hincapié al hecho de que se necesitan generar al menos 2500 conjuntos de datos artificiales para que p tenga un valor aceptable en dos cifras significativas. Entonces nuestros valores de p sólo pueden tener dos decimales calculados con certidumbre. **Si $p > 0.1$ concluimos que $M_i(m_i)$ modela adecuadamente la DCe de los puntajes, si $p < 0.1$ entonces el modelo se desecha.** Lo que es equivalente a decir que **Si $p > 0.1$ concluimos que $m_i(k)$ modela adecuadamente la DRe de los puntajes, si $p < 0.1$ entonces el modelo se desecha.** Lo único que debemos hacer para corroborar que el modelo $m_i(k)$ funciona es realizar el proceso que acabamos de describir.

Para ilustrar mejor el procedimiento de arriba las figuras A.1–A.12 grafican el procedimiento para todos los deportes y juegos de esta tesis y para todos los modelos propuestos. La primera columna de las figuras ilustran el paso 1, donde se ajusta una determinada m_i a la DRe del sistema en una rodaja temporal específica. La segunda columna representa la equivalencia de la primera columna con distribuciones de probabilidad; las líneas rojas representan $M_i(m_i)$ y las líneas negras es la DCe de los puntajes, se proporciona también la correspondiente estadística de Kolmogorov entre estas dos distribuciones. La tercera columna representa un ejemplo de la DCe de datos generados aleatoriamente conforme a la distribución de probabilidad $M_i(m_i)$ obtenida de los ajustes a los datos reales. También se obtiene la M_{art} ajustada a esos datos generados, se proporciona también la estadística de Kolmogorov D_{art} entre dichas distribuciones. Las rodajas temporales utilizadas por disciplina son: FIDE-F (Abril 2016), FIDE-M (Abril 2016), FCWR-C (Semana 53 del 2014), FIFA (Junio 2017), FCWR-G (Semana 33 de 2017), OWGR (21/05/2017), NASCAR-B (2015), NASCAR-W (2013), GPI (31/05/2017), WSD (26/03/2018), ATP (27/12/2010), ESE (2016).

El criterio de Kolmogorov-Smirnov sólo es aplicable cuando estamos comparando distribuciones de probabilidad, es por ello que hicimos hincapié en la equivalencia entre DRe y DCe, donde la DCe sí está bien definida como una distribución de probabilidad. La medida p nos permite considerar el caso en que una base de datos pequeña tendrá ruido no deseado por tener una estadística pobre, elimina esas fluctuaciones estadísticas no deseadas.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

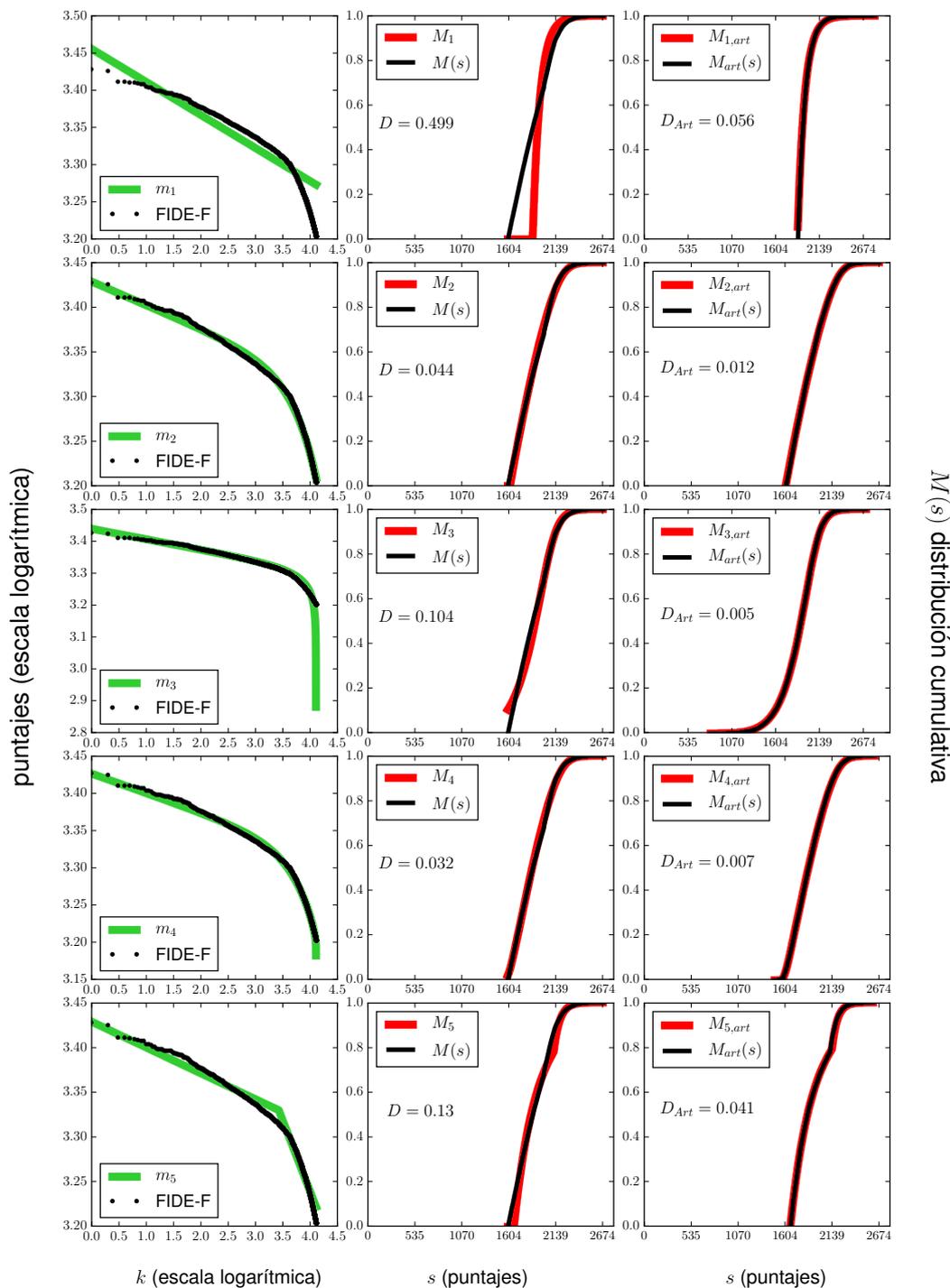


Figura A.1: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIDE-F. Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCE). Tercer columna: distribución cumulativa de datos generados artificialmente.

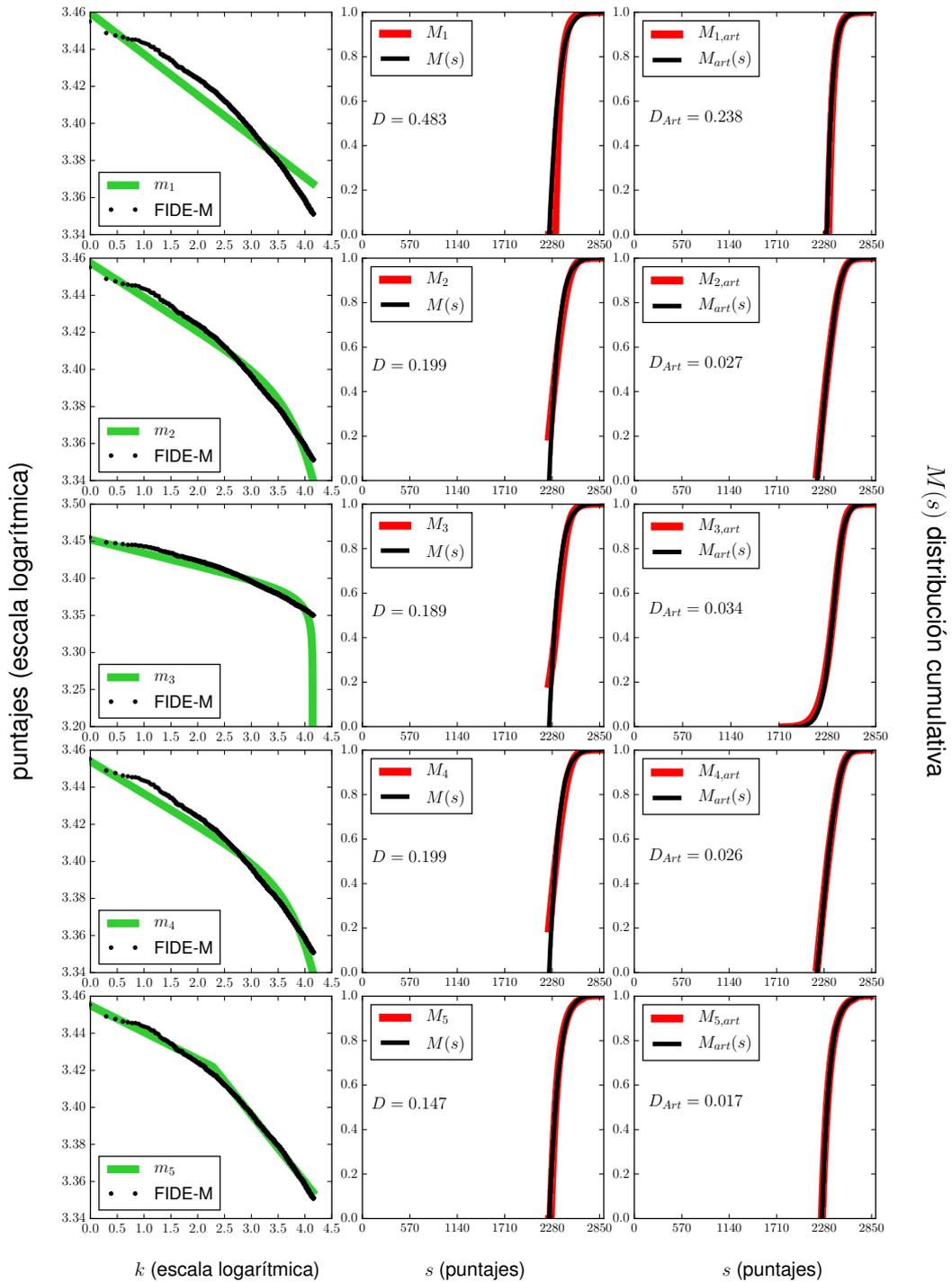


Figura A.2: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIDE-M Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución acumulativa empírica (DCe). Tercer columna: distribución acumulativa de datos generados artificialmente.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

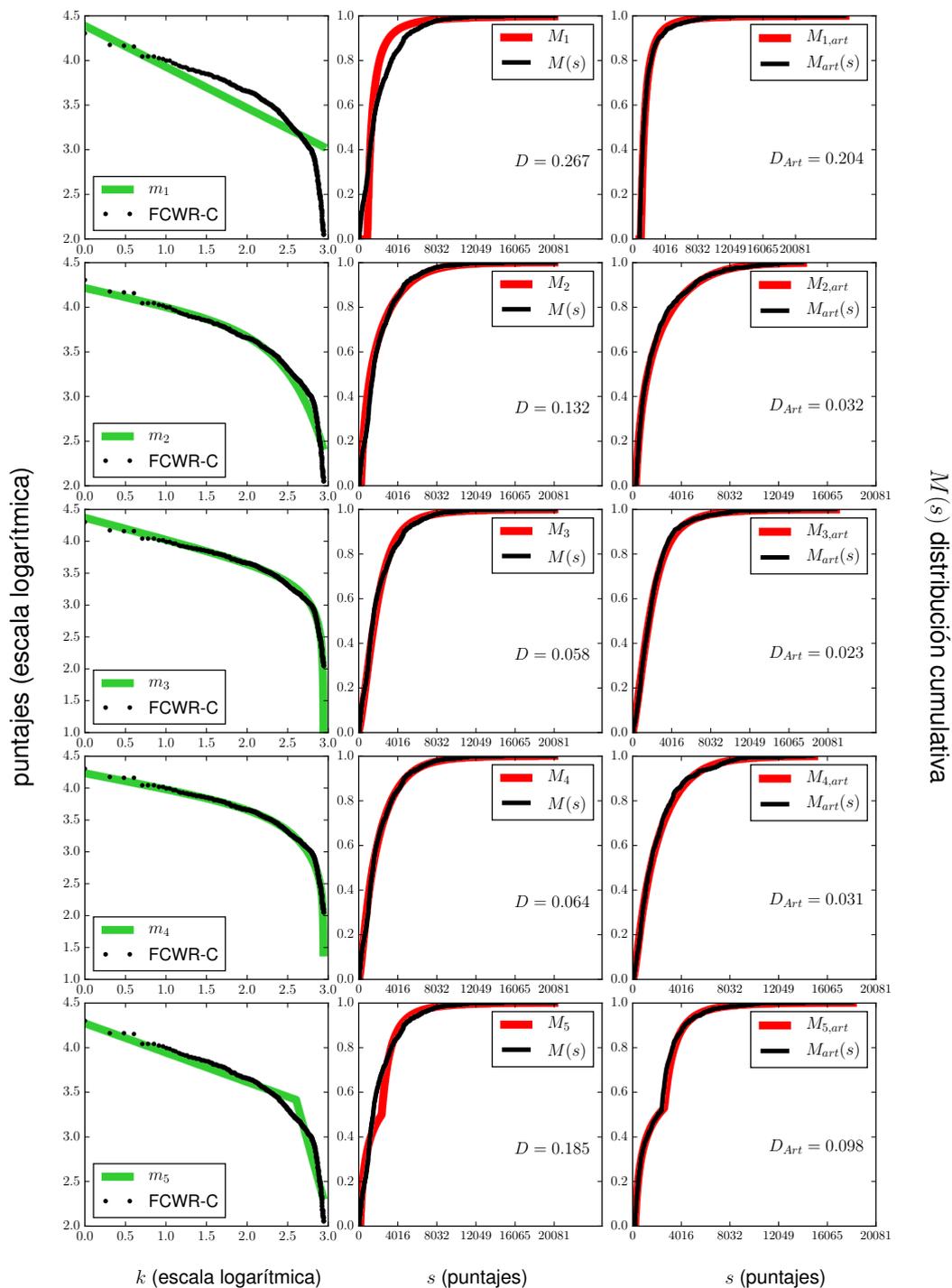


Figura A.3: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FCWR-C. Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución acumulativa empírica (DCE). Tercer columna: distribución acumulativa de datos generados artificialmente.

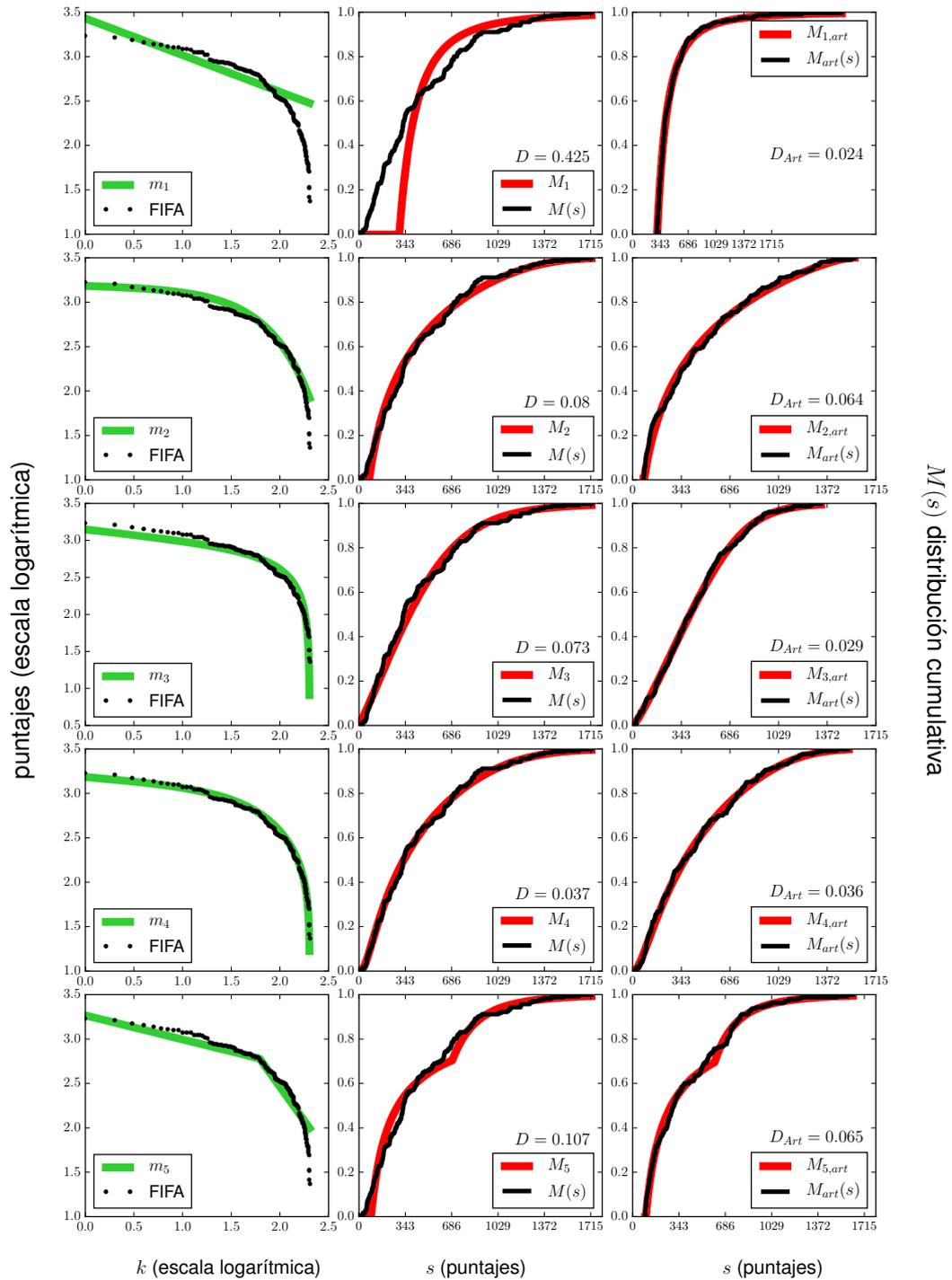


Figura A.4: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FIFA Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

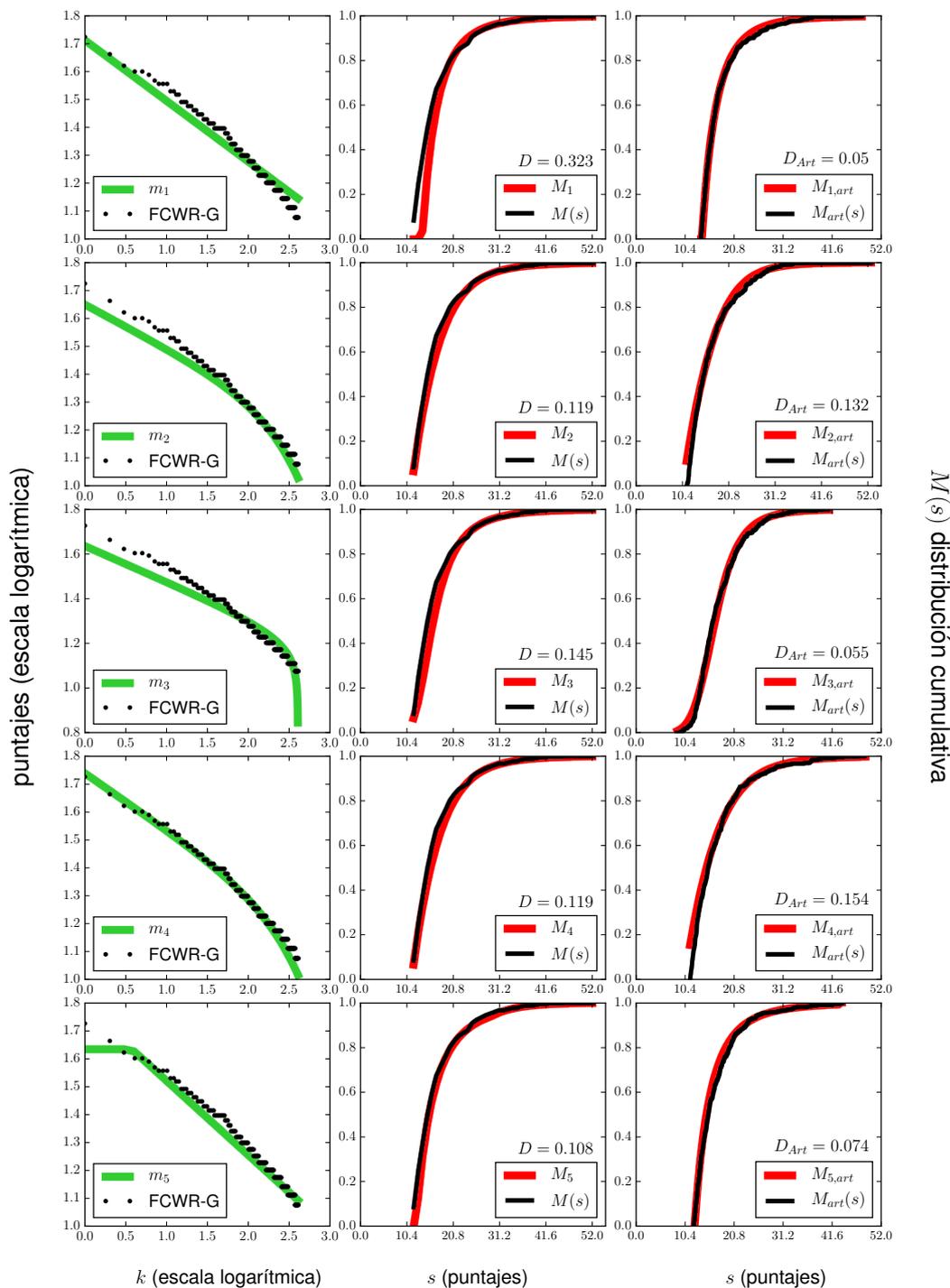


Figura A.5: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de FCWR-G. Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.

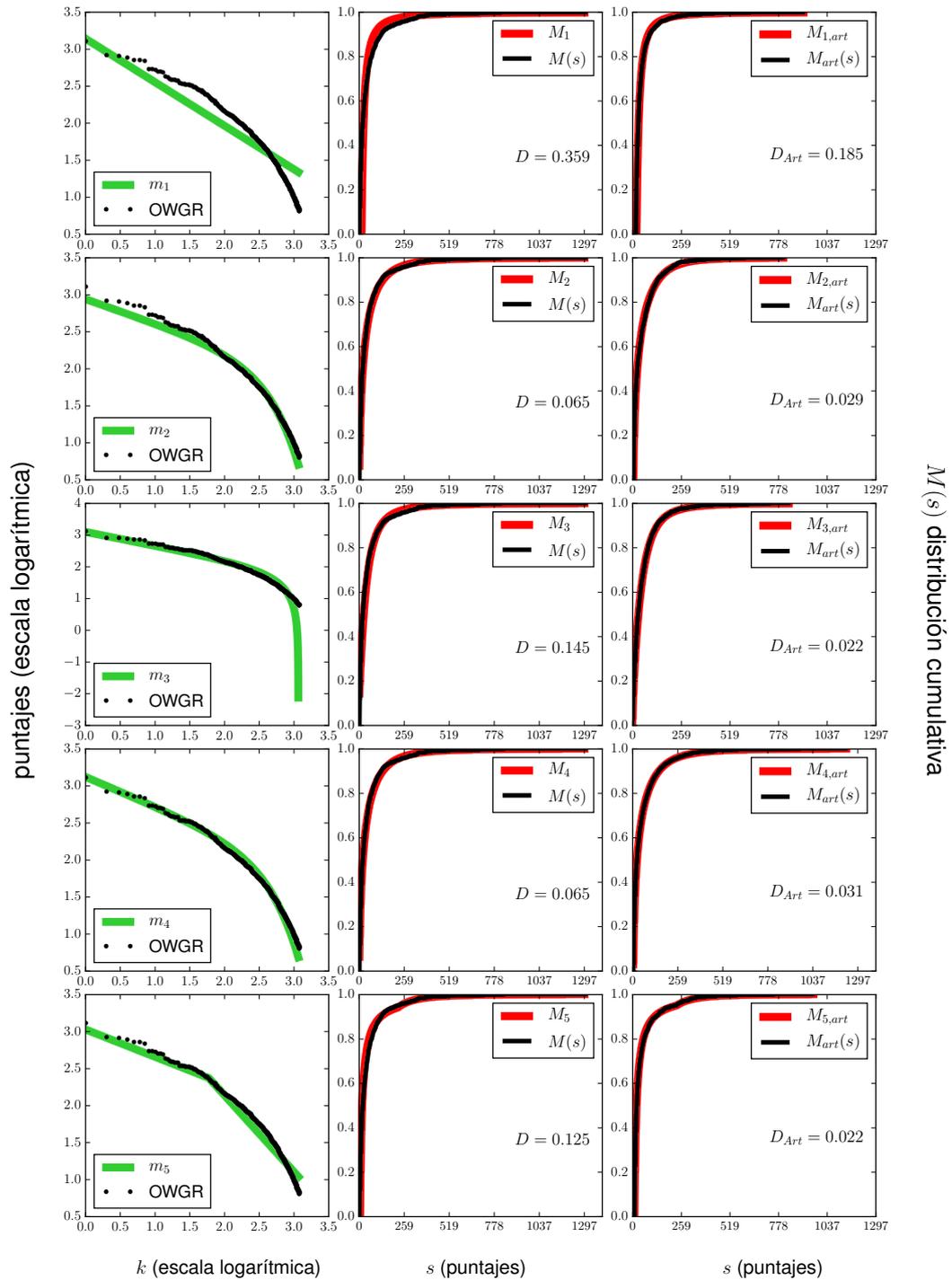


Figura A.6: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de OWGR. Primera columna: Distribución de rango empírica (DR) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

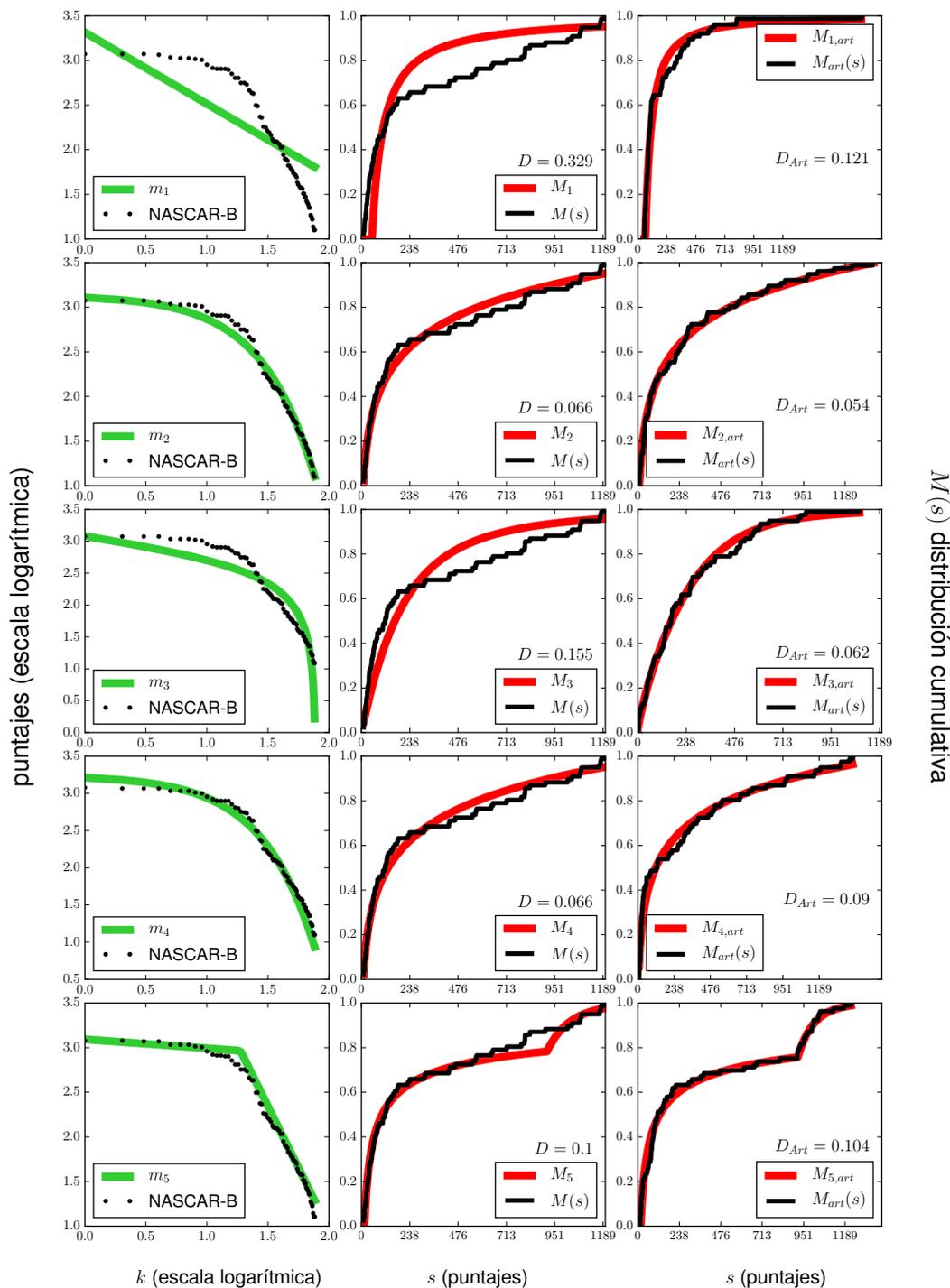


Figura A.7: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de NACAR-B Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCE). Tercer columna: distribución cumulativa de datos generados artificialmente.

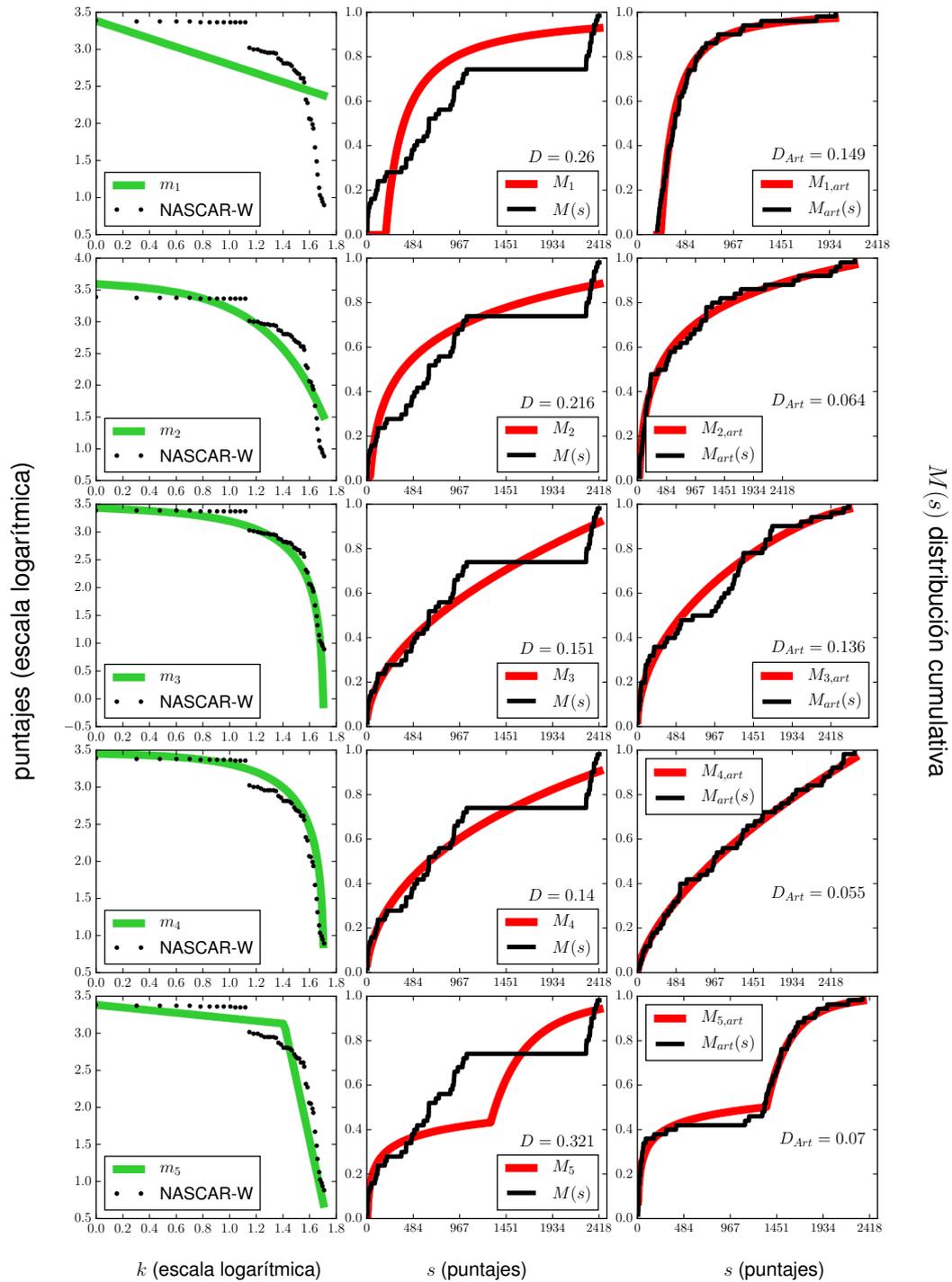


Figura A.8: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de NASCAR-W Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

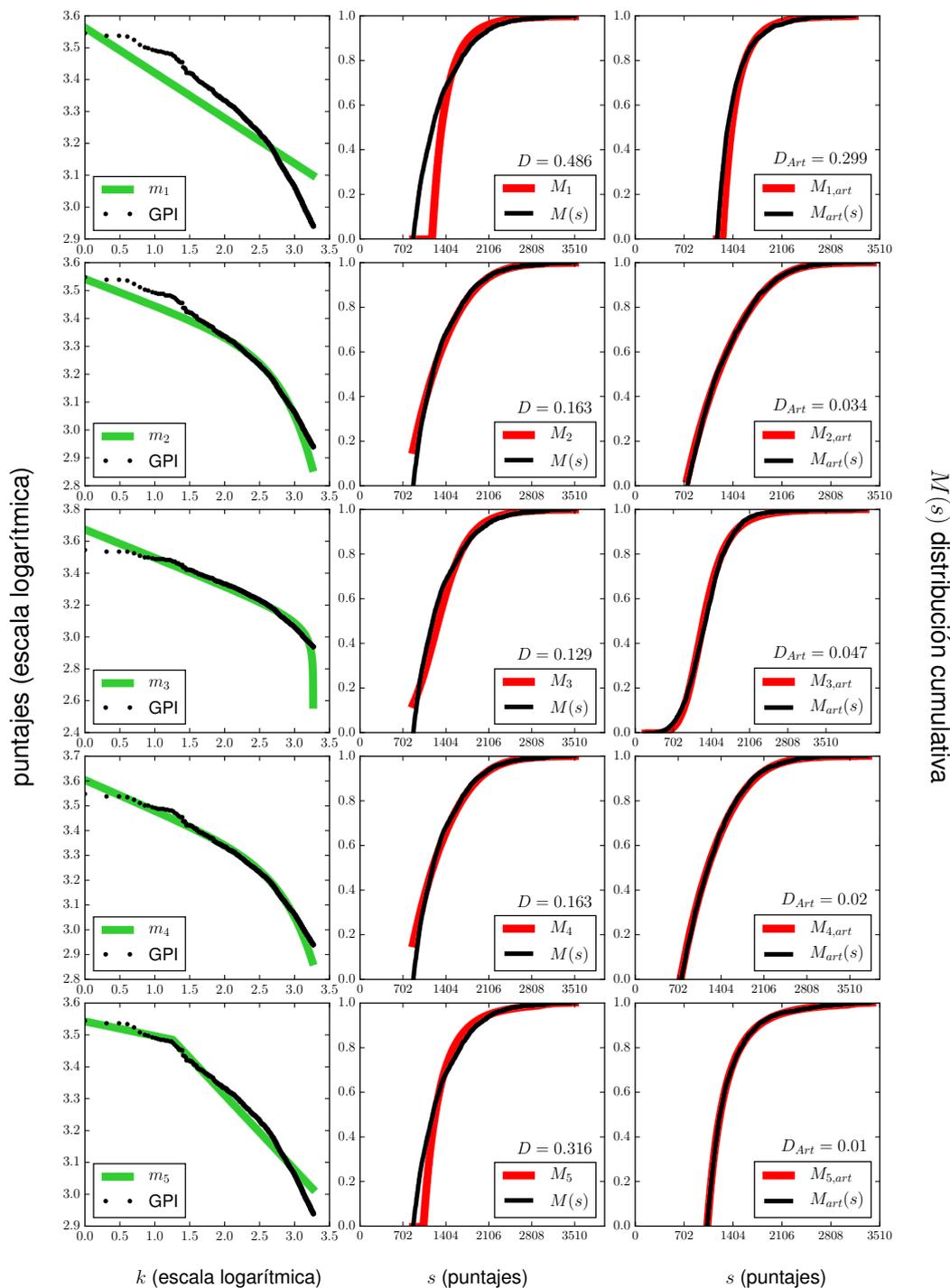


Figura A.9: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de GPI. Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución acumulativa empírica (DCE). Tercera columna: distribución acumulativa de datos generados artificialmente.

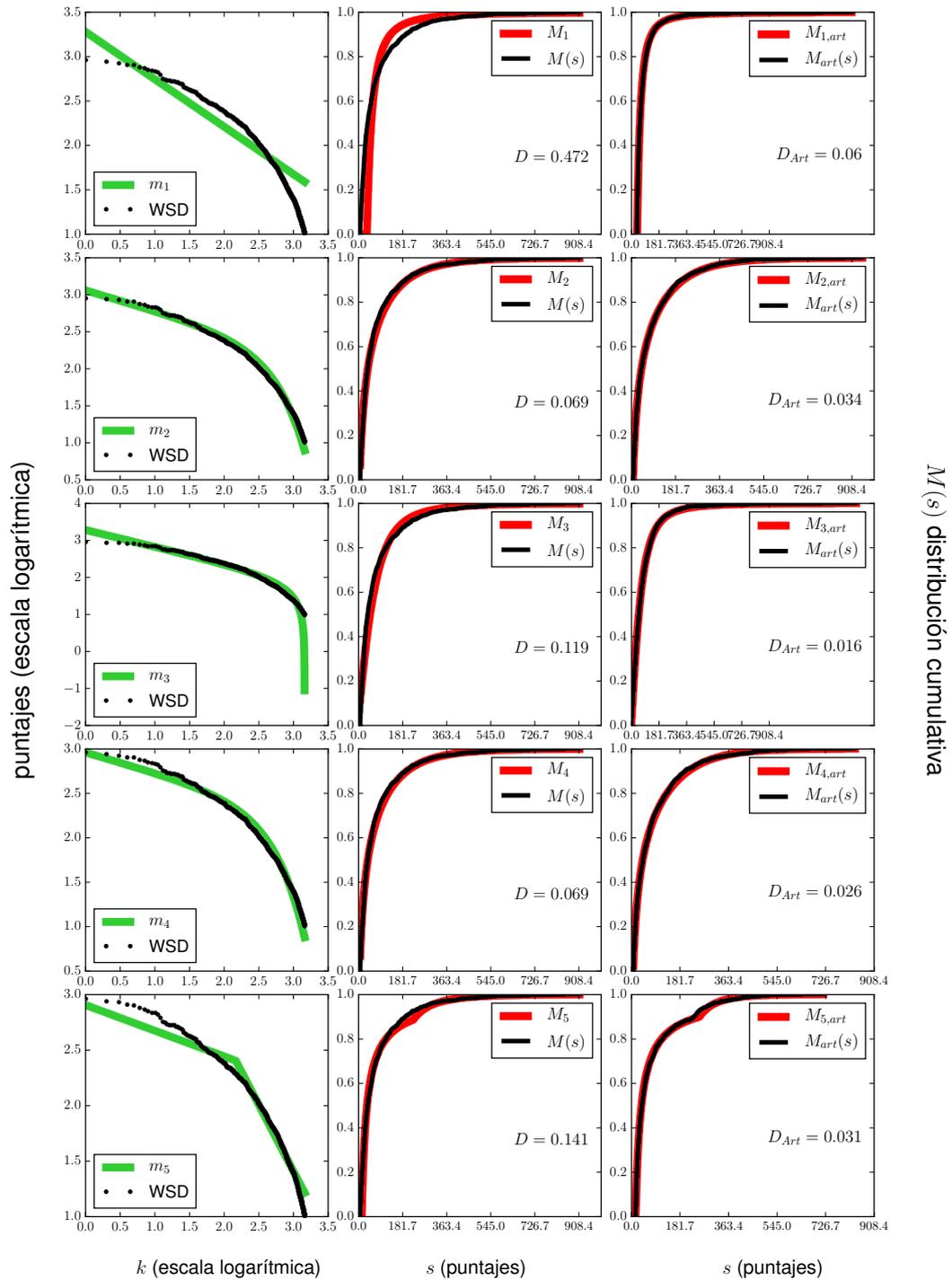


Figura A.10: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de WSD Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución cumulativa empírica (DCe). Tercer columna: distribución cumulativa de datos generados artificialmente.

A. CÁLCULO DEL ÍNDICE P DE KOLMOGOROV-SMIRNOV PARA DISTRIBUCIONES DE RANGO.

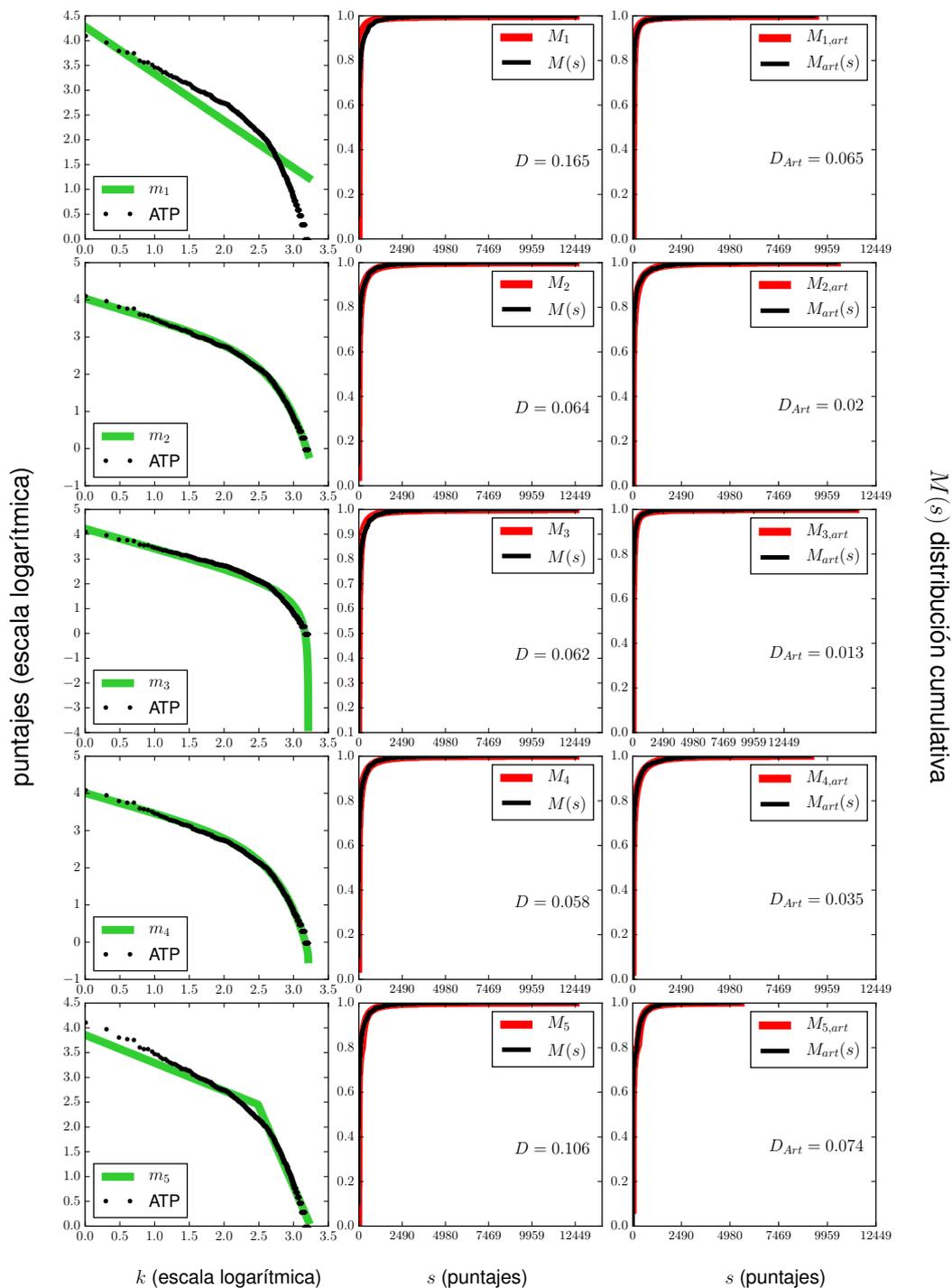


Figura A.11: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de ATP. Primera columna: Distribución de rango empírica (DRe) y su ajuste. Segunda columna: distribución acumulativa empírica (DCE). Tercera columna: distribución acumulativa de datos generados artificialmente.

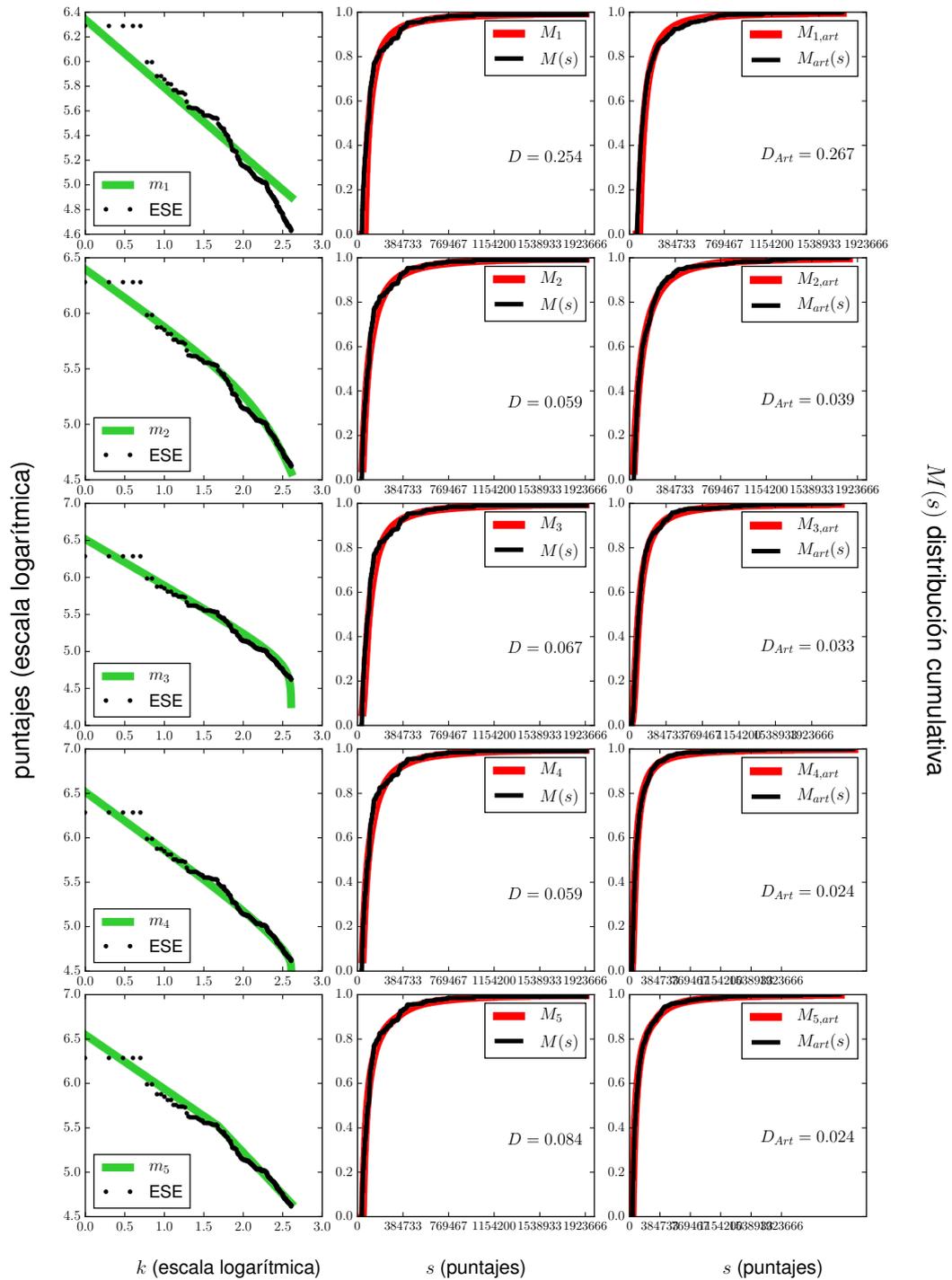


Figura A.12: Proceso de cálculo del índice p de Kolmogorov-Smirnov para el caso de ESE Primera columna: Distribución de rango empírica (DRE) y su ajuste. Segunda columna: distribución acumulativa empírica (DCe). Tercer columna: distribución acumulativa de datos generados artificialmente.

Espaguetis ejemplos.

Para ejemplificar la forma en que ciertos rangos son ocupados por los elementos de los sistemas. Se grafican los espaguetis (concepto definido en el [Capítulo 4](#)) que pasan por los rangos $k = 1$, $k = N/2$ y $k = N$. Se hace para todos los sistemas trabajados (ver [Capítulo 2](#)), donde N es el número de elementos por ranqueo en cada rodaja temporal de las respectivas bases de datos. Se incluirá junto al rango k , la n que simboliza el número de elementos diferentes que ocuparon el dicho rango en todos los tiempos.

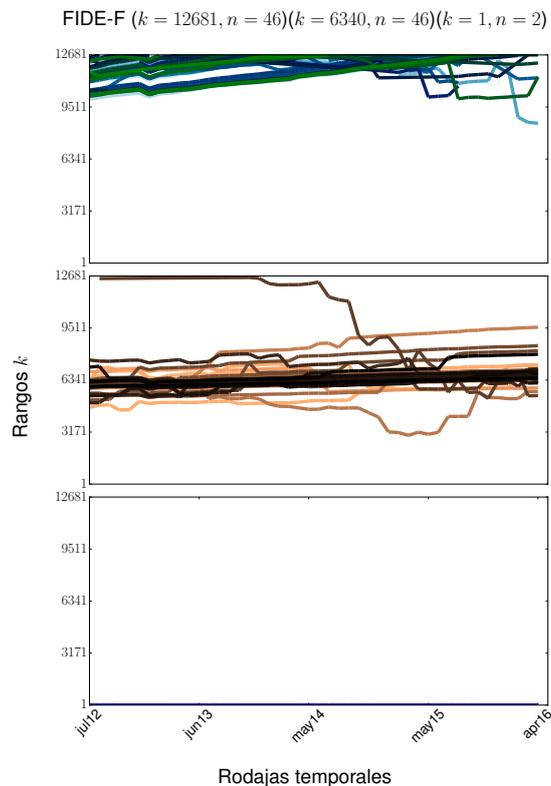


Figura B.1: Espaguetis para FIDE-F para los rangos $k = 12681, 6340, 1$

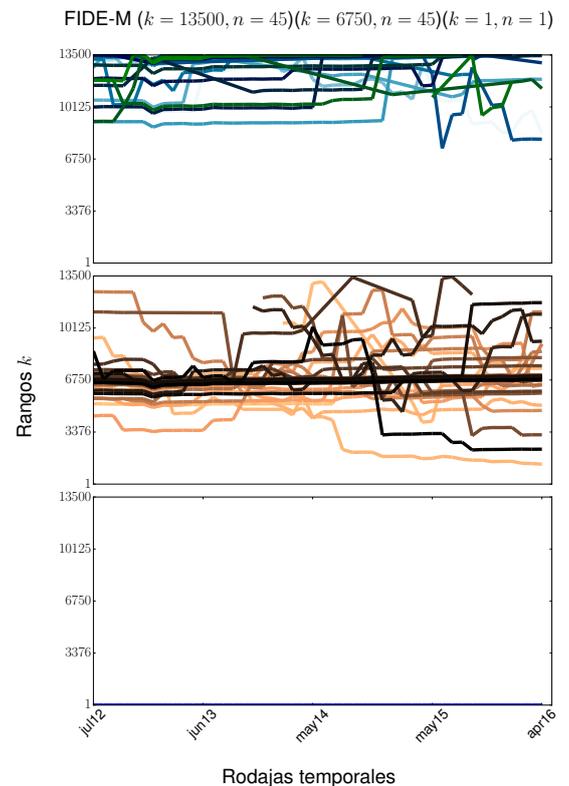


Figura B.2: Espaguetis para FIDE-M para los rangos $k = 13500, 6750, 1$

B. ESPAGUETIS EJEMPLOS.

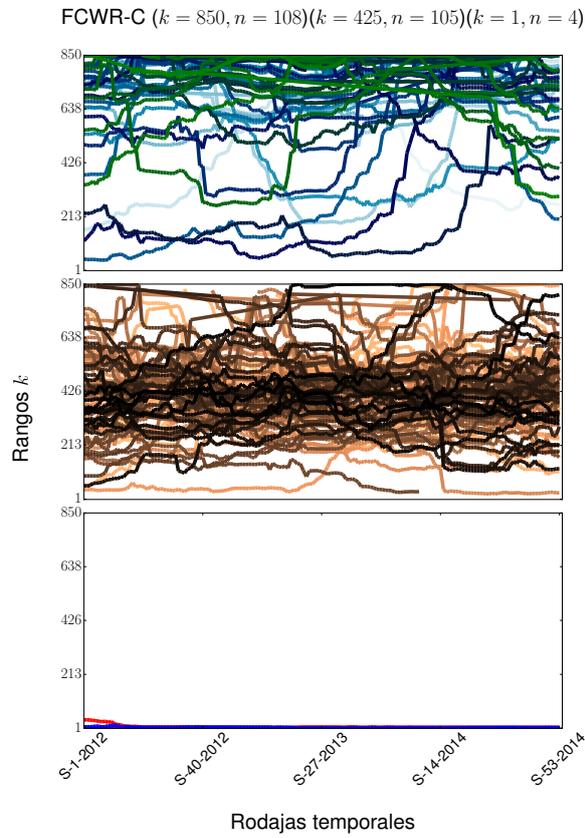


Figura B.3: Espaguetis para FCWR-C para los rangos $k = 850, 425, 1$

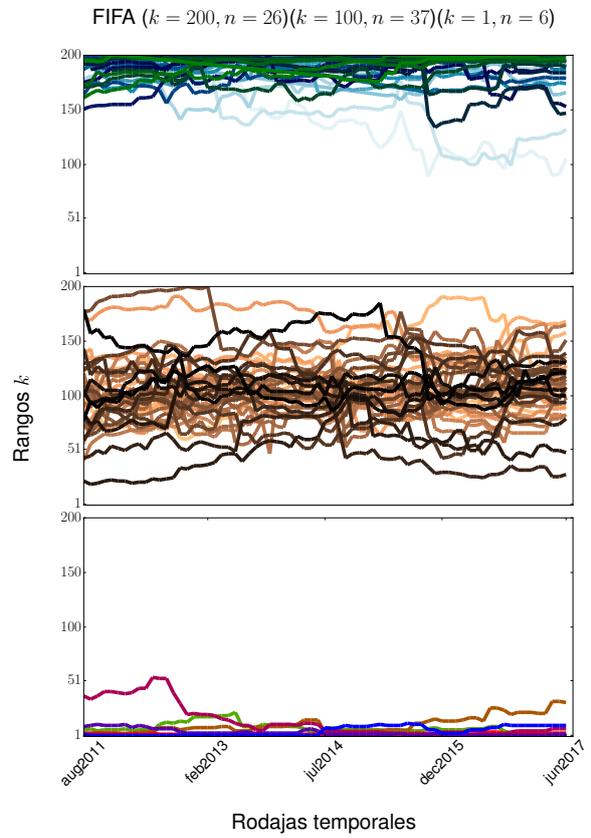


Figura B.4: Espaguetis para FIFA para los rangos $k = 200, 100, 1$

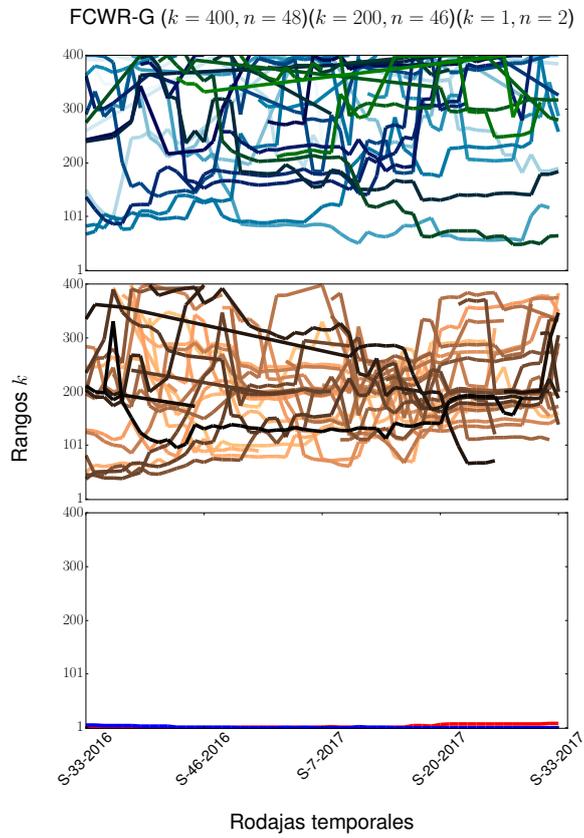


Figura B.5: Espaguetis para FCWR-G para los rangos $k = 400, 200, 1$

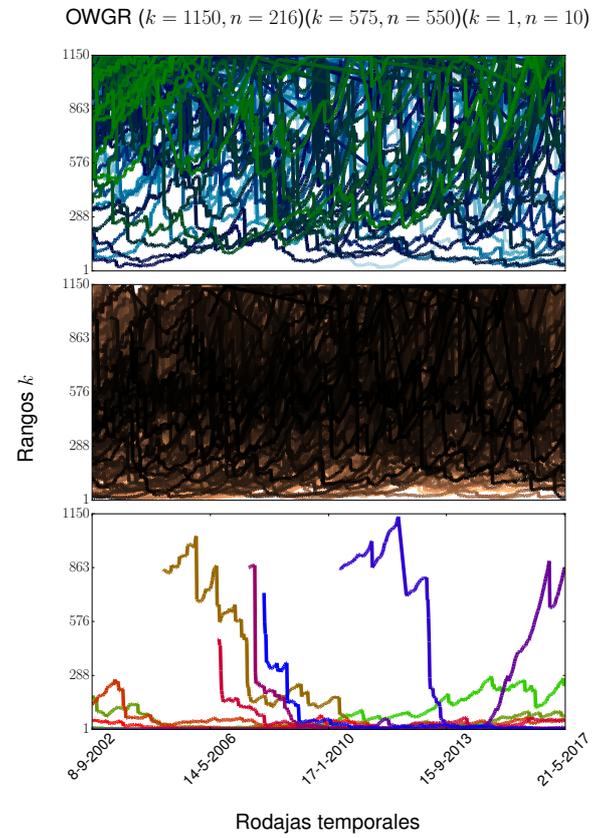


Figura B.6: Espaguetis para OWGR para los rangos $k = 1150, 575, 1$

B. ESPAGUETIS EJEMPLOS.

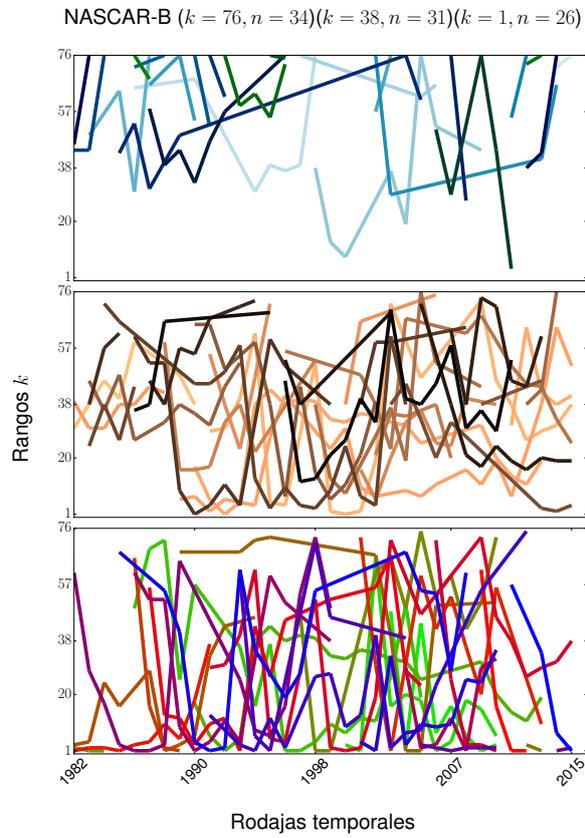


Figura B.7: Espaguetis para NASCAR-B para los rangos $k = 76, 38, 1$

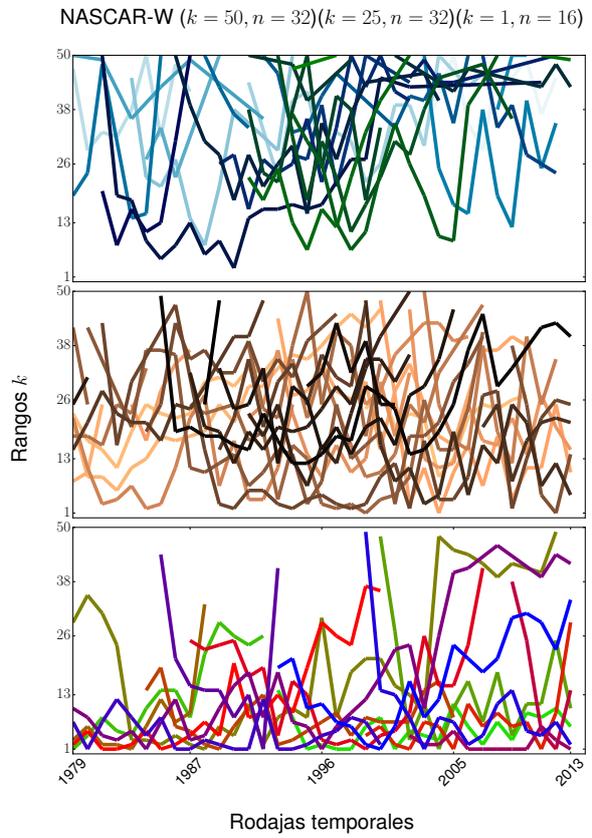


Figura B.8: Espaguetis para NASCAR-W para los rangos $k = 50, 25, 1$

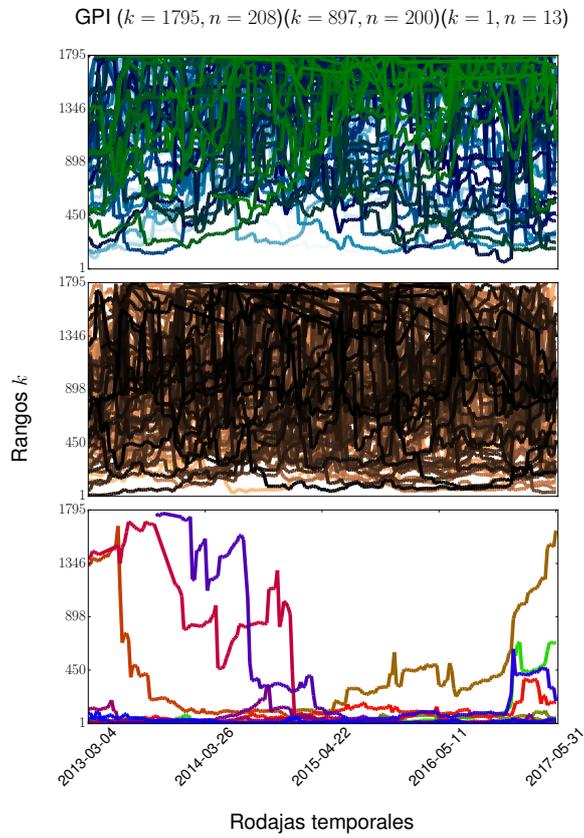


Figura B.9: Espaguetis para GPI para los rangos $k = 1795, 897, 1$

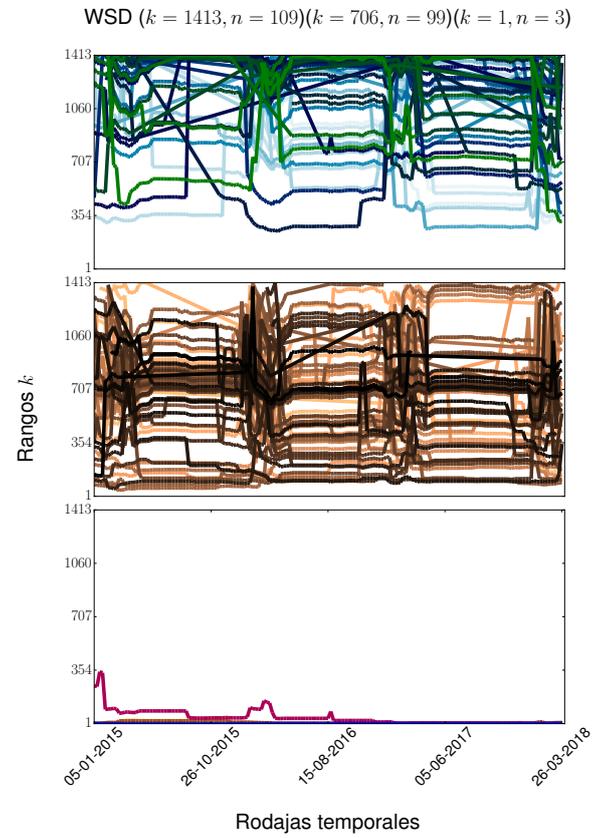


Figura B.10: Espaguetis para WSD para los rangos $k = 1413, 706, 1$

B. ESPAGUETIS EJEMPLOS.

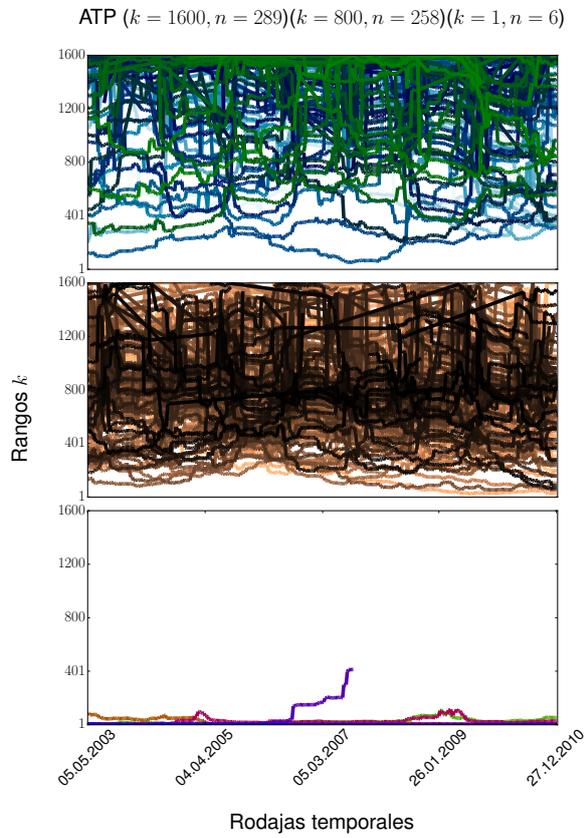


Figura B.11: Espaguetis para ATP para los rangos $k = 1600, 800, 1$

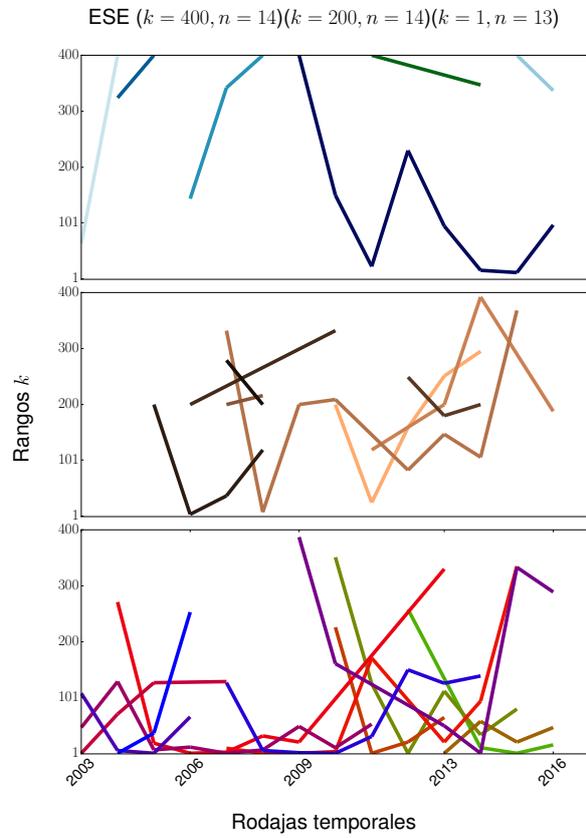


Figura B.12: Espaguetis para ESE para los rangos $k = 400, 200, 1$

Figuras Adicionales.

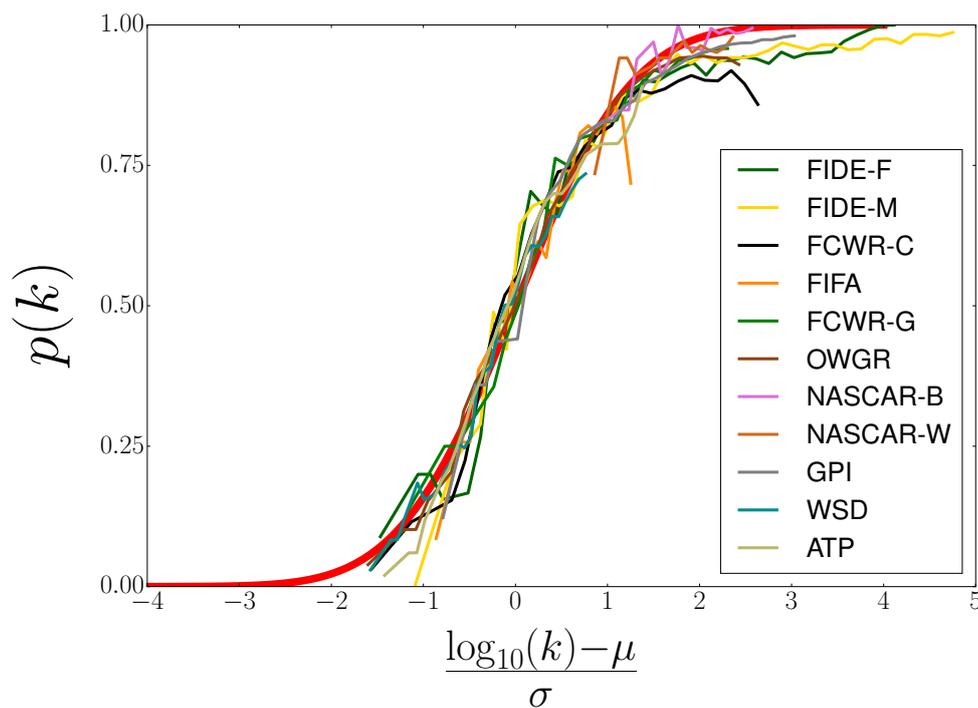


Figura C.1: Similaridad en la probabilidad de cambio normalizada para deportes y juegos. Gráfica que muestra una comparación de la probabilidad de cambio $p(k)$ para todas las actividades consideradas. Con los valores de μ y σ obtenemos el ajuste de Φ , hemos reescalado la abscisa como hicimos en el caso de la diversidad de rango. Como referencia concluimos que forma básica de la [Ecuación 4.3](#) (línea roja en la gráfica), con $\mu = 0$ y $\sigma = 1$. Estos resultados indican que todas las actividades tienen la misma forma funcional para la probabilidad de cambio.

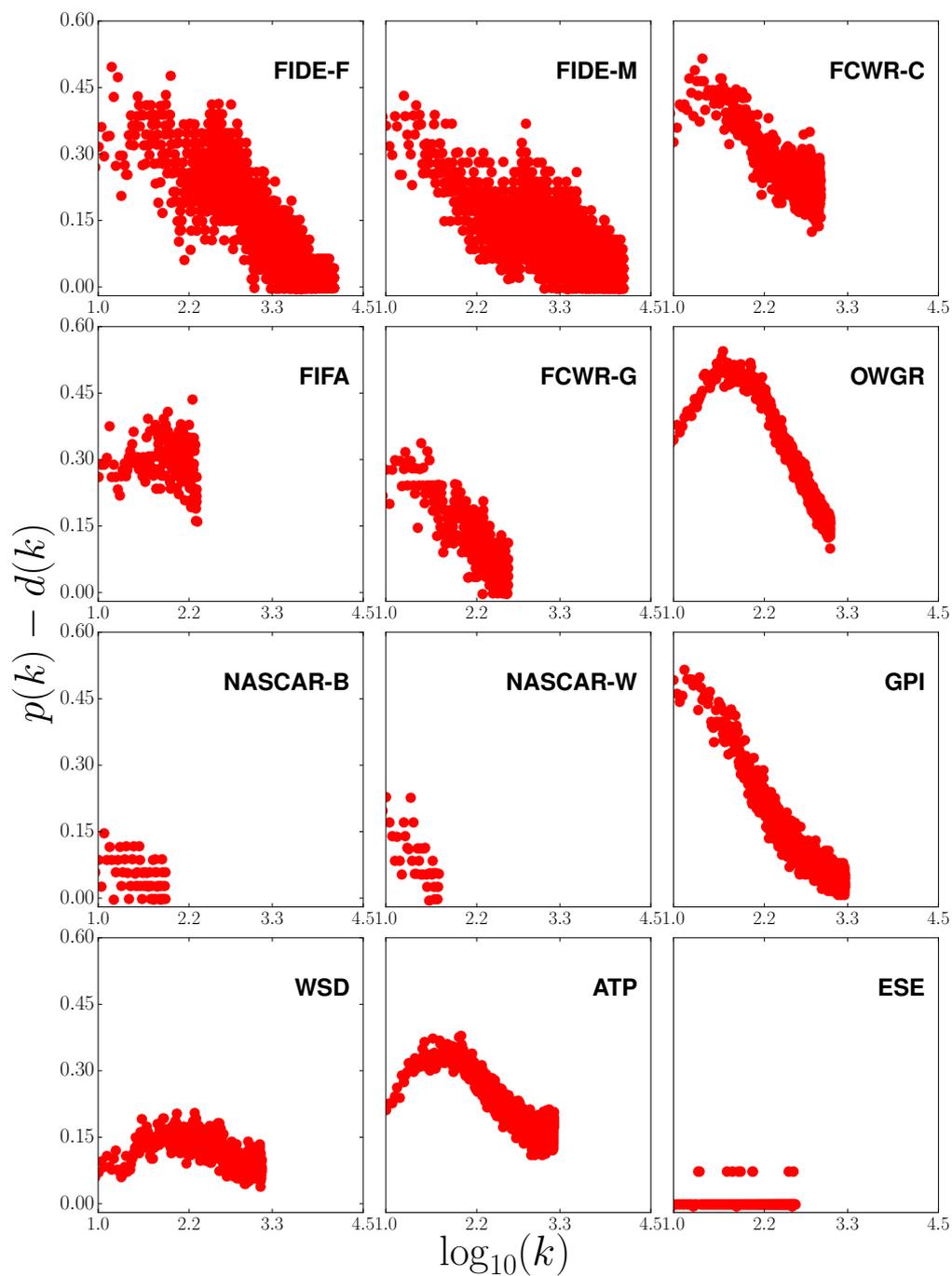


Figura C.2: Diferencias entre la probabilidad de cambio y la diversidad de rango $p(k) - d(k)$ en escala semilogarítmica. En la figura podemos apreciar que todos los valores de la diferencia entre estas dos medidas es siempre mayor o igual a cero, lo que comprueba que $p(k) \geq d(k)$.

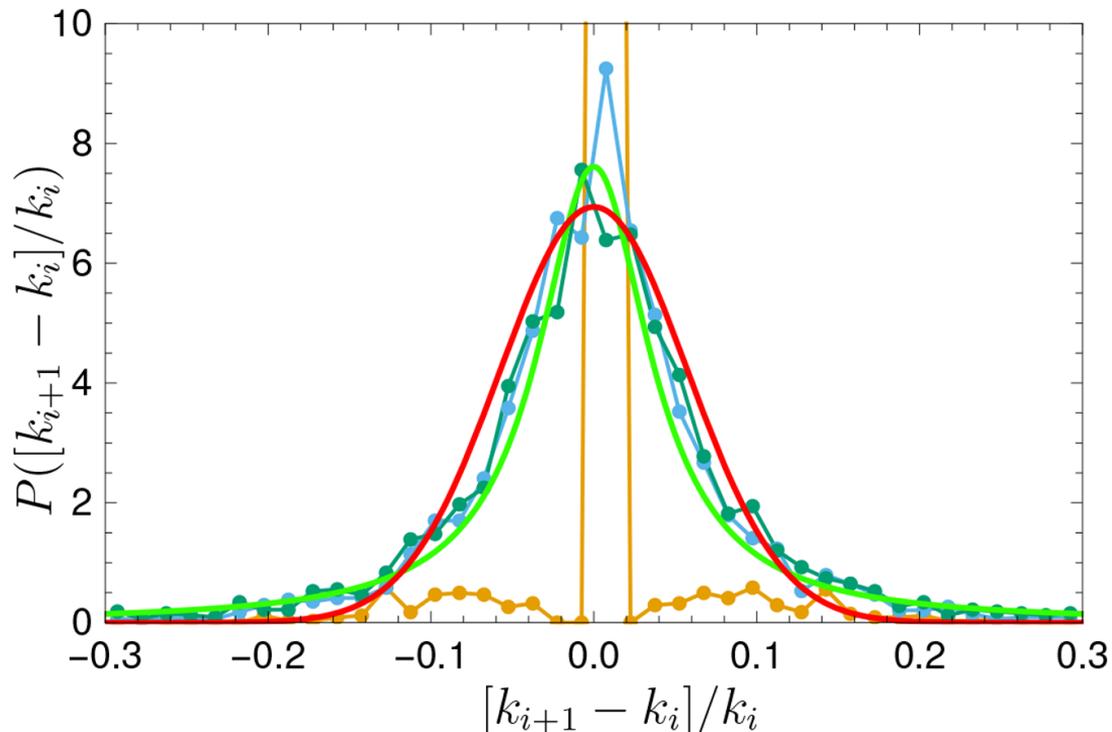


Figura C.3: Distribución de los tamaños relativos del cambio de frecuencias $[k_{t+1} - k_t] / k_t$, para el caso de las palabras en inglés. Las palabras correspondientes a las cabezas de la diversidad (en dorado), las que están en el cuerpo de la diversidad (azul) y las que están en la cola (en verde). La distribución gaussiana con una desviación estándar $\sigma = 0.0575$ está graficada en rojo para hacer una comparativa. En verde se grafica una curva estrecha que representa una distribución Lorentziana que mejor se ajusta con el promedio de las tres distribuciones empíricas que se muestran aquí (las de los tres grupos de palabras). Notemos que las palabras que están en la cabeza, los saltos relativos no se asemejan a la distribución Gaussiana. Para el caso de las distribuciones de las palabras en la cola y cabeza tenemos mucha similitud mutua. Notemos que, en promedio, los saltos relativos parecen muy independientes de los valores de k , por lo que parece que reproducir los rangos requiere de un modelo invariante de escala. Las curvas de la distribución gaussiana y lorentziana centradas en cero son las que mejor se ajustaron a los datos presentados aquí. Aunque la distribución lorentziana parece ajustar mejor a los datos que la gaussiana, usamos la distribución gaussiana en el modelo del caminante aleatorio, pues las largas colas de la lorentziana implicaría grandes saltos de los rangos de las palabras (algo que no se observa en las bases de datos, pues $d(k)$ indica lo contrario). Tal vez una lorentziana truncada en las colas funcionaría mejor, pero complicaría bastante el modelo; optaremos por utilizar la distribución gaussiana. Imagen y descripción tomados de [2].

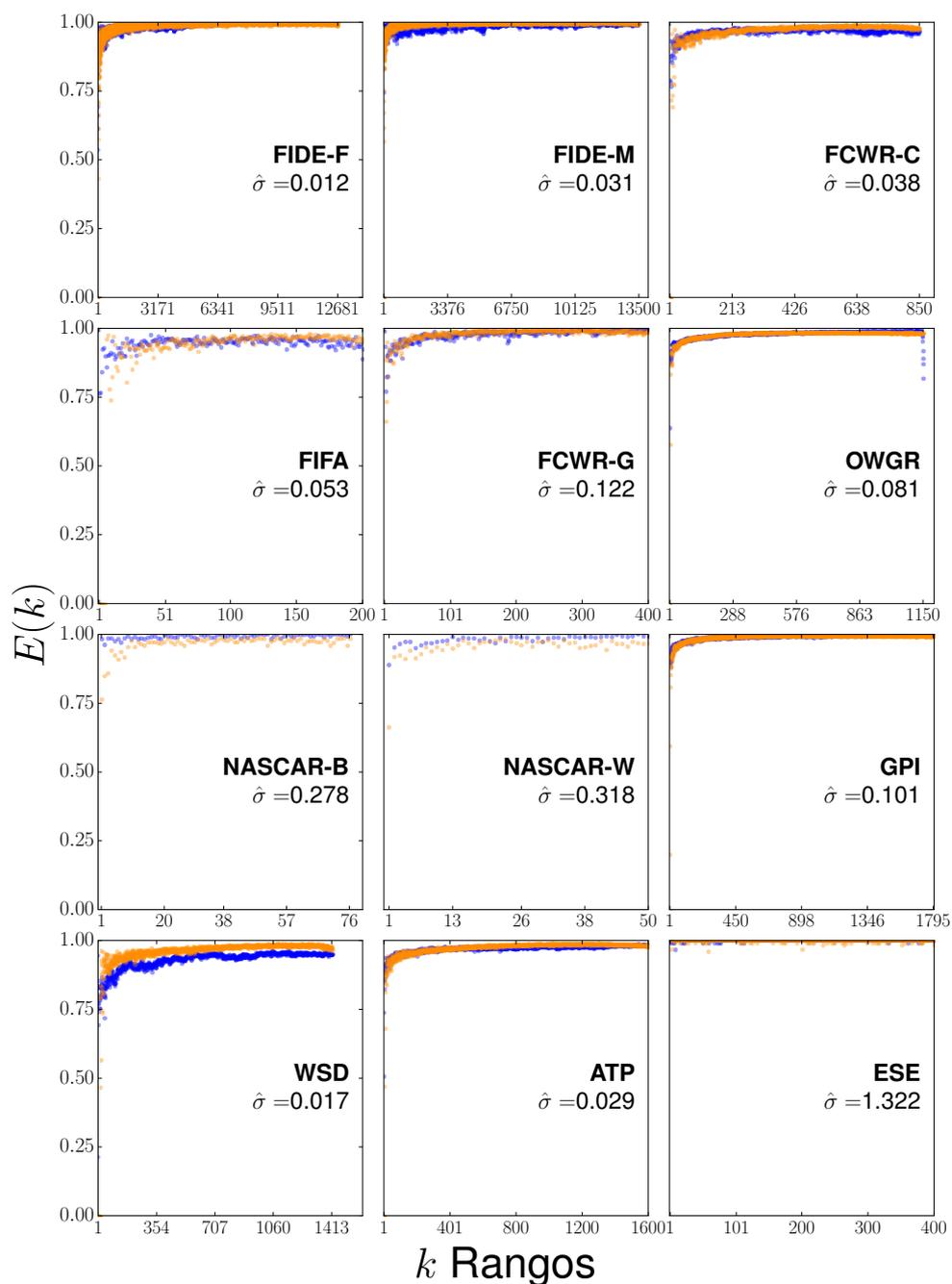


Figura C.4: Comparación entre las entropías de rango empíricas y simuladas con el Modelo del Caminante Aleatorio. Gráfica que muestra la entropía de rango $E(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados).

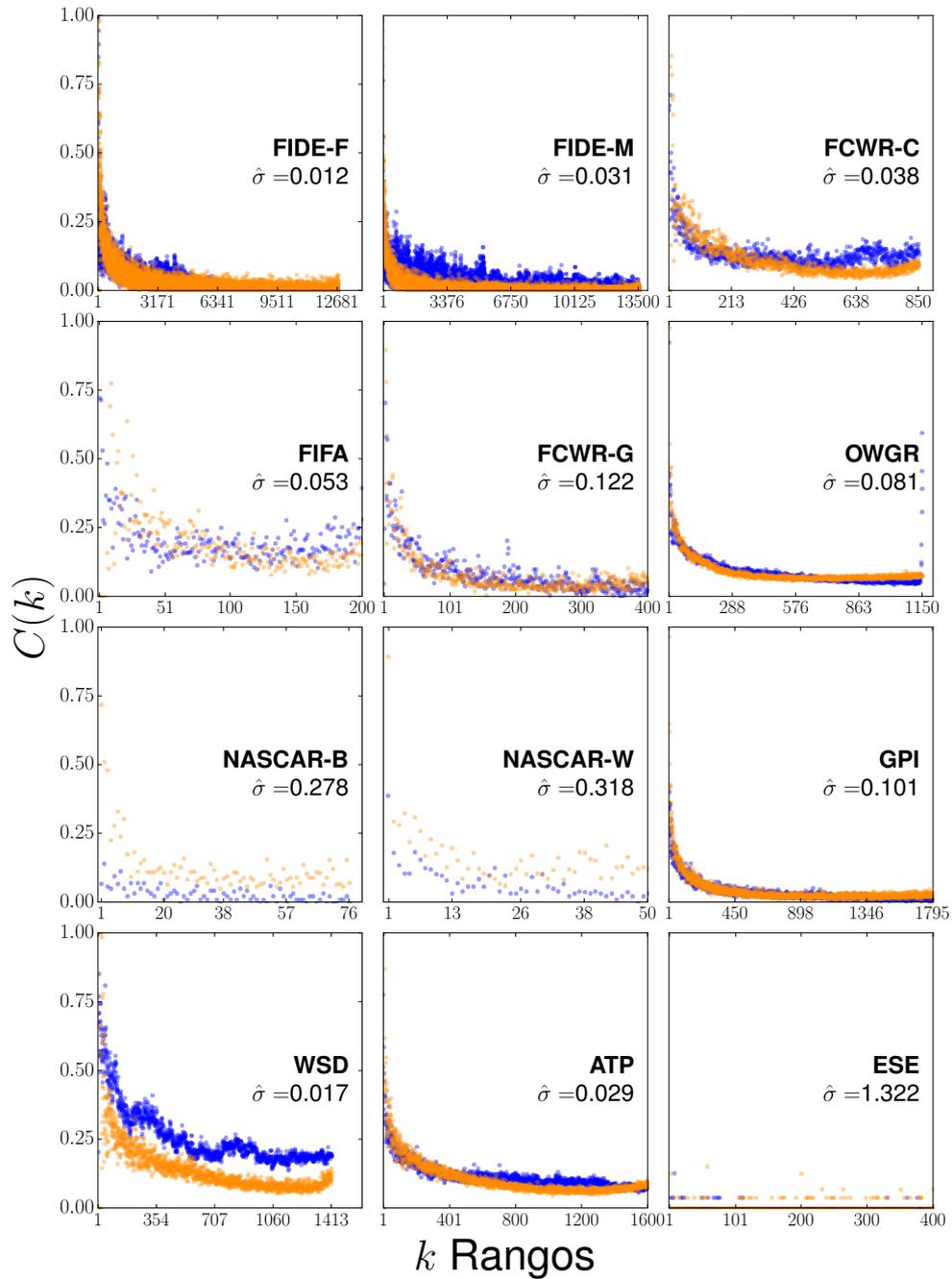


Figura C.5: Comparación entre las complejidades de rango empíricas y simuladas con el Modelo del Caminante Aleatorio. Gráfica que muestra la complejidad de rango $C(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo del caminante aleatorio (puntos anaranjados).

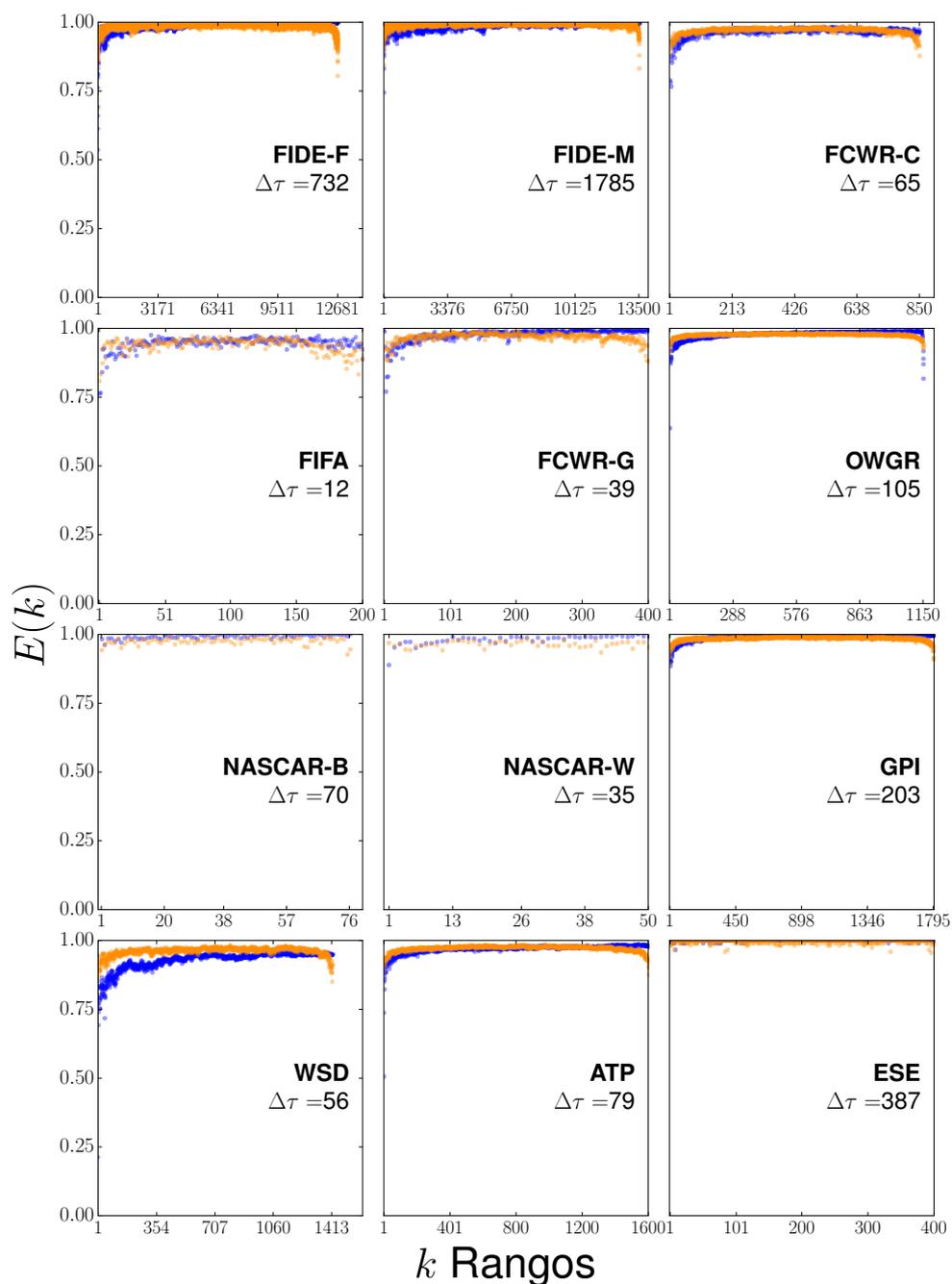


Figura C.6: Comparación entre las entropías de rango empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la entropía de rango $E(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados).

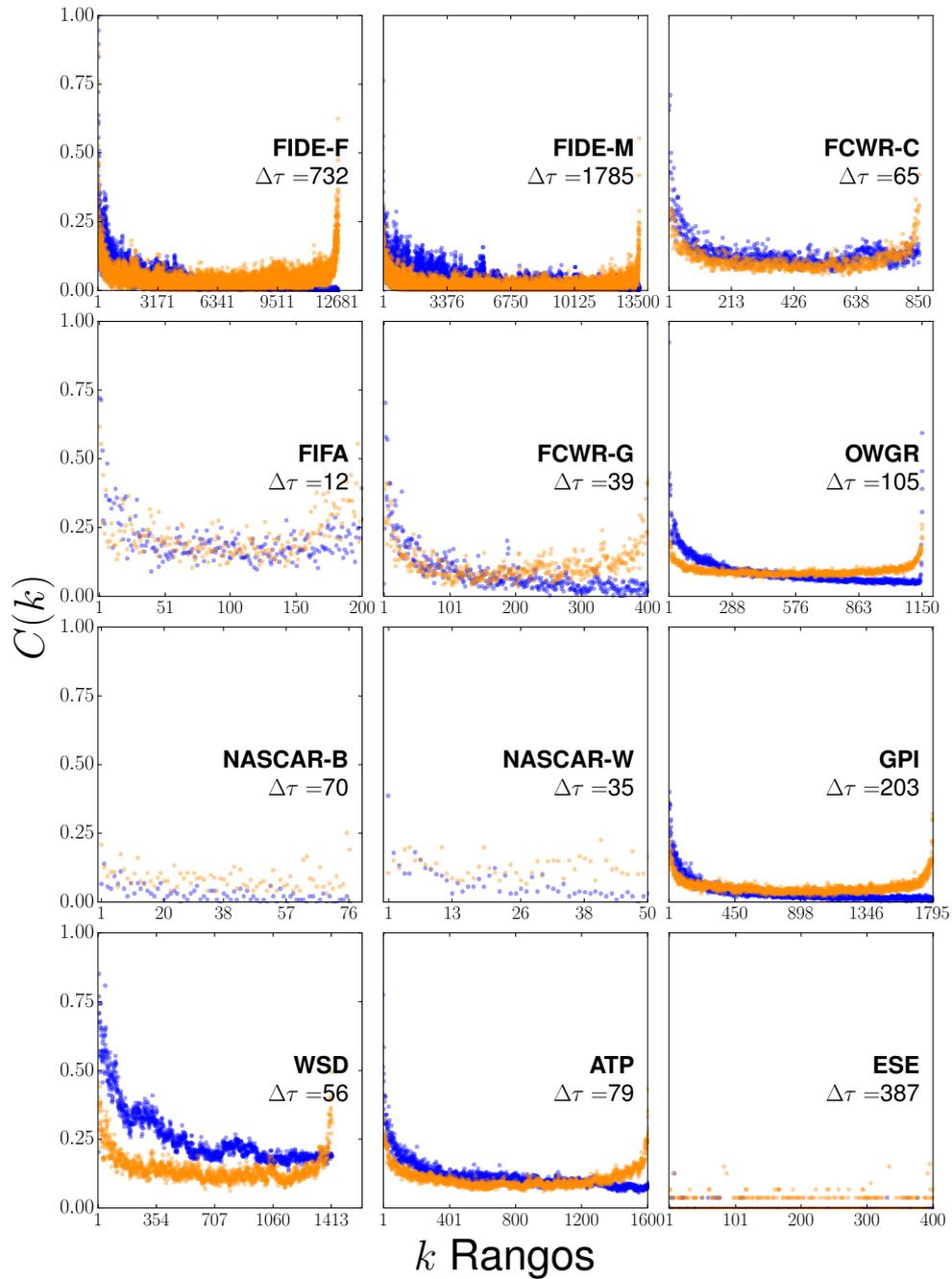


Figura C.7: Comparación entre las complejidades de rango empíricas y simuladas con el Modelo Nulo. Gráfica que muestra la complejidad de rango $C(k)$ de los datos empíricos de los sistemas reales (puntos azules) y los datos simulados con nuestro modelo nulo (puntos anaranjados).

Artículo original publicado.

Aquí se anexa el artículo publicado en la revista *EPJ Data Science* que es una revista de acceso libre en línea. El artículo fue publicado el 25 de Noviembre de 2016 y fui uno de los autores del mismo. Este trabajo de tesis está basado en esa publicación, sin embargo, en el presente se hizo un análisis más extenso, se añadieron conceptos y más casos de estudio. A continuación describo algunas de las cosas adicionales que contiene este trabajo de tesis. También añado la cita a dicha publicación [12].

En la publicación se trabajaron 6 deportes/juegos: Golf, Póquer, Clubes de Fútbol, Equipos nacionales de Fútbol, Jugadores masculinos de ajedrez y jugadores de Tenis. Para efectos de este trabajo se incluyeron estos mismos sistemas con más rodajas temporales en algunos casos y se adicionaron 6 disciplinas más: Jugadores de tabla sobre nieve, Entrenadores de clubes de fútbol, ganancias por jugar videojuegos, jugadores de ajedrez femenino y dos categorías de corredores de NASCAR.

En este trabajo no sólo se calculó el índice de Kolmogorov-Smirnov p para una rodaja temporal de los sistemas, sino que se calculó para todas las rodajas temporales con las que cuentan cada uno de los sistemas, para así proporcionar un promedio de dicha bondad de ajuste. Como vimos en el [Apéndice A](#), el cálculo de p para una rodaja temporal requiere de crear 2500 conjuntos artificiales de datos, por lo que calcular dicha cantidad para todas las rodajas temporales implica un gran procesamiento de datos.

En la publicación aquí anexada se trabaja la diversidad de rango como medida dinámica de los sistemas, sin embargo, en esta tesis en el [Capítulo 4](#) se introducen 3 medidas dinámicas adicionales: la probabilidad de cambio $p(k)$, la entropía de rango $E(k)$ y la complejidad de rango $C(k)$ que describen distintos aspectos de la evolución temporal de un sistema.

Asímismo, aparte del Modelo del Caminante Aleatorio que se describe en la publicación, agregamos en esta tesis el Modelo Nulo de manera superficial, esto para intentar reproducir la evolución de sistemas y generar bases de datos sintéticas que intenten modelar las bases con las que aquí contamos. Los detalles se explican en el [Capítulo 5](#).



Generic temporal features of performance rankings in sports and games

José A Morales¹, Sergio Sánchez¹, Jorge Flores², Carlos Pineda^{2,4*} , Carlos Gershenson^{3,4,5,6,7}, Germinal Cocho², Jerónimo Zizumbo¹, Rosalío F Rodríguez² and Gerardo Iñiguez^{3,8,9}

*Correspondence:

carlospgmat03@gmail.com

²Instituto de Física, Universidad Nacional Autónoma de México, México D.F., 01000, Mexico

⁴Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, México D.F., 04510, Mexico

Full list of author information is available at the end of the article

Abstract

Many complex phenomena, from trait selection in biological systems to hierarchy formation in social and economic entities, show signs of competition and heterogeneous performance in the temporal evolution of their components, which may eventually lead to stratified structures such as the worldwide wealth distribution. However, it is still unclear whether the road to hierarchical complexity is determined by the particularities of each phenomena, or if there are generic mechanisms of stratification common to many systems. Human sports and games, with their (varied but simple) rules of competition and measures of performance, serve as an ideal test-bed to look for universal features of hierarchy formation. With this goal in mind, we analyse here the behaviour of performance rankings over time of players and teams for several sports and games, and find statistical regularities in the dynamics of ranks. Specifically the rank diversity, a measure of the number of elements occupying a given rank over a length of time, has the same functional form in sports and games as in languages, another system where competition is determined by the use or disuse of grammatical structures. We use a Gaussian random walk model to reproduce the rank diversity of the studied sports and games. We also discuss the relation between rank diversity and the cumulative rank distribution. Our results support the notion that hierarchical phenomena may be driven by the same underlying mechanisms of rank formation, regardless of the nature of their components. Moreover, such regularities can in principle be used to predict lifetimes of rank occupancy, thus increasing our ability to forecast stratification in the presence of competition.

Keywords: complex systems; sports; data analysis; rank distribution; rank diversity

1 Introduction

Sports and games can be described as hierarchical complex systems due to the myriad of factors influencing the dynamics of competition and performance in them, including networked interactions, human and environmental heterogeneities, and other traits at the individual and group levels [1–4]. In particular, the performance of players and teams is influenced by a variety of causes: Economical, political and geographical conditions determine their rankings and may thus be used for predicting performance. Moreover, the (relatively) simple rules of competition and measures of performance associated with sports and games allow us to explore basic mechanisms of interaction leading to hierarchy for-

mation, which may be common to many systems driven by competition, not only leisure activities but other social, biological and economic systems. With this goal in mind, the availability of a large corpus of data related to sports, teams, and players allows researchers to perform multiple statistical analyses, in particular with respect to the structure and dynamics of performance rankings [5–7].

Data availability has made it possible not only to study the distribution of scores determining rankings, but also its time evolution [8]. In a recent paper, Deng *et al.* present a statistical analysis of 12 sports and report a universal scaling in rankings, despite the fact that the sports considered have very different ranking systems [9]. Here, we focus on the temporal trajectories of player and team performances, meaning the evolution of rank, with the objective of finding statistical regularities that indicate how competition shapes hierarchies of players and teams. In principle, rankings may be affected in time by events as apparently insignificant as a bad breakfast prior to an important event, or the weather during a competition [10]. Since these factors are inherently present for all activities, we would expect the evolution of rank to have generic features across sports and games.

We propose to quantify such evolution by means of a recently introduced measure, the *rank diversity*. With the help of the Google *n*-gram dataset [11], rank diversity has been used before to study how vocabulary changes in time [12]. That work shows that rank diversity has the same functional form for all languages studied, and is able to discriminate the size of the core of each language. Thus, here we concentrate on the temporal features of rank distributions corresponding to several sports and games with different ranking schemes. We consider data where an appropriate time resolution is available, and limit the analysis to six activities only: tennis, chess, golf, poker and football (both national teams and clubs). We find that all rank diversities have the same functional form as languages, despite having differences in their rank frequency distributions. Finally, we introduce a random walk model that, tuned by the parameter values of each dataset, reproduces qualitatively the diversity of all sports and games considered. Overall, our goal is to use rank diversity as a tool to understand rank dynamics in sports, games, and other hierarchical complex systems, thus enabling us to identify the dependence on rank of a change in the hierarchy of the system. By using this analysis, we may be able to estimate how well can a change in rank be predicted, regardless of the particularities of the phenomenon under study.

The article is organized as follows. In Section 2 we describe the datasets used. We then analyse ranking distributions in Section 3 and compare them with several models. In Section 4 we study the rank diversity for each sporting activity and compare it with a random walk model. The main conclusions of our analysis are included in Section 5. In Appendix A we discuss in detail the Kolmogorov-Smirnov index, which measures the goodness of fit for a given dataset. Finally, in Appendix B we describe the generic relation between rank diversity and the cumulative rank distribution in the random walk model.

2 Ranking data

We use ranking data on players and teams from six sports and games: (a) Tennis players (male), ranked by the Association of Tennis Professionals (ATP) [13]; (b) Chess players (male), ranked by the Fédération Internationale des Échecs (FIDE) [14]; (c) Golf players, ranked by the Official World Golf Ranking (OWGR) [15]; (d) Poker players, ranked by the Global Poker Index (GPI) [16]; (e) Football teams, ranked by the Football Club World

Table 1 Summary of ranking data for each sport and game considered in this study

Sport/game	Data source	Time period	Ranking resolution	#players/teams
Tennis players (male)	Association of Tennis Professionals (ATP) [13]	May 5 2003-Dec 27 2010	Weekly	1,600
Chess players (male)	Fédération Internationale des Échecs (FIDE) [14]	Jul 2012-Apr 2016	Monthly	13,500
Golf players	Official World Golf Ranking (OWGR) [15]	Sept 10 2000-Apr 19 2015	Weekly	1,000
Poker players	Global Poker Index (GPI) [16]	Jul 25 2012-Jun 10 2015	Weekly	1,799
Football teams	Football Club World Ranking (FCWR) [17]	Feb 1 2012-Dec 29 2014	Weekly	850
National football teams	Fédération Internationale de Football Association (FIFA) [18]	Jul 2010-Dec 2015	Monthly	150

Table listing the main properties of the ranking data used here (including data source, time period, ranking resolution, and number of players/teams). In order to have a homogeneous distribution of ranking snapshots and the same number of players/teams in each snapshot for a given activity, we disregard some data for the ATP, FIDE, OWGR, GPI, and FIFA datasets, as explained in the main text.

Ranking (FCWR) [17]; and (f) national football teams, ranked by the Fédération Internationale de Football Association (FIFA) [18].

The ranking procedure varies among sports. In ATP, for example, tennis players are ordered according to the number of points they have up to the date of publication of the ranking. The number of points depends on the tournaments players have participated in (and how well they have performed), but not all tournaments are taken into account. FIDE uses the Elo system [19] to rank players, which considers the number of matches, their results, and the opponent ranking. The FIFA ranking takes into account official matches between countries. The number of points depends on the confederation and classification of each team, as well as the importance and result of the match. Table 1 summarises the main properties of the ranking data considered in this study, including the time resolution used to measure rankings (*i.e.* the time difference between two snapshots of the ranking in a sport or game). In order to have a homogeneous distribution of ranking snapshots and the same number of players/teams in each snapshot for a given activity, we disregard some data for the ATP, FIDE, OWGR, GPI, and FIFA datasets: In all of these cases, the time elapsed between the publication of two rankings varies greatly (from less than a week to more than a month), and the number of players/teams across ranking snapshots may change as well. Therefore, for each dataset we choose a constant time resolution of rankings (weeks or months, as shown in Table 1) that maximises and keeps constant the number of ranked players/teams throughout time. All datasets, filtered as explained above, are included in Additional file 1.

3 Comparison with ranking models

Player or team performance is usually measured by a *score* that varies with time. This score results in a time-dependent *rank* with a rather complex behaviour, as we will explain below. We first focus on the distribution of scores versus ranks (*i.e.* a *rank distribution*) for a given time. Particularly, we are interested in seeing if this distribution can be reproduced by a single ranking model for all sports and games considered. We select five ranking models to fit the data, four of which are particular cases of

$$f(k) = \mathcal{N} \frac{(N+1-k)^a \exp(-bk)}{k^a}, \quad (1)$$

where f is the score associated with rank k , a is an exponent that dominates most of the curve, b an exponent controlling its exponential decay, and q an algebraic decay that regulates a sharp drop of the curve for large k . Finally, N is the total number of elements (*i.e.* players or teams) in the system, and \mathcal{N} is a normalization constant.

The first four models are

$$\begin{aligned}
 m_1(k) &\propto \frac{1}{k^a}, & m_2(k) &\propto \frac{\exp(-bk)}{k^a}, \\
 m_3(k) &\propto \frac{(N+1-k)^q}{k^a}, & m_4(k) &= f(k),
 \end{aligned}
 \tag{2}$$

whereas the fifth model is a double Zipf law [20],

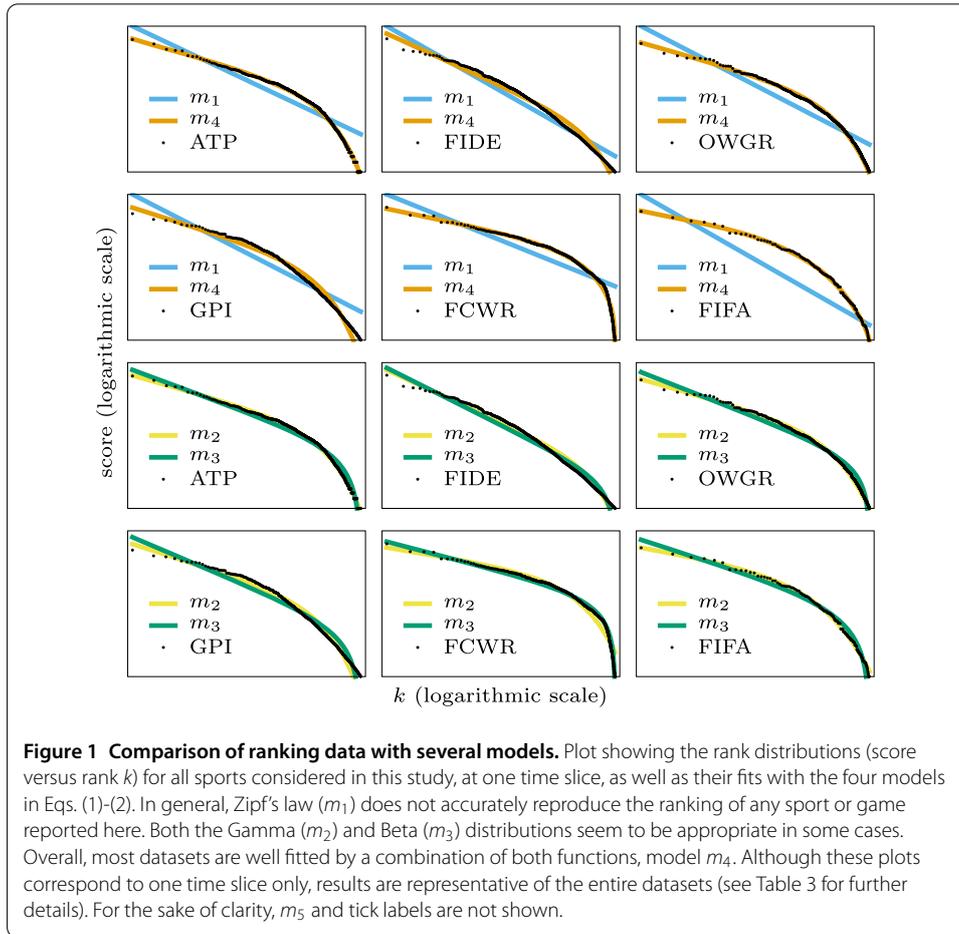
$$m_5(k) = \mathcal{N} \begin{cases} \frac{1}{k^a}, & k \leq k_c, \\ \frac{k_c^{a'-a}}{k^a}, & k > k_c, \end{cases}
 \tag{3}$$

with a' an alternative exponent that regulates the behaviour of the curve after a critical rank k_c . Model m_1 is obtained by setting $q = b = 0$ in Eq. (1), and has been considered in a vast amount of studies, both in the realm of sports [21, 22] and in other studies of ranking behaviour [23], including the famous Zipf’s law of languages where the particular case $a = 1$ has drawn a lot of attention (see, *e.g.*, [24] and references therein). The Gamma (m_2) and Beta (m_3) distributions have been useful in many disciplines for decades; a quick look at their Wikipedia entries provides numerous examples [25, 26]. Model m_4 , being a more general expression than the previous ones, tends to provide a better fit at the expense of more parameters, and will serve as benchmark for the comparison between the rest of the models. Finally, model m_5 in Eq. (3) has been used with success in several contexts [20, 27], prompting us to test it in the area of sports and games.

The results of the fitting process between data and Eqs. (1)-(3) are shown in Figure 1, while Table 2 summarises the parameter values obtained. Data corresponds to a single time snapshot for all sports and games: Dec 27 2010 (ATP); Sept 2014 (FIDE); Mar 18 2015 (GPI); Apr 19 2015 (OWGR); Dec 29 2014 or Week 53 2014 (FCWR); and Dec 18 2014 (FIFA). The following results are, however, representative of all time snapshots (see Table 3 and the text below for further details). Both models and data show variation in their goodness of fit. From Figure 1 it is clear that Zipf’s law (m_1) is not adequate. On the other hand, the Gamma distribution (m_2) fits some datasets rather well, particularly those that do not show an abrupt fall of score as a function of rank. Datasets with an abrupt decay of frequency are well fitted by the Beta distribution (m_3) instead. However, most sports and games seem to be an intermediate case where both functions capture global behaviour accurately, and thus the fit is considerably better for a combination of both models, *i.e.* m_4 . We also see that the double Zipf law (m_5) is a good fit for FIDE and GPI, as seen from Table 3.

In order to objectively compare goodness of fit between models, we consider several measures: The coefficient of determination R^2 [28], the maximum deviation between theory and observation D , and the Kolmogorov-Smirnov index p [28, 29]. The coefficient R^2 is calculated from the 2-norm with respect to the data coming from a single time snapshot,

$$R^2 = 1 - \frac{\sum_k [m_i(k) - y_k]^2}{\sum_k [y_k - \langle y \rangle]^2},
 \tag{4}$$



for a given model $m_i(k)$, $i = 1, \dots, 5$, and data y_k , where $\langle y \rangle$ is the expectation value of y_k . The closer R^2 is to one, the better the fit is. To calculate D , we consider the cumulative of both the proposed distribution $m_i(k)$ and a dataset with N data points (the equivalence between an empirical rank-value distribution and the empirical cumulative distribution corresponding to scores is discussed in [28]). There it is also shown that for a given theoretical rank distribution $m_i(k)$, the cumulative is simply $M_i(m_i) = [N + 1 - k(m_i)]/[N + 1]$, where $k = k(m_i)$ is the inverse function of m_i that implicitly depends on scores. Whereas for a dataset, $M_{\text{data}}(s) = (1/N) \sum_j \theta(s - s_j)$, with θ a step function and $\{s_1, \dots, s_N\}$ the set of scores in the data [28]. We then define D (the so-called Kolmogorov statistics) as the maximum vertical difference between the two curves, $D = \sup_s |M_i(m_i) - M_{\text{data}}(s)|$. The calculation of the Kolmogorov-Smirnov index p is more involved so we discuss it in Appendix A. The measure p allows us to consider that a small dataset will have some noise due to poor statistics. Thus, if a model is consistent with a dataset, but we have poor statistics, we might still have a good (large) p . Usually, a 'good' fit is required to have $p > 0.1$, see e.g. [30].

Table 3 shows the mean values $\langle R^2 \rangle$ and $\langle D \rangle$ (and their associated standard deviations σ_D and σ_{R^2}), averaged over all time slices available, for the fitting process between the six datasets and five models m_i used here. We also include values of p for the single time slice of Figure 1. Higher $\langle R^2 \rangle$ and lower $\langle D \rangle$ imply better fits. Since σ_D and σ_{R^2} are small, the fits shown in Figure 1 are representative of the entire datasets. We observe that none

Table 2 Parameter values for fitting process between sports data and ranking models

	Model m_1		Model m_2		Model m_3		Model m_4		Model m_5		k_c	
	$\log \mathcal{N}$	a	$\log \mathcal{N}$	a	b	$\log \mathcal{N}$	a	q	$\log \mathcal{N}$	a		a'
ATP	4.51	1.04	4.11	0.626	3.18×10^{-3}	-1.46	0.816	1.79	3.13×10^{-3}	3.12×10^{-2}	3.004	3.19×10^2
FIDE	3.46	0.0252	3.46	2.01×10^{-2}	6.25×10^{-6}	3.32	2.21×10^{-2}	3.28×10^{-2}	0.0202	5.05×10^{-9}	0.016	1.97×10^2
OWGR	1.35	0.702	1.05	0.383	2.68×10^{-3}	-0.928	0.53	0.703	0.385	1.18×10^{-2}	0.452	2.03×10^2
GPI	3.75	0.234	3.66	0.144	6.63×10^{-4}	2.54	0.193	0.358	0.144	3.64×10^{-9}	0.133	1.08×10^2
FCWR	4.52	0.529	4.24	0.218	3.06×10^{-3}	2.19	0.341	0.732	0.269	0.458	3.47	4.26×10^2
FIFA	3.37	0.473	3.24	0.142	1.02×10^{-2}	2.30	0.262	0.456	0.148	4.16×10^{-2}	0.227	33.86

Table listing parameter values for all models in Eqs. (1)-(3), obtained in the fitting process with empirical data. These values correspond to the model curves in Figure 1 (model m_5 not shown there), but are representative of the entire datasets (see Table 3 for further details).

Table 3 Goodness of fit measures

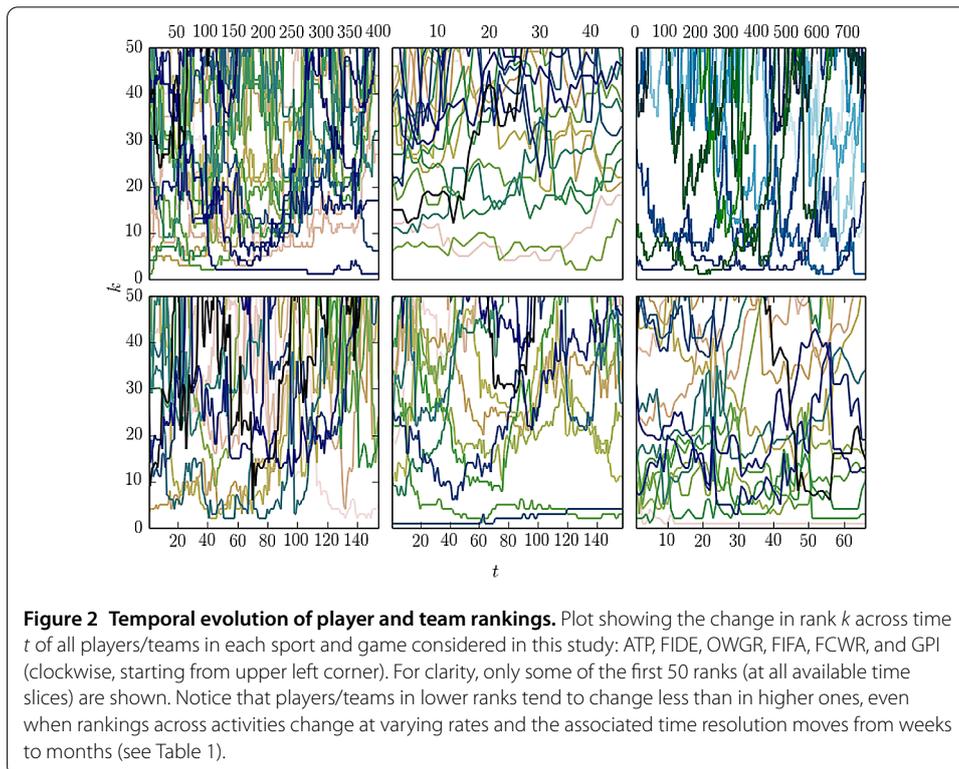
		m_1	m_2	m_3	m_4	m_5
ATP	$\langle R^2 \rangle$	0.222	0.982	0.879	0.982	0.964
	$\langle D \rangle$	0.433	0.044	0.08	0.038	0.077
	σ_{R^2}	0.0969	0.01652	0.009	0.0124	0.0288
	σ_D	0.211	0.0126	0.0672	0.0128	0.0287
	p	0.01	0.17	0.0	0.12	0.0
FIDE	$\langle R^2 \rangle$	0.777	0.936	0.657	0.936	0.991
	$\langle D \rangle$	0.477	0.2	0.188	0.2	0.141
	σ_{R^2}	0.0071	0.0053	0.0028	0.0054	0.0035
	σ_D	0.0072	0.0048	0.0166	0.0048	0.0005
	p	0.0	0.0	0.0	0.0	0.0
OWGR	$\langle R^2 \rangle$	0.631	0.981	0.943	0.982	0.97
	$\langle D \rangle$	0.316	0.046	0.088	0.043	0.088
	σ_{R^2}	0.0264	0.0388	0.0138	0.0381	0.0391
	σ_D	0.1292	0.0165	0.0192	0.0152	0.0104
	p	0.0	0.92	0.0	0.89	0.0
GPI	$\langle R^2 \rangle$	0.791	0.978	0.937	0.978	0.985
	$\langle D \rangle$	0.531	0.201	0.149	0.201	0.202
	σ_{R^2}	0.01029	0.0115	0.0044	0.0115	0.0459
	σ_D	0.01612	0.0039	0.0048	0.0039	0.00533
	p	0.0	0.0	0.0	0.0	0.0
FCWR	$\langle R^2 \rangle$	0.727	0.986	0.981	0.997	0.947
	$\langle D \rangle$	0.295	0.115	0.057	0.055	0.172
	σ_{R^2}	0.0186	0.0183	0.0098	0.0112	0.0268
	σ_D	0.02833	0.0046	0.0052	0.00128	0.0104
	p	0.0	0.0	0.0	0.0	0.0
FIFA	$\langle R^2 \rangle$	0.833	0.993	0.981	0.996	0.979
	$\langle D \rangle$	0.387	0.076	0.071	0.041	0.155
	σ_{R^2}	0.0277	0.0324	0.0135	0.0114	0.0413
	σ_D	0.02888	0.004	0.007	0.002	0.0147
	p	0.0	0.99	0.0	0.99	0.02

Table listing mean values $\langle R^2 \rangle$ and $\langle D \rangle$ (and their associated standard deviations σ_D and σ_{R^2}), averaged over all time slices available, for the fitting process between the six sports and five theoretical rank distributions used here. We also include values of the Kolmogorov-Smirnov index p for the single time slice of Figure 1. Higher $\langle R^2 \rangle$ and lower $\langle D \rangle$ imply better fits. Since σ_D and σ_{R^2} are small, the fits shown in Figure 1 are representative of the entire datasets. The best fits for each sport are shown in bold.

of the models are a good fit for all sports and games, although m_4 and m_5 are the most appropriate (in terms of R^2). However, in three cases (FIDE, GPI, and FCWR) we have $p = 0$ for model m_4 , and no model fits well, meaning that the theoretical distribution is not followed by the data. We stress again that Zipf’s law (m_1) is the worst fit among all considered, except for FIDE. It is interesting to notice that R^2 and p lead to different criteria of what is a ‘good’ or a ‘bad’ model. This is due to both the amount of available data and the number of parameters in the model. The larger the data, the easier it is to distinguish the best available model from a good (but not accurate enough) approximation. On the other hand, the more parameters the model has, the easier it is to fit any data. Both of these aspects are taken into account in the definition of p , but not in R^2 .

4 Rank diversity in sports and games

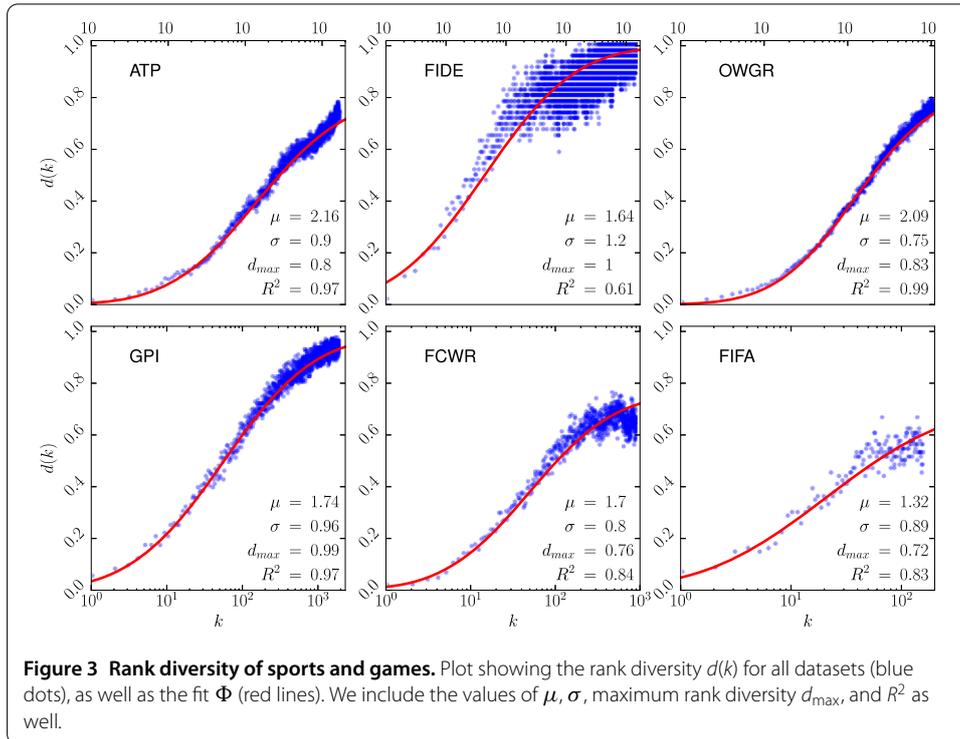
The previous analysis of the functional form of the rank distribution in several sporting activities (even when the goodness of fit has been averaged over time) is restricted by the fact that the rank distribution is inherently an *instantaneous* measure, in the sense that it captures ranking at a given point in time and does not take into account the dynamics of



players and teams changing rank as time goes by. In order to overcome this issue, here we contribute to the analysis of ranking in sports and games by computing the rank diversity, a measure of the number of elements occupying a given rank over a length of time. From previous [12] and current work, it appears that rank diversity has the same functional form, not only for sports but also for other complex systems, such as countries classified by their economic complexity, the 500 leading enterprises ranked by *Fortune* magazine, or a set of millions of words in six Indo-European languages.

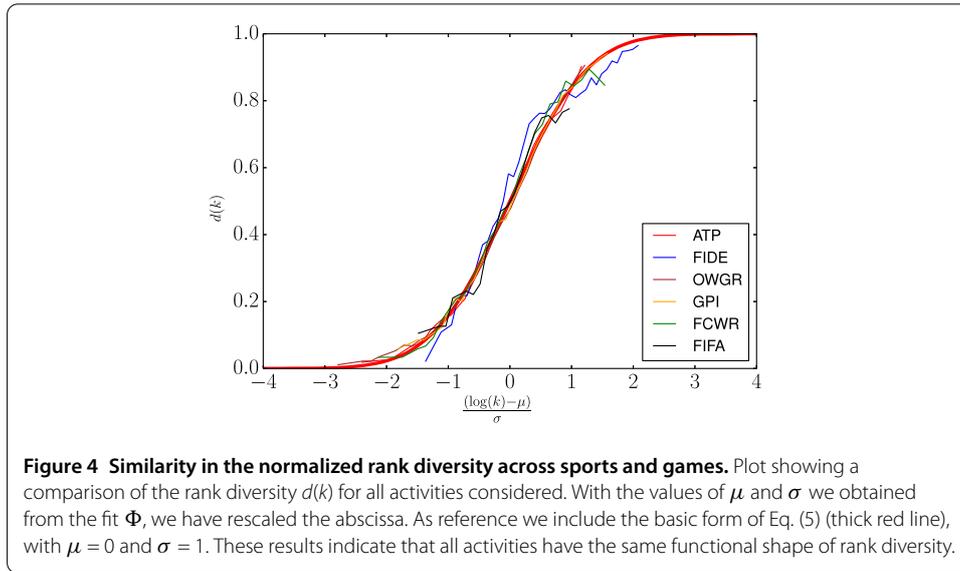
The rank diversity $d(k)$ is defined as the number of distinct elements in a complex system that occupy the rank k at some point during a given length of time. In other words, we choose to focus on the time dependence of ranks, rather than on the static (*i.e.* defined for a single time) rank distribution $f(k)$. An example of the change of ranks in time for the sports and games studied here can be seen in Figure 2. These so-called ‘spaghetti’ curves show how elements - individuals or teams - change their rank in time. The rank diversity $d(k)$ is simply the normalized number of different elements (curves) that spend at least one time interval at a given rank k . The rank diversity for the various sports and games considered here is shown in Figure 3.

We should stress that $d(k)$ and $f(k)$ measure different aspects of the hierarchical structure of a complex system. First of all, the rank diversity includes information on how elements change rank throughout time in a single function, while the rank distribution captures the hierarchy in the system for a single time interval. Secondly, the rank diversity disregards any information on the scores of elements beyond their order, and thus the same $d(k)$ may be obtained for several shapes of $f(k)$ (power-law, Gamma, Beta, etc.). As an example, consider any transformation in time of the scores of elements in the system, such that their ranking order stays the same; then $f(k)$ could interpolate between differ-



ent functional shapes as time goes on, while $d(k)$ would stay constant. The inverse case is also possible, and any rank distribution may produce a wide variety of rank diversities. For example, we could construct several dynamics of scores that keep the number of elements with a given score constant, but that change the amount of time an element holds certain score, thus keeping $f(k)$ fixed and changing $d(k)$. Overall, both $d(k)$ and $f(k)$ measure some aspects of the structure and dynamics of hierarchy in a complex system, but only the rank diversity captures the way elements change their positions in the hierarchy, beyond minor changes in scores that could be attributed, for example, to different ways of measuring performance.

From Figure 3 we see that the empirical curves for rank diversity are (roughly) monotonic and have a single shoulder. The cumulative of a square-integrable function with a single bump would have these properties, and a Gaussian is arguably the simplest choice. Moreover, an analytical argument (see Appendix B) suggests that this may be an appropriate ansatz under very general conditions, at least qualitatively. In a large variety of physical systems composed of alike elements with similar interactions between them, the macroscopic response of the system is usually determined by general laws such as equations of state. However, in different empirical realisations of the same dynamics there may be differences associated to the law of the large numbers or the central limit theorem. These differences across realisations follow a normal Gaussian distribution, according to the Gaussian theory of errors. However, for complex systems with competitive dynamics, there may be generic features described by the Gamma (m_2) and Beta (m_3) distributions [31, 32], and there may also be differences across realisations that follow a multiplicative dynamics. This is indeed the case for several Indo-European languages [12] and for the games and sports datasets considered here (See Figure 1). In Appendix B we introduce the non-trivial idea that there are two different dynamics associated with so-called generic and contingent fea-



tures, which may be described in terms of a one-step Markovian, Gaussian process. This allows us to establish an explicit relation between the diversity $d(k)$ and the cumulative of the rank distribution, $S(t)$.

In fact, studying $d(k)$ for six Indo-European languages [12], we found that the observed rank diversity closely follows the cumulative of a Gaussian (*i.e.* a sigmoid)

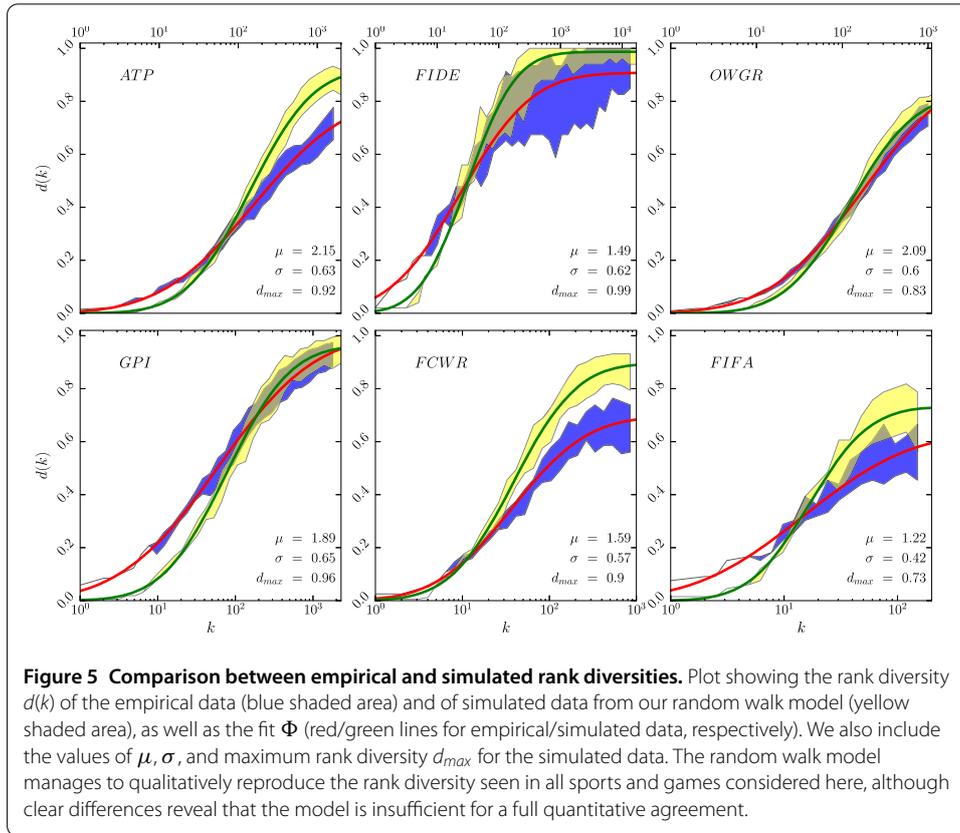
$$\Phi_{\mu,\sigma}(\log k) = \frac{\max_i d(k_i)}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\log k} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) dy. \tag{5}$$

The mean value μ is set as the smallest k_0 for which $d(k_0) = \frac{\max_i d(k_i)}{2}$, while the width σ is fitted and gives the scale for which $d(k)$ gets close to its extreme values. If k_{\pm} are given by $\log_{10} k_{\pm} = \mu \pm 2\sigma$, the bulk of the changes in the values of diversity lies between k_- and k_+ . In Figure 3 we show the fit Φ for all sports and games considered here (R^2 values for the Φ curves are shown there as well). We do not consider neither D nor p , since these measures are only meaningful for distributions, which $d(k)$ is not. To compare different rank diversity curves, their rank can be normalised to $\frac{\log(k)-\mu}{\sigma}$, as shown in Figure 4. Since all the cases considered can be fitted with the sigmoid curve of Eq. (5), we argue that the rank diversity of sports seems to have a generic shape.

4.1 A random walk model

From Figure 2 and Figure 4 we see that players and teams with low ranks change very slowly or not at all, while those with higher k have a larger rank variation in time. This intuition is clear from recent experience in sports like tennis and football: According to the analysed datasets, Hewitt, Nadal, Roddick, Ferrero, Agassi and Federer have been the only number one tennis players from May 2003 till December 2010. The same holds for football clubs: Real Madrid, Atlético Madrid, Barcelona, and Bayern München have been the best-ranked teams from January 2012 till December 2014. In other words, players and teams with small k tend to have a small rank diversity.

In what follows we propose a simple model [12] that captures such intuition (*i.e.* a variation approximately proportional to the current rank), and whose rank diversity resembles



the data presented here. We call this model a scale-invariant random Gaussian walk, since a member with rank k_t , at the discrete time t , is converted to rank k_{t+1} according to the following procedure: We define an auxiliary variable l_{t+1} , which we call pre-rank, at time $t + 1$ by the relation

$$l_{t+1} = k_t + G(k_t \hat{\sigma}), \tag{6}$$

where $G(k_t \hat{\sigma})$ is a Gaussian-distributed random number with standard deviation $k_t \hat{\sigma}$ and mean 0. This means that the random variable l_{t+1} has a width distribution proportional to k_t , and thus will, for small k_t , have small changes as well. Once the values of the pre-ranks l_{t+1} for all members are obtained, we order them according to their magnitude. This new order gives new rankings, *i.e.* the k values at time $t + 1$. The only parameter left in the model is the relative width $\hat{\sigma}$, which we fit by using a least-squares method over a smoothed version of the empirical rank diversity. In Figure 5 we show the rank diversity for systems with the same number of elements as those of Figure 3, but generated with the random model. We see that these two sets of plots are qualitative similar, although clear differences reveal that the model is insufficient for a full quantitative agreement. The fact that both the empirical and simulated rank diversities have a sigmoid shape suggests that rank changes in real systems may be the result of a large number of multiplicative processes. We discuss some analytical ideas supporting this insight in Appendix B. However, the mismatch between model and data seen in Figure 5 shows that not all characterizing features of the empirical process are captured by our model, and further investigation is needed.

5 Discussion and conclusions

Competition and heterogeneous performance are characteristic of the elements of many complex systems in biological, social and economic settings. Despite the fact that these systems show a large variation in the definitions of their constituents and in the relevant interactions between them, it remains to be seen whether the emergence of hierarchical structure is mostly determined by the particularities of each phenomenon, or if there are mechanisms of stratification common to the temporal evolution of many systems. We have explored this notion by considering a set of relatively controlled and simplified systems driven by competition: Human sports and games, where the rules of engagement and measures of performance are well defined, in contrast to, say, the ranking of physicists (the question of whom is the 'best' physicist would have an ambiguous answer, to say the least). This allows us to characterise the emergence of hierarchical heterogeneity by comparing the temporal features of rankings of individuals and teams across activities in a clear way. Explicitly, we analysed the statistical properties of rank distributions in six sports and games, each with different number of members and rules for calculating scores (and, therefore, ranks). By comparing rank distributions with several ranking models, we find that the Zipf law (model m_1) does not provide a suitable fit for the empirical data. Even if the more generic ranking model m_4 (a combination of the Gamma and Beta distributions) tends to offer good fits, it is not always the best.

Furthermore, we studied the temporal features of rankings explicitly by calculating the rank diversity $d(k)$, a measure of the number of individuals or teams occupying a given rank over a length of time. We found that $d(k)$ has the same sigmoid-like functional form, even for relatively small systems like FIFA (with only 150 elements per time slice). Coupled to the fact that a sigmoid rank diversity has also been found in the way vocabulary changes in time [12], our results suggest that the emergence of hierarchical complexity - as measured by $d(k)$ - may have traits common to many systems. This claim is underlined by the fact that a simple model (the scale-invariant random Gaussian walk) can reproduce the diversity of the sports and games studied here, and also of languages [12]. One could initially suspect that rank changes depend on the intrinsic strength or qualities of players and teams. However, given the fact that our random walk model reproduces relatively well the rank dynamics of several sports and games, it seems that rank change can instead be characterised as a random process. This does not imply that rank change is random, but that the specific mechanisms associated with each activity and ranking system are irrelevant for the calculation of rank diversity.

A natural direction to follow in the near future is to study the behaviour of rank diversity in other competitive phenomena beyond sporting activities and language, such as physical, social and economic processes of stratification. If indeed a certain universality in the temporal features of rankings is present in other complex settings, it would indicate that hierarchical phenomena may be driven by the same underlying mechanisms of rank formation, regardless of the nature of their components. Potentially, we may exploit such regularities to predict lifetimes of rank occupancy, thus increasing our ability to forecast stratification in the presence of competition.

Appendix A: Explicit calculation of Kolmogorov-Smirnov p -value

The Kolmogorov-Smirnov p -value is a way to quantify the goodness of fit of some theoretical distribution to the empirical distribution of a dataset. For a given dataset $\{s_1, s_2, \dots, s_N\}$,

the corresponding empirical distribution is defined as

$$M_{\text{data}}(s) = \frac{1}{N} \sum_j \theta(s - s_j), \tag{7}$$

where θ is a step function and the variable s represents scores. The goodness of fit is obtained via such empirical distribution and a theoretical cumulative distribution (CCD). Thus, we need to define the rank distribution in terms of a CCD in order to use this criterion. Ref. [28] shows that there is an equivalence between an empirical rank-value distribution and the empirical cumulative distribution of scores (or frequencies) available from the data. The formula that relates these two functions is

$$M_i(m_i) = \frac{N + 1 - k(m_i)}{N + 1}, \tag{8}$$

where m_i is the value of the theoretical rank distribution, and k the rank related to score s . So, to obtain the corresponding CCD of a rank distribution m_i , it is enough to apply Eq. (8). Note that $k = k(m_i)$, *i.e.* k is the inverse function of m_i . The p -value will then measure how good $M_i(m_i)$ fits the empirical distribution of scores. Indirectly, we are obtaining a measure of goodness of fit of m_i to the empirical rank-value distribution, due to the equivalence stated in Eq. (8). In our case, the theoretical m_i is given by Eq. (2) and Eq. (3).

Next we define the Kolmogorov statistic D as the maximum distance between the empirical distribution of scores and the theoretical cumulative distribution,

$$D = \sup_s |M_i(m_i) - M_{\text{data}}(s)|. \tag{9}$$

We stress that when we talk about m_i , value means a score in the system.

Finally, we describe the process used to calculate the p -value:

1. Compute the parameters of fit m_i for the empirical rank-value distribution (scores).
2. Obtain the empirical distribution of scores and the $M_i(s)$ with Eq. (7) and Eq. (8).
3. Calculate the Kolmogorov statistic D between M_i and M_{data} .
4. Generate (*e.g.* 2,500) artificial datasets of scores, distributed according to the fitted M_i . For each of them, fit to an artificial $M_{i,\text{art}}$ in order to obtain a value D_{art} .
5. Count how many of the 2,500 D_{art} values are larger than the D value of the real dataset and divide it by 2,500. The result is the p -value.

Appendix B: Diversity and cumulative distribution

In previous work we have shown that, under very general conditions in which dynamic competition exists between positive and negative mechanisms, like birth and death processes, the rank distribution is given by the ratio of two power laws [32]. In this Appendix we analyse the difference between the data associated with different realisations of such competitive dynamics and the adjustments to real data in terms of stochastic models such as $m_2(k)$, $m_3(k)$, and $m_4(k)$ given by Eq. (2). Specifically, we adopt the more general point of view that the data (obtained for Indo-European languages [12] and several sports and games) may be represented by a one-step Markovian stochastic process for the allocation of ranks.

The difference between the data associated with several realisations of the competitive dynamics and the adjustments to the real data may be analysed by treating k as a continuous variable. In this case, the time evolution of the probability density distribution of ranks $P(k, t)$ is described by a Fokker-Planck equation (FPE),

$$\frac{\partial}{\partial t} P(k, t) = -\frac{\partial}{\partial k} [A(k)P] + \frac{\partial^2}{\partial k^2} [B(k)P], \tag{10}$$

where $A(k)$ and $B(k)$ are rank-dependent drift and diffusion coefficients, respectively.

Note that in Figures 1, 3 and 4 the abscissa is not the rank k , but $x = \log k$. In other words, the systems exhibit a simpler behaviour in terms of the variable x , a fact that suggests a multiplicative behaviour and, in turn, a log-normal process. This process is the statistical realisation of the multiplicative product of many independent positive random variables, a feature that is justified by considering the central limit theorem in the logarithmic domain, and thus obeys the log-normal distribution. As a consequence, $P(k, t)$ can be expressed in the general form

$$P(x, t) = P^{st}(x) + P_1(x, t), \tag{11}$$

with $x = \log k$. The explicit form of the stationary distribution $P^{st}(x)$ is well known [33, 34], and the time dependent solution $P_1(x, t)$ may be determined as follows. We first note that Eq. (10) may be rewritten as

$$\frac{\partial}{\partial t} P(x, t) = \frac{\partial}{\partial x} [B(x)P_x] + \alpha P_x + \beta P, \tag{12}$$

where $\alpha = -A + B_x$, $\beta = -A_x + B_{xx}$, and each subscript \bullet_x denotes a partial derivative with respect to x . This equation can be further simplified by introducing the variable $v(x, t) \equiv B(x)P(x, t)$. Moreover, in order to simplify the discussion and the resulting equations, we consider the particular case where the drift and diffusion coefficients $A(x)$ and $B(x)$ are proportional to the same function $g(x)$, *i.e.*, $A(x) = \lambda_A g(x)$ and $B(x) = \lambda_B g(x)$. If $\tau \equiv B(x)t$, then Eq. (12) reduces to

$$\frac{\partial}{\partial \tau} v(x, \tau) = -\Lambda \frac{\partial v}{\partial x} + \frac{\partial^2 v}{\partial x^2}, \tag{13}$$

with $\Lambda \equiv \lambda_A/\lambda_B$. Let us now introduce the multiplicative character mentioned above by introducing $u(x, \tau)$ through the following change of variables,

$$\log \frac{v(x, \tau)}{u(x, \tau)} = \Lambda x - \frac{\Lambda^2}{4} \tau. \tag{14}$$

As a result Eq. (13) reduces to the diffusion equation

$$\frac{\partial}{\partial \tau} u(x, \tau) = \frac{\partial^2 u}{\partial x^2}, \tag{15}$$

whose formal solution is a Gaussian,

$$u(x, \tau) = \frac{1}{\sqrt{4\pi\tau}} \int_{-\infty}^{+\infty} e^{-(x-x')^2/4\tau} u(x', 0) dx'. \tag{16}$$

Starting from some initial state x_0 , the distribution of the amount of time required for a stochastic process to encounter a threshold for the first time is known as the first passage time distribution (FPTD). We will now exhibit the relation between the diversity and the diffusion equation (15). To this end and to simplify the notation, in what follows we shall again use the symbol t to denote τ .

Consider the absorbing boundary $u(x_c, t) = 0$, where the subscript c identifies the absorption point x_c , and let $u(x, t; x_0, x_c)$ denote the probability density satisfying this boundary condition for $x < x_c$. The survival probability $S(t, x_c)$ that the particle has remained at a position $x < x_c$ for all times up to t , is given by

$$S(t, x_c) \equiv \int_{-\infty}^{x_c} u(x, t; x_0, x_c) dx, \quad (17)$$

which is also the cumulative distribution of x at time t . Let the probability that a particle has reached the absorption point between times t and $t + dt$ be $h(t) dt = S(t) - S(t + dt)$. If we use a first order Taylor approximation, the first passage time distribution $h(t)$ is then given by

$$h(t) = -\frac{\partial S(t)}{\partial t}, \quad (18)$$

and the relation between the cumulative distribution $S(t)$ and the FPTD (between two arbitrary times t_1 and t_2) is [35, 36]

$$S(t_1) - S(t_2) = \int_{t_1}^{t_2} h(t') dt'. \quad (19)$$

Clearly, as shown in Figure 3, the diversity $d(k)$ (that counts events having achieved rank k in a fixed time window) may be identified with the right hand side of Eq. (19). This equation shows, firstly, the relation between diversity and the diffusion equation (15). Secondly, since there is a relation between the solutions of the diffusion equation and random walks, there is also one between $d(k)$ and the random walk model given by Eq. (6). We have already studied a particular case of these models in [12].

Additional material

Additional file 1: The datasets supporting the conclusions of this article. (zip)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors made substantial contributions to the conception and design of the paper and interpretation of data. They were all involved in drafting the manuscript by contributing with relevant content. JAM and SS also contributed with the acquisition and analysis of data. All authors read and approved the final manuscript.

Author details

¹Facultad de Ciencias, Universidad Nacional Autónoma de México, México D.F., 01000, Mexico. ²Instituto de Física, Universidad Nacional Autónoma de México, México D.F., 01000, Mexico. ³Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México D.F., 01000, Mexico. ⁴Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, México D.F., 04510, Mexico. ⁵SENSEable City Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁶MoBS Lab, Network Science Institute, Northeastern

University, Boston, MA 02115, USA. ⁷ITMO University, St. Petersburg, 199034, Russian Federation. ⁸Centro de Investigación y Docencia Económicas, Consejo Nacional de Ciencia y Tecnología, México D.F., 01210, Mexico. ⁹Department of Computer Science, Aalto University School of Science, Aalto, 00076, Finland.

Acknowledgements

Financial support from CONACyT under projects 212802, 221341, and UNAM-PAPIIT IN111015 is acknowledged.

Received: 15 June 2016 Accepted: 17 November 2016 Published online: 25 November 2016

References

- Duch J, Waizman JS, Amaral LAN (2010) Quantifying the performance of individual players in a team activity. *PLoS ONE* 5(6):10937
- Ben-Naim E, Vazquez F, Redner S (2007) What is the most competitive sport? *J Korean Phys Soc* 50:124
- Merritt S, Clauset A (2013) Environmental structure and competitive scoring advantages in team competitions. *Sci Rep* 3:3067
- Merritt S, Clauset A (2013) Social network dynamics in a massive online game: network turnover, non-densification, and team engagement in halo reach. Eprint. arXiv:1306.4363
- Albert J, Bennett J, Cochran JJ (2005) *Anthology of statistics in sports*, vol 16. SIAM, Philadelphia
- Radicchi F (2011) Who is the best player ever? A complex network analysis of the history of professional tennis. *PLoS ONE* 6(2):17249
- Yucesoy B, Barabási A-L (2016) Untangling performance from success. *EPJ Data Sci* 5:17
- Merritt S, Clauset A (2014) Scoring dynamics across professional team sports: tempo, balance and predictability. *EPJ Data Sci* 3:4
- Deng W, Li W, Cai X, Bulou A, Wang QA (2012) Universal scaling in sports ranking. *New J Phys* 14(9):093038
- Klaassen FJ, Magnus JR (2001) Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *J Am Stat Assoc* 96(454):500-509
- Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Team TGB, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden EL (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176-182
- Cocho G, Flores J, Gershenson C, Pineda C, Sánchez S (2015) Rank diversity of languages: generic behavior in computational linguistics. *PLoS ONE* 10(4):0121898
- ATP World Tour. <http://www.atpworldtour.com/>. Accessed 4 April 2016
- World Chess Federation. <http://ratings.fide.com/>. Accessed 6 April 2016
- Official World Golf Ranking. <http://www.owgr.com/>. Accessed 6 April 2016
- Global Poker Index. <http://www.globalpokerindex.com/>. Accessed 6 April 2016
- Football Club World Ranking. <http://www.clubworldranking.com/ranking-clubs.aspx>. Accessed 6 April 2016
- Fédération Internationale de Football Association. <http://www.fifa.com/>. Accessed 6 April 2016
- Elo AE (1978) *The rating of chessplayers, past and present*. Arco Pub., London
- Gerlach M, Altmann EG (2013) Stochastic model for the vocabulary growth in natural languages. *Phys Rev X* 3:021006
- Katz JS, Katz L (1999) Power laws and athletic performance. *J Sports Sci* 17(6):467-476
- Alvarez-Ramirez J, Rodriguez E (2006) Scaling properties of marathon races. *Physica A* 365(2):509-520
- Visser M (2013) Zipf's law, power laws and maximum entropy. *New J Phys* 15(4):043021
- Baek SK, Bernhardsson S, Minnhagen P (2011) Zipf's law unzipped. *New J Phys* 13(4):043004
- Wikipedia: Gamma distribution. https://en.wikipedia.org/wiki/Gamma_distribution. Accessed 1 March 2016
- Wikipedia: Beta distribution. https://en.wikipedia.org/wiki/Beta_distribution. Accessed 1 March 2016
- Jóhannesson G, Björnsson G, Gudmundsson EH (2006) Afterglow light curves and broken power laws: a statistical study. *Astrophys J Lett* 640(1):L5-L8
- Li W, Miramontes P, Cocho G (2010) Fitting ranked linguistic data with two-parameter functions. *Entropy* 12(7):1743
- Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. *G Ist Ital Attuari* 4(1):83-91
- Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661-703
- Alvarez-Martinez R, Cocho G, Rodríguez RF, Martínez-Mekler G (2014) Birth and death master equation for the evolution of complex networks. *Physica A* 402:198-208
- Martínez-Mekler G, Martínez RA, del Río MB, Mansilla R, Miramontes P, Cocho G (2009) Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE* 4(3):4791
- Van Kampen NG (2007) *Stochastic processes in physics and chemistry*. North Holland, Amsterdam
- Wheeler JC, Gordon RG, Baker GA, Gammel JL (1970) *The Padé approximant in theoretical physics*. Academic, New York
- Perline R (1996) Zipf's law, the central limit theorem, and the random division of the unit interval. *Phys Rev E* 54(1):220
- Perline R, Perline R (2016) Two universality properties associated with the monkey model of Zipf's law. *Entropy* 18(3):89

Bibliografía

- [1] S. H. Strogatz, “Exploring complex networks,” *nature*, vol. 410, no. 6825, p. 268, 2001. [xv](#), [3](#), [4](#)
- [2] G. Cocho, J. Flores, C. Gershenson, C. Pineda, and S. Sánchez, “Rank diversity of languages: generic behavior in computational linguistics,” *PloS one*, vol. 10, no. 4, p. e0121898, 2015. [xxi](#), [5](#), [24](#), [25](#), [27](#), [28](#), [29](#), [49](#), [54](#), [58](#), [61](#), [63](#), [78](#), [81](#), [94](#), [121](#)
- [3] “Sistemas complejos..” https://en.wikipedia.org/wiki/Complex_system. Último acceso 10 de Mayo de 2018. [2](#)
- [4] S. Havlin, D. Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. Kantelhardt, J. Kertész, S. Kirkpatrick, J. Kurths, *et al.*, “Challenges in network science: Applications to infrastructures, climate, social systems and economics,” *The European Physical Journal Special Topics*, vol. 214, pp. 273–293, 2012. [2](#)
- [5] A.-L. Barabási, Z. Dezsó, E. Ravasz, S.-H. Yook, and Z. Oltvai, “Scale-free and hierarchical structures in complex networks,” in *AIP Conference Proceedings*, vol. 661, pp. 1–16, AIP, 2003. [2](#), [3](#)
- [6] A. Vespignani, “Predicting the behavior of techno-social systems,” *Science*, vol. 325, no. 5939, pp. 425–428, 2009. [2](#)
- [7] S. Motegi and N. Masuda, “A network-based dynamical ranking system for competitive sports,” *Scientific reports*, vol. 2, p. 904, 2012. [5](#)
- [8] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting, “Ranking rankings: an empirical comparison of the predictive power of sports ranking methods,” *Journal of Quantitative Analysis in Sports*, vol. 9, no. 2, pp. 187–202, 2013. [5](#)
- [9] W. Deng, W. Li, X. Cai, A. Bulou, and Q. A. Wang, “Universal scaling in sports ranking,” *New Journal of Physics*, vol. 14, no. 9, p. 093038, 2012. [5](#), [35](#)
- [10] F. J. Klaassen and J. R. Magnus, “Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model,” *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 500–509, 2001. [5](#)

BIBLIOGRAFÍA

- [11] <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>. Último acceso 13 de Marzo de 2018. 5, 24
- [12] J. A. Morales, S. Sánchez, J. Flores, C. Pineda, C. Gershenson, G. Cocho, J. Zizumbo, R. F. Rodríguez, and G. Iñiguez, “Generic temporal features of performance rankings in sports and games,” *EPJ Data Science*, vol. 5, no. 1, p. 33, 2016. 7, 53, 54, 127
- [13] “World chess federation.” <http://ratings.fide.com/>. Último acceso 10 de Enero de 2018. 10, 11
- [14] “Football club world ranking.” <http://www.clubworldranking.com/ranking-clubs.aspx>. Último acceso 10 de Enero de 2018. 10, 11, 14
- [15] “Fédération internationale de football association.” <http://www.fifa.com/>. Accessed 6 April 2016. 10, 11, 16
- [16] “Official world golf ranking.” <http://www.owgr.com/>. Último acceso 10 de Enero de 2018. 10, 11, 17
- [17] “National association for stock car auto racing.” <http://www.espn.com/racing/standings> Último acceso 10 de Enero de 2018. 10, 11
- [18] “Global poker index.” <http://www.globalpokerindex.com/>. Último acceso 10 de Enero de 2018. 10, 11, 18
- [19] “World snowboarding.” <http://www.worldsnowboarding.org/> Último acceso 3 de Abril de 2018. 10, 11, 19
- [20] “Atp world tour.” <http://www.atpworldtour.com/>. Último acceso 10 de Enero de 2018. 10, 11, 19
- [21] “Videogame earnings.” <https://www.esportsearnings.com/history> Último acceso 10 de Enero de 2018. 10, 11, 20
- [22] “*La FIDE*.” <https://en.wikipedia.org/wiki/FIDE>. Último acceso 14 de Marzo de 2018. 12
- [23] “*El sistema de ranqueo Elo*.” https://en.wikipedia.org/wiki/Elo_rating_system. Último acceso 15 de Marzo de 2018. 12
- [24] “*El sistema de ranqueo Elo*.” https://es.wikipedia.org/wiki/Sistema_de_puntuación_Elo. Último acceso 15 de Marzo de 2018. 12, 13, 14
- [25] “*Sistema de puntos para NASCAR*.” https://en.wikipedia.org/wiki/List_of_NASCAR_points_scoring_systems. Último acceso 15 de Marzo de 2018. 17
- [26] “*The Ranking that changed Tennis*.” <http://www.atpworldtour.com/en/news/heritage-1973-atp-rankings-celebration-part-ii>. Último acceso 13 de Marzo de 2018. 19

-
- [27] “Reglas de ranqueo para miembros de la atp.” https://en.wikipedia.org/wiki/ATP_Rankings. Último acceso 13 de Marzo de 2018. 19, 20
- [28] “*Deportes electrónicos.*” https://es.wikipedia.org/wiki/Deportes_electrónicos. Último acceso 15 de Marzo de 2018. 20
- [29] G. K. Zipf, “Selected studies of the principle of relative frequency in language,” 1932. 24, 30, 31
- [30] R. Ferrer-i Cancho and B. Elvevåg, “Random texts do not exhibit the real zipf’s law-like rank distribution,” *PLoS One*, vol. 5, no. 3, p. e9411, 2010. 24
- [31] W. Li, P. Miramontes, and G. Cocho, “Fitting ranked linguistic data with two-parameter functions,” *Entropy*, vol. 12, no. 7, pp. 1743–1764, 2010. 25, 32, 35, 97, 98
- [32] G. Altmann, “Prolegomena to menzerath’s law,” *Glottometrika*, vol. 2, no. 2, pp. 1–10, 1980. 25
- [33] R. Alvarez-Martinez, G. Cocho, R. Rodríguez, and G. Martínez-Mekler, “Birth and death master equation for the evolution of complex networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 402, pp. 198–208, 2014. 25, 61
- [34] “Ecuación maestra..” https://en.wikipedia.org/wiki/Master_equation. Último acceso 14 de Abril de 2018. 27
- [35] “Aproximación de padé..” https://en.wikipedia.org/wiki/Padé_approximant. Último acceso 14 de Abril de 2018. 28
- [36] M. Gerlach and E. G. Altmann, “Stochastic model for the vocabulary growth in natural languages,” *Physical Review X*, vol. 3, p. 021006, May 2013. 30, 31
- [37] “Ley de zipf-mandelbrot.” https://en.wikipedia.org/wiki/Zipf-Mandelbrot_law. Último acceso 14 de Abril de 2018. 31
- [38] S. K. Baek, S. Bernhardsson, and P. Minnhagen, “Zipf’s law unzipped,” *New Journal of Physics*, vol. 13, no. 4, p. 043004, 2011. 31
- [39] “Distribución beta.” https://en.wikipedia.org/wiki/Beta_distribution. Último acceso 14 de Abril de 2018. 31
- [40] “Distribución gamma.” https://en.wikipedia.org/wiki/Gamma_distribution. Último acceso 14 de Abril de 2018. 31
- [41] “Distribución empírica acumulativa.” https://en.wikipedia.org/wiki/Empirical_distribution_function. Último acceso 14 de Abril de 2018. 32

BIBLIOGRAFÍA

- [42] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data,” *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009. [36](#), [98](#), [99](#)
- [43] “La función sigmoide.” https://en.wikipedia.org/wiki/Sigmoid_function. Último acceso 1 de Mayo de 2018. [58](#)
- [44] “Distribución normal.” https://en.wikipedia.org/wiki/Normal_distribution. Último acceso 1 de Mayo de 2018. [58](#), [59](#)
- [45] S. Redner, “Random multiplicative processes: An elementary tutorial,” *American Journal of Physics*, vol. 58, no. 3, pp. 267–273, 1990. [59](#), [61](#)
- [46] G. Martínez-Mekler, R. A. Martínez, M. B. del Río, R. Mansilla, P. Miramontes, and G. Cocho, “Universality of rank-ordering distributions in the arts and sciences,” *PLoS One*, vol. 4, no. 3, p. e4791, 2009. [61](#)
- [47] J. A. Morales, E. Colman, S. Sánchez, F. Sánchez-Puig, C. Pineda, G. Iñiguez, G. Cocho, J. Flores, and C. Gershenson, “Rank dynamics of word usage at multiple scales,” *Frontiers in Physics*, vol. 6, p. 45, 2018. [63](#), [67](#), [70](#)
- [48] N. Fernández, C. Maldonado, and C. Gershenson, “Information measures of complexity, emergence, self-organization, homeostasis, and autopoiesis,” in *Guided self-organization: Inception*, pp. 19–51, Springer, 2014. [70](#), [71](#)