TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Qualitat metodològica i report de les tècniques *clustering* en ciències de l'esport: una revisió sistemàtica

Autor: Montse Plensa Santallúsia

Director: Daniel Fernández Martínez i Martí Casals Toquero

Departament: DEIO

Convocatòria: Juliol 2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM I PARAULES CLAU

L'estadística descriptiva i concretament les tècniques clustering són de les més aplicades en les ciències de l'esport. Aquest treball presenta una revisió sistemàtica (RS) d'articles originals que utilitzen les tècniques clústering en l'àmbit de les ciències de l'esport seguint la guia PRISMA. La cerca es va realitzar en diferents bases de dades amb les paraules clau i booleans següents*: "clustering" and "sport"*. Un total de 54 articles van ser seleccionats, on el 85% d'ells es van publicar els últims 7 anys, i el 50% d'aquests estudiaven el rendiment esportiu. Els participants d'aquests estudis són majoritàriament atletes professionals (61,1%), relacionats amb esports com el futbol (*soccer*) (16,7%), bàsquet (7,4%) i tennis (11,1%), respectivament. El 35% dels articles utilitzaven el *clustering* jeràrquic mentre que el mètode més utilitzat va ser el k-means (*clustering* particional). En el 31,3 % dels articles no es va reportar el mètode *clustering* utilitzat, ni el mètode per decidir el nombre de clústers (35,2%), ni en més d'un 70% aspectes computacionals i relacionats amb la reproductibilitat de les dades. A més a més, en aquest treball es posa en pràctica una de les tècniques de *clustering* particional més popular per estudiar el perfil de les jugadores de la lliga professional de bàsquet femenina dels Estats Units des de l'any 1997 fins al 2019.

Paraules clau: *clustering*, *esports*, *revisió sistemàtica*, *rendiment, k-means*

RESUMEN Y PALABRAS CLAVE

La estadística descriptiva y concretamente las técnicas *clustering* son de las más aplicadas en las ciencias del deporte. Este trabajo presenta una revisión sistemática (RS) de artículos originales que utilizan las técnicas *clustering* en el ámbito de las ciencias del deporte siguiendo la guía PRISMA. La búsqueda se realizó en diferentes bases de datos con las palabras clave y booleanos siguientes*: "clustering" and "sport"*. Un total de 54 artículos fueron seleccionados, donde el 85% de ellos se publicaron los últimos 7 años, y el 50% de estos estudiaban el rendimiento deportivo. Los participantes de estos estudios son mayoritariament atletas profesionales (61,1%), relacionados con deportes como el fútbol (*soccer*) (16,7%), baloncesto (7,4%) y tennis (11,1%), respectivamente. El 35% de los artículos utilizaban el *clustering* jerárquico mientras que el método más utilizado fue el *k-means* (*clustering* particional). En el 31,3% de los artículos no se reportó el método *clustering* utilizado, ni el método para decidir el número de clústeres (35.2%), ni en más de un 70% aspectos computacionales y relacionados con la reproductibilidad de los datos. Además, en este trabajo se pone en práctica una de las técnicas de *clustering* particional más popular para estudiar el perfil de las jugadoras de la liga profesional de baloncesto femenina de los Estados Unidos desde el año 1997 hasta el 2019.

Palabras clave*: clustering, deportes, revisión sistemática, rendimiento, k-means*

ABSTRACT AND KEYWORDS

Descriptive statistics and specifically clustering techniques are among the most applied in the sports sciences. This work presents a systematic review (RS) of original articles using cluster techniques in the field of sports sciences following the PRISMA guide. The search was performed in different databases with the following keywords and booleans: *"clustering" and "sport"*. A total of 54 articles were selected, where 85% of them were published over the past seven years, and 50% of these were studying sports performance. The participants of these studies are mostly professional athletes (61,1%), related to sports such as football (soccer) (16,7%), basketball (7,4%) and tennis (11,1%), respectively. 35% of the articles used hierarchical clustering while the most commonly used method was k-means (partitional clustering). In 31,3 percent of the articles, the clustering method used was not reported, nor the method for deciding the number of clusters (35,2%), nor in more than 70% computational and related aspects of data reproducibility. In addition, one of the most popular partitional clustering techniques is put into practice to study the profile of the female professional basketball league players from 1997 to 2019.

Keywords: *clustering*, *sports*, *systematic review*, *performance, k-means*

CLASSIFICACIÓ AMS

62H30 Classification and discrimination; cluster analysis

ÍNDEX DE CONTINGUTS

ÍNDEX DE FIGURES

ÍNDEX DE GRÀFICS

ÍNDEX DE TAULES

1. Introducció

L'estadística esportiva és un camp en la ciència que està causant interès i alhora creixent en els últims anys. Les constants innovacions tecnològiques disponibles, la repercussió d'algunes pel·lícules (ex: *Moneyball* (Miller, 2011), *Concussion* (Landesman, 2015)), l'aparició de podcasts (ex: *Measurables*, *Counterpoints*,..) i la creació de nous departaments de *sports analytics* en universitats (ex: Harvard, Simon Fraser, Carnegie Mellon, California, Berkeley,..) són alguns dels factors que han provocat aquest increment. També, és cada vegada més freqüent trobar publicacions sobre aquest camp en revistes d'estadística reconegudes per l'*American Statitical Association* (ASA) com el *Journal of Quantitative Analysis in Sports*, *Chance*, *The American Statistician* i més conferències obertes com les organitzades per *MathSport International* o les impartides al *Joint Statistical Meeting* (JSM). De fet, en una de les sessions del JSM l'any 1992 es va fundar la *Section on Statistics in Sports* (SIS) de l'ASA per tal de respondre a la necessitat de fomentar el desenvolupament de l'estadística i les seves aplicacions a l'esport (*ASA Community*, s.d.).

La professió de científic de dades o estadístic esportiu té cada vegada més oportunitats (Smyth, 2022). Les habilitats d'aquests professionals es centren principalment en extreure informació útil a través d'un gran volum de dades i en conèixer el patró de combinacions de jugadors, equips, lligues i lesions entre d'altres mitjançant habilitats de *statistical and computational thinking* (Alamar & Oliver, 2013; Miller, 2015) . Les tècniques estadístiques més utilitzades són sovint les de *machine learning* enfocant-se principalment en anàlisi descriptiu multivariant (Musa et al., 2018; Chandran et al., 2019).

Les tècniques d'aprenentatge automàtic o *Machine learning* es classifiquen entre supervisades i no supervisades i la principal diferència entre les dues és que la primera utilitza dades etiquetades per ajudar a predir els resultats mentre que l'altra no. Les més usades entre els professionals de l'estadística esportiva són les no supervisades; que empren algorismes d'aprenentatge automàtic per a detectar patrons entre les observacions de tal forma que les que són similars s'agrupen (Kubat, 2021). Entre les tècniques no supervisades més populars es troben el *Principal Component Analysis* o PCA, que té com principal tasca reduir la dimensionalitat de les dades, i el *clustering*, que tracta de trobar grups, inicialment desconeguts, de perfils similars entre les observacions. El *clustering* té com a finalitat principal aconseguir l'agrupament de conjunts d'objectes no etiquetats i similars entre si per a construir subconjunts de dades coneguts com a clústers (Everitt et al., 2011). Es poden distingir tres tipus de mètodes *clustering*: els de partició, dels que es pot destacar els algoritmes *k-means* (MacQueen, 1967; Lloyd, 1982) i *k-medoids* (Koutroumbas & Theodoridis, 2008), el probabilístic, on ressalta el *model-based clustering* (Fraley & Raftery, 2002; McNicholas, 2016), i el jeràrquic

En la literatura científica de l'àmbit esportiu destaca per sobre de tot el llibre *Handbook of Statistical Methods and Analyses in Sports* (Albert et al., 2017), que ofereix el panorama més actual de la investigació en l'àmbit de l'analítica esportiva. En aquest llibre s'esmenta com s'utilitzen diversos mètodes *clustering* per analitzar algunes característiques d'esports com el beisbol (Albert et al., 2017, p.39) i el bàsquet (Albert et al., 2017, p.245). En el primer s'empren els mètodes *k-means* i *model-based clustering* per identificar els diferents tipus de llançaments

amb diferents velocitats i moviments i si aquests estaran ben colpejats o no; mentre que al bàsquet, els clústers permeten entendre els tipus de jugadors estudiant l'espai per on es mouen a pista per així identificar els emparellaments defensius.

L'objectiu principal d'aquest estudi és realitzar per primera vegada una revisió sistemàtica (RS) de l'aplicació de les tècniques *clustering* i avaluar la qualitat dels resultats i la informació reportada en articles originals en el camp de les ciències de l'esport a través d'una guia coneguda com PRISMA.

2. Metodologia

2.1. *Tipus de Revisions*

En les últimes dècades hi ha hagut un increment molt ràpid de tot tipus d'avenços científics que han motivat l'elaboració d'una gran quantitat de publicacions científiques degut a la necessitat de compartir i actualitzar constantment la informació. Aquesta allau d'informació reportada ha provocat que progressivament hagin cobrat major importància les revisions de la literatura científica (Grant & Booth, 2009). Les revisions permeten sintetitzar de forma lògica i objectiva la informació, justificar la realització d'una investigació, explorar-ne la metodologia i analitzar de forma crítica els resultats del tema d'interès.

Per fer una revisió és important saber quines són les necessitats per les que es planteja realitzar un estudi ja que existeixen diferents tipus de revisions de la literatura científica. S'ha de tenir en compte el tipus de pregunta de la revisió, el tipus i la quantitat d'estudis disponibles i la seva finalitat, entre d'altres factors. Algunes de les diferents classes de revisions de la literatura científica són la *narrative review*, la *umbrella review*, la *scoping review*, la *rapid review* i la *systematic review* (Taula 2.1)

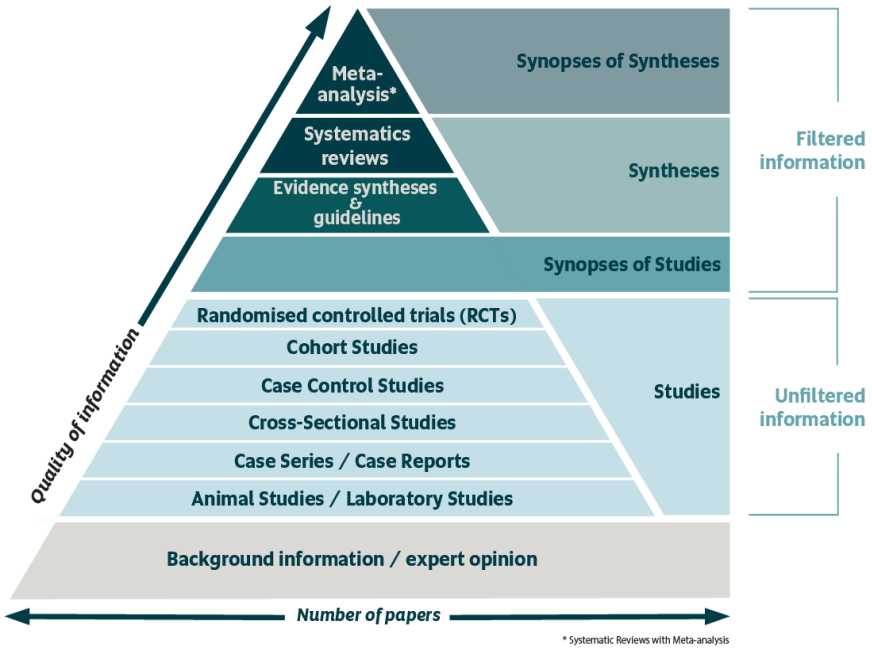
Taula 2.1.: Taula descriptiva de les diferents classes de revisions de la literatura científica mencionades. Font: Grant & Booth, 2009

Tipus	Descripció	Avantatges	Inconvenients
Narrative review	Es tracta de material publicat que proporciona una revisió de la literatura recent o actual	Reuneix el que s'ha aconseguit sense repeticions i identificant omissions	No hi ha un mètode establert que s'asseguri que es consideri tota la literatura d'un tema
Umbrella review	Es refereix específicament a la revisió de l'evidència de múltiples vistes en un document accessible i utilitzable	Compila els resultats de múltiples revisions per a respondre a una pregunta específica. Crea un equilibri entre les opinions generals i les opinions que estan fragmentades a causa de la seva especificitat	Depèn del fet que ja hi hagi una revisió de components més restringida
Scoping review	Té com a objectiu identificar la naturalesa i l'abast de l'evidència de recerca	S'utilitza per a determinar si serà necessària una revisió sistemàtica completa per a arribar a una conclusió	No és un producte financer i corre un major risc de ser esbiaixat
Rapid review	És una avaluació del que ja se sap sobre una qüestió política o pràctica, mitjançant l'ús de mètodes de revisió sistemàtica per a buscar i avaluar críticament la recerca existent	Està dissenyat per a fer-ho ràpidament utilitzant estratègies de cerca menys sofisticades, mirant altres revisions, sense incloure la matèria grisa, i fent avaluacions de qualitat limitades	La reducció del temps d'avaluació de la qualitat augmenta el risc d'utilitzar estudis esbiaixats o de mala qualitat
Systematic review	Busca sistemàticament la cerca, avaluació i síntesi de proves de recerca	Cerca incloure tots els coneixements sobre un tema	És restrictiu per a enfocar un determinat mètode utilitzat en els estudis

2.1.1. Revisió Sistemàtica

Una *systematic review* o revisió sistemàtica (RS) és una revisió exhaustiva de documents de recerca primària centrada en una única pregunta concisa i clara, formulada amb una estructura definida com PICO (*Population, Intervention, Comparison and Outcome*). Es tracta de buscar, sintetitzar i avaluar sistemàticament les proves disponibles corresponents a aquesta pregunta, utilitzant una metodologia rigorosa i documentada de forma clara tant en la cerca com en la selecció dels estudis per tal de minimitzar el biaix en els resultats i permetre, així, prendre decisions informades basades en l'evidència. Tant les RS com els meta anàlisis formen part del primer dels cinc grups dels anomenats nivells d'evidència, que classifiquen de més alta a més baixa qualitat els diferents tipus d'estudis (Figura 2.1). Una RS ben efectuada permet a altres reproduir-la i actualitzar-la. (*Guides: Systematic Reviews: Protocol*, s. d.)

Figura 2.1.: Piràmide dels nivells d'evidència. Font: (Guides: Systematic Reviews: Protocol, s. d.)



Per aconseguir una revisió sistemàtica de qualitat és important disposar de criteris clars i fiables que permetin avaluar-la, tenint en compte el rigor de l'estratègia de cerca i la manera de descriure els mètodes utilitzats i els resultats. La guia PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*), abans QUOROM (*Quality of Reporting of Meta-Analyses*), és un conjunt d'elements de base empírica per la presentació d'informes en revisions sistemàtiques i meta-anàlisis (Figura 2.2). La declaració PRIMSA, a diferència de QUOROM, presenta un document on es detalla l'explicació de cadascun dels 27 ítems que la componen i el procés d'elaboració d'aquestes directrius. Aquesta llista de comprovació també conté un diagrama de flux que representa el flux d'informació a través de les diferents fases d'una RS, des de la identificació inicial dels estudis potencialment rellevants fins la selecció definitiva (Moher, 2009; Rethlefsen et al., 2021).

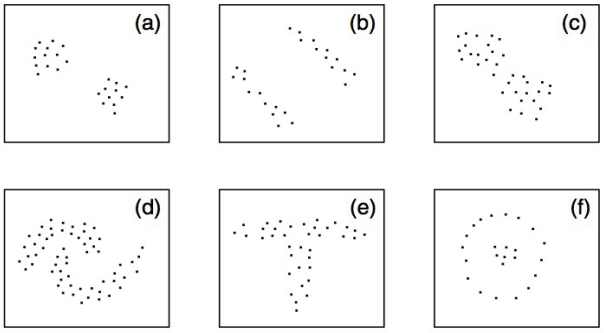
Figura 2.2.: Taula comparativa entre les guies QUOROM i PRISMA. Font: (Moher, 2009)

Section/Topic	Item	QUOROM	PRISMA	Comment
Abstract		√	√	QUOROM and PRISMA ask authors to report an abstract. However, PRISMA is not specific about format.
Introduction	Objective		√	This new item (4) addresses the explicit question the review addresses using the PICO reporting system (which describes the participants, interventions, comparisons, and outcome(s) of the systematic review), together with the specification of the type of study design (PICOS); the item is linked to Items 6, 11, and 18 of the checklist.
Methods	Protocol		√	This new item (5) asks authors to report whether the review has a protocol and if so how it can be accessed.
Methods	Search	√	√	Although reporting the search is present in both QUOROM and PRISMA checklists, PRISMA asks authors to provide a full description of at least one electronic search strategy (Item 8). Without such information it is impossible to repeat the authors' search.
Methods	Assessment of risk of bias in included studies	√	√	Renamed from "quality assessment" in QUOROM. This item (12) is linked with reporting this information in the results (Item 19). The new concept of "outcome-level" assessment has been introduced.
Methods	Assessment of risk of bias across studies		√	This new item (15) asks authors to describe any assessments of risk of bias in the review, such as selective reporting within the included studies. This item is linked with reporting this information in the results (Item 22).
Discussion		√	√	Although both QUOROM and PRISMA checklists address the discussion section, PRISMA devotes three items (24–26) to the discussion. In PRISMA the main types of limitations are explicitly stated and their discussion required.
Funding			√	This new item (27) asks authors to provide information on any sources of funding for the systematic review.

2.2. *Mètodes de clustering*

El *clustering* és la tècnica d'aprenentatge no supervisat més coneguda i utilitzada i té com a finalitat classificar un conjunt d'observacions en grups o clústers amb perfils desconeguts similars. Aquests grups venen determinats segons les característiques d'aquests objectes, de manera que les observacions d'un mateix grup seran molt similars entre si i les que pertanyen en clústers diferents seran molt divergents entre elles. L'estructura dels grups pot variar quan s'utilitzen algoritmes distints en un mateix grup de dades i poden ser difícils de trobar segons l'estructura que tingui el conjunt d'observacions a analitzar (Figura 2.3).

Figura 2.3.: Alguns dels possibles patrons de punts quan hi ha dos clústers*. Font: (Fernández, s.d.)*

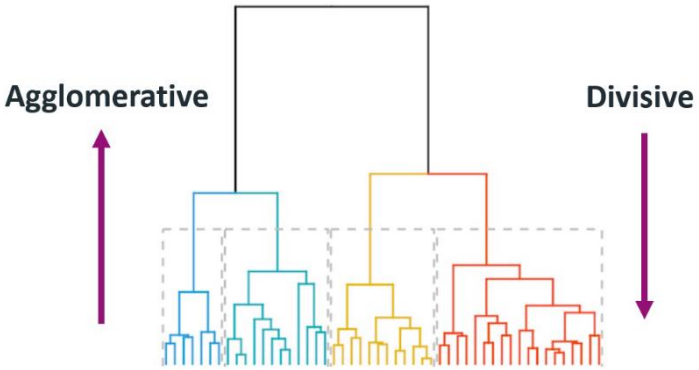


En termes generals, el *clustering* es pot dividir en dos subgrups la informació que es té per assignar una observació a un clúster. Així quan l'agrupament es basa en distàncies matemàtiques es parla de *hard clustering* ja que un objecte pertany exclusivament a un sol grup. D'altra banda, el *soft clustering* es basa en probabilitats i, per tant, per cada observació hi ha una probabilitat d'estar classificat en els diversos clústers resultants. Dels tres mètodes de *clustering* principals, el *clustering* jeràrquic i el particional pertanyen al primer subgrup i el *model-based clustering* forma part del *soft clustering*. A les següents seccions es passarà a explicar cadascun d'aquests mètodes.

2.2.1. *Clustering* jeràrquic

El *clustering* jeràrquic és un mètode que té com a objectiu crear una estructura en forma d'arbre anomenada dendrograma que representi enllaços jeràrquics entre les observacions d'un conjunt de dades (Figura 2.4). En el *clustering* jeràrquic aglomerat, la part inferior del dendrograma mostra com en un principi tots els objectes estan distribuïts en grups d'un i s'observa com, a mesura que es segueix l'eix d'ordenades, aquests objectes es van aglomerant amb els que estan més a prop a través del càlcul de distàncies matemàtiques entre dos punts. Com més llarga és la línia vertical, més gran és la distància entre dos grups. El procés acaba quan totes les observacions formen part d'un sol grup. El *clustering* jeràrquic divisiu, pel contrari, va en la direcció inversa ja que tothom comença en un sol grup i es van disgregant fins formar els grups individuals.

Figura 2.4.: Exemple d'un dendrograma i direcció del procés del clustering jeràrquic aglomerat vs divisiu. Font: (Fernández, s.d.)



Un dels passos més importants en aquests tipus d'algoritmes és l'elecció de la mesura de similitud. Aquesta defineix com es calcula la semblança de qualsevol parell d'observacions i a més pot influir en la forma dels clústers. Existeixen diverses formes de calcular la distància entre dos punts, no és el mateix mesurar la similitud entre dos objectes i mesurar la similitud entre dos grups d'observacions.

Per calcular la similitud entre dos observacions es pot utilitzar:

- Distància Euclidiana

És la distància ordinària en línia recta entre dos punts en l'espai euclidià.

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Distància Manhattan

És similar a l'anterior però ara la distància es calcula sumant el valor absolut de la diferència entre les dimensions.

$$d_{man}(x,y) = \sum_{i=1}^{n}|(x_i - y_i)|$$

Per mesurar la similitud entre dos grups d'observacions es pot usar:

- Enllaç simple

Calcula la distància mínima entre els clústers abans d'aglomerar-se, sent aquesta distància la més curta entre dos elements de dos grups diferents.

$$\triangle (G_1, G_2) = min\{d(x_i, x_i'): x_i \in G_1, x_i' \in G_2\}$$

- Enllaç complet

Calcula la distància màxima entre els clústers abans d'aglomerar-se, volent minimitzar la distància de les observacions més allunyades de dos grups distints.

$$\triangle (G_1, G_2) = max\{d(x_i, x_i'): x_i \in G_1, x_i' \in G_2\}$$

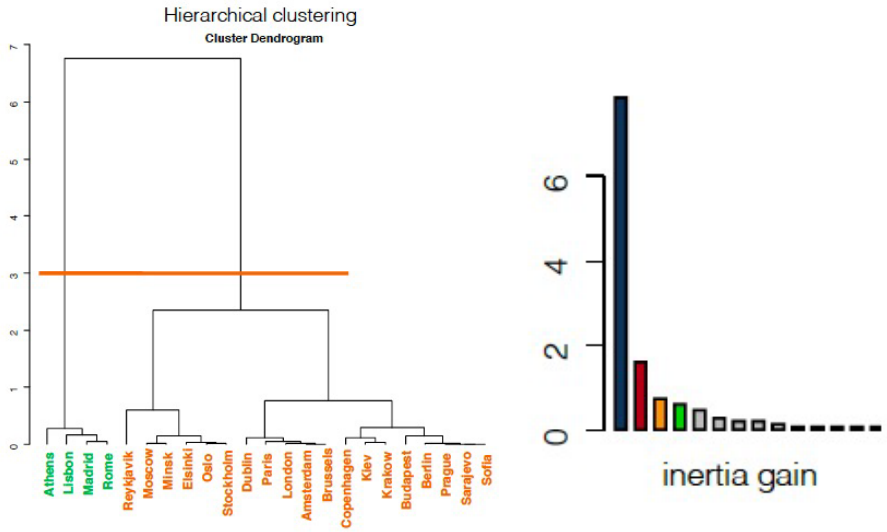
- Criteri de Ward

Ward va proposar que la pèrdua d'informació que es produeix a l'integrar diferents individus en clústers es pot mesurar a través de la suma total de quadrats de les desviacions entre cada punt (individu) i la mitjana del clúster en el que es troba. Per tant, en cada iteració del mètode de Ward s'han d'unir els grups que menys incrementin la suma de quadrats de les desviacions.

$$\triangle (G_1, G_2) = \frac{n_1 * n_2}{n_1 + n_2} d^2(\mu_1, \mu_2)$$

Un cop construït el dendrograma, el que queda és decidir el nombre definitiu de clústers. En el *clustering* jeràrquic es pot fer de tres maneres: 1) tallant el dendrograma a una altura determinada de manera que el nombre final de classes serà el total de línies verticals que talla la línia horitzontal; 2) utilitzant un *inertia gain plot*, que és un gràfic que ajuda a entendre la pèrdua de variabilitat quan es passa de *k* clústers a *k* + 1 clústers de manera que com més diferència entre dues barres més pèrdua hi ha i, per tant, s'escull una *k* on la pèrdua no és molt alta (Figura 2.5); i 3) utilitzant criteris de coneixement dels experts en l'àrea d'aplicació.

Figura 2.5.: A l'esquerra, exemple d'un dendrograma tallat per definir el nombre òptim de clústers i a la dreta, un inertia gain plot per definir el mateix. Font: (Fernández, s.d.)



2.2.2. *Clustering* particional

El *clustering* particional requereix una especificació prèvia del nombre de clústers (*k*) que es vol crear. L'algoritme més popular d'aquest tipus de *clustering* és el *k-means*.

El mètode *k*-means (MacQueen, 1967), que es podria definir com un problema d'optimització, té com a objectiu la partició d'un conjunt *n* d'observacions en *k* grups en els quals cada observació pertany al clúster més proper a la mitjana, de manera que la variació interna dels clústers es mínima. La variació interna dels grups es pot mesurar com:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

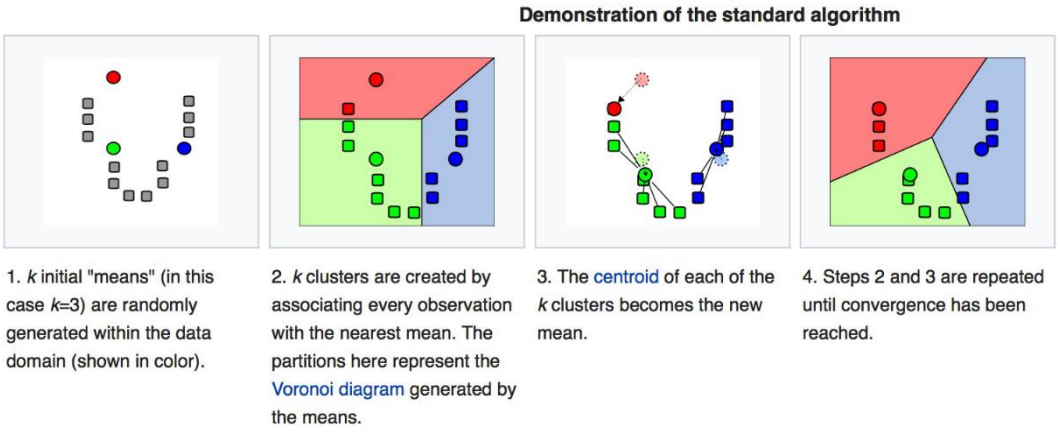
sent x_i un punt pertanyent al clúster C_k i situat el més a prop possible del punt cèntric d'aquest grup i μ_k el valor mitjà dels punts assignats al clúster C_k. La suma de $W(C_k)$ mesura la compactació del clúster i interessa que sigui el més petita possible.

Aquest algoritme consta de tres passos (Figura 2.6):

i) En primer lloc, i després d'haver escollit els *k* grups, s'estableixen *k* centroides aleatòriament.
ii) Seguidament, cada objecte s'assigna al centroide més proper.
iii) Finalment s'actualitza la posició del centroide de cada clúster utilitzant com a nou centroide la mitjana dels elements de cada grup.

Es van repetint els passos dos i tres fins que s'aconsegueix l'estabilitat amb un nombre predeterminat de grups.

Figura 2.6.: Procés de formació de clústers amb l'algoritme k-means. Font: (Wikipedia contributors, 2022)



Per determinar el nombre més òptim de clústers es poden utilitzar tres mètodes diferents:

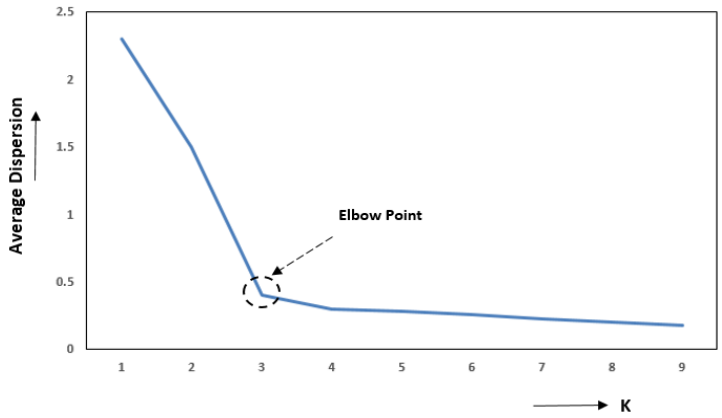
- *Elbow method*

Aquest mètode consta de quatre fases (Thorndike, 1953). La primera etapa consisteix en calcular l'algoritme *k-means* per diferents valors de *k*, amb $k = 1, \ldots, 10$. A continuació, per cada valor de *k* es calcula el quadrat de la suma intergrupal total (wss):

$$wss = \sum_{k=1}^{k} W(C_k)$$

A partir d'aquest valor, es grafica la corba de wss d'acord amb el nombre de clústers k i aquell punt on es pugui observar el colze de la corba serà el nombre apropiat de grups (Figura 2.7).

Figura 2.7.: Exemple gràfic de l'elbow method. Font: (Dangeti, s. d.)



- *Silhouette method*

El *silhouette method* (Kaufman & Rousseeuw, 1990) calcula l'*average silhouette* de les observacions pels diferents valors de *k*. El nombre òptim de clústers *k* és el que maximitza aquest *average silhouette* en un rang de valors possibles de *k*.

El *silhouette coeficient* per un punt particular es calcula:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

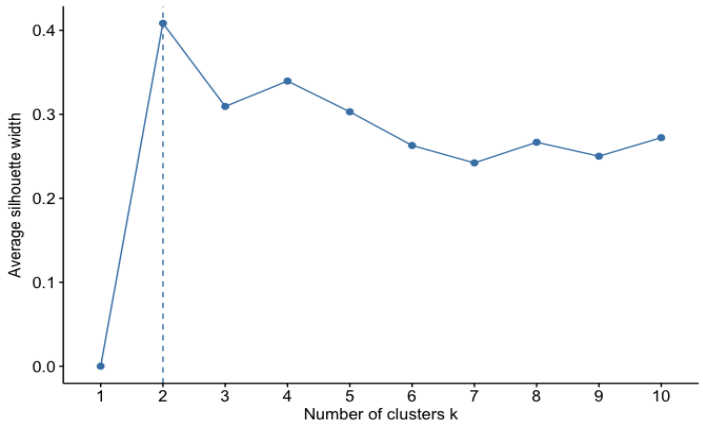
on $a(i)$ és la distància mitjana entre i i la resta de punts del clúster on es troba i i $b(i)$ és la distància mitjana des d'i fins a tots els clústers on no es troba i. Així doncs, l'*average silhouette* es calcula com:

$$\bar{S}(i) = \frac{\sum_{i=1}^{n} S(i)}{n}$$

amb *n* com el nombre total de punts pertanyents al clúster *k*.

Igual que en el mètode anterior, aquests valors es grafiquen en forma de corba per cada *k* i la localització del màxim es considera el nombre apropiat de grups (Figura 2.8).

Figura 2.8.: Exemple gràfic del silhouette method. Font: (K-Means Cluster Analysis · UC Business Analytics R Programming Guide, s. d.)



- *Gap statistic method*

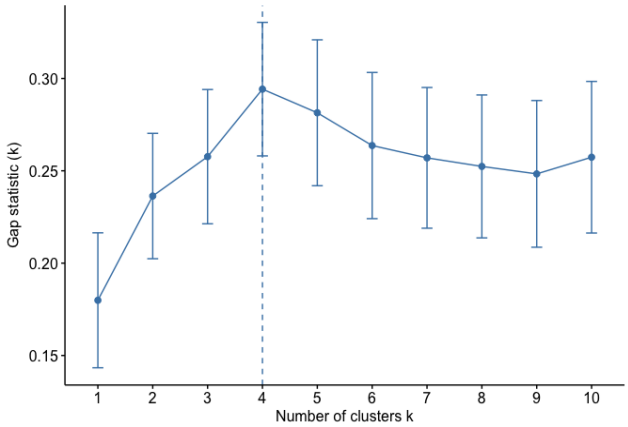
El *gap statistic method* (Tibshirani et al., 2001) es basa en comparar l'evidència amb la hipòtesis nul·la. El *gap statistic* compara la variació total interna dels clústers per diferents valors de *k* amb el seu valor esperat sota la distribució de referència nul·la de les dades. Aquest estadístic es calcula:

$$Gap_n(k) = E_n^* * log(W_k) - log(W_k)$$

on E_n^* és definit via *bootstrapping* generant B copies de la base de dades de referència. Aquesta base de dades ha estat prèviament generada utilitzant simulacions de Monte Carlo del procés de mostreig.

De la mateixa manera que l'*elbow method* i l'*average silhouette method*, es representen gràficament els valors de l'estadístic per cada grup *k*, en el que el nombre apropiat de clústers és el *k* on es troba el màxim valor del *gap statistic* (Figura 2.9).

Figura 2.9.: Exemple gràfic del gap statistic method. Font: (K-Means Cluster Analysis · UC Business Analytics R Programming Guide, s. d.)



2.2.3. *Model-based Clustering*

En el *model-based clustering* (MBC), a diferència dels dos mètodes anteriors, cada observació té una probabilitat de pertànyer a cada clúster i a més permet identificar automàticament el nombre òptim de grups. L'enfocament més popular és el Model Mixt Gaussià (GMM), que utilitza una combinació de distribucions de probabilitat normals i requereix l'estimació dels paràmetres de desviació mitjana i estàndard (Figura 2.10). El model general es pot definir com:

$$p(x, \psi) = \pi_1 * f(x; \mu_1, \sigma_1^2) + \cdots + \pi_G * f(x; \mu_G, \sigma_G^2)$$

amb

$$f(x; \theta_g) = f(x; \mu_g, \sigma_g^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

i $g = 1, \ldots, G$, $\pi_g > 0$ i $\pi_1 + \cdots + \pi_G = 1$, on π_g són els pesos del que contribueix cada component.

La seva funció de versemblança corresponent es calcula:

$$L(\psi) = \prod_{n=1}^{N} p(x_n;\psi) = \prod_{n=1}^{N} \left[\sum_{g=1}^{G} \pi_g f(x_{n;}\theta_g) \right]$$

tot i que a la pràctica és més comú usar la funció log-versemblança:

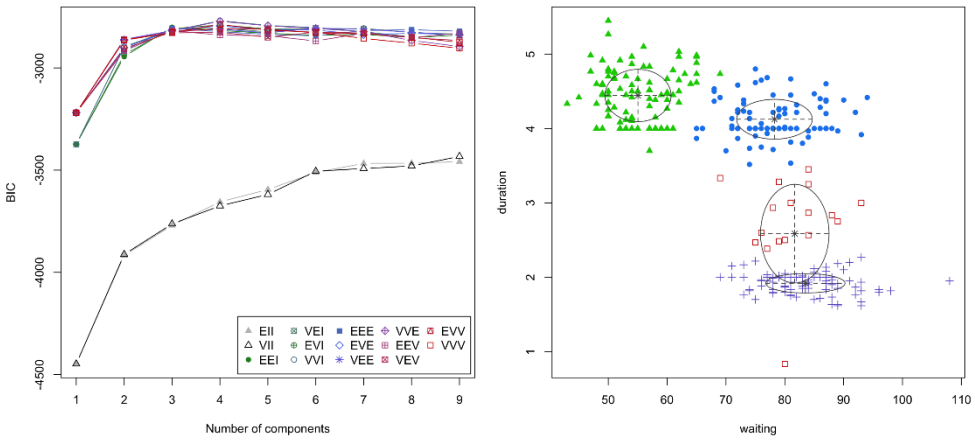
$$\mathcal{L}(\psi) = log(L(\psi))$$

L'algorisme EM (*expectation-maximization*) permet estimar aquests paràmetres de les distribucions i determinar les particions. Els models generats es comparen a partir del *Bayesian Information Criterion* (BIC) i es prefereix el que presenta un BIC menor. L'expressió que defineix aquest criteri és:

$$BIC = -2\mathcal{L}(\psi) + mlog(n)$$

on $\mathcal{L}$ és la funció de log-versemblança , *m* és el nombre de paràmetres lliures a estimar i *n* és el nombre de punts de dades.

Figura 2.10.: Identificació el model GMM òptim i el nombre de clústers (esquerra) i classificació de les observacions en els diferents clústers identificats. Font: (Boehmke & Greenwell, 2019)



3. Cos del treball

3.1. *Criteris de la revisió sistemàtica*

Disseny

S'ha dut a terme una revisió sistemàtica de l'aplicació de les tècniques *clustering* en el camp de les ciències de l'esport seguint les recomanacions de la declaració PRISMA (Moher, 2009; Rethlefsen et al., 2021).

Estratègia de cerca

La cerca de la revisió sistemàtica es va realitzar el 7 de maig del 2022 utilitzant les següents bases de dades: *Web of Science* (WoS), *PubMed*, *SportDiscus* i *CINAHL* (*Cumulative Index to Nursing and Allied Health Literature*). En totes les bases de dades es va fer una recerca utilitzant els mateixos operadors booleans: *"clustering" and "sport"*.

Selecció dels estudis

Com a criteris d'elegibilitat dels estudis, es van definir els següents criteris d'inclusió i exclusió:

- Criteris d'inclusió

Els articles inclosos en la revisió sistemàtica havien de ser articles originals, escrits en angles i que apliquessin tècniques *clustering*. A més, aquest estudis havien de formar part de revistes de ciències de l'esport incloses al *Journal Citation Reports* (JCR) a l'any 2020 més dues revistes que no apareixien en aquesta llista com són el *Journal of Quantitative Analysis in Sports* (JQAS) i el *Journal of Sports Analytics* (JSA), on s'hi poden trobar una gran quantitat de publicacions d'estadística esportiva i també citades a la secció d'esports de l'ASA (Swartz, 2018).

- Criteris d'exclusió

Es van excloure de la RS els treballs de tesi, tesina, conferència i articles no originals (*special issue*, *opinion*, *review*, ...). També tots aquells articles escrits en un idioma diferent a l'anglès i que s'hagin publicat en una revista no inclosa al JCR en l'àmbit de ciències de l'esport ni a les revistes JQAS i JSA.

Per al procés de selecció es van fer dos passos. En primer lloc, es van seleccionar els estudis en funció del títol i del resum. En una segona etapa, es van revisar els texts complets dels articles per tal d'avaluar si aquests complien els criteris d'elegibilitat en situacions en les que no es podia prendre una decisió a partir de la informació proporcionada pel títol i l'*abstract*.

Extracció de les dades

La informació recollida en els estudis seleccionats es van agrupar en tres categories: característiques generals dels articles seleccionats, característiques generals de l'esport i característiques de les tècniques *clustering*. La primera categoria va permetre donar una idea general i bàsica de l'article, amb informació com l'autor, el país d'on provenen les dades, l'any i la revista de publicació (amb l'*Impact Factor* (IF) i el quartil corresponents), el nombre de participants i la seva edat i el propòsit de l'estudi. Les altres dues categories van donar una visió més específica sobre els tòpics pels que es realitzava la RS (Taula 3.1). Entre les

característiques generals de l'esport es van trobar variables com el tipus d'esport en el que s'enfocava l'article, el gènere en el que es centrava (homes, dones o ambdós), la categoria dels participants (professional, amateur o ambdós), la font de les dades i el camp del que tractava l'article dins les ciències de l'esport (*Sports performance analysis*, *Sports technology*, *Moviment Integration*, *Health* i *eSports*) (Taula 3.2). Per l'última categoria es va recollir la descripció dels *outcomes*, el tipus de mètode o tècnica *clustering* i el nom del mètode usat, el nombre i la mida dels clústers, el mètode per decidir el nombre de clústers (*elbow, silhouette, gap, ...)*, el *software* i el *package* utilitzat i si el propi article comparteix el codi o les dades usades (Taula 3.3).

Les dades es van recollir i emmagatzemar en una base de dades. A continuació, es van comprovar les dades per trobar discrepàncies entre els tres autors (MP, DF i MC). Les discrepàncies es van resoldre per consens després de revisar de nou els articles conflictius.

Taula 3.1.: Taula descriptiva de les característiques generals dels articles definitius.

Variable	***Description***
General characteristics of the selected articles	
Author (citation)	Citació en format APA dels autors
Country	País de les dades de l'article
Publication Year	Any de publicació de l'article
Journal Name	Nom del *Journal* on està publicat l'article
IF (Quartile)	*Impact Factor* i *Quartile* del *Journal*
Longitudinal study	Si es tracta o no d'un estudi longitudinal
N (participants)	Nombre de participants de l'estudi
Age (participants)	Mitjana o interval d'edat dels participants
Principal Aim	Descripció de l'objectiu principal de l'estudi

Taula 3.2.: Taula descriptiva de les característiques generals de l'esport.

General characteristics of the sport	
Author (citation)	Citació en format APA dels autors
Sport	Esport/s en el/s que es centra l'article
Gender	Sexe dels participants en el que es centra l'article *(Male, Female, Both*)
Category participants	Categoria dels participants (*Professional, Amateur, Both*)
Name of source data (League, Association, Organitzation, Federation,....)	Nom de la font de les dades de l'estudi
Category classification	Cinc categories: 1) *Sports Performance Analysis;* 2) Sports technology; 3) Movement integration; 4) Health; 5) eSports

Taula 3.3.: Taula descriptiva de les característiques de les tècniques clústering utilitzades en els articles.

Characteristics of clustering technique	
Author (citation)	Citació en format APA dels autors
Description of Outcomes	Descripció dels *outcomes* tal com està escrita a l'article
Type of clustering method	Tipus de mètode *clustering* que s'utilitza a l'article (*Hierarchical clustering*, *Partitioning methods*, *Model-based Clustering*, *Fuzzy Clustering*, *Density-based clustering*, ...)
Name of the clustering method used	Nom del mètode *clustering* utilitzat en l'article (PCA, *k-means*, *k-medoids*, *Kamila*, *hclust*, *Gaussian Mixture Models*, ...)
Number of clusters	Nombre de clústers
Cluster size per cluster	Mida de cada clúster
Method to decide the number of clusters	Mètode per decidir el nombre òptim de clústers (*Elbow*, *Gap*, *Sillhoutte*, *Dendrogram*, ...)
Software used	*Software* utilitzat
Package used	Paquet utilitzat del *software*
Data shared	Si l'article comparteix o no les dades que utilitza en l'estudi
Code shared	Si l'article comparteix o no el codi del *software*
Repository of Data or Code shared	Si l'article comparteix un repositori amb les dades o el codi utilitzat
Implicació practica results	Breu resum de la implicació pràctica dels resultats de l'estudi

Identificació dels estudis

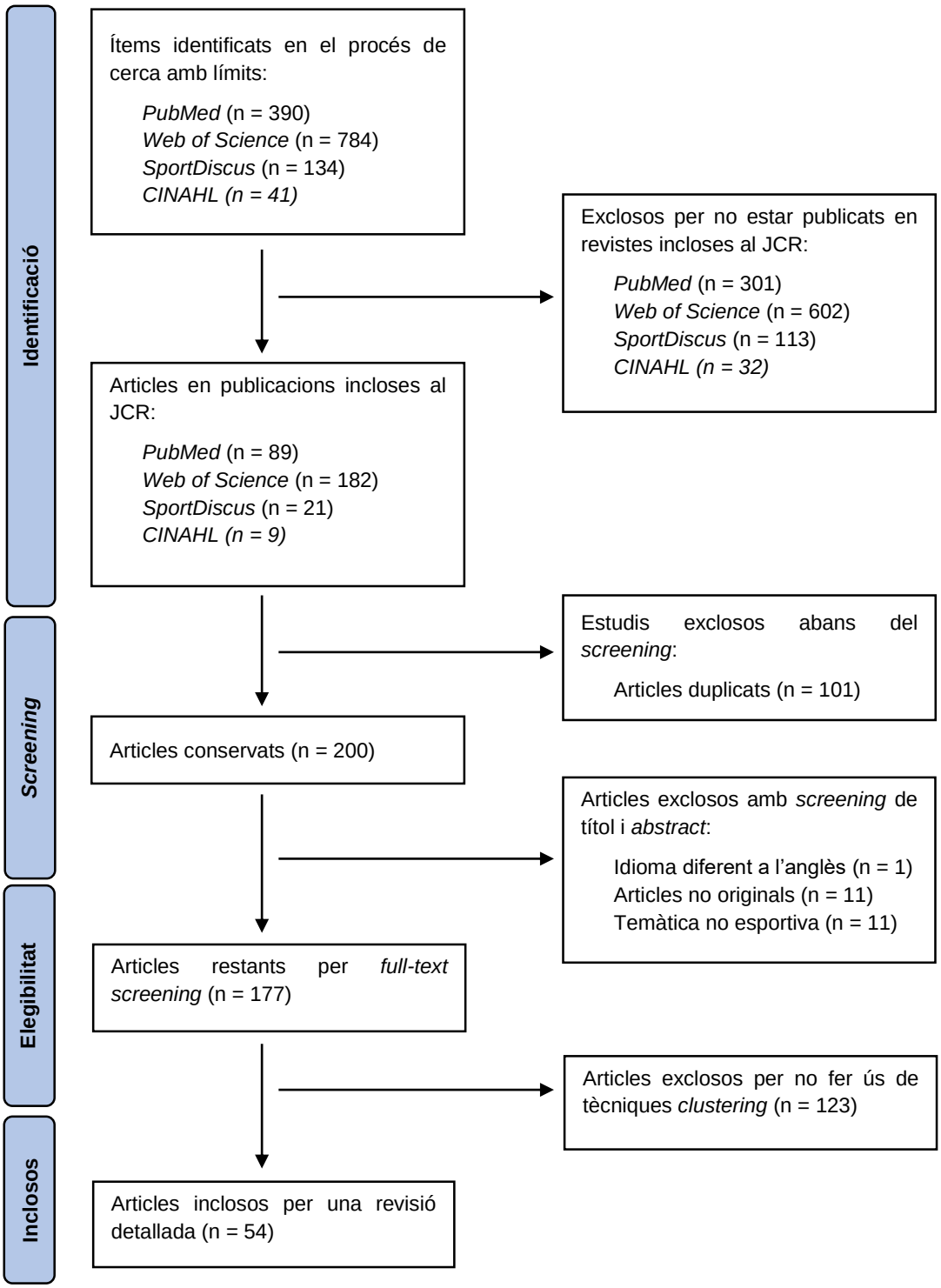
La Figura 3.1 mostra el diagrama de flux PRISMA per resumir totes les etapes del procés de selecció. En el primer procés de cerca, on ja es va filtrar en les bases de dades (BBDD) que ho permetien per idioma (*PubMed*, *WoS* i *CINAHL*) i per tipus de literatura científica (*WoS*), es van recollir un total de 390 ítems a *PubMed*, 784 a WoS, 134 a *SportDiscus* i 41 a *CINAHL*. Al revisar si tots aquests articles formaven part d'algun dels camps de les ciències de l'esport es van excloure tots aquells que no estaven publicats en revistes incloses al JCR, al JQAS i JSA; quedant d'aquesta manera 89 articles a *PubMed*, 182 a WoS, 21 a *SportDiscus* i 9 a *CINAHL* i d'aquest total de 301 articles se'n van excloure 101 per estar duplicats, obtenint així un total de 200 articles.

Després de la inspecció del títol i l'*abstract*, es van excloure 11 articles no originals, un article escrit en francès que no s'havia pogut filtrar en la primera cerca i 11 articles de temàtica no esportiva.

A partir dels 177 articles seleccionats fins al moment, es va procedir a la segona fase de la revisió on es va revisar el text complet de tots els articles restants per tal d'observar si utilitzaven algun mètode *clustering* en els seus estudis. Dels 177 articles que havien quedat de la primera fase de la revisió, es va poder observar que en 123 articles solament utilitzar la paraula *clustering* per referir-se a l'agrupació de les dades en la seva estructura inicial i no per descriure un mètode *clustering* en l'anàlisi estadístic de les dades.

Finalment, tan sols 54 articles van ser inclosos per una revisió més detallada. La Figura 3.1 resumeix el nombre d'articles identificats i les raons d'exclusió en cada fase.

Figura 3.1.: PRISMA: Diagrama de flux de la revisió sistemàtica de l'aplicació de les tècniques clustering en articles originals en el camp de les ciències de l'esport.



4. Resultats

4.1. *Resultats de la revisió sistemàtica*

En aquest apartat, es mostren els resultats de l'anàlisi descriptiu de les variables descrites a les taules de l'apartat anterior referents als diferents tipus de característiques dels articles definitius de la revisió sistemàtica. En primer lloc, es presenta una taula de freqüències dels resultats principals de les característiques de cada taula.

Taula 4.1.: Freqüències de les característiques generals dels articles per grup.

Variable (N = 54)	***Category***	***n (%)***
General characteristics of the selected articles		
Country	*International*	4 (7,4%)
	Australia	5 (9,3%)
	USA	7 (13,0%)
	UK	7 (13,0%)
	No reported	12 (22,0%)
	*Others**	19 (35,2%)
Publication Year	2022	4 (7,4%)
	2021	12 (22,0%)
	2020	4 (7,4%)
	2019	8 (14,9%)
	2018	4 (7,4%)
	2017	8 (14,9%)
	2016	6 (11,1%)
	2015	2 (3,7%)
	2013	2 (3,7%)
	2010	1 (1,8%)
	2009	1 (1,8%)
	2007	1 (1,8%)
	1981	1 (1,8%)
Longitudinal study	*No*	50 (92,6%)
	Yes	3 (5,6%)
	Unclear	1 (1,8%)

A la variable Country, "others" reporta els països de menor freqüència: Canada, Denmark, Germany, Ireland, Malaysia, Singapur, Australia | New Zealand, China, Finland, Italy, Slovenia, Brazil, France, Japan i Spain.*

Taula 4.2.: Freqüències de les característiques generals de l'esport dels articles per grup.

General characteristics of the sport		
Sport	*Multidisciplinar*	9 (16,7%)
	Physical Activity	8 (14,9%)
	Soccer	9 (16,7%)
	Tennis	6 (11,1%)
	Basketball	4 (7,4%)
	Golf	3 (5,6%)
	No reported	2 (3,7%)
	*Others***	13 (23,9%)
Gender	*Male*	22 (40,7%)
	Female	6 (11,1%)
	Both	20 (37,1%)
	No reported	6 (11,1%)
Category participants	*Professional*	33 (61,1%)
	Amateur	16 (29,7%)
	Both	2 (3,7%)
	No reported	3 (5,6%)
Category classification	*Sports Performance Analysis*	27 (50,0%)
	Health	9 (16,7%)
	Movement Integration	16 (29,7%)
	Sports Technology	1 (1,8%)
	Sports NGOs	1 (1,8%)

*A la variable Sport, "others**" reporta els esports de menor freqüència: Australian Rules Football, Cross-country sit-ski, Swimming, Waterpolo, Field Hockey, Rugby, Netball, Cricket, Running, Volleyball, Wheelchair Basketball i Synchronized swimming | Handball.*

Taula 4.3.: Freqüències de les característiques de les tècniques clústering dels articles per grup.

Characteristics of clustering technique		
Type of clustering method	*Hierarchical clustering*	19 (35,2%)
	Partitional clustering	16 (29,7%)
	Model-based clustering	3 (5,6%)
	Density-based Spatial clustering	3 (5,6%)
	High-dimensional clustering	3 (5,6%)
	*More than one type**	5 (9,3%)
	*Others****	3 (5,6%)
	No reported	2 (3,7%)
Name of the clustering method used	*k-means*	20 (37,1%)
	Spectral clustering	3 (5,6%)
	PCA	3 (5,6%)
	*Others*****	11 (20,4%)
	No reported	17 (31,3%)

Method to decide the number of clusters	*Silhouette*	6 (11,1%)
	Dendrogram	7 (13,0%)
	BIC	4 (7,4%)
	Elbow	2 (3,7%)
	Gap	2 (3,7%)
	Majority Rule	1 (1,8%)
	Cubic Clustering Criterion (CCC)	1 (1,8%)
	No reported	19 (35,2%)
	*Others******	12 (22,0%)
Software used	*R*	14 (25,9%)
	MATLAB	11 (20,4%)
	SPSS	6 (11,1%)
	Python	2 (3,7%)
	*Others*******	13 (24,1%)
	No reported	8 (14,9%)
Package used	*No reported*	39 (72,2%)
	*Reported**	15 (27,8%)
Data shared	*No*	54 (100%)
Code shared	*No*	54 (100%)
Repository of Data or Code shared	*No*	54 (100%)

*A la variable Type of clustering method, "others***" reporta els tipus de tècniques clustering de menor freqüència: Trajectory-based Longitudinal Clustering i Two-step Clustering; mentre que "Moret han one type*" reporta la combinació de dos tipus de tècniques clústering: Hierarchical Clustering | Partitional Clustering i Partitional Clustering | Model-based Clustering.*

*A la variable Name of the clustering method used, "others****" reporta els noms dels mètodes clustering usilitzats de menor freqüència: Gaussian mixture models, Johnson's hierarchical procedure, Latent class analysis, Mixed Data Method, hclust, kmlShape, Louvain method i combinacions d'alguns dels mètodes ja nombrats.*

*A la variable Method to decide the number of clusters, "others*****" reporta els mètodes utilitzats per decidir el nombre de clusters de menor freqüència: AIC, Calinski-Harabasz, Fréchet mean, Clustergram, Point biseral correlation, CH Index i combinacions d'alguns dels mètodes ja nombrats.*

*A la variable Software used, "others******" reporta els softwares utilitzats de menor freqüència: NCSS, Excel, Statica | Modalisa, Graphpad Prism i combinacions d'alguns dels mètodes ja nombrats.*

A la variable Package used, "Reported" els paquets que sí que es reporten però en menor freqüència: Boruta, kml-shape, SPM, hclust, acc, lm, bios2mds, NbClust, limma, pvclust, mclust, RPostgreSQL, stats, rpart, kmeans.m, k-means++, fpc, pawacc i combinacions dels anteriors.*

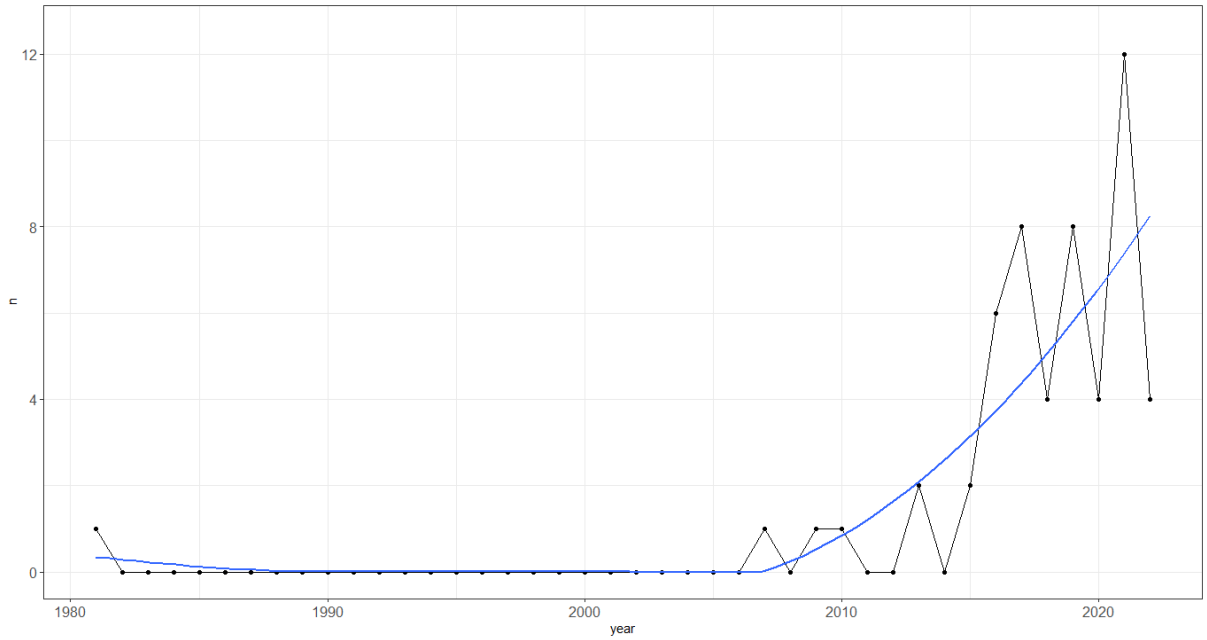
4.1.1. Característiques generals dels articles seleccionats

Dels 54 articles seleccionats, 12 (22,0%) no han reportat el país d'on provenen les dades, en 4 articles (7,4%) aquestes dades són internacionals i en 5 (9,3%), 7 (13,0%) i 7 (13,0%) articles provenen d'Austràlia, Estats Units i el Regne Unit, respectivament. Per la resta, gairebé en el 52% dels articles les dades venen de diferents països europeus, en el 5% venen de països americans i el 7,4% dels articles provenen d'Àsia.

La Taula 4.1 mostra com la gran majoria d'articles originals publicats sobre l'ús de tècniques *clustering* en el camp de les ciències de l'esport es concentren en els últims 7 anys. Es pot observar que el nombre d'articles originals publicats oscil·la entre els 4 i els 8 per any, destacant l'any 2021 en el que es van publicar 12 articles que representen un 22,0% dels

totals. En la resta d'anys en que se'n va publicar algun, el nombre varia entre un i dos articles per any, representant solament un 1,8 i un 3,7% respectivament.

Gràfic 4.1.: Tendència de publicació d'articles originals sobre l'ús de tècniques clustering en el camp de les ciències de l'esport per any.



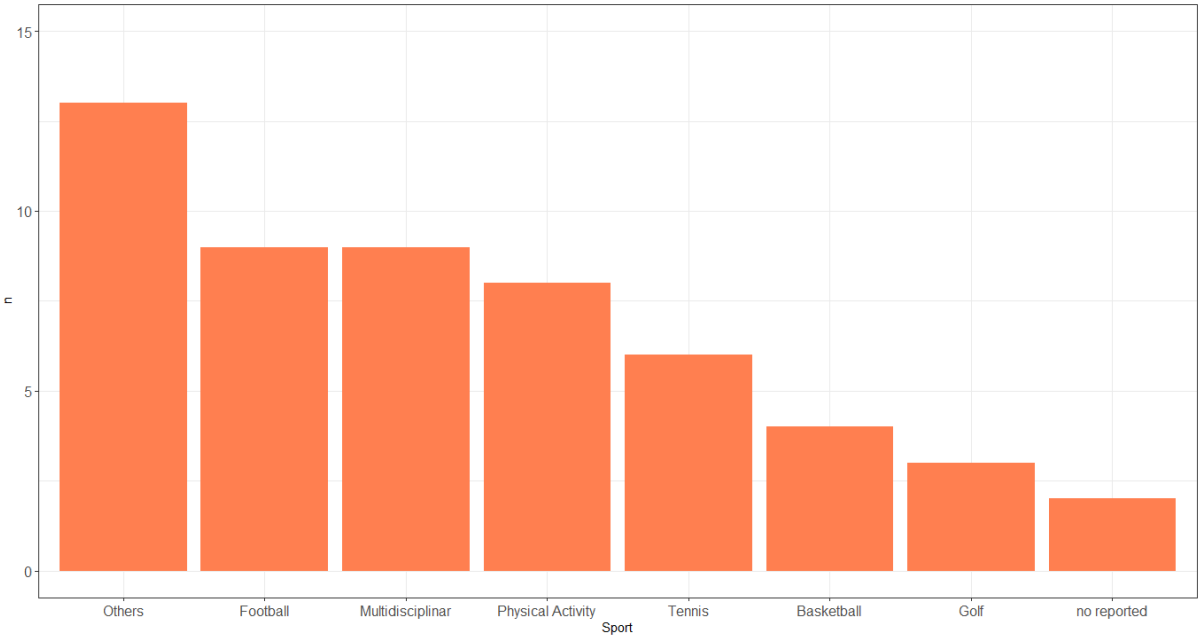
En el Gràfic 4.1 es pot veure com el primer article es va publicar l'any 1981 i després no se'n va publicar cap més fins l'any 2007. A partir d'aquell any, el nombre d'articles publicats en revistes del JCR, al JQAS i JSA ha anat augmentant provocant una tendència positiva.

L'última característica d'aquesta taula reporta que 50 dels 54 articles seleccionats no són estudis longitudinals, representant un 92,6% del total, i només 3 sí que ho són. S'ha trobat un article que no deixa clar si es tracta d'un estudi longitudinal o no i que s'ha classificat a la categoria *unclear*.

4.1.2. Característiques generals de l'esport

A la Taula 4.2 es pot observar que 9 dels 54 articles seleccionats, que representen un 16,7% del total, no tracten dades d'un sol esport sinó que els participants són atletes de diferents disciplines esportives. Aquests articles s'agrupen en la categoria *multidisciplinar*. En 8 articles (14,9%) s'utilitza alguna tècnica clústering per estudiar l'activitat física dels participants sense tenir en compte si aquests practiquen un esport concret i 2 articles, que són un 3,7% del total, no reporten el tipus d'esport. En els esports de pilota destaquen el futbol (*soccer*), present en 9 articles (16,7%) i el bàsquet, present en 4 (7,4%). L'esport de raqueta més representat és el tennis, que es pot trobar en 6 dels 54 articles (11,1%) i finalment, esports com el rugby, el volley, el cricket o esports de piscina com el Waterpolo o la natació entre d'altres es troben agrupats en la categoria *others*, ja que només es troben en articles solts i en total representen el 23,9% dels articles (Gràfic 4.2).

Gràfic 4.2.: Nombre d'articles originals que fan ús de tècniques clustering per esport.



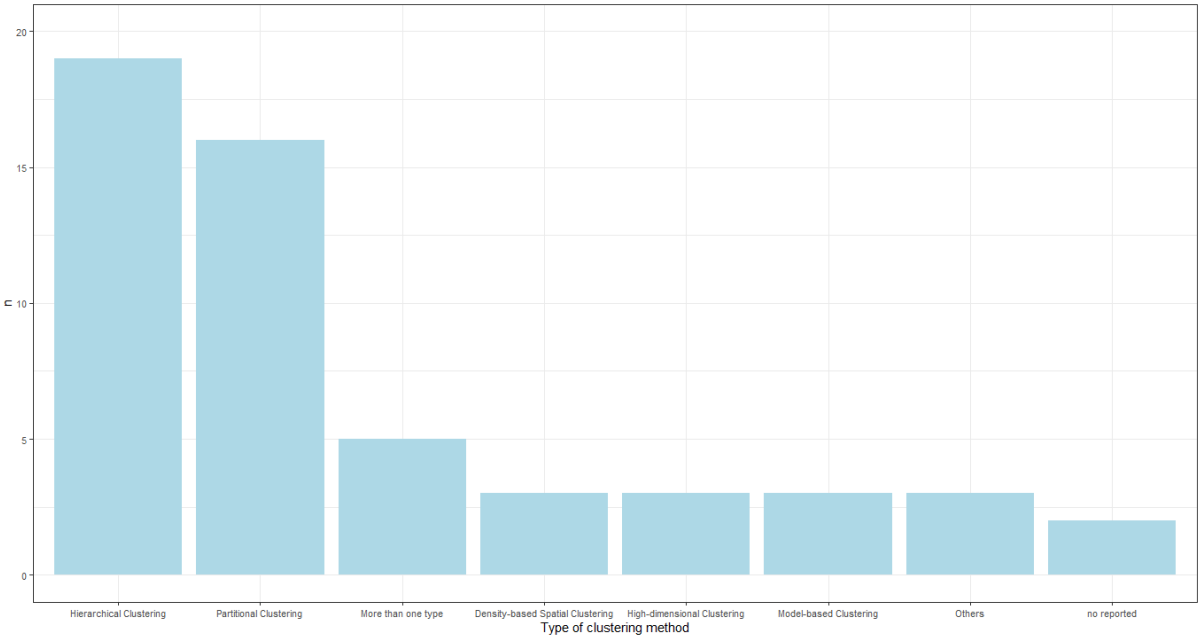
En el 40,7% dels articles els participants són de sexe masculí, mentre que només en l'11,1% són únicament dones. En 20 dels 54 articles definitius, que representen un 37,1% del total, els participants són d'ambdós sexes. Del total de participants dels articles, el 61,1% practiquen esport de manera professional mentre que el 29,7% són amateurs. Només en 2 dels 54 articles hi ha participants d'ambdues categories i en 2 articles no es reporta a quin dels dos grups pertanyen.

La Taula 4.2 també reporta la freqüència d'articles classificats segon la seva categoria. A l'inici, es van definir 5 categories diferents: *Sports performance analysis*, *Sports technology*, *Moviment Integration*, *Health* i *eSports*, però finalment només s'han classificat els articles segons les 4 primeres ja que no se'n ha trobat cap que pugui pertànyer a la última categoria. El 50% dels articles s'han classificat per la categoria *Sports performance analysis*, el 29,7 % per *Movement Integration* i el 16,7% per *Health*. Solament s'ha trobat un article que es pugui classificar com a *Sports Technology* i un recollit en la categoria *sports NGOs* (*non-governmental organizations*) i ambdós representen l'1,8% del total dels articles seleccionats.

4.1.3. Característiques de les tècniques *clustering*

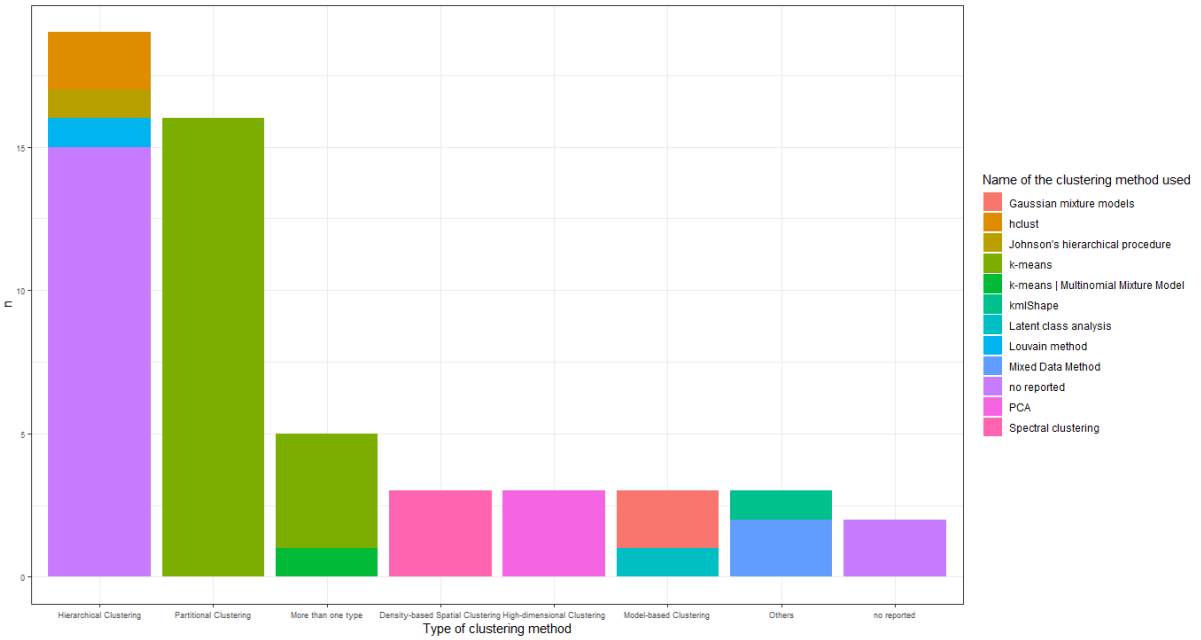
En la última taula (Taula 4.3) es reporten els resultats de les característiques de les tècniques *clustering* utilitzades en els diferents articles. En primer lloc, es pot observar com els tipus de tècniques més usades són el *clustering* jeràrquic, utilitzat en 19 dels 54 articles (35,2%), i el *clustering* particional, utilitzat en 16 articles (29,7%). Altres tipus com el *Model-based clustering*, el *Density-based Spatial clustering* i el *High-dimensional clustering* es troben amb menys freqüència, en el 5,6% d'articles cada un. L'ús de més d'un tipus de *clustering* dels anteriors en el mateix article s'ha agrupat en la categoria *More than one type* i represente el 9,3% dels articles. En molts pocs casos s'utilitzen altres tipus de mètodes *clustering* que no siguin cap dels anteriors i en 2 articles (3,7%) no es reporta el tipus de tècnica *clustering* utilitzat (Gràfic 4.4).

Gràfic 4.4.: Nombre d'articles segons el tipus de mètode clustering utilitzat.



Tot seguit, es reporta la freqüència dels mètodes utilitzats en els articles dels diferents tipus de tècniques *clustering*. El *k-means* és el mètode que més s'usa, es troba en almenys 18 dels 54 articles (33,3%) ja que també s'utilitza juntament a altres mètodes en alguns articles i aquests queden englobats en la categoria *others*. Tant l'*Spectral clustering* com el PCA s'utilitzen en 3 articles, representant un 5,6% dels articles cada un. A la categoria *others* (20,4%) s'hi poden trobar mètodes com el *Louvain method,* els *Gaussian Mixture Models* i el *Latent class analysis*, entre d'altres. El 31,3% dels articles no reporten el mètode utilitzat.

Gràfic 4.3.: Nombre d'articles segons el tipus de tècnica clustering i el nom del mètode clustering utilitzat.



El Gràfic 4.3 relaciona els mètodes utilitzats amb el tipus de tècniques *clustering* usades en els articles. Es pot observar com el *partitional clustering* és, en aquest cas, íntegrament el mètode *k-means*, mentre que pels altres tipus s'utilitzen diferents mètodes d'agrupació de les observacions.

Els *softwares* que més s'utilitzen en els articles seleccionats són l'R i el MATLAB, que representen un 25,9 i un 20,4% respectivament. L'ús d'únicament l'SPSS representa un 11,1% i la combinació d'aquests tres *softwares* entre si o amb altres l'Excel o el Python, que es troba en 2 (3,7%) articles, s'engloba en la variable *others* (24,1%). En 8 dels 54 articles (14,9%) no es reporta el *software* utilitzat. Tot i reportar el *software* utilitzat en la majoria de casos, en 39 articles (72,2%) no es reporta el *package* i la resta utilitzen paquets com *Boruta*, *hclust*, *NbClust* i *kmeans.m* entre molts altres, cap d'ells trobats en més d'un article.

Per últim, en cap article d'aquesta revisió sistemàtica es comparteixen les dades, el codi o el repositori de les dades o del codi que s'usa a cada article.

4.2. *Estudi de cas*

Els resultats de la revisió sistemàtica d'articles que utilitzen tècniques *clustering* en les ciències de l'esport han mostrat que el mètode de *clustering* particional més per estudiar dades esportives és el mètode *k-means*. Per aquest motiu i per posar en pràctica un dels mètodes revisats en la RS, s'ha decidit utilitzar l'algoritme *k-means* en una BBDD amb característiques i estadístiques de les jugadores de la Women's National Basketball Association (WNBA), que és la lliga professional de bàsquet femenina dels Estats Units, des de la temporada 1997 fins la temporada 2019. La WNBA és la lliga professional femenina de bàsquet als Estats Units i compta amb 12 equips i 36 partits per temporada de 40 minuts de duració.

Aquesta base de dades es troba en el repositori *Github* i ha estat compartida per la web americana *FiveThirtyEight*, que utilitzen les dades i la evidència per fer créixer el coneixement públic (https://github.com/fivethirtyeight/WNBA-stats). En aquesta base de dades es poden trobar característiques de cada jugadora com el seu nom, la seva edat en cada temporada jugada i la seva posició en el joc; i estadístiques de *performance* per temporada com les que es descriuran a continuació. Les variables numèriques que s'han considerat més rellevants per trobar el perfil de les jugadores i que s'utilitzaran pel *k-means* i la seva interpretació són:

Taula 4.4. Taula descriptiva de les variables que s'utilitzaran en el k-means.

Variable	**Descripció**	**Unitat**
PER	Puntuació de la eficiència dels jugadors. Suma tots els èxits positius d'un jugador, resta els èxits negatius i retorna la valoració per minut del rendiment d'una jugadora	Puntuació/minut
TS_pct	Volum de tir real. Recull l'encert en tirs de 1, 2 i 3 punts	% tirs encertats/temporada
ThrPAr	Taxa d'intents de 3 punts (intents de 3p/intents tir totals)	Intents/temporada
FTr	Taxa de tirs lliures (intents tirs lliures/intents tir totals)	Intents/temporada

ORB_pct	Volum de rebots ofensius. Calcula el % de rebots ofensius disponibles que una jugadora agafa mentre és a pista	% rebots ofensius/temporada
TRB_pct	Volum de rebots totals. Calcula el % de rebots totals disponibles que una jugadora agafa mentre és a pista	% rebots totals/temporada
AST_pct	Volum d'assistències. Calcula el % d'assistències disponibles que una jugadora dona mentre és a pista	% assistències/temporada
STL_pct	Volum de pilotes robades. Calcula el % de pilotes robades disponibles que una jugadora roba mentre és a pista	% pilotes robades/temporada
BLK_pct	Volum de taps. Calcula el % de taps disponibles que una jugadora fa mentre és a pista	% taps/temporada
TOV_pct	Volum de pilotes perdudes. Calcula el % de pilotes perdudes que una jugadora perd cada 100 jugades.	% pèrdues/100 jugades
USG_pct	Volum d'ús. Estima el percentatge de jugades d'equip utilitzades per una jugadora mentre està a pista.	% jugades/temporada
OWS	Volum de victòries defensives. Estima el nombre de victòries que una jugadora produeix pel seu equip gràcies a la seva habilitat ofensiva.	% victòries defensives/temporada
DWS	Volum de victòries ofensives. Estima el nombre de victòries que una jugadora produeix pel seu equip gràcies a la seva habilitat defensiva.	% victòries ofensives/temporada
WS	Volum de victòries totals (*OWS – DWS*)	% victòries/temporada

Cal destacar que totes les variables numèriques referents al rendiment esportiu de les jugadores excepte el *PER* fan referència al total de la temporada i la variable *TOV_pct* que fa referència a les pilotes perdudes per cada 100 jugades. Per tal d'igualar les mètriques de totes les variables i aprofitant que la base de dades reporta el total de minuts jugats per temporada, s'han dividit tots els valors d'aquestes tres variables. Tot seguit s'han agrupat per jugadora les estadístiques de totes les temporades utilitzant la mediana amb l'objectiu de reduir la influència d'*outliers* i d'aquesta manera s'ha obtingut un *data frame* final de 953 jugadores i 14 variables.

4.2.1. *Clustering amb k-means*

Pel mètode de *clustering k-means* s'han utilitzat només les variables numèriques de l'últim *data frame* generat. En primer lloc, es fixa una llavor amb la funció *set.seed()*, ja que el *clustering* fa agrupacions que poden variar depenent de l'assignació aleatòria inicial dels centroides. Fixar una llavor ajuda a la reproductibilitat. Tot seguit, es calcula l'índex de Hopkins per mirar si les dades són clusteritzables amb la funció *get_clust_tendency()* (Hopkins & Skellam, 1954). S'obté un valor major a 0,98 i per tant sí que són clusteritzables.

Tal com s'ha explicat en la metodologia (apartat 2), el mètode *k-means* no estima el número de clústers i per tant obliga a fixar el nombre de clústers que l'algorisme determinarà a priori.

Per determinar-ho, s'ha utilitzat la funció de R *NbClust()*, que proporciona 30 índexs per a determinar el nombre de clústers i proposa a l'usuari el millor esquema de clustering a partir dels diferents resultats obtinguts variant totes les combinacions de nombre de clústers, mesures de distància i mètodes de clustering. S'ha obtingut que el nombre òptim de clústers podria ser *k = 2* i *k* = 3. La Figura 4.1 mostra la sortida de la funció *NbClust()*.

Figura 4.1.: Sortida 1 de la funció NbClust() del software

```
*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 7 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 4 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

Quan s'ha estudiat el repartiment de les dades per clústers, s'ha vist que l'observació 220 representava una dada bastant aïllada en *k* = 2 i un clúster aïllat en *k* = 3. Aquesta observació pertany a la jugadora Cori Chambers i degut a la falta de minuts jugats i els seus valors en les variables *PER* i *TOV_pct* no s'ajusta bé a cap clúster. Degut a la sensibilitat de l'algorisme *k-means* als *outliers*, s'ha decidit excloure aquesta jugadora de l'anàlisi. A partir d'aquí, s'ha tornat a començar el procés de nou des del càlcul de l'índex de Hopkins, que pren un valor de 0,94 i per tant les dades segueixen sent clusteritzables. La Figura 4.2 mostra la sortida de la funció *NbClust().*

Figura 4.2.: Sortida 2 de la funció NbClust() del software

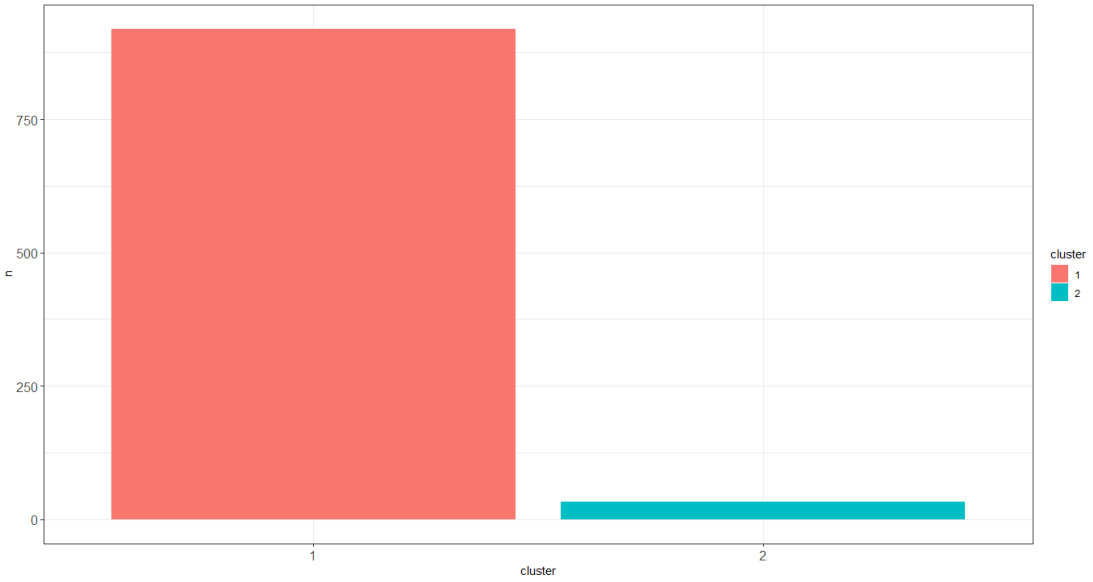
```
*******************************************************************
* Among all indices:
* 9 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 8 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2
```

Una vegada decidit el nombre de clústers (*k* = 2), s'ha creat una matriu amb les mitjanes de les mètriques per clúster de les variables utilitzades per construir el *clustering*, un vector que indica el clúster en el que es troba cada jugadora i la mida de cada clúster. El repartiment de les jugadores entre els dos grups determinats ha estat de 919 jugadores agrupades en el clúster 1 i 33 jugadores incloses en el clúster 2. Es presenta una taula de freqüències (Taula 4.5) i el Gràfic 4.5, que representa la diferència de tamany entre ambdós grups.

Taula 4.5. Taula de freqüències dels clústers.

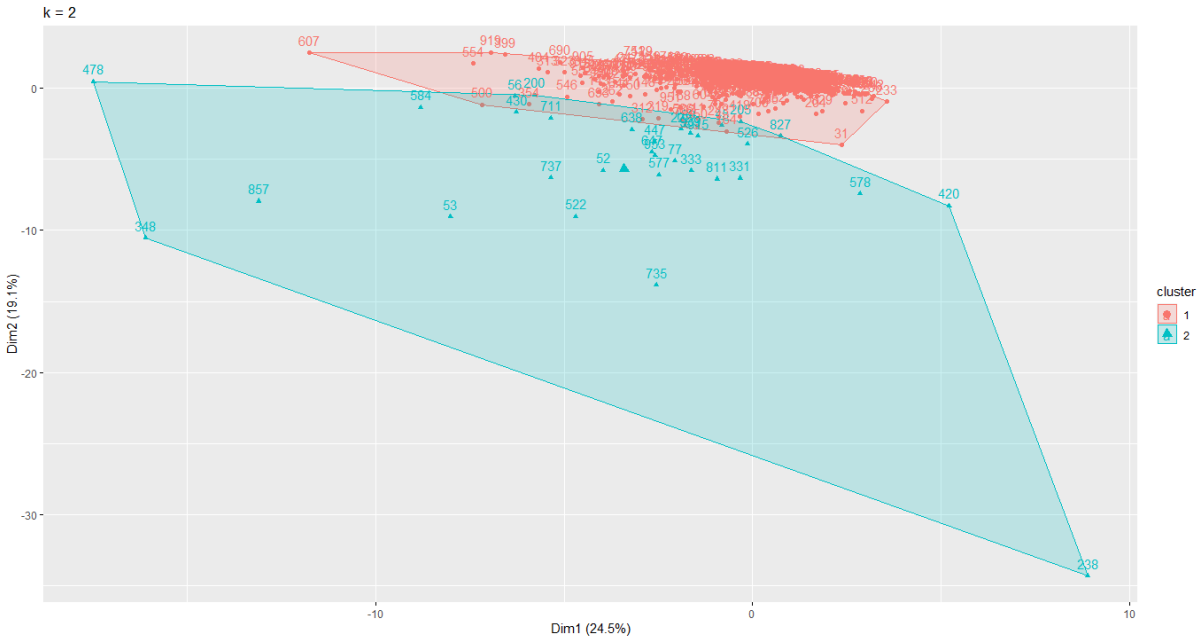
clúster	n	%
1	919	96,5%
2	33	3,5%

Gràfic 4.5.: Representació gràfica de la mida dels clústers per k = 2.



El Gràfic 4.6 mostra el repartiment de les dades en els clústers determinats representats en un pla de dos dimensions mitjançant components principals, que mostren solament un 43,6% de la variància explicada. Cal destacar que els números que apareixen al Gràfic 4.6 són els nombres de les files on es troben les dades de cada jugadora en la BBDD per tal de facilitar la seva visualització en el gràfic, ja que ficar els noms de les jugadores al gràfic el faria més confús.

Gràfic 4.6.: Gràfic definit per components principals per k = 2.



Com a complement a la representació gràfica dels clústers en el pla de dos dimensions, la Taula 4.6 mostra la correlació de les mètriques respecte la primera i la segona component principal. D'aquesta manera es pot acabar de veure la diferència entre els dos clústers abans d'interpretar amb més detall els perfils de les jugadores.

Taula 4.6.: Correlació entre les variables i la primera i segona component principal.

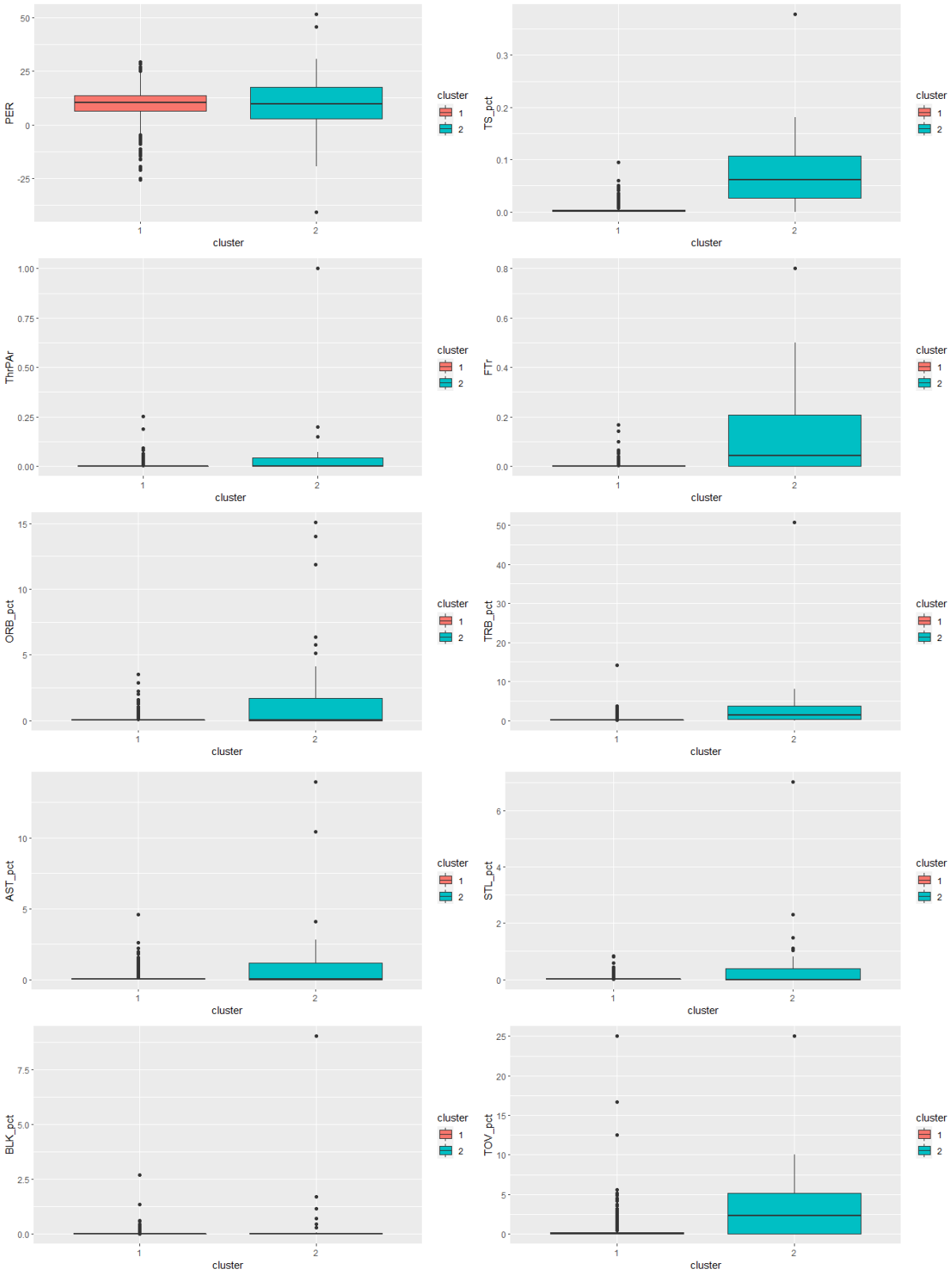
	PC1	**PC2**
PER	0,070724693	0,87909598
TS_pct	-0,829047854	0,26607995
ThrPAr	-0,277741185	-0,36431428
FTr	-0,448714286	-0,01868467
ORB_pct	-0,705235152	0,06131445
TRB_pct	-0,416834463	-0,04825278
AST_pct	-0,317005442	-0,31339763
STL_pct	-0,210390026	-0,03164831
BLK_pct	-0,657336866	0,39747111
TOV_pct	-0,323520091	-0,42368527
USG_pct	-0,521773495	-0,37504834
OWS	0,001848329	0,84903706
DWS	0,349310585	0,43613503
WS	-0,002588109	0,88073804

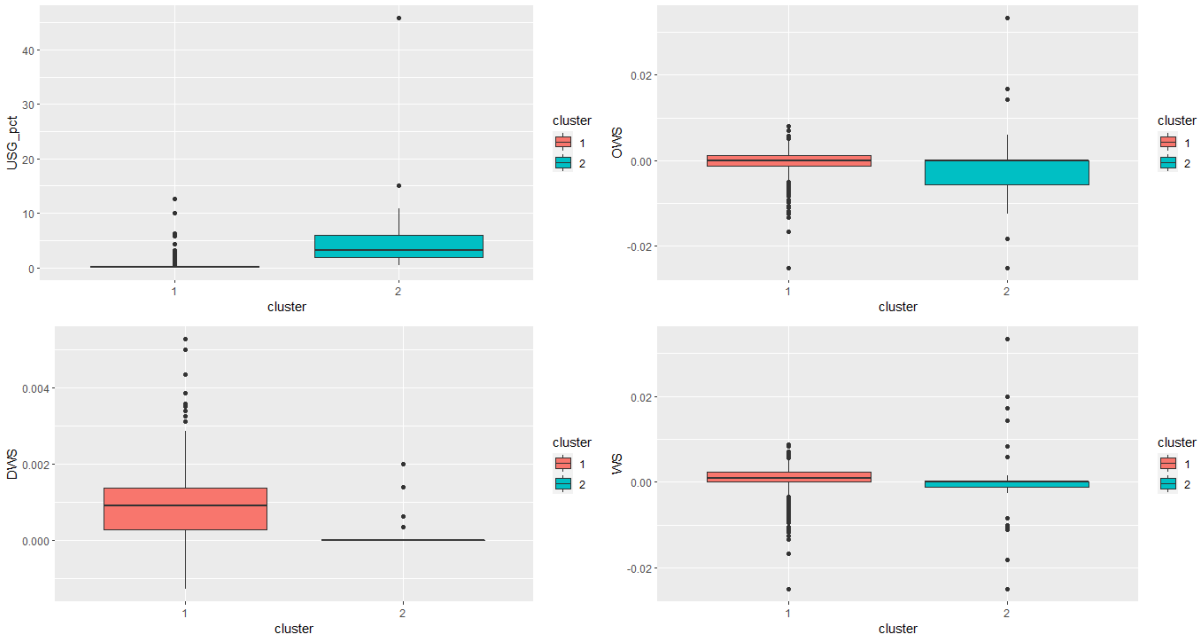
En aquesta taula es pot observar que gairebé totes les variables tenen una correlació negativa respecte la primera component i en la majoria de casos no és forta (excepte per les variables *TS_pct* i *ORB_pct*), per això les jugadores que es trobin més a l'esquerra i al voltant del valor 0 de l'eix d'abscisses presentaran un millor rendiment. Pel que fa a la segona component, els valors negatius no s'allunyen molt del 0 mentre que dels valors positius, les variables *PER, ows* i WS presenten una forta correlació positiva Per tant, les jugadores que es trobin a la part superior del Gràfic 4.6 i al voltant del 0 en l'eix d'ordenades tindran millors estadístiques. Amb aquests resultats sembla que les jugadores agrupades sobretot en el clúster 1 presentaran un millor rendiment.

Per veure les diferències entre els dos clústers i per tant definir-ne les característiques, s'ha utilitzat el *Mann-Whitney test* per les variables numèriques perquè aquestes no segueixen una distribució gaussiana (comprovat amb el *Shapiro-Wilk test*) i el *Chi-squared test* per les categòriques de les dades sense estandarditzar. Els resultats d'ambdós tests han mostrat que no hi ha diferències significatives entre clústers en les variables numèriques *PER, ThrPAr, ORB_pct, AST_pct* i *STL_pct*; mentre que sí que es presenten diferències en la resta de variables (Taula 4.4).

Juntament amb els resultats anteriors, el Gràfic 4.7 mostra un resum de les variables segons els diferents clústers obtinguts que pot servir d'ajuda per observar el perfil de joc de les jugadores.

Gràfic 4.7.: Distribució de les variables en cada clúster.





Nota: El Gràfic 4.7 s'ha hagut de dividir en dos pàgines.

En línies generals i observant els resultats dels *boxplots*, sembla que les jugadores agrupades en el clúster 2 tenen un perfil més de jugadores interiors, amb percentatges més alts d'encert en tirs de camp, amb un volum més gran de tirs lliures intentats per minut i amb un tant per cent de rebots totals agafats per minut més elevat que les jugadores incloses en el clúster 1. A més, També tenen una incidència en el joc més alta, és a dir, intervenen més en les jugades de l'equip mentre són a pista i això pot dur a un major volum de pèrdues de pilota tal com mostra el *boxplot* de la variable *TOV_pct*. Pel que fa a les jugadores agrupades en el clúster 1, presenten millors estimacions en el nombre de victòries tant ofensives com defensives que produeixen pel seu equip gràcies a les seves habilitats. En la resta de variables mostren valors practicament nuls excepte alguns valors atípics ubicats per sobre del màxim i això fa més difícil definir-lo.

Els resultats d'una revisió més concreta per cada variable sense estandaritzar són:

Puntuació de la eficiència de les jugadores

A partir del *Mann-Whitney test* s'ha demostrat que la variable *PER* no presenta diferències significatives entre ambdós clústers, és a dir, que les jugadores agrupades en cada clúster no tenen una puntuació per minut estadísticament diferent entre elles en referència a la seva eficiència ofensiva.

Volum de tir real per minut

La variable *TS_pct* mostra valors molt diferents entre ambdós clústers. Les jugadores incloses en el clúster 1 estan agrupades gairebé al 0 excepte algun valor atípic més elevat. Les jugadores incloses en el clúster 2 presenten un volum de tir real més elevat, és a dir, les jugadores tenen un percentatge d'encert en tirs de camp més alt. Aquests valors estan distribuïts entre el 0 i el 0,18 aproximadament, amb una mediana de gairebé 0,6. Destaca en aquest clúster la jugadora Dana Wynne, amb un tant per cent d'encert de tirs reals per minut de gairebé el 38%.

Taxa d'intents de 3 punts per minut

A partir del *Mann-Whitney test* s'ha demostrat que la variable *ThrPAr* no presenta diferències significatives entre ambdós clústers, és a dir, que les jugadores agrupades en cada clúster no tenen un percentatge de tirs intentats de 3 punts per minut estadísticament diferent entre elles.

Volum de tirs lliures per minut

Les jugadores agrupades en el clúster 1 de la variable *FTr* presenten un volum de tirs lliures intentats per minut de pràcticament 0, mentre que en el clúster 2 el 50% de les jugadores encerten entre un 0 i un poc més del 20% dels tirs lliures que tiren per minut, amb un màxim d'aproximadament el 50% d'intents per minut. El valor atípic de la part superior de la gràfica pel clúster 2 representa a la jugadora Renee Robinson, que tira el 80% dels seus tirs per minut des de la línia de tir lliure.

Volum de rebots ofensius per minut

A partir del *Mann-Whitney test* s'ha demostrat que la variable *ORB_pct* no presenta diferències significatives entre ambdós clústers, és a dir, que les jugadores agrupades en cada clúster no tenen un percentatge de rebots ofensius disponibles agafats per minut estadísticament diferent entre elles.

Volum total de rebots per minut

En els *boxplots* de la variable *TRB_pct* s'observa que les jugadores incloses en el primer clúster presenten valors de gairebé 0 excepte algun valor atípic com en les variables anteriors; mentre que en segon clúster, el 75% de les jugadores tenen un volum total de rebots per minut més gran que 0, amb un màxim d'aproximadament el 8% del total de rebots disponibles. En aquest clúster destaca la jugadora Ambrosia Anderson, que va agafar el 51% dels rebots totals disponibles per minut mentre era a pista.

Volum d'assistències per minut

A partir del *Mann-Whitney test* s'ha demostrat que la variable *AST_pct* no presenta diferències significatives entre ambdós clústers, és a dir, que les jugadores agrupades en cada clúster no tenen un volum d'assistències disponibles repartides per minut estadísticament diferent entre elles.

Volum de pilotes robades per minut

A partir del *Mann-Whitney test* s'ha demostrat que la variable *STL_pct* no presenta diferències significatives entre ambdós clústers, és a dir, que les jugadores agrupades en cada clúster no tenen un tant per cent de pilotes per minut robades diferent estadísticament entre elles.

Volum de taps per minut

Ambdós *boxplots* de la variable *BLK_pct* presenten gairebé el 100% de les observacions al voltat del 0 tot i que sembla que el clúster 1 mostra valors lleugerament més elevats que el 2. Tot i això, en el clúster 2 destaca un valor atípic que pertany a la jugadora Dana Wynne, que fa un 9% dels taps disponibles per minut.

Volum de canvis de possessió per minut

La variable *TCV_pct* mostra valors molt diferents entre ambdós clústers. Les jugadores incloses en el clúster 1 estan agrupades al voltant del 0 igual que en les variables anteriors. Les jugadores agrupades al clúster 2 presenten més canvis de possessió, és a dir, perden més pilotes per minut. Aquests valors estan distribuïts entre el 0 i el 10. En ambdós clústers destaquen dues jugadores amb un volum de pèrdues de pilota per minut bastant elevat. Pel primer clúster és Kim Gessig i pel segon la jugadora és Kellie Jolly Harper, ambdues amb un percentatge de 25 pilotes disponibles perdudes per minut.

Volum d'ús per minut

Les jugadores incloses en el clúster 1 de la variable *USG_pct* estan agrupades al voltant del zero excepte alguna jugadora amb un valor atípic superior al màxim. El clúster 2 mostra un volum d'ús més gran. En aquest clúster, les jugadores tenen una major incidència en la resolució de jugades, que presenta un mínim lleugerament superior a 0 i un màxim de poc més del 10%. Com més elevat sigui aquest percentatge d'ús, més incidència té la jugadora. La que més destaca en aquesta estadística és Imani Wright, amb un 45% d'incidència per minut en el joc de l'equip.

Volum de victòries ofensives per minut

En el *boxplot* del clúster 1 de la variable *OWS* es pot observar que la meitat de les jugadores no sumen en el nombre de victòries que produeixen pel seu equip amb la seva habilitat defensiva. Tot i això, tots els valors són pràcticament 0. En el clúster 2, són el 75% de les jugadores les que tenen un percentatge de victòries ofensives negatiu i lleugerament inferior comparat amb el primer clúster i per tant el que fan és perjudicar a l'equip. Destaca amb una aportació positiva en el clúster 2 la jugadora Dana Wynne i amb una aportació negativa les jugadores Tanae Davis-Cain agrupada al clúster 2 i Martina Weber inclosa en el clúster 1.

Volum de victòries defensives per minut

El clúster 1 de la variable DWS presenta que almenys el 75% de les jugadores s'estima que produeixen un nombre de victòries defensives positu pel seu equip, això significa que ajuda positivament al seu equip gràcies a la seva habilitat defensiva. En el clúster 2, no sembla que les jugadores no sumin ni restin gràcies a la seva habilitat defensiva. Tot i això, en ambdós casos els valors són molt propers a 0.

Volum de victòries per minut

La variable WS mostra la diferència entre els valors de les dues variables anteriors. Per aquest motiu, el boxplot del clúster 1 presenta valors positius en el 75% de les observacions i *boxplot* del clúster 2 presenta valors negatius en el mateix percentatge de jugadores.

A partir d'aquests resultats desglossats i resumits dels clústers per cada variable, es pot acabar de definir les característiques dels clústers i el perfil de les jugadores incloses en cada grup:

Clúster 1

Les jugadores incloses en el clúster 1 són les jugadores que s'estima que produeixen més victòries tant ofensives com defensives pel seu equip gràcies a les seves habilitat amb i sense la pilota. A part d'això, no presenten cap més diferència significativa amb les jugadores agrupades al clúster 2 i per tant és difícil definir-ne un perfil concret. En aquest clúster hi ha incloses jugadores com Diana Taurasi i Sue Bird, considerades dos de les millors bases del món en la última dècada i Breanna Stewart, considerada la millor jugadora del món actualment.

Clúster 2

Les jugadores agrupades en el clúster 2 són jugadores amb un percentatge d'encert en tirs de camp per minut més gran que la resta, la majoria d'aquest encert segurament vingui de la quantitat de tirs lliures que tiren per minut. Aquest fet, juntament amb el volum de rebots totals agafats per minut i la incidència en el joc, fa pensar que són jugadores més grans físicament, capaces de jugar en posicions interiors per guanyar avantatge en la posició de tir i en les faltes rebudes que portin a la línia de tir lliure. També és habitual en aquest perfil de jugadores tenir un percentatge més elevat de pèrdua de pilotes. En aquest clúster es poden trobar jugadores com Angelina Wolvert i Kym Hampton, jugadores grans amb joc interior.

5. Discussions

5.1. *Discussió dels resultats de la RS*

Els 54 articles seleccionats en aquesta revisió sistemàtica mostren que la primera publicació de la literatura científica que utilitza una tècnica *clustering* en algun camp de les ciències de l'esport és de l'any 1981 i que fins 26 anys després no es va tornar a publicar res. A partir del 2007, la publicació d'aquest tipus d'articles ha estat pràcticament ininterrompuda i la tendència de publicació amb tècniques de *clustering* en el camp de l'esport té una tendència creixent. Aquesta tendència no sembla coincidir per exemple amb l'interès de cerca dels mètodes *clustering* proporcionada per *Google Trends*, que mostra un decreixement l'any 2006 i posteriorment es mostra estable però amb un interès baix.

La meitat dels articles publicats lligats a les ciències de l'esport estudien el rendiment esportiu (*Sports performance analysis*) dels participants, en 16 articles (29,7%) s'estudia algun aspecte relacionat amb el moviment (*Movement integration*) i s'ha trobat amb menys freqüència (16,7%) algun article que relacioni l'esport amb les ciències de la salut. Tal com mostra al seu llibre Albert et al. (2017), en alguns tipus d'esports com el futbol (*soccer*) (16,7%), el bàsquet (7,4%) i el tennis (11,1%) entre d'altres es poden estudiar aspectes que ajudin a millorar el rendiment dels atletes ja sigui descrivint el patró de les posicions al camp o descrivint la prevenció de lesions (Siedlik et al., 2016; Anıl Duman et al., 2021).

Del total d'articles seleccionats, en un 22,0% no es reporta el país de provinença de les dades i els que sí són majoritàriament d'Estats Units (13,0%), Austràlia (9,3%) i Regne Unit (13%). Els participants són amb més freqüència atletes professionals (61,1%) que amateurs (29,7%), segurament degut a la major facilitat per aconseguir les dades per dur a terme els estudis i la influència del camp *Sports performance analysis* o *Sports analytics* més desenvolupat tant a nivell acadèmic com en la indústria de l'esport (Alamar, 2013; Glickman, 2017).

En 22 dels 54 articles seleccionats (40,7%) els participants són homes mentre que en el 37,1% dels articles hi ha individus d'ambdós sexes, fet que sembla interessant destacar ja que tant les característiques físiques com l'estil de joc dels dos sexes són diferents i per tant els resultats d'un sexe no són completament extrapolables per l'altre (Thibault et al., 2010).

Pel que fa a les tècniques *clustering*, tal com s'ha vist als resultats, la majoria d'estudis utilitzen el *clustering* jeràrquic (35,2%), amb diversos algoritmes, i el *clustering* particional (29,7%), íntegrament amb l'algoritme *k-means*, per agrupar les observacions i extreure'n resultats. Tot i això, els autors d'alguns articles opten per nombrar el tipus de tècnica *clustering* que utilitzen però no reporten ni el nom del mètode utilitzat (31,3%) ni el mètode per decidir el nombre de clústers (35,2%) i això pot complicar l'avaluació de l'adequació dels enfocaments utilitzats pels anàlisis dels estudis, ja que existeixen diferents tipus de mètodes més o menys eficients segons el tipus de dades i que poden reportar resultats divergents (Singh & Gosain, 2013).

Els *softwares* més utilitzats, ja sigui de forma única o combinant-ne més d'un en un mateix estudi, són l'R (25,9%) i el MATLAB (20,4%) i gairebé el 15% dels articles no reporten el *software*. De la mateixa manera que passa amb les tècniques *clustering*¸ tot i reportar-se el *software* en la majoria d'estudis, en el 72,2% dels articles no es reporta el *package* utilitzat i

com a conseqüència tampoc es pot validar l'adequació dels enfocaments. A més, el fet de reportar aquest tipus d'informació pot ajudar a conèixer quins són els *softwares* i els *packages* més utilitzats i segurament més útils i eficients per aquest tipus de tècniques.

Cal destacar que en cap article es comparteixen les dades utilitzades en els estudis, els codis dels *softwares* o el repositori de les dades o el codi. Aquest succés comporta que sigui encara més difícil reproduir o replicar els estudis i validar els anàlisis i els resultats, i és un problema a l'alça en la literatura científica. Tan gran és el problema del *data sharing* que en diversos camps científics es parla d'una crisi de la reproductibilitat de la investigació científica que no només afecta aspectes científics sinó que també ètics i que és important solucionar (Resnik & Shamoo, 2016; Schwab et al., 2022).

Una de les conclusions principals que s'extreu d'aquest estudi és la possibilitat de millora en reportar la informació en els articles tant de les tècniques d'anàlisi *clustering* i les seves característiques (tipus de mètode *clustering* utilitzat (3,7%), nom del mètode *clustering* utilitzat (31,3%), mètode per decidir el número òptim de clústers (35,2%)) com d'aspectes computacionals i científics (*software* utilitzat (14,9%), *package* utilitzat (72,2%)); a més de compartir les dades i el codi utilitzat en els articles; sobretot amb la tendència creixent de publicació d'articles que utilitzen mètodes *clustering* en les ciències de l'esport.

Limitacions de la RS

Com a limitació d'aquesta revisió sistemàtica es pot destacar el fet que solament s'han mirat els articles publicats en revistes del JCR més el JQAS i el JSA. Això fa que la revisió només s'hagi centrat en revistes esportives que poden aplicar tècniques *clustering* i per tant es descarten revistes de *Machine Learning*, de tècniques no supervisades o de mètodes *clustering* escrites per estadístics o *data scientists* que poden aplicar-se utilitzant dades reals o simulades en algun dels camps de les ciències de l'esport. Tot i això, es creu que amb totes les revistes incloses en la revisió s'ha cobert la majoria de revistes en aquest camp que on es poden utilitzar tècniques *clustering.*

Tot i així, cal destacar que és la primera vegada segons es coneix que s'ha fet una revisió sistemàtica d'articles que apliquen tècniques *clustering* en les ciències de l'esport. Això pot ser un fet molt interessant per el desenvolupament i la millora dels articles amb aquestes mateixes característiques que es puguin publicar posteriorment, a més de poder servir d'ajuda a fer créixer l'estadística en els diferents camps de les ciències de l'esport.

5.2. *Discussió dels resultats de l'estudi de cas*

A partir de la base de dades que mostra estadístiques per temporada de les jugadores dels diferents equips de la WNBA, s'ha realitzat el mètode *clustering k-means* per agrupar les jugadores segons les seves estadístiques i extreure'n els seus perfils.

Com a resultat del mètode, s'han agrupat les jugadores en dos clústers de dimensions bastant diferents. El clúster 1, representat amb jugadores que ajuden a sumar victòries als seus equips gràcies a la seva habilitat ofensiva i defensiva, representa el 96,5% del total de jugadores de la BBDD; mentre que el clúster 2 representa solament l'3,5% del total i tenen un perfil de jugadores grans de caràcter ofensiu. Aquest repartiment, tot i ser molt desequilibrat, sembla

relativament acurat amb l'estil de joc del bàsquet americà actual ot ti ser més propi de la lliga masculina; ja que veure un partit de bàsquet d'alguna de les lligues nord-americanes és veure un partit ple de possessions curtes i ràpides, amb bastant tir exterior i de sota l'aro i una defensa que deixa jugar fàcilment al jugador o jugadora de l'equip contrari. Per aquest motiu, variables com el percentatge d'encert, el volum de pèrdues de pilota i la incidència tenen valors més elevats.

Limitacions de l'estudi de cas

Podria ser que amb estadístiques més concretes del joc com per exemple els punts per partit, l'encert d'un, dos i tres punts per separat i no solament l'intentat, els rebots defensius i les pilotes recuperades s'aconseguís definir perfils de jugadores més acurats i més adequats a l'estil de joc d'aquest tipus de lliga, ja que les estadístiques reportades en la base de dades utilitzada per l'estudi de cas representaven de manera bastant general les estadístiques reals recollides en un partit de bàsquet.

6. Conclusions

Els últims anys han estat anys de creixement en el camp de l'estadística esportiva, provocant l'increment de publicacions en revistes esportives d'articles que tracten de l'aplicació de mètodes estadístics en els diferents camps de les ciències de l'esport. En aquesta revisió sistemàtica s'han volgut estudiar tots els articles escrits en anglès i publicats en revistes incloses en el JCR o les revistes JQAS i JSA que utilitzen algun tipus de tècnica clústering en qualsevol camp de les ciències de l'esport.

De la RS realitzada es pot concloure que gran part dels articles publicats que compleixen les característiques anteriors han estat publicats en els últims 15 anys i que bàsicament busquen realitzar estudis sobre el rendiment esportiu o la salut dels atletes utilitzant mètodes de *hierarchical* o *partitional clustering*. Tot i això, tots els articles pertanyents a aquesta revisió tenen bastant marge de millora, ja sigui per la transparència de les dades o el codi utilitzat o per la claredat amb la que es reporten les tècniques utilitzades i els resultats, donat que hi ha molts articles que no reporten una o més característiques relacionades amb les tècniques *clustering* i això dificulta la reproductibilitat i la validació dels estudis.

Finalment, tal com s'ha mostrat en l'estudi de cas realitzat, les tècniques *clustering* com el *k-means* són una eina molt útil per estudiar les estadístiques d'esports com el bàsquet i les característiques dels seus jugadors i jugadores i això pot beneficiar als propis jugadors, a entrenadors, equips i lligues a millorar la seva qualitat i el seu rendiment físic i esportiu.

7. Bibliografia

[1] Alamar, B., & Oliver, D. (2013). *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers* (Illustrated ed.). Columbia University Press.

[2] Alamar, B. (2013). Sports analytics. In *Sports Analytics*. Columbia University Press.

[3] Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (2017). *Handbook of Statistical Methods and Analyses in Sports* (1.ª ed.). Chapman and Hall/CRC.

[4] Anıl Duman, E., Sennaroğlu, B., & Tuzkaya, G. (2021). A cluster analysis of basketball players for each of the five traditionally defined positions. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 175433712110620. https://doi.org/10.1177/17543371211062064

[5] *ASA Community*. (s. d.). Higher Logic, LLC. https://community.amstat.org/sis/journals

[6] Boehmke, B., & Greenwell, B. (2019). *Hands-On Machine Learning with R (Chapman & Hall/CRC The R Series)* (1.ª ed.). Chapman and Hall/CRC.

[7] Chandran, A., Brown, D., Nedimyer, A. K., & Kerr, Z. Y. (2019). Statistical Methods for Handling Observation Clustering in Sports Injury Surveillance. *Journal of Athletic Training*, *54*(11), 1192–1196. https://doi.org/10.4085/1062-6050-438-18

[8] *Concussion (2015)*. (2016, 17 marzo). IMDb. https://www.imdb.com/title/tt3322364/?ref_=nv_sr_srsg_0

[9] Dangeti, P. (s. d.). *Numerical Computing with Python*. O'Reilly Online Learning. https://www.oreilly.com/library/view/numerical-computing-with/9781789953633/381f85ce-024d-4480-b7e7-39e0fbc6a0fd.xhtml

[10] Fernández, D. (s.d.). Material del *Master in Innovation and Research in Informatics*. UPC.

[11] Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611–631. https://doi.org/10.1198/016214502760047131

[12] Glickman, M. E. (2017). Discussion of practical problems in Sports Analytics. In *JQAS Invited Session 490 at the Joint Statistical Meetings*. Baltimore, MD.

[13] Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

[14] *Guides: Systematic Reviews: Protocol*. (s. d.). UNISA. https://guides.library.unisa.edu.au/SystematicReviews/Protocols

[15] Hopkins, B., & Skellam, J. G. (1954). A New Method for determining the Type of Distribution of Plant Individuals. Annals of Botany, 18(2), 213–227. https://doi.org/10.1093/oxfordjournals.aob.a083391

[16] *K-means Cluster Analysis · UC Business Analytics R Programming Guide*. (s. d.). University of Cincinnati. https://uc-r.github.io/kmeans_clustering#fn:kauf

[17] Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. *Technometrics*, *34*(1), 111. https://doi.org/10.2307/1269576

[18] Koutroumbas, K., & Theodoridis, S. (2008). *Pattern Recognition* (4.ª ed.). Academic Press.

[19] Kubat, M. (2021). *An Introduction to Machine Learning* (3rd ed. 2021 ed.). Springer.

[20] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/tit.1982.1056489

[21] MacQueen, J. B. & Western Management Science Inst Univ of California Los Angeles. (1966). *Some Methods for Classification and Analysis of Multivariate Observations*. Defense Technical Information Center.

[22] McNicholas, P. D. (2016). *Mixture Model-Based Classification*. Amsterdam University Press.

[23] Miller, T. W. (2015). *Sports Analytics and Data Science: Winning the Game With Methods and Models (FT Press Analytics)*. Ft Pr.

[24] Miscellaneous Clustering Methods. (2011). *Cluster Analysis*, 215–255. https://doi.org/10.1002/9780470977811.ch8

[25] Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[26] *Moneyball (2011)*. (2011, 17 noviembre). IMDb. https://www.imdb.com/title/tt1210166/?ref_=fn_al_tt_1

[27] Musa, M. R., Taha, Z., Majeed, P. A., & Abdullah, M. R. (2018). *Machine Learning in Sports: Identifying Potential Archers (SpringerBriefs in Applied Sciences and Technology)* (1st ed. 2019 ed.). Springer.

[28] Resnik, D. B., & Shamoo, A. E. (2016). Reproducibility and Research Integrity. Accountability in Research, 24(2), 116–123. https://doi.org/10.1080/08989621.2016.1257387

[29] Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., & Koffel, J. B. (2021). PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews*, *10*(1). https://doi.org/10.1186/s13643-020-01542-z

[30] Schwab, S., Janiaud, P., Dayan, M., Amrhein, V., Panczak, R., Palagi, P. M., Hemkens, L. G., Ramon, M., Rothen, N., Senn, S., Furrer, E., & Held, L. (2022). Ten simple rules for good research practice. *PLOS Computational Biology*, *18*(6), e1010139. https://doi.org/10.1371/journal.pcbi.1010139

[31] Siedlik, J. A., Bergeron, C., Cooper, M., Emmons, R., Moreau, W., Nabhan, D., Gallagher, P., & Vardiman, J. P. (2016). Advanced Treatment Monitoring for Olympic-Level Athletes Using Unsupervised Modeling Techniques. *Journal of Athletic Training*, *51*(1), 74–81. https://doi.org/10.4085/1062-6050-51.2.02

[32] Singh, D., & Gosain, A. (2013, August). A comparative analysis of distributed clustering algorithms: A survey. In 2013 International Symposium on Computational and Business Intelligence (pp. 165-169). IEEE.

[33] Smyth, D. (2022, 2 febrero). *The Job Market for Sports Analysts*. Work - Chron.Com. https://work.chron.com/job-market-sports-analysts-24524.html

[34] Thibault, V., Guillaume, M., Berthelot, G., Helou, N. E., Schaal, K., Quinquis, L., Nassif, H., Tafflet, M., Escolano, S., Hermine, O., & Toussaint, J. F. (2010). Women and Men in Sport Performance: The Gender Gap has not Evolved since 1983. Journal of sports science & medicine, 9(2), 214–223.

[35] Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/bf02289263

[36] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

[37] Wikipedia contributors. (2022, 28 abril). *K-means clustering*. Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering

8. Annex

8.1. *Codi R*

Link GitHub: https://github.com/MontsePlensaSantallusia/TFG_codiR.git