UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dep. d'Estadística i Investigació Operativa Aplicades i Qualitat

Anàlisi del Balanç Competitiu en Bàsquet Femení Espanyol.

Treball Fi de Màster

Màster Universitari en Enginyeria d'Anàlisi de Dades, Millora de Processos i Presa de decisions

AUTOR/A: Cercós Navarro, Beatriz

Tutor/a: Debón Aucejo, Ana María

Cotutor/a extern: Briz Redón, Álvaro

CURS ACADÈMIC: 2023/2024

Agraïments

Volia agrair a totes aquelles persones que han estat, d'una manera o una altra, presents en aquest treball:

En primer lloc, als meus tutors, Ana i Álvaro, per la seua inestimable ajuda, dedicació i suport al llarg d'aquest treball. Gràcies per la vostra paciència, per compartir el vostre coneixement i per guiar-me en aquest treball. La vostra experiència i consells han enriquit no només aquest treball, sinó també el meu creixement personal i acadèmic.

Als meus amics, que han estat un suport constant durant aquest període. Gràcies per escoltar-me i ajudar-me a desconnectar per poder recarregar energies.

Finalment, a la meua família, tant els que estan dia rere dia, com aquells que ja no estan entre nosaltres. Gràcies especialment als meus pares i al meu germà; ja saps que eres el meu pilar. Sense la vostra comprensió i encoratjament constant, aquest treball no hauria sigut possible.

Gràcies a tots.

Resumen

El balanç competitiu és un element crucial a l'esport, ja que reflecteix el grau d'igualtat i rivalitat entre els equips participants en una lliga. La literatura evidencia que la incertesa en els resultats de les competicions esportives és essencial per mantindre l'interés dels aficionats, ja que genera emoció, i augmenta la demanda d'entrades i de drets de transmissió de televisió. En aquest Treball Final de Màster s'analitza el balanç competitiu a la lliga femenina espanyola de bàsquet (Lliga Femenina Endesa), entre les temporades 1997/98 i 2023/24. S'estudien diversos índexs de competitivitat per temporada i jornada, per tal d'analitzar-ne l'evolució en el temps. La metodologia aplicada es basa en models estadístics avançats, com ara els models multinivell, per capturar l'estructura jeràrquica de les dades i les variacions entre temporades i equips. D'altra banda, s'utilitzen altres eines estadístiques, com l'Anàlisi de Components Principals o el *clustering*, per analitzar com afecten les estadístiques de joc al balanç competitiu i, alhora, identificar els factors clau que influeixen en el rendiment dels equips. Aquest treball contribueix a l'enteniment del balanç competitiu al bàsquet femení espanyol i proporciona una base sòlida per a futures investigacions i polítiques dirigides a millorar l'equitat a la lliga, assegurant-ne també la viabilitat econòmica.

Paraules clau: balanç competitiu, bàsquet femení, HICB, models multinivel.

Abstract

Competitive balance is a crucial element in sport, since it reflects the degree of equality and rivalry between the participating teams in a league. The literature shows that uncertainty in the results of sports competitions is essential to maintain the interest of fans, as it generates excitement, increases the demand for tickets and television transmission rights. This Master's Final Project analyzes the competitive balance in the Spanish women's basketball league (Liga Femenina Endesa), between the 1997/198 and 2023/24 seasons. Various competitiveness indices are studied by season and day, in order to analyze their evolution over time. The applied methodology is based on advanced statistical models, such as multilevel models, to capture the hierarchical structure of the data and variations between seasons and teams. On the other hand, other statistical tools are used, such as Analysis of Principal Components or clustering, in order to analyze how game statistics affect the competitive balance and at the same time identify the key factors that influence the teams' performance. This work contributes to the understanding of the competitive balance in Spanish women's basketball and provides a solid basis for future research and policies aimed at improving equity in the league, also ensuring its economic viability.

Key words: competitive balance, women's basketball, HICB, multilevel models.

Índex

Índex de figures

Índex de taules

Capítol 1

Introducció

En els darrers anys, hem assistit a una transformació profunda en l'esport, on la tecnologia i l'anàlisi de dades han adquirit un protagonisme destacat. El **balanç competitiu**, que mesura el nivell d'igualtat entre els equips d'una competició, s'ha convertit en un tema d'investigació essencial. Els aficionats estan clarament preocupats pel balanç competitiu, i molts tenen opinions formades sobre el nivell d'equilibri que perceben en el seu esport preferit, tant en comparació amb altres lligues com amb les temporades anteriors. Al mateix temps, els comentaristes esportius mostren una gran preocupació pel grau de balanç competitiu en les lligues professionals. Així doncs, l'anàlisi del balanç competitiu desperta un interés notable tant entre economistes com entre aficionats als esports.

En l'actualitat, el **bàsquet** s'ha convertit en un reclam mundial, sobretot de les anomenades "grans lligues", com la NBA. El bàsquet modern s'ha transformat en una indústria mundial, arribant així a convertir alguns clubs en empreses molt rendibles, que mouen milions d'euros anualment; com ara amb les entrades, els drets televisius, la venda de samarretes o els transports a les ciutats on es disputen els partits, entre altres.

No obstant això, tot i els avanços recents i l'augment de l'interés per part del públic i els mitjans de comunicació, el **bàsquet femení** encara no ha arribat al mateix nivell de reconeixement i rendibilitat. La desigualtat en els salaris, la menor cobertura mediàtica, i la falta d'inversió en infraestructures i formació són algunes de les barreres que persisteixen en aquest àmbit. Per aquest motiu, en aquest Treball Final de Màster, s'analitza el balanç competitiu a la Lliga Femenina Endesa, entre les temporades de 1997/98 a 2023/24. S'ha dut a terme un enfocament plenament matemàtic; deixant a banda els termes econòmics; en el qual s'estudien en profunditat els índexs de la LB Endesa, en comptes de comparar els índexs entre lligues.

En la secció següent s'aprofundeix més en el tema principal del treball; el balanç competitiu i per què té tanta importància. En la secció 3, s'exposen els objectius que es volen assolir. A continuació, en la secció 4, s'explica com s'ha obtés la base de dades que s'ha utilitzat, a

més del pretractament necessari que s'ha dut a terme. Posteriorment, en la secció 5, es descriuen amb precisió les metodologies aplicades; les quals inclouen PCA (anàlisi de components principals), *clustering*, i els índexs del balanç competitiu (que poden ser estàtics o dinàmics), a més dels models que es van a ajustar, en els quals destaquen els models multinivell. Tot seguit, en la secció 6, s'exposen amb detall els resultats obtinguts, els quals es troben il·lustrats amb gràfiques i taules, per a una millor comprensió. Finalment, en base als resultats obtinguts, expliquem els aspectes més remarcables del treball a mode de conclusions a la secció 7.

Aquest estudi no només contribueix a una millor comprensió del bàsquet femení espanyol, sinó que també estableix una base sòlida per a futures investigacions i estratègies destinades a millorar l'equitat i la competitivitat en la lliga.

Capítol 2

Marc teòric

2.1 L'economia de l'esport: el balanç competitiu

Mentre que les organitzacions industrials desitjarien la menor quantitat de competició possible, les organitzacions esportives necessiten un cert equilibri competitiu per ser viables. Qualsevol competició esportiva, siga de bàsquet o d'altra modalitat d'equip professional, resulta més atractiva com més equilibrada estiga la competició, ja que el resultat serà més incert. Segons aquest fet, que es coneix com la hipòtesi de la incertesa del resultat, els espectadors prefereixen assistir a partits igualats i seguir un campionat disputat, que no anar a partits, el guanyador dels quals és conegut per tots abans de començar. No obstant això, només la incertesa del resultat a llarg termini, o que un mateix equip acabe guanyant la lliga any rere any, demostren tindre un impacte negatiu a l'assistència d'espectadors [1]. Existeixen equips dominants com els *Golden State Warriors* (NBA), o l'*UConn Women's Basketball Program* (WNBA), els quals es consideren "súper equips" que banalitzen la competició i guanyen partits de manera senzilla. Òbviament, tot aficionat prefereix que el seu equip guanye el partit, però una llarga ratxa ininterrompuda de victòries es considera avorrida per a un aficionat, perquè elimina la incertesa binomial de l'esdeveniment [2].

Per **balanç competitiu** s'entén la igualtat de forces que hi ha entre els equips en una lliga esportiva [3]. Que tots els equips tinguen les mateixes probabilitats de guanyar (o perdre) és essencial per mantindre jugadors i aficionats compromesos al llarg de temporades senceres de manera constant. Són moltes les polítiques i regulacions econòmiques que les lligues esportives d'equips ja han introduït per tal de millorar el balanç competitiu, com el sistema de repartiment d'ingressos, els límits salarials i les restriccions a la mobilitat dels jugadors [1]. No obstant això, aquestes polítiques són pròpies de les lligues d'Amèrica del Nord, les quals són tancades en la seua majoria, i no obertes com en la majoria de països d'Europa[1].

[1]Les lligues tancades són aquelles on existeixen un cert nombre d'equips que participen totes les temporades, com la NBA o la WNBA. Les lligues obertes, per contra, són lligues d'ascens-descens on participen només aquells equips que han obtés bons resultats en la temporada anterior.

Tot i que no existeixen articles concrets sobre el bàsquet femení espanyol, des de fa un temps són diversos els estudis i anàlisis que indiquen conclusions contradictòries sobre l'evolució del balanç competitiu en els últims anys. Per una banda, segons es coneix, s'està produint un continu descens de la competitivitat a les diverses lligues de **futbol** tant a nivell nacional [4], com a nivell europeu [5]. Per l'altra banda, en altres esports, com ara el **beisbol**, el balanç competitiu està augmentant en els últims anys [3].

Per respondre preguntes com si l'equilibri competitiu a la lliga augmenta amb el temps o no, es necessiten mesures de balanç competitiu que siguen comparables en el temps i entre lligues. El principal problema és que no existeix una definició precisa i comuna sobre què s'entén per balanç competitiu, i per aquest motiu, són tants els índexs de balanç competitiu que existeixen i que es calculen de maneres diverses, que analitzar-los correctament és una tasca complexa. Cada índex ofereix una perspectiva diferent sobre el grau d'igualtat entre els equips d'una lliga, i la interpretació dels resultats pot variar segons el context o l'objectiu de l'anàlisi. A més, l'aplicació d'aquests índexs requereix una comprensió profunda de les dades i de les dinàmiques pròpies de cada competició esportiva.

2.2 Bàsquet femení: Lliga Femenina Endesa

Històricament, la participació de la dona ha sigut sempre menor que la dels hòmens en l'àmbit laboral, cultural, polític, etc. Actualment, en el sector esportiu, la dona ha hagut de superar barreres creades pels estereotips socials i culturals. Han hagut de lluitar contra idees masclistes, com ara l'existència d'esports per a hòmens i esports per a dones o amb prejudicis com que les dones són inferiors als hòmens en les activitats esportives, entre altres.

La Lliga Femenina Endesa és la màxima competició d'equips femenins de bàsquet que es juga a Espanya. Actualment consta de 16 equips, que juguen un format de tots contra tots, classificant-se els 8 primers per als *playoffs* de quarts, semifinals i final al millor de 3 partits cadascun. Les darreres temporades 2022-23 i 2023-24, el València Bàsquet ha sigut l'equip que ha guanyat la Lliga [6]. La LF Endesa, va ser per primera vegada disputada en el 1964, mentre que la lliga de bàsquet masculina espanyola, coneguda com a Lliga ACB, va ser fundada al 1923, més de 40 anys abans.

La falta d'articles i estudis sobre esport femení, i en concret sobre el bàsquet femení espanyol, reflecteix una realitat preocupant: la invisibilitat i la falta de reconeixement que les esportistes han patit (i pateixen) al llarg del temps. Aquesta falta d'investigació i cobertura mediàtica perpetua la desigualtat de gènere en l'esport. En aquest context, resulta fonamental analitzar el balanç competitiu de la LF Endesa per tal de no només comprendre millor les dinàmiques internes de la competició, sinó també per tal d'incrementar l'interés i l'anàlisi acadèmica en l'àmbit de l'esport femení, amb la finalitat de donar-li la importància que mereix.

Capítol 3

Objectius

L'objectiu principal d'aquest treball és analitzar i proporcionar una comprensió més profunda del **balanç competitiu de la Lliga Femenina Endesa** a les temporades d'estudi, la qual cosa podria ser útil per avaluar les desigualtats de la competició. Per a això, es fa necessari identificar els factors que influeixen en l'equilibri entre els equips i com aquests han evolucionat al llarg del temps. A més, com l'anàlisi de dades als esports està a l'ordre del dia, també volem realitzar una anàlisi completa de les dades, per tal d'aconseguir una millor comprensió de les mateixes.

Per aconseguir aquests propòsits s'han concretat els objectius específics següents:

- **Objectiu 1:** Realitzar una anàlisi exploratòria completa de les estadístiques de joc de les temporades d'estudi.

  Es realitza una PCA exploratòria (Anàlisi de Components Principals), on es redueix la dimensionalitat (passem de 25 variables a 4 components principals), per tal de facilitar la comprensió de com les variables es relacionen.

- **Objectiu 2:** Categoritzar les observacions (partits) segons certes característiques específiques; principalment en equips guanyadors i perdedors.

  Mitjançant el *clustering*, una eina de *machine learning*, es creen grups sense una etiquetació prèvia. Els grups obtinguts s'utilitzaran per tal d'explicar el balanç competitiu al llarg del temps. Es pretén trobar un patró dels equips al llarg de les temporades per saber quines coses han de buscar els entrenadors en el seu equip per tal de guanyar.

- **Objectiu 3:** Calcular els índexs de competitivitat **per temporada** per avaluar si el balanç competitiu al bàsquet femení espanyol ha disminuït recentment com en altres esports.

  S'estudia la formulació de les mesures de balanç competitiu de la literatura seleccionada;

i es calculen els índexs per temporada, els quals seran utilitzats únicament de manera descriptiva ja que només es tenen dades de 27 temporades.

- **Objectiu 4:** Calcular els índexs de competitivitat **per jornada** per tal d'estudiar els factors que afecten a l'augment o disminució del balanç competitiu a la lliga.

  S'utilitzaran models multinivell, entre altres, per estudiar principalment si els efectes aleatoris tenen significativitat. És a dir, si existeixen diferències als índexs de competitivitat a nivell de temporada o de jornada.

Capítol 4

Materials

Per a la realització d'aquest TFM, utilitzarem una base de dades on tenim les estadístiques de cada equip en cada jornada de les temporades d'estudi de la Lliga Femenina Endesa. La base de dades s'ha obtés mitjançant *webscrapping* des de la web BueStats [7]. En aquesta base, tenim mesurades les variables d'interés per al nostre estudi, con el nombre de punts o de victòries dels equips en la lliga, únicament per a la fase de grups de la lliga, també coneguda com Lliga Regular única o *round-robin*[1].

S'han extret les dades des de la temporada 1997/98 fins la temporada 2023/24, ja que són les que es troben disponibles a la pàgina web oficial de la Lliga Femenina Endesa [6]. Cada fila de les dades representa un equip en una jornada en específic. D'aquesta manera, tenim un total de 9848 observacions, ja que el nombre d'equips i de jornades canvia en funció de l'any, com podem observar en la Taula 4.1:

Taula 4.1: Nombre d'equips i jornades per temporada

Temporada	**Equips**	**Jornades**	**Temporada**	**Equips**	**Jornades**	**Temporada**	**Equips**	**Jornades**
1997/98	12	22	2006/07	14	26	2015/16	14	23
1998/99	14	26	2007/08	14	26	2016/17	14	26
1999/00	14	26	2008/09	14	26	2017/18	14	26
2000/01	14	26	2009/10	14	26	2018/19	14	26
2001/02	14	26	2010/11	14	26	2019/20	14	22
2002/03	14	26	2011/12	14	26	2020/21	16	30
2003/04	14	26	2012/13	11	20	2021/22	16	30
2004/05	14	26	2013/14	12	22	2022/23	16	30
2005/06	14	26	2014/15	14	26	2023/24	16	30

[1]El sistema de tots contra tots o *round-robin* és un sistema de tornejos de competició, generalment en l'àmbit esportiu, en què cada equip del torneig s'enfronta contra tots els altres equips i en un nombre constant de partits (habitualment un o dos).

A la base de dades obtinguda tenim un total de 30 variables d'estudi, les quals s'expliquen en profunditat en la Taula 4.2:

Taula 4.2: Descripció de les variables de la base de dades.

Variable	**Descripció**
Nombre	Nom de l'equip (text).
Jornada	Nombre de jornada a la que fa referència.
Partidos	Nombre de partits a que fa referència cada fila.
Minutos	Minuts que va durar el partit.
Puntos	Punts totals anotats.
T2 Anotados	Nombre de canastes de 2 anotades.
T2 Lanzados	Nombre de canastes de 2 llançades.
% T2	Proporció de canastes de 2 anotades.
T3 Anotados	Nombre de triples anotats.
T3 Lanzados	Nombre de triples llançats.
% T3	Proporció de triples anotats.
T1 Anotados	Nombre de tirs lliures anotats.
T1 Lanzados	Nombre de tirs lliures llançats.
% T1	Proporció de tirs lliures anotats.
Reb. Ofensivos	Nombre de rebots ofensius.
Reb. Defensivos	Nombre de rebots defensius.
Rebotes	Total de rebots.
Asistencias	Nombre d'assistències.
Robos	Nombre de pilotes furtades a l'equip rival.
Perdidas	Nombre de pilotes furtades per l'equip rival.
Tapones	Nombre de tapons fets a l'equip rival.
Tapones Recibidos	Nombre de tapons fets per l'equip rival.
Mates	Nombre de mates fets.
Faltas Cometidas	Nombre de faltes dels jugadors de l'equip.
Faltas Recibidas	Nombre de faltes dels jugadors de l'equip contrari.
Valoracion	Valoració de l'equip en el partit. (*)
Local	Nom de l'equip que juga com a local (text).
Visitante	Nom de l'equip que juga com a visitant (text).
Victoria	Binària amb valor TRUE si l'equip guanya i FALSE si perd.
Diferencia	Nombre de punts de diferència entre el guanyador i el perdedor.

(*) La variable Valoració segueix la següent fórmula:

Valoracion = (Puntos + Rebotes + Asistencias + Robos + Faltas Recibidas + Tapones) - (*Tiros Fallados* + Perdidas + Tapones Recibidos + Faltas Cometidas).

On, *Tiros Fallados* = T1 Lanzados + T2 Lanzados + T3 Lanzados - (T1 Anotados + T2 Anotados + T3 Anotados).

4.1 Pretractament de les dades

Abans de començar amb l'anàlisi, durem a terme una revisió i transformació de les dades. En primer lloc, eliminem la variable **Partidos**, ja que té valor 1 per a totes les observacions i per tant no aporta informació; i la variable **Minutos**, ja que és quasi constant per a totes les observacions. Eliminem també les variables **Local** i **Visitante**. A continuació, recategoritzem la variable **Victoria** de booleana a binària, és a dir, TRUE = 1 i FALSE = 0. Recategoritzem també la variable **Jornada**, de format text a format numèric: passem de tindre Jornada 1 a només 1.

Després, introduïm noves variables:

- **Temporada**, que indica a quina temporada pertany cada observació.
- **Local**, que serà una binària on 1 significa que l'equip juga com a local i 0, que juga com a visitant.
- **Guanyats**, que és la suma acumulativa de les victòries al llarg de les jornades dins de cada temporada.
- **Porc_guanyats**, que és la proporció de partits guanyats, que es calcula com la divisió de les variables **Guanyats** entre **Jornada**.
- **Diferencia_acum**, que és la suma de les diferències de punts entre l'equip i el seu rival al llarg de les jornades dins de cada temporada.
- **Posicio**, que indica la posició de l'equip en cada jornada. Té en compte el nombre de partits guanyats, és a dir, la variable **Guanyats**; i en cas d'empat, té en compte la quantitat de punts de diferència acumulats, és a dir, la variable **Diferencia_acum**.
- **Nombre2**, que és una recategorització dels noms dels equips.[2] Podem trobar el patró de recategorització emprat en l'Annex 1 (Taules 8.1, 8.2 i 8.3).

Després d'aquestes transformacions, comptem amb dues variables de tipus text (Nombre i Nombre2), tres variables de tipus factor (Temporada, Jornada i Posició), dues variables binàries (Victoria i Local) i 27 variables numèriques, les quals corresponen a les estadístiques de joc per a cada equip en la jornada corresponent.

[2]Alguns equips han canviat el seu nom per motius de patrocini, entre altres. Aquesta recategorització té com a finalitat poder estudiar l'evolució dels equips al llarg dels anys, independentment del seu nom o patrocinadors. Hem passat de tindre 153 noms a només 48.

Capítol 5

Metodologia

En aquesta secció es descriuen en profunditat els processos metodològics i eines emprades en l'anàlisi estadística. Per a la resolució dels objectius prèviament plantejats, es duen a terme diferents tècniques de *Machine Learning* a partir del programa informàtic R [8].

5.1 Software emprat: R

El programa R [8], és un software lliure que, mitjançant la definició de funcions, ofereix una gran varietat de tècniques estadístiques i gràfiques. Per la seua part, RStudio és un entorn de desenvolupament integrat (IDE) per a R que permet establir un espai de treball, crear gràfics i emprar molts paquets estadístics. RStudio inclou una consola, un editor amb ressaltat de sintaxi que permet l'execució directa de codi, i eines per al traçat, historial, depuració i gestió de l'espai de treball [9]. RStudio no pot funcionar sense R.

5.2 Anàlisi de components principals (PCA)

Una de les tècniques emprades és l'Anàlisi de Components Principals o PCA (llibreria *factoextra* [10]). L'objectiu principal de la PCA és condensar la informació continguda d'un nombre elevat de variables correlacionades en un nombre reduït de components no observables directament i incorrelacionades, caracteritzades per ser combinació lineal de les originals [11]. L'objectiu d'aquestes nous components és explicar la màxima variància de les dades amb restricció d'ortogonalitat, és a dir, busquen relacions entre els individus (en el nostre cas, partits de bàsquet) o entre variables (en el nostre cas, les estadístiques de cada partit).

És important ressaltar la necessitat de centrar i escalar a variància unitària cadascuna de les variables numèriques de la base de dades ja que aquesta tècnica és sensible a l'escala.

Així mateix, la fórmula característica de la PCA per a N observacions, J variables (auto-escalades) i A components principals és la següent:

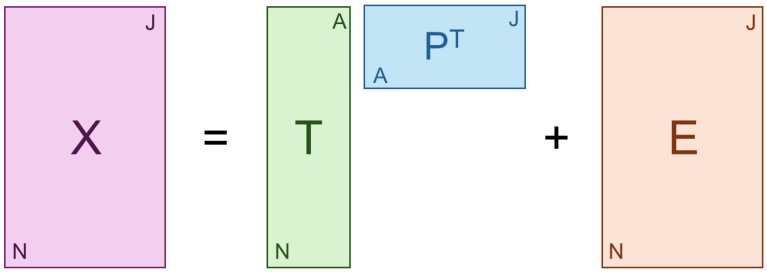

Figura 5.1: Estructura matricial PCA

On X és la matriu de dades auto-escalada, T la matriu dels *scores*, P^T la matriu dels *loadings* i E la matriu residual [11]. Els *scores* es defineixen com les puntuacions que es donen a cada observació en l'espai latent i els *loadings* com el pes que tenen les variables en els diferents components.

Per a la validació prèvia de les observacions s'utilitzaren la suma de quadrats residuals o **SPE** (Squared Predicted Error) i l'estadístic **T^2 de Hotelling**. El SPE és la distància (al quadrat) Euclídea de l'observació i a la seua projecció en l'hiperplà format pels nous A components del model. El T^2-Hotelling és la suma dels quadrats dels *scores* tipificats, que representen la distància (al quadrat) estimada de Mahalanobis de la projecció a la mitjana.

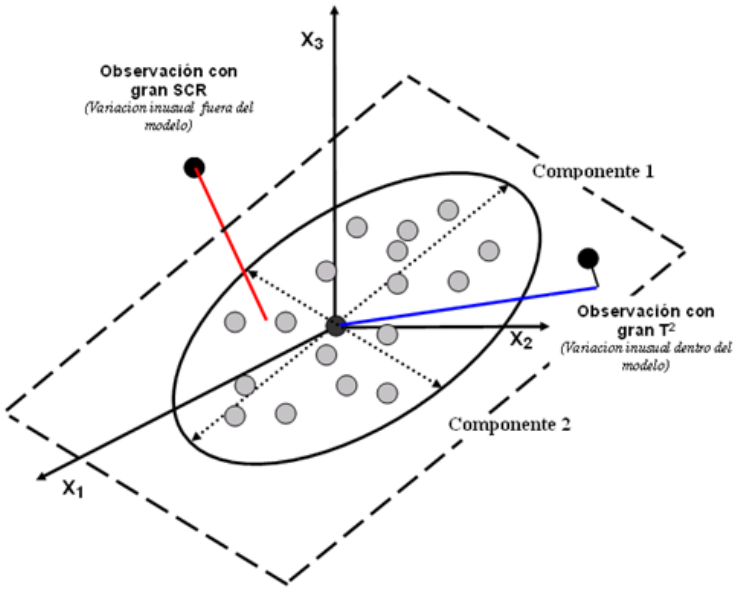

Figura 5.2: Distàncies que mesuren els estadístics SPE i T^2-Hotelling

Com podem observar a la Figura 5.2, el SPE mesura la distància a l'hiperplà que formen els A components principals del model, mentre que el T^2-Hotelling mesura la distància al centre del núvol de punts format per les projeccions de les observacions. Valors elevats de SPE (línia roja en la Figura 5.2), suposen observacions atípiques, que són **outliers moderats** per al model. De manera contrària, valors elevats de T^2-Hotelling (línia blava en la Figura 5.2), suposen observacions extremes, que són **outliers severs** per al model. Aleshores, per a cada observació es calculen aquests indicadors:

$$T_i^2 = \sum_{a=1}^{A} \frac{t_{i,a}^2}{\lambda_a}, \quad SPE_i = \sum_{j=1}^{J} e_{ij}^2, \quad i = 1, ..., N. \tag{5.2.1}$$

On, A és el nombre de components principals que considerem en l'anàlisi PCA, $t_{i,a}$ és el valor dels *scores* del component $a = 1, ..., A$ per a l'observació $i = 1, ..., N$, i e_{ij} és el valor del residu de l'observació $i = 1, ..., N$ per a la variable $j = 1, ..., J$.

Gràcies a la tècnica de la PCA es pot realitzar una anàlisi exploratòria de les dades, així com es pot estudiar la correlació de les variables, la detecció de dades d'anòmales i la predicció d'algunes variables resposta.

5.3 *Clustering*

El *clustering* és una tècnica d'aprenentatge no supervisat del *machine learning* utilitzada principalment per agrupar elements amb certes característiques similars. A diferència dels mètodes supervisats, el *clustering* no requereix etiquetes predefinides per a les dades, la qual cosa el fa particularment útil per a explorar estructures subjacents en dades on no tenim coneixement previ de les categories existents. L'objectiu principal del *clustering* en aquest Treball de Fi de Màster és identificar grups naturals dins del conjunt de dades estudiat, revelant relacions intrínseques i comportaments similars entre els elements agrupats. Aquest enfocament ens ha permés identificar grups de partits que comparteixen característiques semblants en termes de punts, forma de joc, victòries, etc.

Per al nostre cas d'estudi, es va aplicar l'**estadístic de Hopkins** per tal de verificar l'existència de tendència a l'agrupament. Aquest estadístic mesura si les dades estan distribuïdes de manera aleatòria o mostren agrupament. Valors propers a 1 en aquest estadístic indiquen altes tendències d'agrupament. Considerem una mostra aleatòria $x_1, \ldots, x_m$ de tamany $m < N$ de la base de dades original. Aleshores, l'estadístic és:

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}, \tag{5.3.1}$$

on, u_i és la distància del punt x_i $(i = 1, \ldots, m)$ al punt més proper del conjunt de dades originals, i w_i $(i = 1, \ldots, m)$ és la distància del punt x_i al punt més proper entre els punts aleatoris.

Per a realitzar les agrupacions de les observacions, en primer lloc es calculen les distàncies o índexs de dissimilitud de les observacions. Com les K variables d'interés tenen unitats de mesura diferents, després d'escalar-les, la **distància Euclidiana** entre les observacions i i j és:

$$E_{ij} = \sqrt{\sum_{k=1}^{K} (X_{ki} - X_{kj})^2} \tag{5.3.2}$$

On, X_{ki} és el valor de la variable k per a la observació i i X_{kj} el valor de la variable k per a la observació j. Utilitzarem aquesta mesura de distància euclidiana per tal d'agrupar els partits en funció de les seues similituds o diferències en un espai multidimensional de les estadístiques de joc de la base de dades.

A continuació, es realitzà l'anàlisi d'agrupament mitjançant la tècnica de clúster *k-means*, amb el seu nombre òptim de clústers. Per obtindre aquest nombre de clústers hem utilitzat el coeficient de Silhouette i el mètode de suma de quadrats dins dels clústers. El nombre òptim de clústers és aquell que obté alhora un major valor al coeficient de Silhouette i una menor suma de quadrats. En aquest treball s'ha implementat la tècnica *clustering* utilitzant la llibreria *cluster* [12].

5.4 Índexs de competitivitat

L'equilibri competitiu entre els equips que participen en un torneig és una de les claus de l'èxit de les lligues esportives [13]. Aquest equilibri no només assegura que els partits siguen emocionants i imprevisibles, sinó que també manté l'interés dels aficionats i assegura la viabilitat econòmica a llarg termini de la lliga. Tots els competidors han de ser igualment capaços de guanyar la lliga per tal de tindre la màxima competitivitat possible. Un equilibri adequat fomenta un entorn on cap equip no domina de manera constant, evitant així la monotonia i augmentant la incertesa dels resultats, cosa que és fonamental per mantenir l'emoció i la passió dels seguidors. Els índexs de balanç competitiu poden ser de dos tipus: estàtics i dinàmics [14].

Els **índexs estàtics** mesuren la dispersió dels percentatges de victòries o de punts en les lligues esportives, proporcionant una visió instantània de l'equilibri competitiu. Aquests índexs són útils per comprendre la paritat en una temporada específica, i per tant, seran útils per estudiar el balanç competitiu al llarg de les temporades. Mostren sensibilitat pel nombre d'equips i de jornades a les lligues, i no reflectixen de manera adequada els canvis relatius en la classificació de la lliga al llarg dels anys [3].

D'altra banda, els **índexs dinàmics** examinen la variabilitat de la competitivitat al llarg del temps, permetent una anàlisi més profunda de les tendències i els canvis en la competitivitat d'una lliga al llarg de diverses temporades. D'aquesta manera, els índexs dinàmics poden revelar si l'equilibri competitiu és sostenible en el temps o si està subjecte a fluctuacions significatives.

5.4.1 Índexs estàtics

Per a aquest treball, farem servir els 6 següents índexs estàtics:

- **RSD**$_{w1}$: Ratio of Standard Deviation (per a 1 temporada).
- **RT**: Record Test.
- **HICB**: Herfindahl Index of Competitive Balance. Utilitzarem les versions de punts, de valoració i de victòries, a més de la versió estandarditzada de l'HICB per victòries, l'**SHICB**: Standardized Herfindahl Index of Competitive Balance.

Tot i que aquests índexs tenen diferents escales en funció de la manera en la que es calculen, tots tenen una cosa en comú; a majors valors dels índexs, menor és el balanç competitiu de la temporada o jornada d'estudi.

5.4.1.1 Ratio of Standard Deviation per a 1 temporada (RSD$_{w1}$)

En primer lloc, definim el percentatge de victòries de l'equip $i = 1, ...N$ en certa temporada:

$$WPCT_i = \frac{w_i}{G_i},$$

on w_i és el nombre de partits guanyats dels G_i partits jugats per l'equip i. En segon lloc, la desviació típica d'aquesta proporció de victòries per a certa temporada és:

$$\sigma_{w1} = \sqrt{\frac{\sum_{i=1}^{N}(WPCT_i - \overline{WPCT})^2}{N}},$$

on $\overline{WPCT}$ és la proporció mitjana de victòries en la temporada d'estudi. En les lligues amb un calendari de competició equilibrat i sense empats (com és el cas d'estudi), $\overline{WPCT}$ serà igual a 0.5 [3]. Les lligues o temporades amb un valor de σ_{w1} més elevat tenen més variabilitat en la distribució dels percentatges de victòries i per tant un menor balanç competitiu.

En general, les mesures de balanç competitiu basades en la desviació típica depenen de les característiques de les lligues, com ara el nombre d'equips que participen i el nombre de partits disputats. Per tant, considerem la **desviació típica idealitzada** per a una temporada on cada equip juga G partits o jornades [3]:

$$\sigma_1 = \frac{0.5}{\sqrt{G}}.$$

D'aquesta manera, obtenim l'**RSD (Ratio of Standard Deviation)** o proporció de Noll-Scully, que es calcula dividint la desviació obtinguda en la temporada d'estudi entre la desviació idealitzada:

$$RSD_{w1} = \frac{\sigma_{w1}}{\sigma_1} = \sqrt{\frac{G \cdot \sum_{i=1}^{N}(WPCT_i - \overline{WPCT})^2}{0.5^2 \cdot N}}. \tag{5.4.1}$$

Si l'índex RSD_{w1} pren el valor 1, indica que la desviació de victòries de la temporada d'estudi és igual a la desviació típica idealitzada. Aleshores, el balanç competitiu de la lliga durant la temporada d'estudi serà major com més prop d'1 es trobe l'RSD$_{w1}$, i menor com més lluny es trobe.

5.4.1.2 Record Test (RT)

L'**RT (Record Test)** és una mesura de la variació de victòries que de manera expressa té en compte el nombre total de partits en una temporada (G) [15].

$$RT = 4G\sum_{i=1}^{N}\left(\frac{w_i}{G} - \frac{1}{2}\right)^2. \tag{5.4.2}$$

Aquest índex està molt relacionat amb σ_{w1}, però en aquest cas es parteix de la hipòtesi que tots els equips tenen el mateix potencial i per tant, la mateixa probabilitat del 50% de guanyar cada partit. Així, els resultats d'aquest índex es distribueixen com una binomial amb mitjana G i variància 0.5.

Valors de $\frac{w_i}{G}$ més propers a 0.5 suposen un major balanç a la temporada, la qual cosa obté un valor menor de l'índex RT. Així, valors menors del Record Test suposen un major balanç competitiu, i pel contrari, valors majors del Record Test s'obtindran en temporades amb un menor balanç competitiu.

5.4.1.3 Mesures basades en l'índex de Herfindahl-Hirschman (HICB i SHICB)

L'**índex de Herfindahl-Hirschman** (HHI) és una mesura de concentració principalment utilitzada en organització industrial per avaluar el grau de competència d'un mercat. En el camp d'economia de l'esport, els usos comuns de l'HHI són la medició del balanç competitiu i de la concentració de les lligues [3]. Aquest índex resulta de la suma de quadrats de les proporcions de certs resultats, tals com la quantitat de campionats guanyats, victòries o punts. En general, per a una lliga on participen N equips, l'HHI es pot expressar com:

$$HHI = \sum_{i=1}^{N} s_i^2,$$

on s_i^2 és la proporció de victòries corresponent a l'equip $i = 1, ..., N$, encara que també es pot calcular en funció d'altres estadístiques de joc, com ara els punts. En aquest treball, es consideren tres tipus d'HHI. El primer, considera per al càlcul de l'HHI, la proporció de victòries de cada equip sobre el total de victòries de la temporada. En una lliga amb calendari equilibrat, amb N equips i G jornades, on cada equip juga el mateix nombre de vegades, el total de victòries és $\frac{NG}{2}$ [16]. Per tant:

$$HHI_v = \sum_{i=1}^{N} \left(\frac{2 \cdot w_i}{NG} \right)^2,$$

on w_i són el nombre de victòries de l'equip $i = 1, ..., N$ en la temporada d'estudi.

El segon i tercer casos, són casos anàlegs que consideren el total de punts i valoració de cada equip sobre els punts totals i valoració acumulada, respectivament, de la temporada. És a dir:

$$HHI_p = \sum_{i=1}^{N} \left(\frac{p_i}{\sum_{n=1}^{N} p_i} \right)^2, \quad HHI_{val} = \sum_{i=1}^{N} \left(\frac{val_i}{\sum_{n=1}^{N} val_i} \right)^2,$$

on p_i i val_i són la quantitat total de punts i valoració acumulada de l'equip $i = 1, ..., N$ en la temporada d'estudi.

L'HHI no és una mesura estandarditzada ja que depén principalment del nombre d'equips de la lliga [16]. Un nombre major d'equips suposa, generalment, un major HHI. Per tant, es considera l'**HICB (Herfindahl Index of Competitive Balance)**, que utilitza un ajust multiplicatiu per al límit inferior de l'HHI, i així té en compte el nombre d'equips que participen per temporada [13]. De nou, es consideren tres índexs, que tenen en compte la proporció de victòries, de punts i de valoració, respectivament.

$$HICB_v = \frac{HHI_v}{\frac{1}{N}} \cdot 100, \quad HICB_p = \frac{HHI_p}{\frac{1}{N}} \cdot 100, \quad HICB_{val} = \frac{HHI_{val}}{\frac{1}{N}} \cdot 100. \tag{5.4.3}$$

Un índex HICB de 100 indica que la temporada té un **balanç perfecte**. A mesura que es fa gran, el balanç competitiu de la lliga disminueix. No obstant això, el límit superior de l'índex $HICB_v$ és sensible al nombre d'equips que formen la lliga. Considerem el següent límit superior (UB, de l'anglés, *upper bound*) per a l'$HICB_v$ en una lliga amb N equips [17]:

$$HHI_v^{UB} = \frac{2(2N-1)}{3N(N-1)} \implies HICB_v^{UB} = \frac{2N(2N-1)}{3N(N-1)} \cdot 100$$

A continuació, tenim calculats aquests límits superiors per als diferents nombres d'equips que tenim en la nostra base de dades, els quals trobem a la Taula 4.1.

Taula 5.1: Límits superiors per als nombres d'equips de les dades d'estudi.

Nre. Equips (N)	**Límit Superior ($HICB_v^{UB}$)**
11	140.00
12	139.40
14	138.46
16	137.78

Com el límit superior de l'índex HICB varia depenent del nombre d'equips de la lliga, considerem una versió estandaritzada d'aquesta mesura, l'**SHICB (Standardized Herfindahl Index of Competitive Balance)**. Aquest es calcula dividint entre el límit superior de l'HICB per al nombre d'equips que té. És a dir:

$$SHICB_v = \frac{HICB_v}{HICB^{UB}} \cdot 100 \tag{5.4.4}$$

Un índex de 100 en l'SHICB indica la lliga amb el **menor** balanç competitiu possible, i a mesura que aquest valor disminueix, el balanç competitiu és major [13].

5.4.2 Índexs dinàmics

Per aquest treball, farem servir els següents índexs dinàmics:

- **RSD**$_{w2}$: Ratio of Standard Deviation (per a T>1 temporades).
- **CBR**: Competitive Balance Ratio.

5.4.2.1 Ratio of Standard Deviation per a T>1 temporades (RSD$_{w2}$)

En primer lloc, hem de calcular la desviació típica dels percentatges de victòries per a múltiples temporades, per tal d'avaluar l'evolució dels equips al llarg dels anys. Com en el nostre cas d'estudi, es tracta d'una lliga d'ascens i descens, on no tots els equips es mantenen en la lliga al llarg dels anys, calcularem l'evolució dels percentatges de victòries en **funció de les posicions finals** dels equips. És a dir, comparem, per exemple, el percentatge de victòries que ha tingut l'equip que ha quedat en primera posició en una temporada, amb l'equip que queda en la mateixa posició la temporada següent.

Aleshores, considerem en aquest cas que N_t és el nombre total d'equips que han participat en la temporada $t = 1, ..., T$. En conseqüència, la desviació típica del percentatge de victòries per a la lliga durant les T temporades serà:

$$\sigma_{w2} = \sqrt{\sum_{t=1}^{T} \sum_{i=1}^{N_t} \frac{(WPCT_{i,t} - \overline{WPCT})^2}{N_t T}},$$

on $WPCT_{i,t}$ és la proporció de victòries de l'equip que **acaba la lliga en la posició i-èssima**, $i = 1, ..., N_t$ en la temporada $t = 1, ..., T$, i $\overline{WPCT}$ és la mitjana de tots els percentatges de victòries en les T temporades.

D'igual forma que ocorre amb σ_{w1} per a una única temporada, el valor de σ_{w2} depén del nombre d'equips (N) i de temporades (T) que es calculen. Per poder comparar el balanç competitiu en lligues amb calendari variable, s'ha de realitzar una normalització del les desviacions, que es realitza mitjançant una desviació típica idealitzada.

Recordem que la desviació típica idealitzada per a una temporada on cada equip juga G partits o jornades és σ_1 [3]. Aleshores, per a un nombre variable de partits o jornades G_t per a la temporada $t = 1, ..., T$, la desviació típica idealitzada serà:

$$\sigma_1 = \frac{0.5}{\sqrt{G}} \implies \sigma_1' = \frac{0.5}{\sqrt{G_t}}.$$

Aleshores, per al cas de l'estudi de la variació de diverses temporades, l'**RSD (Ratio of Standard Deviation)** que s'obté per a T temporades, es calcula dividint la desviació obtinguda entre la idealitzada σ_1'.

$$RSD_{w2} = \frac{\sigma_{w2}}{\sigma_1'} = \sqrt{\sum_{t=1}^{T}\sum_{i=1}^{N_t} \frac{G_t \cdot (WPCT_{i,t} - \overline{WPCT})^2}{0.5^2 \cdot N_t T}}. \tag{5.4.5}$$

Aquesta és una mesura que té en compte més d'una temporada, es tracta d'un **índex dinàmic**. Com es tracta d'una proporció, un valor d'1 en l'índex RSD_{w2} suposa un balanç competitiu perfecte, i aquest balanç disminueix com més lluny es trobe l'índex d'1.

5.4.2.2 Competitive Balance Ratio (CBR)

Donat el fet que les mesures de balanç competitiu estàtiques no tenen en compte la variació en les posicions relatives al llarg del temps (si un equip es manté en la lliga molts anys, o no), es proposa com a mesura dinàmica, el CBR o proporció del balanç competitiu. Aquest índex expressa l'equilibri competitiu com un únic nombre que compara la variació dels percentatges de victòries d'un equip (intraequip) amb la variació dels percentatges de victòries de la resta de la lliga (intralliga) [3]. Cal tindre en compte que si un equip no participa en una temporada, el seu percentatge de victòries serà 0, de manera que la seua mitjana al llarg de les temporades d'estudi disminuirà.

Es defineix la variació en els percentatges de victòries **per a l'equip** $i = 1, ..., N$[1] al llarg de T temporades com:

$$\sigma_{T,i} = \sqrt{\frac{\sum_{t=1}^{T}(WPCT_{i,t} - \overline{WPCT}_i)^2}{T}}, \quad i = 1, ...N_t.$$

Aquesta mesura reflexa l'evolució d'un equip individual al llarg de les temporades. Per l'altra banda, es defineix la variació en els percentatges de victòries **intratemporada**, és a dir, en la temporada $t = 1, ..., T$ per a N equips com:

$$\sigma_{N,t} = \sqrt{\frac{\sum_{i=1}^{N}(WPCT_{i,t} - 0.5)^2}{N}}, \quad t = 1, ...T.$$

Aquesta mesura obté un únic valor per temporada, i per tant, reflexa l'evolució de la lliga al llarg de les temporades.

[1] En aquest cas, N és el nombre d'equips **diferents** que participen durant les T temporades.

L'índex CBR és el quocient de les mitjanes d'aquests dos tipus de variacions, que aconsegueix així una mesura de variació mitjana dels percentatges de victòries de les temporades considerades.

$$CBR = \frac{\overline{\sigma}_T}{\overline{\sigma}_N} = \frac{\frac{\sum_{i=1}^{N} \sigma_{T,i}}{N}}{\frac{\sum_{t=1}^{T} \sigma_{N,t}}{T}} = \frac{T \cdot \sum_{i=1}^{N} \sigma_{T,i}}{N \cdot \sum_{t=1}^{T} \sigma_{N,t}}. \tag{5.4.6}$$

Els beneficis del CBR davant altres índexs com l'HICB, són entre altres, que el CBR facilita la comparació sobre períodes de temps diferents, ja que no s'ha de comparar amb un valor de variació ideal. Donat el fet que el denominador del CBR inclou la σ_{w2}, aquestes dues mesures es troben inversament relacionades. El CBR reflexa no només part de la mateixa informació que la desviació típica del percentatge de victòries, sinó que també reflexa la mitjana de la variació del percentatge de partits guanyats/perduts de cada equip, cosa que no es veu reflectida en σ_{w2} [3].

A més com es tracta d'una proporció, els seus valors es troben entre 0 i 1. Valors propers a 1 indiquen major balanç competitiu, mentre que valors més xicotets impliquen menor equilibri competitiu.

5.5 Models

En aquesta secció, s'expliquen els models realitzats per tal d'assolir el quart dels objectius exposats prèviament. En aquests models volem trobar quins efectes fixos i aleatoris tenen relació amb l'augment o disminució del balanç competitiu.

5.5.1 Models Lineals Generalitzats

Els **models lineals generalitzats** (GLM) són una extensió del model lineal $Y = X\beta + \epsilon$, que permet assumir distribucions diferents a la normal per a la variable Y, com recomptes o proporcions; cosa que ens permet cert grau de **no linealitat** en l'estructura del model [18].

Els GLM segueixen l'estructura següent:

$$g(\mu_i) = \alpha + X_i\beta, \quad i = 1, \ldots, n, \tag{5.5.1}$$

on $\mu_i \equiv E(Y_i)$, β és un vector de paràmetres desconeguts i g és una funció monòtona "suau" anomenada **funció d'enllaç**, la qual descriu com l'esperança d'Y està relacionada amb una combinació lineal dels predictors [19].
El GLM assumeix que els Y_i són independents i segueixen alguna distribució de la família de les exponencials, per tant la seua funció de densitat tindrà la forma:

$$f_\theta(y) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right], \tag{5.5.2}$$

on a, b i c són funcions arbitràries, ϕ un paràmetre d'escala arbitrari, i θ és conegut com el paràmetre canònic de la distribució.

5.5.2 Models multinivell o d'efectes aleatoris

En certs camps d'investigació, com ara l'epidemiologia, entre altres, és molt comú que l'estructura dels individus estiga organitzada en forma jeràrquica. És a dir, els individus es troben agrupats dins d'unitats de nivell més alt, i alhora també agrupats en altres unitats [20].

El fet d'establir aquesta jerarquia entre les variables suposa certes conseqüències a l'hora d'analitzar els resultats, ja que els individus amb un mateix context tendiran a ser més pareguts entre ells, i a tindre comportaments similars. Aquesta similitud entre els individus dins d'un mateix grup impedeix el compliment de la hipòtesi d'independència, hipòtesi necessària per als models de regressió tradicionals. Aquest fet, suposa estimacions molt més xicotetes dels errors estàndard per l'alta correlació entre els individus, i aquests resultats són, a sovint, falsament significatius [20]. Recentment, s'ha intentat adaptar aquestes estructures jeràrquiques als models lineals generalitzats, per donar pas als **models multinivell o models jeràrquics**.

Per al nostre cas d'estudi, els individus són els índexs de competitivitat per jornada, per a cadascuna de les temporades. És a dir, tenim **2 nivells** de jerarquia, com podem observar en la Figura 5.3:

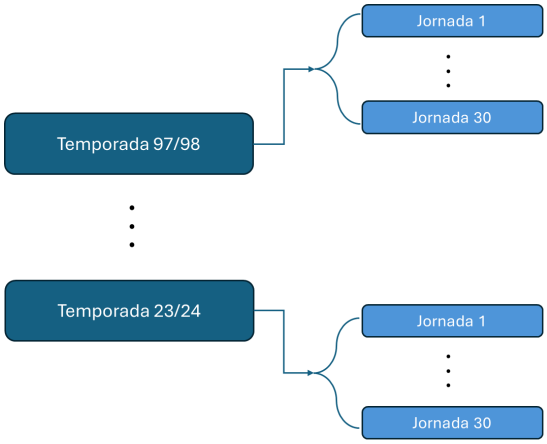

Figura 5.3: Esquema jeràrquic de les dades

El model de regressió multinivell complet assumeix que existeix un conjunt de dades jeràrquiques, amb una única variable depenent, que és mesurada en el nivell més baix de la jerarquia; i variables explicatives que existeixen a tots els nivells. Aquest model es pot considerar com un sistema jeràrquic d'equacions de regressió.

Siga Y la variable depenent (numèrica) que volem modelitzar en el nostre estudi, i siguen $X_1, \ldots, X_I$, les I variables explicatives del nostre model. Com hem comentat prèviament, considerem un model multinivell a 2 nivells; temporada i jornada. Aleshores, per a la temporada $j = 1, \ldots, T$ i per a la jornada $k = 1, \ldots, J$ (on J és el nombre màxim de jornades que es juguen en alguna de les temporades d'estudi, les quals podem veure en la Taula 4.1); l'equació del model és la següent:

$$\beta_0 + \sum_{i}^{I} \beta_i X_{ijk} + u_j + v_k, \quad \begin{cases} j = 1, \ldots, T, \\ k = 1, \ldots, J. \end{cases} \tag{5.5.3}$$

On $\beta_0, \beta_1, \ldots, \beta_I$ són els coeficients usuals que coneixem per a les variables explicatives del model, per a totes les observacions (independentment de la jornada o temporada); u_j per a $j = 1, ..., T$ són els coeficients de l'efecte aleatori temporada, i v_k per a $k = 1, ..., J$ són els coeficients de l'efecte aleatori jornada.

Existeixen molts avantatges d'utilitzar models multinivell, com obtindre millors estimacions dels coeficients de regressió, entre altres. Cal assenyalar que existeix una teoria molt més complexa sobre aquests models, i que la interpretació dels resultats d'aquests no sempre és molt evident, especialment quan es tracten estructures complexes [20]. Per a la implementació en R, farem ús de la llibreria *lme4* [21].

Capítol 6

Resultats

En aquesta secció s'exposen els resultats obtinguts després d'aplicar les tècniques explicades a la nostra base de dades, per tal d'assolir els objectius proposats. S'inclouen taules i gràfiques obtingudes mitjançant el paquet *ggplot2* [22] amb la finalitat d'aconseguir una millor visualització i comprensió de les conclusions.

6.1 PCA exploratòria

En primer lloc, es mostra el gràfic de correlacions (Figura 6.1, obtés mitjançant la llibreria *corrplot* [23]) entre les variables numèriques de la base de dades d'estudi, per tal de reflectir una idea general de les relacions ja existents entre aquestes variables.

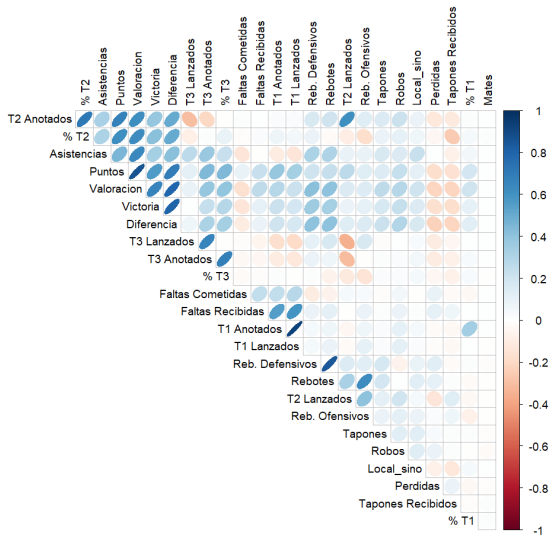

Figura 6.1: Gràfic de correlacions entre les variables numèriques

Com era d'esperar, existeix correlació positiva entre les variables que indiquen la victòria de l'equip, variables com ara Punts, Valoració, Victòria o Diferència; i alhora, existeix correlació negativa entre aquestes i les variables relacionades amb les estadístiques negatives, com ara les Pèrdues o els Tapons Rebuts.

A continuació, es realitza la validació de les observacions mitjançant l'estadístic T2-Hotelling i el SPE (Squared Prediction Error).

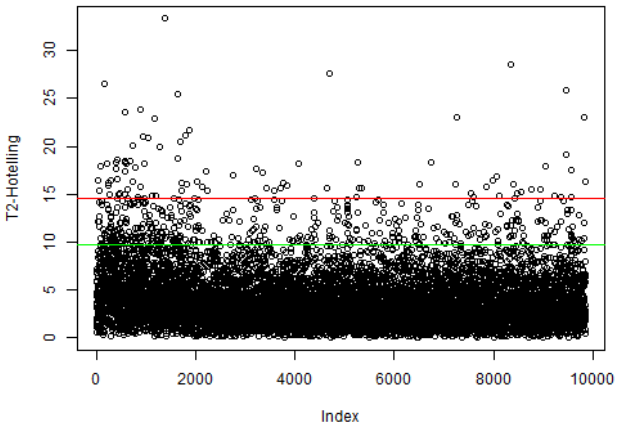

Figura 6.2: Validació PCA: T^2 - Hotelling

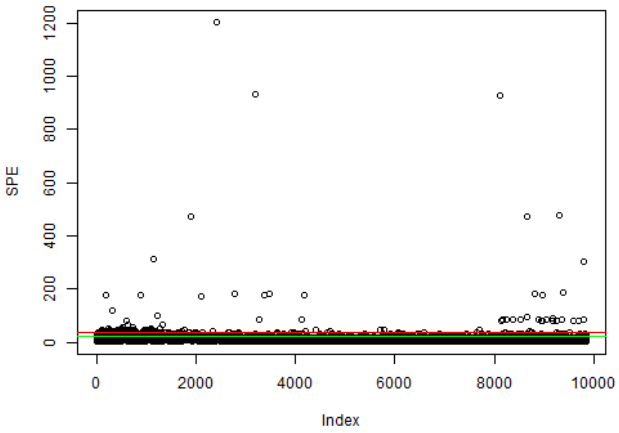

Figura 6.3: Validació PCA: SPE

Com podem observar, no trobem cap observació (partit) per damunt del triple del percentil 99 de l'estadístic T^2 de Hotelling (línia roja en la Figura 6.2). No obstant això, són diverses les observacions que es troben per damunt del triple del percentil 99 de la suma dels errors al quadrat (línia roja en la Figura 6.3). En concret, de les 9850 observacions d'estudi, aquelles que podem considerar com observacions atípiques, i per tant, com **outliers moderats**, són un total de 15 observacions, el resum de les quals podem trobar a l'Annex 2 (Taula 8.4). Aleshores, com es tracten d'outliers moderats i no severs, no eliminem aquestes observacions de la nostra base de dades per realitzar l'anàlisi.

No obstant això, es deixa l'anàlisi d'aquests outliers com a futura línia d'investigació, ja que l'estudi d'aquests és un aspecte important per a aprofundir en la comprensió de les causes que poden influir en la presència d'aquestes observacions atípiques. L'anàlisi detallada dels outliers moderats podria proporcionar informació rellevant sobre possibles factors desconeguts o no considerats en el model inicial, així com identificar patrons específics en els equips o partits que poden estar associats a aquestes anomalies.

Després d'escalar les variables numèriques d'estudi, per tal que tinguen mitjana nula i variància unitària, considerem que el nombre òptim de components a considerar per al PCA són 4, ja que en el *Scree Plot* de la Figura 6.4 podem observar que entre el component 4 i 5 es forma l'anomenat *colze* que indica que, efectivament, 4 és el nombre òptim. Entre els 4 primers components s'explica un **51.9%** de la variància de les dades.

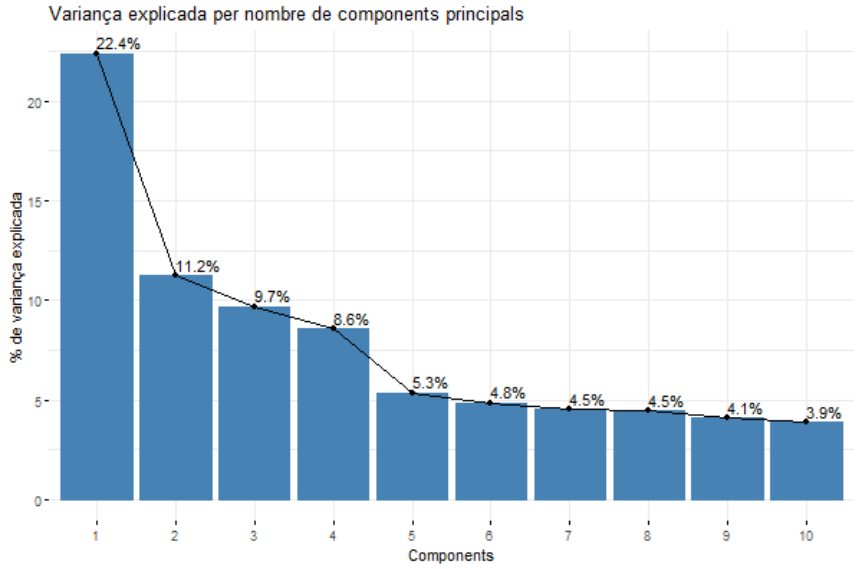

Figura 6.4: *Scree Plot*: variància explicada per nombre de components principals

En la Taula 6.1, podem veure quines variables contribueixen més a cada component, és a dir, aquelles variables que tenen un valor en el *loading* major que 0.29 (en valor absolut), per a cadascun dels 4 components principals.

Taula 6.1: Variables que més contribueixen a cada component

1r component	**2n component**	**3r component**	**4t component**
Valoracion	T1 Lanzados	T2 Lanzados	Rebotes
Puntos	T1 Anotados	Rebotes	T2 Anotados
Diferencia	T3 Anotados	Reb. Ofensivos	% T2
Victoria	Faltas Recibidas	T3 Anotados	T3 Lanzados
	T3 Lanzados	T1 Anotados	Reb. Ofensivos
			Reb. Defensivos

A continuació, realitzem un *Loading Plot*, en el qual es pinten les variables en funció de la seua contribució, per tal de trobar les relacions entre les variables per a la seua posterior anàlisi. En la Figura 6.5, trobem els components 1 i 2 enfrontats, mentre que a la Figura 6.6 trobem els components 3 i 4.

En aquests gràfics podem observar que existeixen variables com ara Punts, Valoració, Diferència, Victòria, Rebots, T1 Llançats o T1 Anotats (totes elles pintades en roig), que tenen una major contribució als components estudiats. De la mateixa manera, existeixen altres variables (pintades de color blau) que tenen una menor contribució als components d'estudi.

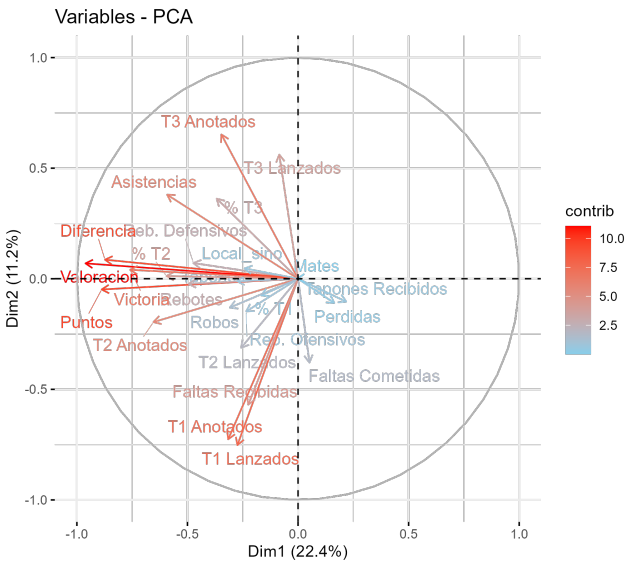

Figura 6.5: Loading plot components 1 i 2

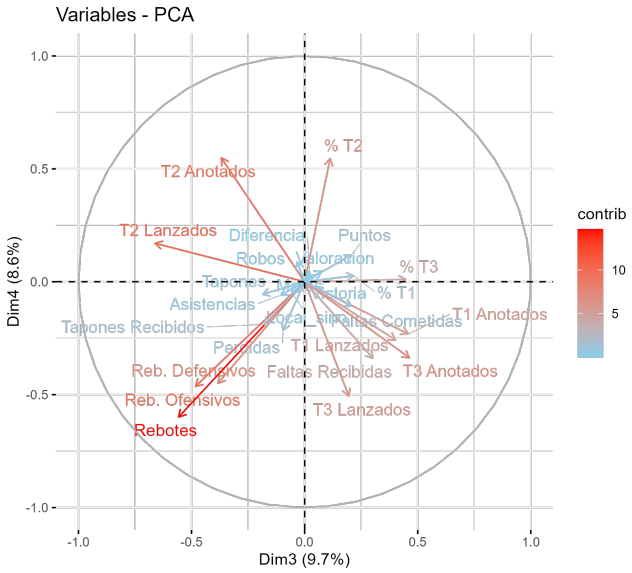

Figura 6.6: Loading plot components 3 i 4

Les variables que més contribueixen al **primer component** són aquelles que es troben pintades de roig a l'eix x de la Figura 6.5. Com podem observar, la variable Victòria té alta contribució a aquest component, a més de coeficient negatiu. Això indica que altres variables amb la mateixa direcció seran estadístiques indicatives d'un bon partit, i per tant, d'una possible victòria, com ara Diferència, Valoració i Punts (amb major contribució), o T2 Anotats, % T2, Assistències, Rebots o Rebots Defensius (variables amb menor contribució). És a dir, estadístiques elevades en aquestes variables són característiques de les victòries. D'altra banda, aquelles variables que es troben a la dreta en l'eix x, com Tapons Rebuts o Pèrdues, seran indicadors que el partit acabarà en derrota.

Les variables que més contribueixen al **segon component** són aquelles que es troben a l'eix y de la Figura 6.5, i principalment les relacionades amb els tirs d'1 i 3 punts. Com podem observar en la Figura 6.5, les variables T3 Llançats i T3 Anotats es troben inversament relacionades a les variables T1 Llançats, T1 Anotats i Faltes Rebudes. Açò podria fer referència a la forma de jugar dels equips, ja que un major nombre de triples indicaria que donen més pes al joc

per fora de l'àrea (jugadors amb posicions de base, escolta i aler), i un major nombre de faltes rebudes, i per tant, tirs lliures, indicaria que donen major pes als jugadors interiors (jugadors amb posicions de pivot i aler-pivot) [24].

Les variables que més contribueixen al tercer i quart components no són tan clares d'identificar com les anteriors. No obstant això, en ambdós components, els rebots tenen un gran pes, rebots tant ofensius, com defensius. En quant al **tercer component**, podem observar en la Figura 6.6 que les variables T2 Llançats, Rebots i Rebots Ofensius es troben inversament relacionats amb les variables T3 Anotats i T1 Anotats. En quant al **quart component**, podem observar que les variables T2 Anotats i % T2 es troben inversament relacionades amb les variables Rebots, T3 Llaçats i Rebots Ofensius i Defensius.

6.2 Clúster

En primer lloc, cal recordar que l'objectiu d'aquesta secció és fer grups d'equips per temporades i jornades, per tal de tindre la proporció d'equips que pertanyen a cadascun d'aquests grups o clústers. Aleshores, calculem primer l'**estadístic de Hopkins** mitjançant la funció `get_clust_tendency()`, de la llibreria *factoextra* [10]. Aquesta anàlisi detallada ens permet obtindre una perspectiva integral sobre la tendència d'agrupació en les dades estandarditzades, on, valors propers a 1 indiquen una major tendència d'agrupació a les dades d'estudi. Cal tindre en compte que per a aquesta secció utilitzarem únicament les estadístiques de joc de la base de dades original, les quals han sigut auto-escalades per tal de tindre mitjana nul·la i variància unitària.

Taula 6.2: Mesures de posició de l'estadístic de Hopkins

Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
0.8073	0.8234	0.8278	0.8275	0.8327	0.8422

Vist en la Taula 6.2 que els valors es troben al rang de 0.8073 a 0.8422, es consideren propers a 1, i per tant, podem concloure que hi ha una clara tendència d'agrupament entre les observacions. A continuació, utilitzem el mètode de Silhouette i el mètode de suma de quadrats dins dels clústers per tal d'elegir el nombre òptim de clústers.

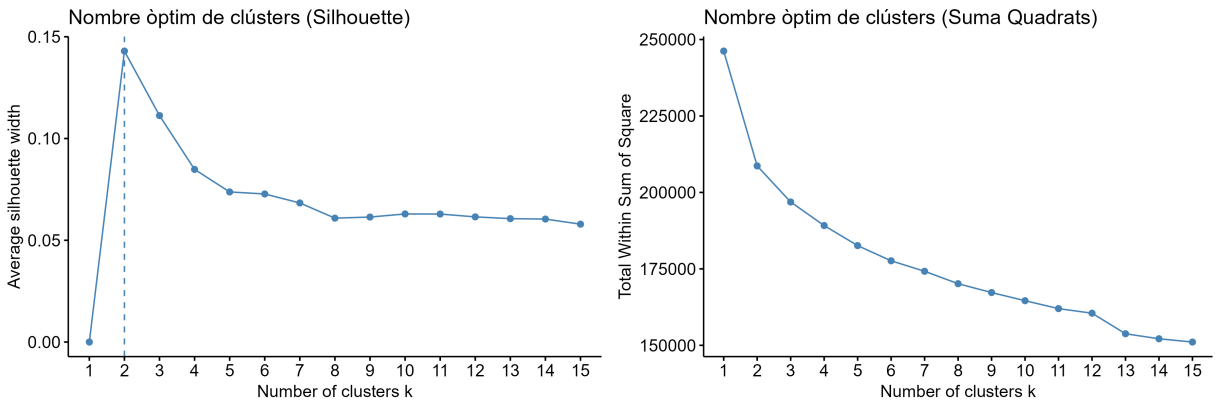

Figura 6.7: Coeficient de Silhouette i Suma de Quadrats dins dels clústers

Els resultats de l'anàlisi del coeficient Silhouette de la Figura 6.7 suggereixen que el nombre òptim de clústers és 2. No obstant això, després d'examinar amb deteniment les dues gràfiques, hem observat que l'òptim se situa en un rang entre 4 i 7 clústers, ja que aquests valors presenten similituds en termes del coeficient de Silhouette. A més, en analitzar la suma de quadrats intracluster, hem identificat un marcat canvi en el criteri del colze al voltant de 4 clústers. Considerant ambdós enfocaments, hem conclòs que el nombre òptim de per a la nostra anàlisi és $k = 4$.

Després d'escollir una llavor d'aleatorietat per tindre reproductibilitat dels resultats, obtenim les següents proporcions d'observacions dins de cada clúster.

Taula 6.3: Repartició de les observacions per a 4 clústers

Clúster	**Nre. d'observacions**	**Proporció**
1	2054	0.209
2	2157	0.219
3	3497	0.355
4	2142	0.217

Com podem observar a la Taula 6.3, els clústers estan prou equilibrats, a excepció del clúster 3, que té quasi el doble d'observacions que la resta de clústers. Generem aleshores el gràfic preliminar de *scores* de la PCA amb el model seleccionat (*k-means*).

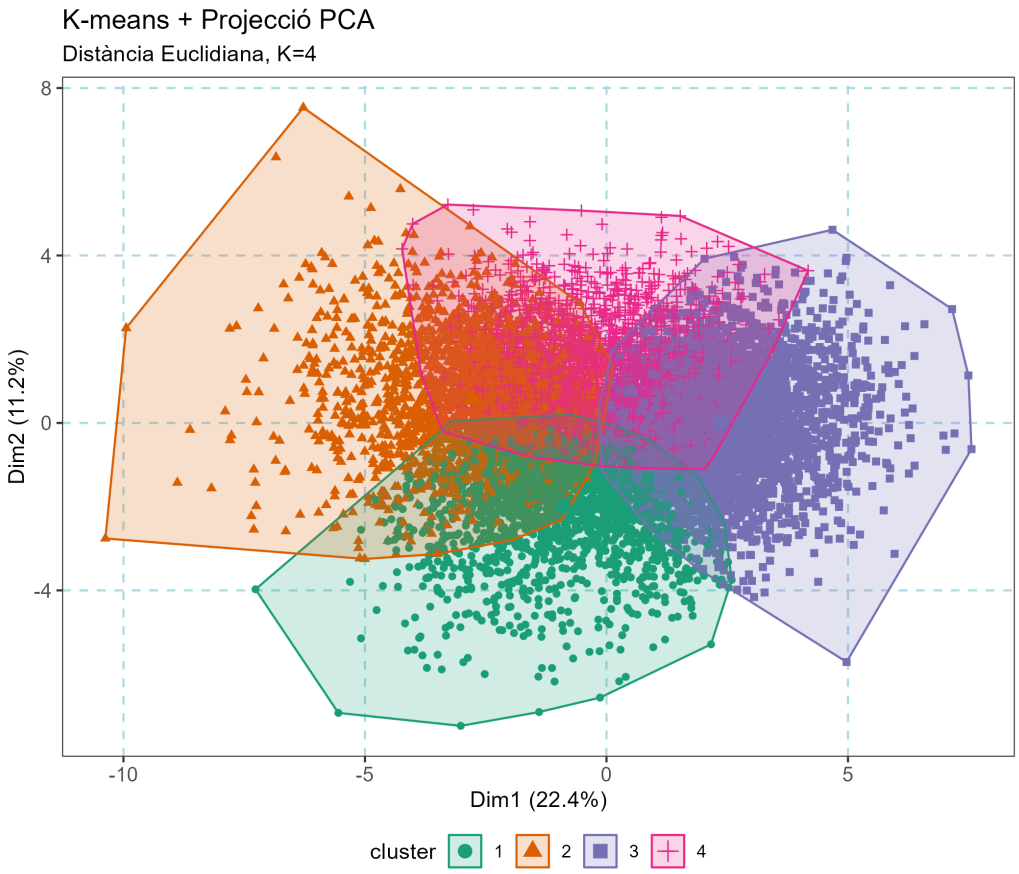

Figura 6.8: Clúster k-means sobre les projeccions de la PCA

En la Figura 6.8 observem que els 4 clústers es solapen per a les observacions amb valors propers a 0 en el primer i segon components de la PCA. En concret, podem observar que existeixen dues tendències, que es troben ben diferenciades pels dos primers components de la PCA. En concret, els clústers 2 (taronja) i 3 (morat) es troben ben diferenciats pel primer component, mentre que els clústers 1 (verd) i 4 (rosa) es troben diferenciats pel segon component.

No obstant això, aquesta superposició als clústers és una senyal negativa en l'anàlisi d'agrupament, ja que indica que l'algorisme no està aconseguint distingir de manera clara les diferències entre grups. Aquest fet el podem corroborar amb un gràfic del coeficient de Silhouette (Figura 6.9). En aquesta figura, s'observa que el criteri elegit (*k-means*) no presenta un bon rendiment. Això es reflecteix en un valor mitjà més xicotet de Silhouette i una major proporció d'equips mal classificats (indicats per coeficients negatius).

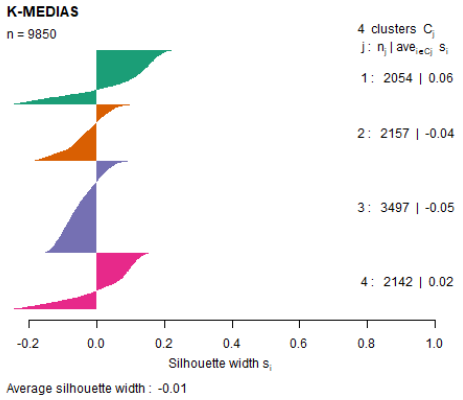

Figura 6.9: Coeficients Silhouette

A continuació, relacionarem l'anàlisi PCA prèvia amb els resultats del *clustering* per tal d'interpretar millor els resultats del clúster.

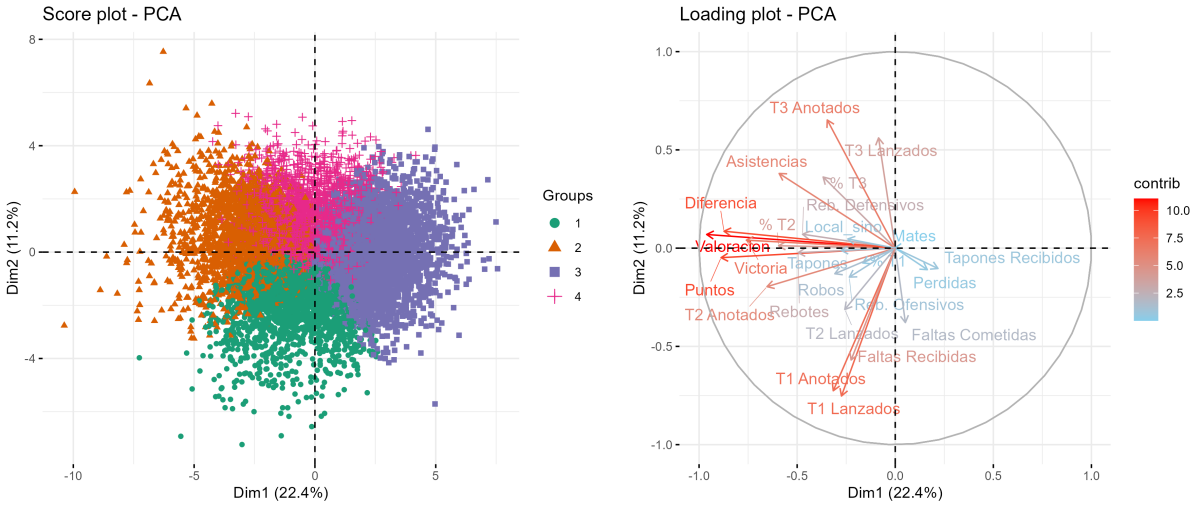

Figura 6.10: *Score plot* amb els clústers (esquerra) i *loading plot* (dreta) de la PCA exploratòria

Per una banda, comparant els gràfics d'observacions i variables, podem veure que el clúster 2 (taronja) està compost per observacions que en general, són partits guanyats per molts punts, ja que tenen majors valors per a les variables punts, valoració, diferència i victòria. Per tant, en

contraposició als primers, aquelles observacions que es troben al clúster 3 (morat) seran partits perduts per molts punts.

Per l'altra banda, per a aquells partits on els resultats han estat més igualats, es té una nova classificació mitjançant els clústers 1 i 4; que es troben diferenciats pel segon component de la PCA. Les observacions que pertanyen al clúster 4 (rosa) tindran valors positius en aquest component, la qual cosa suposa majors valors a les variables T3 llançats, T3 anotats i assistències. És a dir, seran partits on el joc exterior ha tingut major pes (jugadors amb posicions de base, escolta i aler han destacat més) [24]. Per contra, les observacions que pertanyen al clúster 1 (verd) obtenen majors valors a les variables T1 llançats, T1 anotats i faltes rebudes, la qual cosa indica un major pes al joc interior (destaquen més els jugadors amb posicions de pivot i aler-pivot).

En concret per al nostre estudi ens interessa saber la proporció d'equips que pertanyen a cada clúster en cada jornada i temporada. En la Figura 6.11 podem observar com van evolucionant en funció de la temporada o jornada, la proporció d'equips que pertanyen a cada clúster.

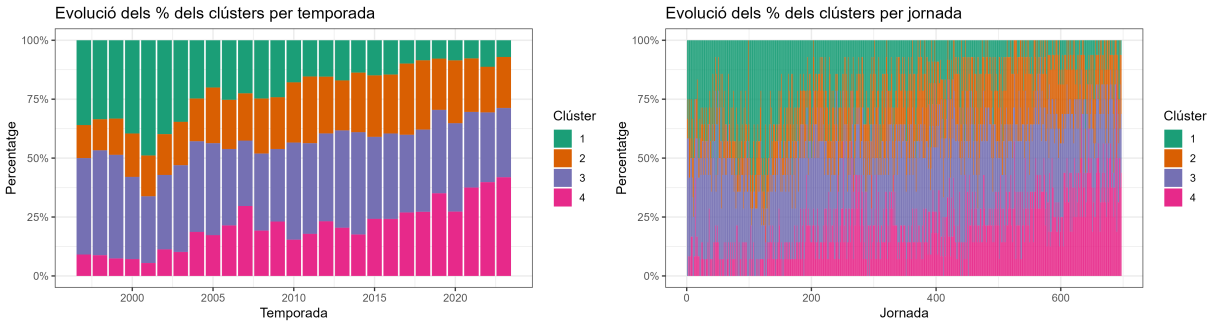

Figura 6.11: Evolució per temporada (esquerra) i jornada (dreta) de les proporcions d'equips que pertanyen a cada clúster

En primer lloc, podem veure com els clústers 2 i 3 són complementaris, i els clústers 1 i 4 també ho són, és a dir, si un d'ells augmenta, el seu complementari disminueix. D'aquesta manera, podem afirmar que la quantitat d'equips que destaquen amb molt bones estadístiques (clúster 2) era menor en els primers anys d'estudi, i que amb el temps ha crescut fins mantindre's estable entre les temporades 2005-2023. De manera contrària, els equips amb pitjors estadístiques (clúster 3) predominaven en els primers anys, i ara s'han reduït.

En quant a la tipus predominant de joc, ha hagut un gran canvi des de les primeres temporades d'estudi fins les últimes. Els equips amb major joc interior (clúster 1) predominaven a finals dels anys 90, i amb el pas del temps han donat lloc a equips on predomina el joc exterior (clúster 4). Aquest és un fet molt comú que ha ocorregut en el bàsquet en general en les últimes dècades, sobretot en la NBA, i són molts els articles i estudis que parlen del tema.

6.3 Índexs de competitivitat per temporada

Per tal d'aconseguir el tercer dels objectius, es calcularen els índexs de competitivitat **per temporada** de les dades disponibles. Per tant, com tenim dades de la Lliga Femenina Endesa des de la temporada 1997/98 fins la 2023/24, tindrem 27 valors dels diversos índexs de competitivitat.

6.3.1 Índexs estàtics

En primer lloc, es desenvoluparen els índexs de competitivitat estàtics, és a dir, aquells que utilitzen les dades d'una única temporada per ser calculats.

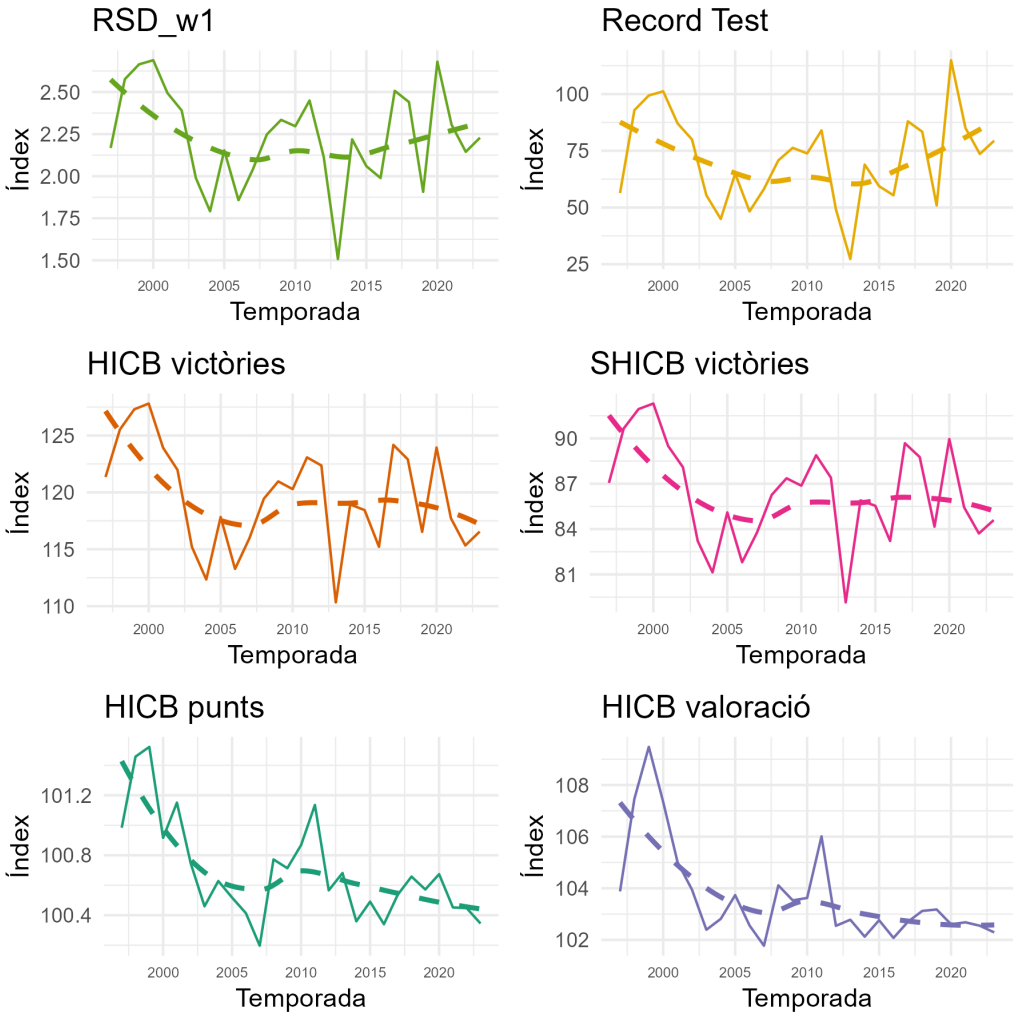

Figura 6.12: Índexs de competitivitat estàtics per temporada

En la Figura 6.12 podem observar els valors dels 6 índexs d'estudi al llarg de les 27 temporades, a més de la suavització de les dades mitjançant la funció `loess()`. El mètode de regressió *loess* és un enfocament *binned* [1] on s'ajusta una funció suau a través dels punts de dades en cada bin o interval. Cal destacar que tots els índexs **estàtics** estudiats tenen diferents escales (com podem observar a la Figura 6.13), però en tots els casos, **menors valors en els índexs indiquen major balanç competitiu**, i a mesura que els índexs augmenten, el balanç competitiu disminueix.

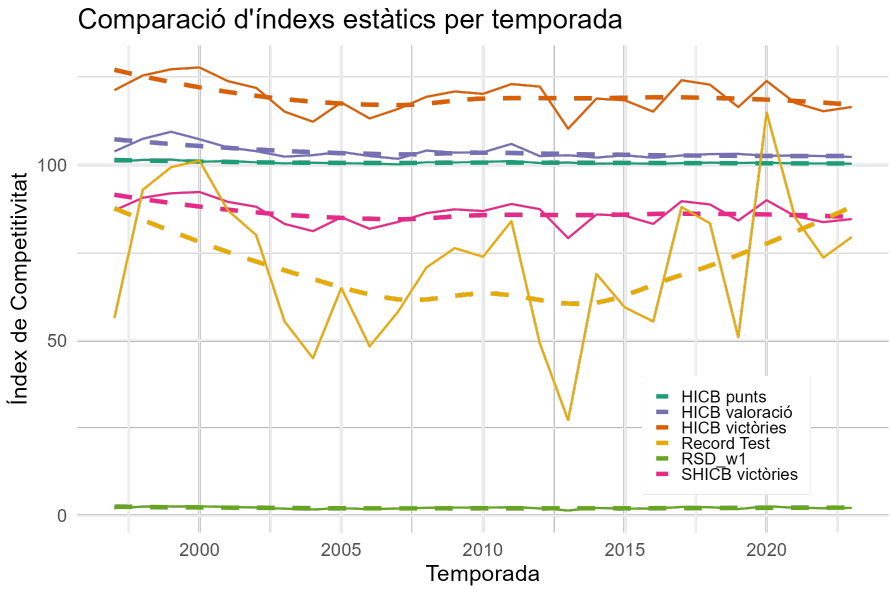

Figura 6.13: Estàtics per temporada

En primer lloc, podem observar a la Figura 6.12 que tots els índexs estàtics calculats coincideixen en que les primeres temporades d'estudi són aquelles que tenen un menor balanç competitiu, és a dir, on els equips o partits estaven més desequilibrats. No obstant això, per a les últimes temporades d'estudi s'obtenen conclusions prou diverses en funció de l'índex que s'estiga estudiant.

En les gràfiques per separat de la Figura 6.12, trobem tres tendències diferents als valors obtinguts, que estan justificades pel mètode pel qual s'han calculat els índexs. Així, trobem diferències entre els índexs basats en la desviació típica del percentatge de victòries (RSD_{w1} i Record Test), i els índexs basats en l'HHI; per una banda aquells que tenen en compte la proporció de victòries ($HICB_v$ i $SHICB_v$), i per l'altra, aquells que tenen en compte la proporció d'altres estadístiques de joc ($HICB_p$ i $HICB_{val}$).

[1] El mètode de regressió *binned* és una tècnica no paramètrica utilitzada per modelitzar relacions no lineals entre variables predictores i la variable resposta [25].

En concret, podem observar que, en els índexs de les desviacions típiques indiquen que el balanç competitiu havia augmentat entre les temporades 2000-2010, però que en els últims anys ha disminuït de nou, és a dir, de nou hi ha més desigualtat en la quantitat de victòries per equip. Els índexs basats en l'HHI i el percentatge de victòries obtenen uns valors prou similars als anteriors, a excepció de les últimes temporades, en les quals consideren que el balanç competitiu no disminueix tant. Per últim, els $HICB_p$ i $HICB_{val}$ consideren que en les últimes temporades s'ha assolit un balanç competitiu quasi perfecte, és a dir, que els punts s'han repartit de forma equitativa.

Tot seguit, es presenten tots els índexs junts per mostrar la comparació dels índexs autoescalats (Figura 6.14), per tal que tots tinguen mitjana nul·la i variància unitària. Cal destacar que podem realitzar la comparació d'aquests índexs d'aquesta manera per què, en tots els casos, menors valors dels índexs indiquen major balanç competitiu.

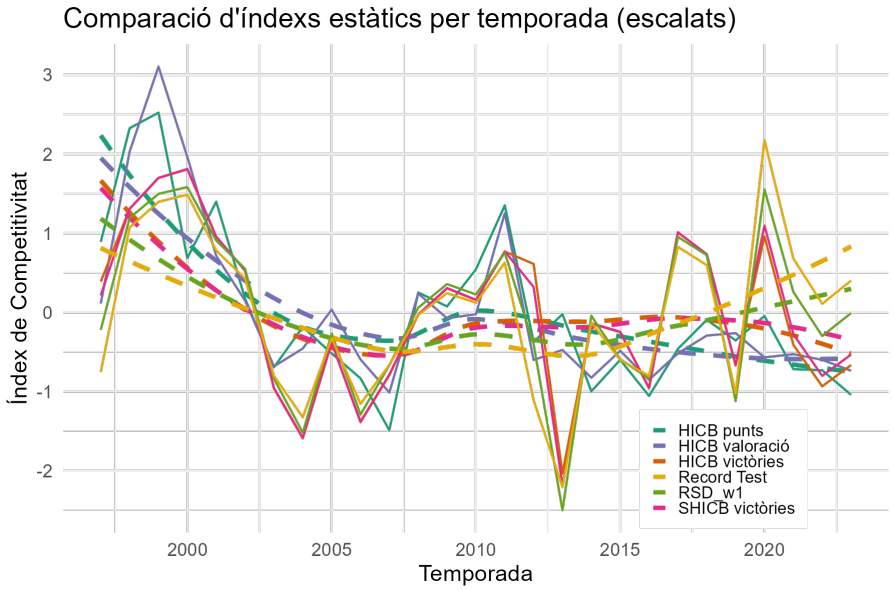

Figura 6.14: Estàtics per temporada escalats

Com podem observar en la Figura 6.14, en general, els índexs d'estudi mostren patrons similars en termes de variació del balanç competitiu de la Lliga Endesa al llarg del temps. Els índexs RSD_{w1} i Record Test mostren una clara tendència ascendent en les últimes temporades, que denota una disminució al balanç competitiu; mentre que la resta mostra una tendència més descendent, la qual indica el augment del balanç competitiu en les darreres temporades.

6.3.2 Índexs dinàmics

Les mesures dinàmiques de balanç competitiu s'han de calcular per a un període de temps de $n > 1$ temporades, ja que incorporen mesures de balanç competitiu tant entre temporades (intertemporada) com dins una mateixa temporada (intratemporada) [3].

D'aquesta manera, a diferència dels índexs estàtics, quan es calculen els índexs dinàmics no s'obté un únic valor per temporada, sinó que es calcula l'índex al llarg de $n > 1$ temporades. D'aquesta manera, si per exemple, es considera $n = 2$, s'obté un valor de l'índex per a les temporades 97-98 i 98-99, un altre per a les temporades 98-99 i 99-00, i així fins arribar a l'index de les temporades 21-22 i 22-23. D'aquesta manera, donat que tenim un total de 27 temporades d'estudi; si considerem l'estudi dels índexs cada n temporades, s'obtenen $27 - n + 1$ valors de l'índex, i no 27 sempre com ocorre amb els índexs estàtics.

Cal destacar que, d'igual manera que ocorre amb els índexs estàtics, ambdós índexs **dinàmics** estudiats tenen diferents escales. En aquests cas, els dos índexs dinàmics no es distribueixen de la mateixa manera. Per una banda, per a l'índex RSD_{w2}, menors valors indiquen major balanç competitiu. Per l'altra banda, i de manera contrària als estàtics, majors valors en l'índex CBR indiquen major balanç competitiu, i a mesura que l'índex disminueix, el balanç competitiu disminueix també. Per aquest motiu, comparem ambdós índexs per separat.

Recordem que, per una banda, amb l'índex RSD_{w2} podem veure l'evolució dels percentatges de victòries en funció de la posició dels equips quan acaba la lliga. Per l'altra banda, l'índex CBR estudia la variació de les victòries intraequip entre la variació intratemporada.

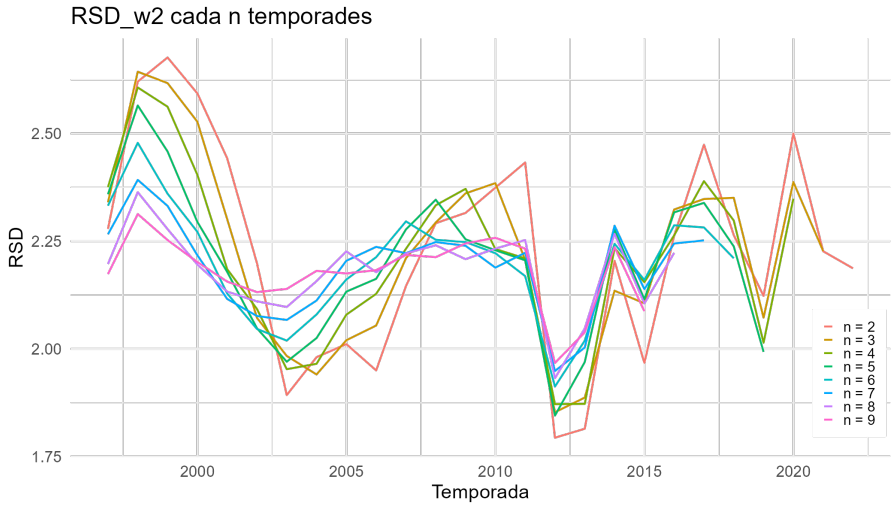

Figura 6.15: Índex de competitivitat dinàmic RSD_{w2} per temporada

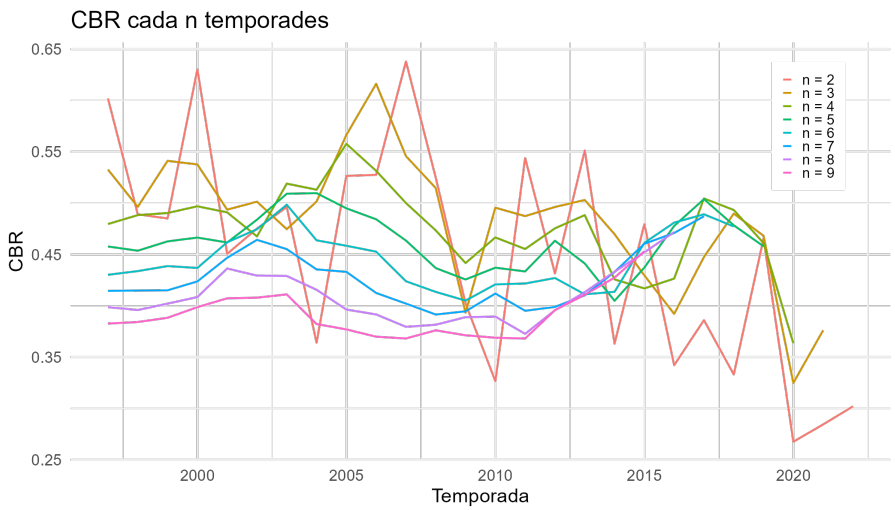

Figura 6.16: Índex de competitivitat dinàmic CBR per temporada

Com podem observar a les Figures 6.15 i 6.16, a mesura que augmenta el tamany n, més es *suavitza* la corba que formen els índexs. Per una banda, la tendència de l'índex RSD_{w2} és amb forma de W, és a dir, es troben tres pics alts en el principi, a meitat, i en les últimes temporades, que correspondrien amb períodes amb menor balanç competitiu, i per tant majors desigualtats entre equips. Per l'altra banda, l'índex CBR obté valors no molt elevats en general, ja que el major dels casos no arriba a 0.65[2]. També es mostra molt inestable per a valors xicotets de n, ja que no podem veure la tendència general fins $n = 4$. No obstant això, es pot interpretar que la tendència general és decreixent, al contrari que l'índex anterior. Aquest fet indica que el balanç competitiu de la Lliga Endesa mai ha sigut molt elevat, i que en les darreres temporades d'estudi ha disminuït més encara.

[2]Cal tindre en compte que aquests valors tan baixos es deuen en part a que es tracta d'una lliga d'ascens-descens, on no tots els equips es mantenen dins de la lliga al llarg dels anys, cosa que disminueix les variacions de victòries intraequip.

6.4 Índexs de competitivitat per jornada

Per tal d'aconseguir el quart dels objectius, es calcularen els índexs de competitivitat **per jornada** de les dades disponibles. Per tant, com tenim dades de la Lliga Femenina Endesa des de la temporada 1997/98 fins la 2023/24, i tenim un nombre de jornades diferent per temporada (podem veure-ho a la Taula 4.1), tindrem un total de 697 valors dels diversos índexs de competitivitat per jornada.

En aquest cas, només es poden calcular els índex $HICB$ per **punts** ($HICB_p$) i per **valoració** ($HICB_{val}$), ja que la resta d'índexs estàtics i dinàmics són mesures de competitivitat basades en com es distribueixen les victòries entre els equips o temporades. En el cas dels índexs per jornada, com el nombre de victòries per equip per jornada serà 0 o 1 en totes les jornades d'estudi, no es poden calcular.

D'igual manera que als índexs per jornada, els $HICB_p$ i $HIBV_{val}$ no tenen la mateixa escala, per tant les compararem en primer lloc per separat. En les Figures 6.17 i 6.18 podem observar els índexs $HICB_p$ i $HIBV_{val}$, respectivament, al llarg de les jornades d'estudi.

En general, podem observar que la seua tendència és molt similar a la que presenten els índexs homòlegs per temporada (Figura 6.12). Com era d'esperar, les dues mesures per jornada coincideixen en què, en general, les primeres temporades d'estudi són els que majors valors tenen per als índexs i per tant, menor balanç competitiu existeix dins de cada jornada.

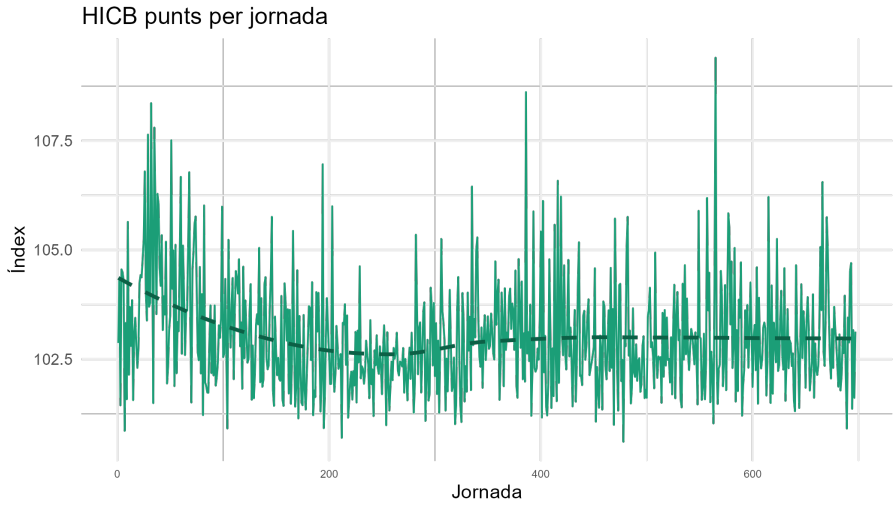

Figura 6.17: Índex de competitivitat estàtic $HICB_{val}$ per jornada

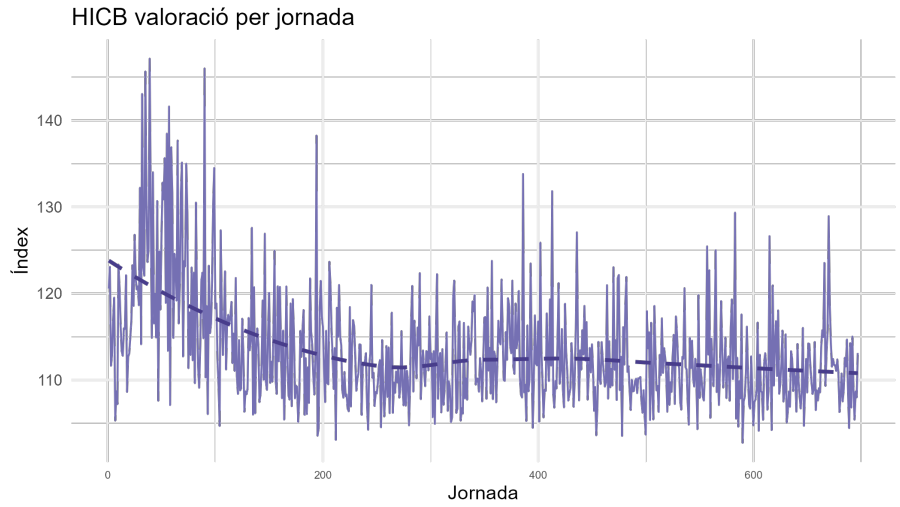

Figura 6.18: Índex de competitivitat estàtic $HICB_p$ per jornada

A continuació, podem observar ambdós índexs junts, per tal de poder veure com, efectivament, tenen diferents escales (Figura 6.19). De nou, podem realitzar la comparació d'aquests índexs d'aquesta manera per què en tots els casos, valors **menors dels índexs indiquen major balanç competitiu**.

Com podem observar a la Figura 6.19, els valors obtinguts de l'índex per punts (blau) són molt més pròxims al valor 100, que es considera el màxim de balanç competitiu, mentre que l'índex per valoració (morat), assoleix valors molt majors, la qual cosa ens recorda més al comportament de l'índex original $HICB_v$, l'index que es basa en les victòries (Figura 6.12). Per tant, a priori, sembla que l'índex $HICB_{val}$ ens va a donar un indicador més fidedigne del balanç competitiu per jornada.

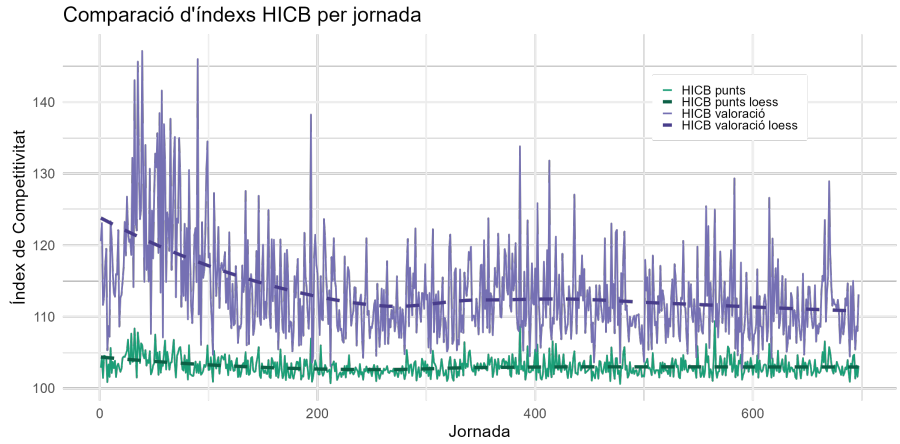

Figura 6.19: Índexs de competitivitat estàtics per jornada

En la Figura 6.20 podem observar els índexs $HICB_p$ i $HICB_{val}$ auto-escalats. Notem que, tot i que a primera vista tenen uns valors escalats molt similars, l'índex $HICB_{val}$ (morat) es superposa a l'índex per punts a les primeres jornades d'estudi, mentre que l'$HICB_p$ (blau), per contra, obté majors valors que l'índex per valoració a les últimes jornades.

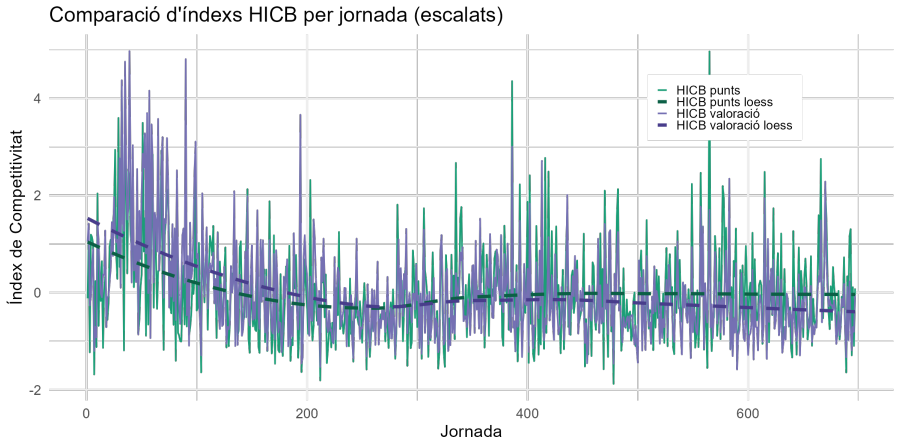

Figura 6.20: Índexs de competitivitat estàtics per jornada escalats

En les seues aproximacions pel mètode *loess* (línies discontínues) podem observar que la tendència general és decreixent. En concret, l'index que té en compte la valoració obté majors valors a les primeres temporades (menor balanç competitiu) i menors a les últimes (major balanç competitiu), en contraposició a l'índex que té en compte els punts. És a dir, que la tendència general decreixent es veu més marcada en l'índex $HICB_{val}$ en comparació amb l'índex $HICB_p$.

Això indica que la disminució en el balanç competitiu per jornada és més notable quan es basa l'índex en la valoració, suggerint que els canvis en el rendiment dels equips reflectits en aquesta valoració podrien ser més significatius que els simples punts anotats. Aquesta discrepància podria suggerir que l'impacte del balanç competitiu en la valoració dels equips és més pronunciat que en els punts totals anotats, fent que **la valoració** siga un millor indicador de les variacions del balanç competitius al llarg de les jornades.

6.5 Models amb els índexs per jornada

Per tal d'assolir el quart dels objectius plantejats en aquest treball, s'aplicaren tant models GLM com multinivell per tal de trobar els factors o variables que afecten als índexs per jornada $HICB_p$ i $HICB_{val}$.

6.5.1 Models lineal amb les variables originals

En primer lloc es realitzà un model amb les variables originals. Per tal d'aconseguir-ho, es calcularen les mitjanes i desviació típiques de les estadístiques de joc per jornada. A més, es va afegir una variable de tipus categòrica, *equip_top1*, que incloïa el nom de l'equip que estava en primera posició en cadascuna de les jornades. Cal destacar que en aquesta nova variable, les classes estaven desequilibrades, ja que la majoria dels equips obtinguts havien estat en primera posició molt poques vegades, com podem veure a la Figura 6.21 (esquerra). Per aquest motiu, es varen unir tots aquells equips que havien estat menys de 20 vegades en la primera posició en una nova categoria ALTRES, com podem veure en la Figura 6.21 (dreta).

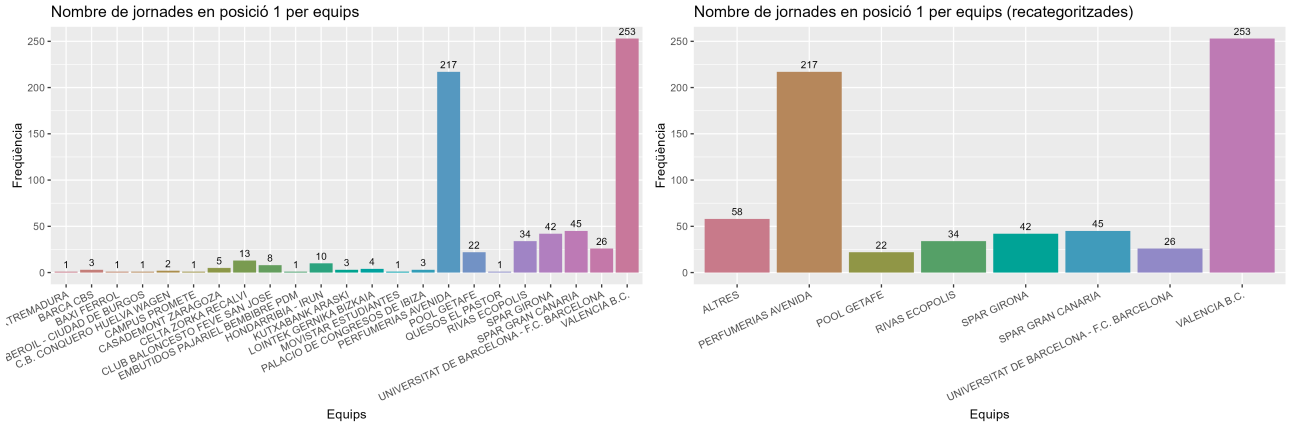

Figura 6.21: Quantitat de jornades que ha estat cada equip en primera posició

Per a la inclusió d'aquesta variable categòrica als models, es considera com a classe de referència l'equip POOL GETAFE. Aquest va ser un equip que va guanyar de manera consecutiva 8 anys (entre les temporades 1991/92-1997/98; de les quals tenim dades només de l'última temporada). És a dir, segons les nostres dades, és un equip que participa únicament una temporada, en la que guanyà.

A continuació trobem els resultats dels models realitzats per a l'índex $HICB_p$ (Taula 6.4) i per a l'índex $HICB_{val}$ (Taula 6.5). Cal destacar que totes les variables numèriques (les variables explicatives i també els índexs de competitivitat) havien sigut escalades per tal de tindre mitjana nul·la i variància unitària.

Taula 6.4: Coeficients del model lineal per a l'$HICB_p$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.064	0.036	-1.780	0.075	.
equip_top1ALTRES	0.113	0.042	2.715	0.007	**
equip_top1PERFUMERIAS AVENIDA	0.061	0.039	1.592	0.112	
equip_top1RIVAS ECOPOLIS	0.070	0.045	1.550	0.122	
equip_top1SPAR GIRONA	0.068	0.044	1.532	0.126	
equip_top1SPAR GRAN CANARIA	0.111	0.042	2.615	0.009	**
equip_top1UNIVERSITAT DE BARCELONA	0.024	0.048	0.509	0.611	
equip_top1VALENCIA B.C.	0.055	0.038	1.469	0.142	
Puntos.mean	-0.259	0.007	-39.007	0.000	***
Puntos.sd	0.967	0.007	135.451	0.000	***
T2 Anotados.sd	0.018	0.008	2.436	0.015	*
% T2.sd	-0.014	0.007	-2.092	0.037	*
Tapones.sd	-0.051	0.013	-3.803	0.000	***
Tapones Recibidos.sd	0.049	0.013	3.680	0.000	***
Faltas Cometidas.sd	0.019	0.007	2.739	0.006	**

Codis de significativitat: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Taula 6.5: Coeficients del model lineal per a l'$HICB_{val}$

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.245	0.054	-4.527	0.000	***
equip_top1ALTRES	0.307	0.059	5.216	0.000	***
equip_top1PERFUMERIAS AVENIDA	0.240	0.060	4.026	0.000	***
equip_top1RIVAS ECOPOLIS	0.206	0.066	3.109	0.002	**
equip_top1SPAR GIRONA	0.227	0.066	3.423	0.001	***
equip_top1SPAR GRAN CANARIA	0.448	0.055	8.104	0.000	***
equip_top1UNIVERSITAT DE BARCELONA	0.202	0.069	2.944	0.003	**
equip_top1VALENCIA B.C.	0.232	0.056	4.113	0.000	***
Puntos.mean	-0.238	0.077	-3.080	0.002	**
T2 Lanzados.mean	0.155	0.048	3.243	0.001	**
% T2.mean	0.147	0.045	3.289	0.001	**
T3 Anotados.mean	0.415	0.076	5.475	0.000	***
T3 Lanzados.mean	-0.168	0.041	-4.127	0.000	***
% T3.mean	-0.115	0.023	-4.903	0.000	***
T1 Lanzados.mean	0.127	0.044	2.872	0.004	**
% T1.mean	0.031	0.014	2.226	0.026	*
Faltas Cometidas.mean	0.130	0.023	5.572	0.000	***
Faltas Recibidas.mean	-0.138	0.016	-8.597	0.000	***
Valoracion.mean	-0.342	0.020	-16.881	0.000	***
T2 Anotados.sd	0.046	0.011	4.069	0.000	***
% T2.sd	-0.027	0.010	-2.761	0.006	**
T3 Anotados.sd	-0.028	0.010	-2.734	0.006	**
Tapones.sd	-0.033	0.008	-4.053	0.000	***
Valoracion.sd	0.833	0.009	89.296	0.000	***

Codis de significativitat: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

En primer lloc, per al **model de l'índex per punts** (Taula 6.4), podem observar que el coeficient del terme independent o intercepte no resulta significatiu, així com alguns dels equips si es troben en primera posició. En canvi, els equips ALTRES i SPAR GRAN CANARIA tenen coeficients positius que resulten significatius en comparació a l'equip POOL GETAFE. Per tant, quan es troben els equips ALTRES i SPAR GRAN CANARIA a la primera posició, el balanç competitiu de la jornada és significativament major que si és el POOL GETAFE qui lidera la jornada.

En quant a com afecten la resta d'estadístiques de joc al balanç competitiu per punts, per una banda, les variables amb coeficient positiu són les que afecten negativament al balanç (és a dir, que si augmenten, disminueix el balanç). Aquestes variables són les desviacions típiques dels Punts, T2 Anotats, Tapons Rebuts i Faltes comeses. Per l'altra banda, les variables amb coeficient negatiu afecten de manera positiva al balanç (és a dir, si augmenten, augmenta també el balanç). Aquestes variables són la mitjana dels Punts i les desviacions típiques de % T2 i dels Tapons.

En segon lloc, pel que fa al **model de l'índex per valoració** (Taula 6.5), com aquesta variable és més complexa que els punts, el model resultant també ho és, ja que són moltes les variables que resulten significatives. És important destacar que en aquest cas l'intercepte té coeficient negatiu i sí que resulta significatiu; i que tots els nivells (equips) de la variable categòrica equip_top1 resulten significatius (i amb coeficients positius) al comparar-los amb l'equip POOL GETAFE. En quant a la resta de coeficients del model, de nou les variables que tenen coeficient positiu augmenten l'índex d'estudi, i per tant, disminueixen el balanç competitiu i viceversa.

6.5.2 Models multinivell amb les variables latents (PCA i *clustering*)

En aquesta secció, es realitzà un model multinivell (5.5.3) considerant 2 nivells: temporada i jornada. Les variables explicatives no són les originals sinó que incloem les mitjanes per jornada de les 4 variables latents obtingudes mitjançant l'anàlisi PCA (*scores*); i els percentatges d'equips que pertanyen a cada clúster en cadascuna de les jornades (només s'inclouen 3 dels 4 clústers ja que al ser proporcions s'han de tractar com a variables *dummies*). S'ha realitzat el model en tres passos; primer, incloem només *scores* de la PCA, després els clústers i finalment, ambdós. Trobem els resultats obtinguts per a l'índex $HICB_p$ en la Taula 6.6, i per a l'índex $HICB_{val}$ en la Taula 6.7.

En quant als **efectes fixos** (aquells que són independents dels nivells temporada i jornada), comparant la Taula 6.6 i la Taula 6.7 podem observar que, en el **Model 1**, els *scores* del primer component són significatius per a ambdós índexs $HICB$ per punts i per valoració. En canvi, els del tercer component només resulten significatius (a un nivell de significativitat del 5%) per a l'índex basat en la valoració. És a dir, tenint el compte el signe dels coeficients; per obtindre

una jornada amb menors índexs $HICB_p$ i $HICB_{val}$ i per tant un major balanç competitiu; s'han d'obtindre menors valors als *scores* del primer component, i majors valors a la del tercer (podem veure les variables que més contribueixen a cada component a la Taula 6.1). Així, tenint en compte el signe negatiu dels *loadings* de les variables que més contribueixen (Figura 6.5), majors valors a les mitjanes de les estadístiques que indiquen victòria (Victòria, Punts, Diferència i Valoració) suposen menors als *scores 1* i per tant un major balanç competitiu. De manera contrària, tenint en compte els coeficients dels *loadings* de les variables que més contribueixen al tercer component (Figura 6.6), valors grans de les mitjanes de T3 Anotats i T1 Anotats, a més de valors xicotets a les mitjanes de T2 Llançats, Rebots i Rebots Ofensius, suposen majors valors als *scores 3* i per tant un major balanç competitiu.

Taula 6.6: Coeficients dels models multinivell per a l'$HICB_p$

	Model 1				**Model 2**				**Model 3**			
Variables	**Estimate**	**Std. Error**	**p value**		**Estimate**	**Std. Error**	**p value**		**Estimate**	**Std. Error**	**p value**	
(intercept)	0.001	0.072	0.986		0.000	0.077	0.999		0.001	0.073	0.990	
scores 1	0.158	0.043	0.000	***					0.104	0.052	0.046	*
scores 2	-0.026	0.043	0.625						-0.149	0.077	0.055	.
scores 3	-0.086	0.054	0.062	.					0.088	0.050	0.081	.
scores 4	0.038	0.050	0.444						-0.104	0.052	0.048	*
clúster 1					-0.250	0.064	0.000	***	-0.412	0.083	0.000	***
clúster 3					0.095	0.055	0.086	.	0.013	0.069	0.851	
clúster 4					-0.234	0.063	0.000	***	-0.307	0.074	0.000	***
	Variance	**Std. Error**	**chisq**		**Variance**	**Std. Error**	**chisq**		**Variance**	**Std. Error**	**chisq**	
Jornada	0.017	0.131	0.124		0.005	0.073	0.560		0.009	0.097	0.336	
Temporada	0.089	0.299	0.000	***	0.122	0.350	0.000	***	0.106	0.325	0.000	***
AIC	1921.8				1871.8				1867.6			

Codis de significativitat: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Taula 6.7: Coeficients dels models multinivell per a l'$HICB_{val}$

	Model 1				**Model 2**				**Model 3**			
Variables	**Estimate**	**Std. Error**	**p value**		**Estimate**	**Std. Error**	**p value**		**Estimate**	**Std. Error**	**p value**	
(intercept)	0.000	0.099	0.999		0.001	0.104	0.993		-0.001	0.101	0.994	
scores 1	0.229	0.039	0.000	***					0.181	0.045	0.000	***
scores 2	-0.051	0.054	0.348						-0.103	0.069	0.133	
scores 3	-0.141	0.042	0.001	***					0.052	0.044	0.242	
scores 4	0.035	0.048	0.462						-0.130	0.048	0.007	**
clúster 1					-0.231	0.056	0.000	***	-0.387	0.071	0.000	***
clúster 3					0.157	0.048	0.001	**	0.014	0.058	0.810	
clúster 4					-0.271	0.055	0.000	***	-0.385	0.063	0.000	***
	Variance	**Std. Error**	**chisq**		**Variance**	**Std. Error**	**chisq**		**Variance**	**Std. Error**	**chisq**	
Jornada	0.001	0.038	0.843		0.000	0.000	1.000		0.000	0.000	1.000	
Temporada	0.239	0.489	0.000	***	0.270	0.519	<2e-16	***	0.253	0.503	<2e-16	***
AIC	1750.5				1675.6				1658.5			

Codis de significativitat: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

El **Model 2** inclou els percentatges d'equips que pertanyen a cada clúster en cada jornada. Observem a les taules 6.6 i 6.7 que els clústers 1 i 4 són significatius per a ambdós models, mentre que el clúster 3 només ho és per model de l'índex $HICB_{val}$. Per tant, tenint en compte el signe dels coeficients, com més partits pertanyen al clúster 1 (joc predominant: interior) i al 4 (joc predominant: exterior); i com menys partits al clúster 3 (perden de molts punts), s'obté un menor índex per jornada i per tant un major balanç competitiu. Aleshores, com els clústers 1 i 4 discriminen en quant a tipus de joc predominant (interior o exterior); i els clústers 2 i 3 discriminen en funció de guanyar o perdre el partit; els resultats indiquen que una jornada on és més important el tipus de joc que si guanyen o no està més balancejada.

En el **Model 3**, es mostren com interactuen les variables de la PCA amb les del *clustering*. Com podem veure, els clústers que romanen significatius són el clúster 1 i el 4, mentre que en els scores; tenint en compte que els coeficients dels *scores* 3 i 4 canvien de signe del Model 1 al Model 3 per a ambdós models; trobem que la incorporació dels clústers competeix amb la significativitat dels *scores* del component 3, la qual cosa ens porta a que siguen els *scores* del component 4 els que siguen significatius. D'aquesta manera, majors valors als *scores* del quart component suposen uns menors índexs de competitivitat i per tant un major balanç competitiu. Tenint en compte el signe dels *loadings* de les variables que més contribueixen a aquest component (Figura 6.6) majors valors a les mitjanes de %T2 i T2 Anotats, i menors valors a les mitjanes de T3 Llançats, Rebots, Rebots Ofensius i Rebots Defensius suposen un major balanç competitiu.

Per tal d'avaluar quin és el millor model per a cadascun dels índexs tenim el **AIC**[3]. Per a cadascun dels índexs, realitzem una comparació entre els tres models mitjançant una ANOVA; i en ambdós casos obtenim que el **Model 3** és significativament millor que els altres dos. És a dir, que les variables que afecten de manera significativa als índexs, i per tant al balanç competitiu, són els components principals 1 i 4, i els clústers 1 i 4.

Finalment, podem observar que dels **efectes aleatoris** d'estudi (Jornada i Temporada), per a ambdós models només és significatiu l'efecte **Temporada**. És a dir, que els coeficients del model multinivell a nivell de Temporada van variant en funció d'aquesta; tant en el model de l'índex de punts com en de valoració. A continuació trobem com canvien aquests coeficients per temporada per als tres models d'estudi per a l'índex $HICB_p$ (Figura 6.22) i per a l'índex $HICB_{val}$ (Figura 6.23).

[3]El criteri d'informació d'Akaike (AIC) és una mesura de la qualitat relativa d'un model estadístic, per a un conjunt donat de dades. Donat un conjunt de models candidats per a les dades, el model preferit és el que té el valor **mínim** a l'AIC.

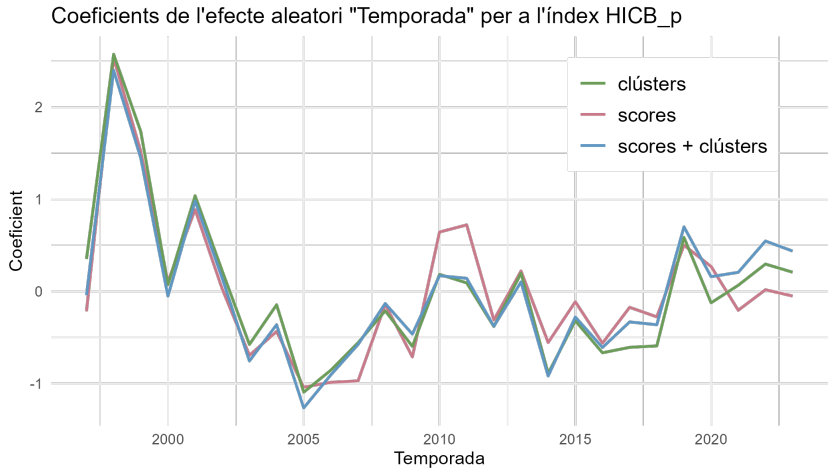

Figura 6.22: Coeficients per temporada del model multinivell per a l'índex $HICB_p$

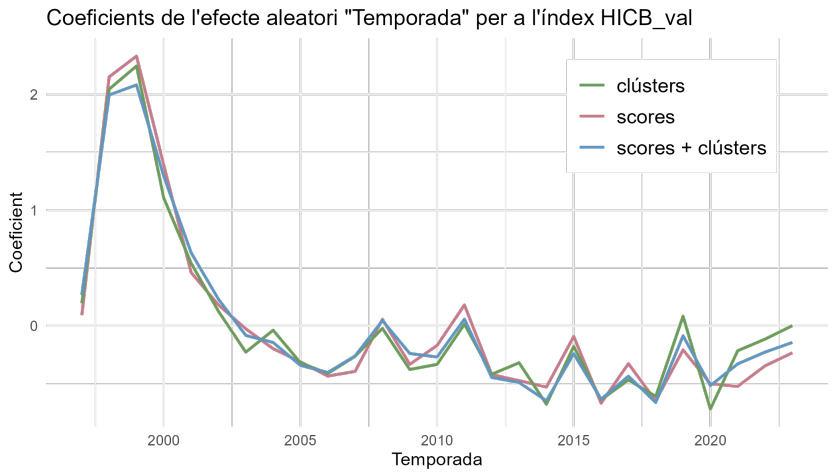

Figura 6.23: Coeficients per temporada del model multinivell per a l'índex $HICB_{val}$

Després d'observar detingudament aquestes gràfiques, podem notar com els coeficients dels models multinivell tenen valors molts similars als índexs $HICB_p$ i $HICB_{val}$ calculats per temporada, els quals podem veure a la Figura 6.12. De nou obtenim el mateix resultat que en els índexs per temporada; les primeres temporades d'estudi (1997-2002) tenen coeficients positius, per tant majors valors als índex i un menor balanç competitiu. En les temporades del mig (2003-2018), s'obtenen coeficients negatius als models multinivell; i per tant són temporades amb un major balanç competitiu. No obstant això, per al últims anys de l'estudi (2019-2023), els índexs per punts i per valoració no concorden. L'índex $HICB_p$ obté coeficients positius i per tant menor balanç competitiu; però l'índex $HICB_{val}$ indica coeficients negatius i per tant que l'alt balanç competitiu es manté.

Capítol 7

Conclusions

Fins on es coneix, aquest és el primer estudi d'aquest tipus que examina el balanç competitiu en una lliga de bàsquet femenina al llarg de diverses temporades, i per tant aquest treball té una clara aportació al món de l'esport i de l'economia de l'esport. El balanç competitiu és un component necessari per avaluar la viabilitat d'una lliga de qualsevol esport, la qual té impacte directe en les fonts d'ingressos de la mateixa.

Mitjançant l'aplicació d'eines de *machine learning* com la PCA i el *clustering*, s'ha realitzat una anàlisi exploratòria exhaustiva i s'han identificat patrons clars que expliquen les dinàmiques competitives de la lliga, tant en termes de rendiment dels equips com del tipus de joc predominant durant la lliga. A més, s'han trobat 4 clústers diferents, i la quantitat de partits que pertanyen a cada clúster ha sigut un dels factors clau per a la caracterització dels índexs de competitivitat.

En aquest sentit, els **índexs de competitivitat**, calculats per **temporada i jornada**, han permés mesurar amb precisió el grau d'equilibri entre els equips, oferint una anàlisi detallada dels factors que contribueixen a aquesta competitivitat. Per estudiar aquest balanç competitiu, s'han examinat diverses de les mesures de balanç competitiu més habituals. No s'han explorat tots aquells mètodes que existeixen, sinó que s'ha centrat en els que tenen major pes a la literatura estudiada.

En l'anàlisi dels índexs per **temporada** s'ha pogut observar com el **balanç competitiu** de la Lliga Femenina Endesa ha experimentat canvis significatius al llarg de les temporades d'estudi. Els resultats obtinguts confirmen que, en general, i de manera contrària al futbol a nivell nacional i europeu, **el balanç competitiu de la Lliga Endesa ha augmentat en els últims anys**. En concret, s'ha vist que els primers anys d'estudi (1997-2000) tenien un menor balanç competitiu, i moltes desigualtats entre els equips de la lliga. En els anys següents (2001-2010) s'ha observat un augment del balanç competitiu a la lliga. No obstant això, pel fet d'estudiar tants índexs de competitivitat diferents, no s'ha arribat a un consens en quant als últims anys

d'estudi (2011-2023). Els índexs que tenen en compte els percentatges de **victòries** han obtés que el balanç està disminuint i s'està arribant a valors de desigualtats tals com a les primeres temporades; mentre que els índexs que estan basats en estadístiques de joc com els **punts** o la **valoració** indiquen que el balanç competitiu està augmentant. És a dir, que en els últims anys, tot i que les estadístiques dels equips s'estan igualant entre ells, continua havent algunes desigualtats respecte als percentatges de victòries.

Pel que fa als índexs per **jornada**, s'ha obtés que el balanç competitiu està augmentant any rere any en les últimes temporades d'estudi. Només s'han pogut estudiar els índexs per **punts** i per **valoració** per la quantitat limitada que hi ha d'índexs de competitivitat per al bàsquet. Pel que fa als models realitzats, s'ha observat que no hi ha un efecte significatiu a nivell de jornada en els models multinivell. És a dir, que la competitivitat d'una certa jornada pot variar entre temporades i, al mateix temps, ser diferent respecte a altres jornades dins d'una mateixa temporada. A més, s'ha vist que l'efecte a nivell de temporada sí és significatiu i coincideix amb els índexs estudiats per temporada. S'ha observat que dels factors analitzats, els que més afecten al balanç competitiu són els *scores 1* i *scores 4*; i els clústers 1 i 4. També s'ha conclòs que l'índex per jornada basat en la valoració és una representació més fidedigna del balanç competitiu de la lliga.

Els obstacles trobats per fer aquest treball (estadístiques de pocs anys i difícils de trobar; la manca de literatura sobre el balanç competitiu en bàsquet,...) mostren una vegada més que la Lliga Femenina Endesa encara s'enfronta a reptes en termes de visibilitat mediàtica i inversió econòmica. Tanmateix, la millora en el balanç competitiu suggereix que, amb les polítiques i estratègies adequades, és possible continuar fomentant la igualtat i l'equitat en el bàsquet femení espanyol, assegurant així la viabilitat econòmica de la lliga.

Aquest treball proporciona una base sòlida per a futures investigacions en aquest camp, especialment en l'anàlisi de dades en la competitivitat esportiva. A més, obri noves vies per explorar el paper de les polítiques de redistribució econòmica i l'organització interna de la lliga en la promoció d'una competició més equilibrada. Polítiques de redistribució com ara el sistema de repartiment d'ingressos, els límits salarials i les restriccions a la mobilitat dels jugadors; polítiques que ja estan en marxa en altres lligues com l'NBA o la WNBA.

Bibliografia

[1] S. Késenne. Cómo puede mejorarse el balance competitivo. *Papeles de Economía Española, Deporte y Economía*, 159:32–42, 2019.

[2] K. Alwell. *Analyzing competitive balance in professional sport*. PhD thesis, University of Connecticut - Storrs, 2020.

[3] B. R. Humphreys. Una guía práctica para medir el balance competitivo. *Papeles de Economia Española*, (159):43–60, 2019.

[4] J. García Villar and P. Rodríguez Guerrero. El balance competitivo en el fútbol español.

[5] CIES Football Observatory Monthly Report. n°40 - December 2018. `https://football-observatory.com/Competitive-balance-a-spatio-temporal-comparison` (último acceso: 13 de septiembre de 2024).

[6] Liga Femenina Endesa - Federación Española de Baloncesto. `https://www.lfendesa.es/inicio.aspx` (último acceso: 13 de septiembre de 2024).

[7] A. Arbues. *BueStats*. `https://www.upf.edu/web/adria-arbues/buestats` (último acceso: 13 de septiembre de 2024).

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[9] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.

[10] A. Kassambara and F. Mundt. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, 2020. R package version 1.0.7.

[11] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[12] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2023. R package version 2.1.6.

[13] S. Mondal. She kicks: The state of competitive balance in the top five women's football leagues in europe. *J. Glob. Sport Manag.*, 8(1):432–454, 2023.

[14] S. Szymanski and R. Smith. Equality of opportunity and equality of outcome: Static and dynamic competitive balance in european and north american sports leagues, 2002.

[15] L. Van Scyoc and K. McGee. Testing for competitive balance. *Empirical Economics*, 50:1029–1043, 2016.

[16] C. A. Depken. Free-agency and the competitiveness of major league baseball. *Review of Industrial Organization*, 14(3):205–217, 1999.

[17] P. D. Owen, M. Ryan, and C. R. Weatherston. Measuring competitive balance in professional team sports using the herfindahl-hirschman index. *Review of Industrial Organization*, 31:289–302, 2007.

[18] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *J. R. Stat. Soc. Ser. A*, 135(3):370, 1972.

[19] J. J. Faraway. *Extending the linear model with R.* Chapman and Hall/CRC, March 2016.

[20] F. De la Cruz. Modelos multinivel. *Revista peruana de epidemiología*, 12(3):1–8, 2008.

[21] D. Bates, M. Mächler, B. Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[22] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[23] T. Wei and V. Simko. *R package 'corrplot': Visualization of a Correlation Matrix*, 2021. (Version 0.92).

[24] M. Á. G. Ruano, A. L. Calvo, E. O. Toro, J. E. Sampaio, and S. J. I. Godoy. Diferencias en las estadísticas de juego entre bases, aleros y pívots en baloncesto femenino. *Cultura, Ciencia y Deporte*, 2(6):139–144, 2007.

[25] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning.* Springer series in statistics. Springer, New York, NY, 2 edition, 2009.

Capítol 8

Annexos

Annex 1. Recategorització dels noms dels equips

Taula 8.1: Patrons de recategorització dels noms dels equips (1 de 3).

Nombre	Nombre2	Provincia
A.D. CORTEGADA	EXTRUGASA	PONTEVEDRA
A.D. UNIVERSIDAD DE OVIEDO	UNIVERSIDAD DE OVIEDO	ASTURIAS
A.D.UNIVERSIDAD DE OVIEDO	UNIVERSIDAD DE OVIEDO	ASTURIAS
ACIS-SUFI LEON	CLUB BALONCESTO FEVE SAN JOSE	LEON
ACIS INCOSA LEON	CLUB BALONCESTO FEVE SAN JOSE	LEON
ALIMENTOS DE ZAMORA	QUESOS EL PASTOR	ZAMORA
ALTER ENERSUN AL-QAZERES EXTREMADURA	ALTER ENERSUN AL-QAZERES EXTREMADURA	CACERES
ALUMINIOS TIZONA	BEROIL - CIUDAD DE BURGOS	BURGOS
ANARES RIOJA ISB	ANARES RIOJA ISB	GUIPUZCOA
ANDALUCIA - AIFOS	ANDALUCIA AIFOS	JAEN
ANDALUCIA AIFOS	ANDALUCIA AIFOS	JAEN
ANDALUCIA SOLO HAY UNA	ANDALUCIA AIFOS	JAEN
ARGON UNI GIRONA	SPAR GIRONA	GERONA
ARRANZ-JOPISA BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
ARRANZ - JOPISA BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
ARRANZ ACINAS	BEROIL - CIUDAD DE BURGOS	BURGOS
ASEFA ESTUDIANTES	MOVISTAR ESTUDIANTES	MADRID
B. CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
BALONCESTO ALCALA	BALONCESTO ALCALA	MADRID
BALONCESTO CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
BANCO SIMEON	CELTA ZORKA RECALVI	PONTEVEDRA
BARCA CBS	BARCA CBS	BARCELONA
BARCELONA BC UNIVERSITARI	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
BARCELONA BC.UNIVERSITARI	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
BASKET ZARAGOZA	CASADEMONT ZARAGOZA	ZARAGOZA
BAXI FERROL	BAXI FERROL	A CORUNA
BEROIL - CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
BF HOSPITALET	BF HOSPITALET	BARCELONA
BIZKAIA GDKO	BIZKAIA GDKO	VIZCAYA
C.B. AL-QAZERES EXTREMADURA	ALTER ENERSUN AL-QAZERES EXTREMADURA	CACERES
C.B. BEMBIBRE PDM	EMBUTIDOS PAJARIEL BEMBIBRE PDM	LEON
C.B. CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
C.B. CONQUERO HUELVA WAGEN	C.B. CONQUERO HUELVA WAGEN	HUELVA
C.B. OLESA	C.B. OLESA	BARCELONA
C.B. OLESA - ESPANYOL	C.B. OLESA	BARCELONA
C.B. PUIG D'EN VALLS SANTA EULALIA	PALACIO DE CONGRESOS DE IBIZA	ISLAS BALEARES
C.B. STA. ROSA DE LIMA	CB.STA.ROSA DE LIMA HORTA	BARCELONA
C.B.N.	CBN	NAVARRA
C.D. ENSINO	DURAN MAQUINARIA ENSINO	LUGO
C.R.E.F. HOLA!	C.R.E.F. HOLA!	MADRID

Taula 8.2: Patrons de recategorització dels noms dels equips (2 de 3).

Nombre	Nombre2	Provincia
CADI - ICG SOFTWARE	CADI LA SEU	LERIDA
CADI LA SEU	CADI LA SEU	LERIDA
CADI LA SEU D URGELL	CADI LA SEU	LERIDA
CAJA RURAL DE CANARIAS	SPAR GRAN CANARIA	LAS PALMAS
CAJA RURAL TINTOS DE TORO	QUESOS EL PASTOR	ZAMORA
CAJACANARIAS	SPAR GRAN CANARIA	LAS PALMAS
CAJASUR LINARES	ANDALUCIA AIFOS	JAEN
CAMPUS PROMETE	CAMPUS PROMETE	LA RIOJA
CAMPUS PROMETE LF1	CAMPUS PROMETE	LA RIOJA
CANAL ISABEL II	CANAL ISABEL II	MADRID
CASADEMONT ZARAGOZA	CASADEMONT ZARAGOZA	ZARAGOZA
CB.STA.ROSA DE LIMA HORTA	CB.STA.ROSA DE LIMA HORTA	BARCELONA
CBN	CBN	NAVARRA
CD ENSINO	DURAN MAQUINARIA ENSINO	LUGO
CD ENSINO UNIVERSIDADE	DURAN MAQUINARIA ENSINO	LUGO
CE UNIVERSITARI	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
CELTA BANCO SIMEON	CELTA ZORKA RECALVI	PONTEVEDRA
CELTA ZORKA RECALVI	CELTA ZORKA RECALVI	PONTEVEDRA
CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
CIUDAD DE LA LAGUNA TENERIFE	TENERIFE	SANTA CRUZ DE TENERIFE
CIUDAD ROS CASARES	VALENCIA B.C.	VALENCIA
CIUDAD ROS CASARES VALENCIA	VALENCIA B.C.	VALENCIA
CLUB BALONCESTO CONQUERO	C.B. CONQUERO HUELVA WAGEN	HUELVA
CLUB BALONCESTO FEVE SAN JOSE	CLUB BALONCESTO FEVE SAN JOSE	LEON
CLUB BALONCESTO SAN JOSE	CLUB BALONCESTO FEVE SAN JOSE	LEON
CORTEGADA-GRUPO 10	EXTRUGASA	PONTEVEDRA
CORTEGADA GRUPO 10	EXTRUGASA	PONTEVEDRA
DURAN MAQUINARIA ENSINO	DURAN MAQUINARIA ENSINO	LUGO
EBE IBIZA-PDV	PALACIO DE CONGRESOS DE IBIZA	ISLAS BALEARES
EBE PROMOCIONES-PDV STA.EULALIA	PALACIO DE CONGRESOS DE IBIZA	ISLAS BALEARES
EBE PROMOCIONES SANTA EULALIA	PALACIO DE CONGRESOS DE IBIZA	ISLAS BALEARES
EMBUTIDOS PAJARIEL BEMBIBRE PDM	EMBUTIDOS PAJARIEL BEMBIBRE PDM	LEON
ENSINO	DURAN MAQUINARIA ENSINO	LUGO
ENSINO YAYA MARIA	DURAN MAQUINARIA ENSINO	LUGO
EXTREMADURA DATO	EXTREMADURA DATO	BADAJOZ
EXTRUGASA	EXTRUGASA	PONTEVEDRA
EXTRUGASA COCINAS CARBALLO	EXTRUGASA	PONTEVEDRA
EXTRUGASA MUEBLES CARBALLO	EXTRUGASA	PONTEVEDRA
FEMENINO TRES CANTOS	FEMENINO TRES CANTOS	MADRID
FILTROS MANN ZARAGOZA	CASADEMONT ZARAGOZA	ZARAGOZA
GERNIKA BIZKAIA	LOINTEK GERNIKA BIZKAIA	VIZCAYA
GIPUZKOA UPV	IDK EUSKOTREN	GUIPUZCOA
GIRONA FC	SPAR GIRONA	GERONA
GRAN CANARIA	SPAR GRAN CANARIA	LAS PALMAS
GRAN CANARIA 2014	SPAR GRAN CANARIA	LAS PALMAS
GRAN CANARIA 2014 LA CAJA DE CANARIAS	SPAR GRAN CANARIA	LAS PALMAS
GRAN CANARIA LA CAJA DE CANARIAS	SPAR GRAN CANARIA	LAS PALMAS
HALCON VIAJES	PERFUMERIAS AVENIDA	SALAMANCA
HONDARRIBIA - IRUN	HONDARRIBIA - IRUN	GUIPUZCOA
HOZONO GLOBAL JAIRIS	HOZONO GLOBAL JAIRIS	MURCIA
IDK EUSKOTREN	IDK EUSKOTREN	GUIPUZCOA
IDK GIPUZKOA	IDK EUSKOTREN	GUIPUZCOA
IDK GIPUZKOA UPV	IDK EUSKOTREN	GUIPUZCOA
INNOVA-TSN LEGANES	INNOVA-TSN LEGANES	MADRID
JOPISA CIUDAD DE BURGOS	BEROIL - CIUDAD DE BURGOS	BURGOS
KUTXABANK ARASKI	KUTXABANK ARASKI	ALAVA
LACTURALE ARASKI	KUTXABANK ARASKI	ALAVA
LACTURALE ART ARASKI	KUTXABANK ARASKI	ALAVA
LOINTEK GERNIKA BIZKAIA	LOINTEK GERNIKA BIZKAIA	VIZCAYA
MANN-FILTER	CASADEMONT ZARAGOZA	ZARAGOZA
MANN-FILTER CASABLANCA	CASADEMONT ZARAGOZA	ZARAGOZA
MANN FILTER ZARAGOZA	CASADEMONT ZARAGOZA	ZARAGOZA
MOTIVA REAL CANOE N.C.	REAL CANOE N.C.	MADRID
MOVISTAR ESTUDIANTES	MOVISTAR ESTUDIANTES	MADRID
NISSAN AL-QAZERES EXTREMADURA	ALTER ENERSUN AL-QAZERES EXTREMADURA	CACERES
P.C. MENDIBIL	P.C. MENDIBIL	ALAVA
PALACIO DE CONGRESOS DE IBIZA	PALACIO DE CONGRESOS DE IBIZA	ISLAS BALEARES

Taula 8.3: Patrons de recategorització dels noms dels equips (3 de 3).

Nombre	Nombre2	Provincia
PERFUMERIAS AVENIDA	PERFUMERIAS AVENIDA	SALAMANCA
POOL GETAFE	POOL GETAFE	MADRID
POPULAR BASQUET GODELLA	VALENCIA B.C.	VALENCIA
QUESOS EL PASTOR	QUESOS EL PASTOR	ZAMORA
QUESOS EL PASTOR DE LA POLVOROSA	QUESOS EL PASTOR	ZAMORA
R.CLUB CELTA BANCO SIMEON	CELTA ZORKA RECALVI	PONTEVEDRA
RC CELTA BALONCESTO	CELTA ZORKA RECALVI	PONTEVEDRA
RC CELTA INDEPO	CELTA ZORKA RECALVI	PONTEVEDRA
REAL CANOE N.C.	REAL CANOE N.C.	MADRID
REAL CLUB CELTA BANCO SIMEON	CELTA ZORKA RECALVI	PONTEVEDRA
REAL CLUB CELTA INDEPO	CELTA ZORKA RECALVI	PONTEVEDRA
REAL CLUB CELTA VIGOURBAN	CELTA ZORKA RECALVI	PONTEVEDRA
RIVAS ECOPOLIS	RIVAS ECOPOLIS	MADRID
RIVAS FUTURA	RIVAS ECOPOLIS	MADRID
ROS CASARES	VALENCIA B.C.	VALENCIA
ROS CASARES VALENCIA	VALENCIA B.C.	VALENCIA
RPK ARASKI	KUTXABANK ARASKI	ALAVA
SALAMANCA HALCON VIAJES	PERFUMERIAS AVENIDA	SALAMANCA
SANDRA GRAN CANARIA	SPAR GRAN CANARIA	LAS PALMAS
SNATT'S FEMENI SANT ADRIA	SNATT'S FEMENI SANT ADRIA	BARCELONA
SOLLER BON DIA!	SOLLER BON DIA!	ISLAS BALEARES
SOLLER JOVENTUT MARIANA	SOLLER BON DIA!	ISLAS BALEARES
SPAR CITYLIFT GIRONA	SPAR GIRONA	GERONA
SPAR GIRONA	SPAR GIRONA	GERONA
SPAR GRAN CANARIA	SPAR GRAN CANARIA	LAS PALMAS
SPAR UNIGIRONA	SPAR GIRONA	GERONA
STAR CENTER-UNI FERROL	BAXI FERROL	A CORUNA
SYMEL TENERIFE	SYMEL TENERIFE	SANTA CRUZ DE TENERIFE
TENERIFE	TENERIFE	SANTA CRUZ DE TENERIFE
TINTOS DE TORO CAJA RURAL	QUESOS EL PASTOR	ZAMORA
TONY ROMA S REAL CANOE NC	REAL CANOE N.C.	MADRID
TOYOTA RECREATIVO CONQUERO	C.B. CONQUERO HUELVA WAGEN	HUELVA
UNB OBENASA	KUTXABANK ARASKI	ALAVA
UNB OBENASA LACTURALE	KUTXABANK ARASKI	ALAVA
UNIVERSIDAD DE OVIEDO	UNIVERSIDAD DE OVIEDO	ASTURIAS
UNIVERSITARIO DE FERROL	BAXI FERROL	A CORUNA
UNIVERSITAT DE BARCELONA	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
UNIVERSITAT DE BARCELONA - F.C. BARCELONA	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
UNIVERSITAT DE BARCELONA B.F.	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	BARCELONA
USP-CEU MMT ESTUDIANTES	MOVISTAR ESTUDIANTES	MADRID
USP CEU - ADECCO ESTUDIANTES	MOVISTAR ESTUDIANTES	MADRID
VALENCIA B.C.	VALENCIA B.C.	VALENCIA
VETUSTA OVIEDO	UNIVERSIDAD DE OVIEDO	ASTURIAS
YAYA MARIA BREOGAN	DURAN MAQUINARIA ENSINO	LUGO
YAYA MARIA PORTA XI	DURAN MAQUINARIA ENSINO	LUGO
ZAMARAT	QUESOS EL PASTOR	ZAMORA

Annex 2. Outliers moderats obtinguts en la PCA exploratòria

Taula 8.4: Outliers moderats en la PCA exploratòria

Id	Nombre	Jornada	Puntos	Temporada	Posición
180	POPULAR BASQUET GODELLA	4	55	1997	10
309	BARCELONA BC UNIVERSITARI	19	57	1998	8
897	R.CLUB CELTA BANCO SIMEON	9	69	1999	2
1139	C.B.N.	17	78	2000	7
1887	FILTROS MANN ZARAGOZA	11	56	2002	3
2115	C.B. PUIG D'EN VALLS SANTA EULALIA	5	66	2003	9
2411	USP CEU - ADECCO ESTUDIANTES	15	72	2003	11
2772	USP CEU - ADECCO ESTUDIANTES	12	81	2004	5
3185	ACIS INCOSA LEON	9	79	2006	9
3373	HONDARRIBIA - IRUN	15	63	2006	4
3489	UNIVERSITAT DE BARCELONA - F.C. BARCELONA	1	94	2006	10
4173	REAL CLUB CELTA INDEPO	9	72	2008	7
8111	EMBUTIDOS PAJARIEL BEMBIBRE PDM	1	65	2020	13
8647	INNOVA-TSN LEGANES	27	66	2021	13
8804	SPAR GRAN CANARIA	4	57	2021	13

Annex 3. Relació del treball amb els Objectius de Desenvolupament Sostenible de la agenda 2030.

Taula 8.5: Grau de relació del treball amb els Objectius de Desenvolupament Sostenible (ODS).

Objectius de Desenvolupament Sostenible		Alt	Mitjà	Baix	No procedeix
ODS 1.	Fi de la pobresa.				✓
ODS 2.	Fam zero.				✓
ODS 3.	Salut i benestar.		✓		
ODS 4.	Educació de qualitat.				✓
ODS 5.	Igualtat de gènere.	✓			
ODS 6.	Aigua neta i sanejament.				✓
ODS 7.	Energia assequible i no contaminant.				✓
ODS 8.	Treball decent i creixement econòmic.		✓		
ODS 9.	Indústria, innovació i infraestructures.				✓
ODS 10.	Reducció de les desigualtats.			✓	
ODS 11.	Ciutats i comunitats sostenibles.				✓
ODS 12.	Producció i consum responsables.				✓
ODS 13.	Acció pel clima.				✓
ODS 14.	Vida submarina.				✓
ODS 15.	Vida d'ecosistemes terrestres.				✓
ODS 16.	Pau, justícia i institucions sòlides.				✓
ODS 17.	Aliances per assolir objectius.				✓

Descripció de l'alineació del TFM amb els ODS amb un grau de relació més alt.

Aquest treball de final de màster està alineat amb diversos Objectius de Desenvolupament Sostenible (ODS), que formen part de l'Agenda 2030. Aquests són els ODS identificats en el treball amb un major grau de relació:

- **ODS 3:** Salut i Benestar.

  L'esport, incloent el bàsquet, és una activitat clau per promoure hàbits de vida saludables, tant en l'àmbit professional com en l'amateur. Participar en activitats físiques contribueix a previndre malalties cròniques com l'obesitat, malalties cardiovasculars i problemes de salut mental.

- **ODS 5:** Igualtat de gènere.

Aquest objectiu busca aconseguir la igualtat entre gèneres i empoderar totes les dones i xiquetes. En el treball, la Lliga Femenina Endesa es converteix en un exemple clar de la promoció de la igualtat de gènere en l'esport. L'anàlisi de la competitivitat en el bàsquet femení contribueix a visibilitzar l'esport practicat per dones, una àrea sovint desatesa en comparació amb les competicions masculines.

- **ODS 8:** Treball decent i creixement econòmic.

  Es posa de manifest el creixement de la indústria esportiva, incloent el bàsquet femení, i la seua capacitat de generar ocupació i riquesa, així com l'impacte positiu que pot tindre la promoció de condicions laborals justes en l'àmbit esportiu.

- **ODS 10:** Reducció de les desigualtats.

  El treball es focalitza en reduir les desigualtats tant dins com entre països, explorant la manera com l'esport pot esdevindre una eina per a la igualtat d'oportunitats, especialment quan es fomenta la igualtat entre hòmens i dones en disciplines com el bàsquet professional.