Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v17n3p653

Reporting of clustering techniques in sports sciences: a scoping review
By Fernández et al.

15 December 2024

Electronic Journal of Applied Statistical Analysis
Vol. 17, Issue 03, December 2024, 653-675
DOI: 10.1285/i20705948v17n3p653

Reporting of clustering techniques in sports sciences: a scoping review

Daniel Fernández[a], Martí Casals*[b,c], Martí Oliver[a], Montse Plensa[a], and Marica Manisera[d]

[a] *Department of Statistics and Operations Research (DEIO), Universitat Politècnica de Catalunya · BarcelonaTech (UPC), Barcelona, Spain.*
[b] *Sport and Physical Activity Studies Centre (CEEAF), Faculty of Medicine, University of Vic-Central University of Catalonia (UVic-UCC), Barcelona, Spain.*
[c] *National Institute of Physical Education of Catalonia (INEFC-UB), Spain*
[d] *BODaI-Lab, Department of Economics and Management, University of Brescia, Brescia, Italy.*

15 December 2024

Multivariate statistical methods are among the most used ones in sports sciences with clustering methods emerging as prominent unsupervised learning techniques. This study presents a scoping review of original articles utilizing clustering techniques in sports sciences, following the PRISMA-SCR guidelines. A comprehensive search across various databases using the boolean "AND" combination of "clustering" and "sport" yielded 278 articles. Notably, 86.7% of these articles were published within the last 14 years, with a predominant focus (66.2%) on sports performance analysis. The majority of studies included professional athletes (56.4%), with football/soccer, basketball, and tennis being the most commonly studied sports, representing 12.2%, 7.5%, and 2.2% of the selected articles, respectively. Hierarchical clustering was the most frequently used method (31.6%), followed by the k-means algorithm for partitional clustering. However, the clustering method was not reported in 26.6% of the articles, and 55.0% did not specify the criterion used for determining the optimal number of clusters. Moreover, more than 85% of the articles lacked computational details related to data reproducibility. These findings underscore the urgent need for substantial improvement in reporting practices regarding the methodology, algorithms, criteria for cluster identification, and software usage in sports science literature.

*Corresponding author: Ctra. de Roda, 70, 08500 Vic, Spain.
emails: marticasals@gmail.com, marti.casals1@umedicina.cat

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

keywords: Cluster analysis, unsupervised learning, scoping review, sports sciences.

1 Introduction

The interest in sports statistics has been steadily increasing over the last years. The current technological innovations available, the impact of some films (e.g.: Miller (2011); Landesman (2015)), the development of podcasts (e.g.: Measurables (`https://shorturl.at/rLQRS`) and Counterpoints (`https://shorturl.at/avEU5`)), and the establishment of new sports analytics departments and laboratories in universities (e.g.: Harvard (`https://shorturl.at/acgBZ`), Simon Fraser (`https://shorturl.at/cglzD`), Carnegie Mellon (`https://www.stat.cmu.edu/cmsac/`), California (`https://shorturl.at/bpwL6`), Brescia (`https://bodai.unibs.it/bdsports/`)) describe some of the main factors that have caused this growth. Additionally, it is relatively common to find peer-reviewed publications related to this field in statistical journals recognized by the American Statistical Association (ASA) such as the Journal of Quantitative Analysis in Sports, Chance, and The American Statistician, and more conferences such as those organized by MathSport International and Joint Statistical Meeting (JSM) or the seminars and webinars organized by the S-Training group (Sports – Training and Research in DAta ScIence Methods for ANalytics and INjury Prevention Group, `https://s-training.eu/Homepage.html`), for instance. In fact, the Section on Statistics in Sports (SIS) of the ASA was founded at one of the JSM's contributed sessions back in 1992 to respond to the need to encourage the development of statistics and its applications in sports (`https://community.amstat.org/sis/journals`). Also, the Special Interest Group in Sports Statistics of the International Statistical Institute (`https://www.isi-web.org/committee/special-interest-group-sports-statistics`) aims to promote the understanding, development and good practice of sports statistics worldwide. Data science and sports statistics are one of the most in-demand careers Smyth (2022).

The skills of the professionals working in those fields are mainly focused on extracting useful information from a large volume of data and on knowing the hidden patterns of combinations of players, teams, leagues, and injuries among others via statistical and computational-thinking skills Alamar (2013); Miller (2015). The most used statistical techniques are those developed at the borderline between computer science and statistics and then related to statistical learning or machine learning methods focusing mainly on descriptive multivariate analysis and, particularly, unsupervised learning approaches Musa et al. (2018); Chandran et al. (2019); James et al. (2021, 2023). Statistical and machine learning techniques are classified as supervised and unsupervised learning approaches. One of the main and relevant differences between them is that supervised learning methods use labeled data (for each observational unit, the predictors measurements are associated with a response measurement), which allow us to obtain predictions and optimal classifiers (see e.g. support vector machines, random forests, neural networks and even linear or logistic regression). On the contrary, the unsupervised learning techniques (e.g., k-means cluster analysis, principal component analysis, and -again-

neural networks) use unlabelled data (for each observational unit, only the predictors measurements are observed) and they are focused on the understanding the relationships among variables, among observations, between variables and observations, on the description of the data, reducing the dimensionality, and finding hidden groups. Among the most popular approaches used in sports statistics, there are unsupervised methods and, in particular, cluster analysis (a.k.a. clustering) Kaufman and Rousseeuw (2009); Everitt et al. (2011). Clustering methods intend to determine the unknown number of groups in data sets in such a way that the algorithm allocates the subjects of interest in homogeneous groups, i.e., with similar profiles within each group and divergent from the rest of the groups Everitt et al. (2011); Kubat and Kubat (2021). Homogeneity, similarity and divergency are evaluated with respect to some variables of interest, measured on the available subjects. For example, in sports analytics, clustering can be used to identify groups of similar players based on their performance statistics. This can help coaches and analysts to understand the strengths and weaknesses of different players and to make more informed decisions about player selection and team strategy. Clustering can also be employed to identify groups of players who exhibit similar playing styles, in order to redefine player roles in team sports. For instance, in basketball, the traditional five positions may no longer adequately represent the evolving styles of play, and clustering can facilitate the emergence of new roles Bianchi et al. (2017). In general, clustering can be used in sports analytics to group players, teams, matches, game fractions for a wide variety of aims (see, among others, (Hodge and Petlichkoff, 2000; Ball and Best, 2007; Murray and Hunfalvay, 2017; Zhang et al., 2018; Núñez et al., 2019; Hautbois et al., 2020; Zuccolotto and Manisera, 2020; Anzer et al., 2021; Cartigny et al., 2021; van der Linden et al., 2021; Dalton-Barron et al., 2022; Muniz and Flamand, 2022; Carpita et al., 2023)).

Three main types of clustering methods can be distinguished (sorted by mathematical complexity): hierarchical clustering (where dendrograms are used to classify), partitional methods (among which the most common algorithms are k-means, (MacQueen et al., 1967; Lloyd, 1982); and k-medoids, (Koutroumbas and Theodoridis, 2008), and probabilistic clustering, in which the model-based clustering (Fraley and Raftery, 2002; McNicholas, 2016)) is the most popular one. Hierarchical and partitional methods are mostly distance-based or dissimilarity-based, i.e. calculates distances or dissimilarities to obtain the allocation of the subjects into the groups. The probabilistic clustering fits underlying probabilistic model distributions to determine that allocation. Another relevant difference is that hierarchical and partitional methods are classified as hard clustering techniques, i.e. one observation can be classified in only a single group. Instead, the probabilistic clustering assigns probabilities that represent the likelihood of a subject belonging to each cluster (soft clustering). Therefore, the latter can be useful when dealing with data that may not clearly belong to a single cluster or when dealing with overlapping clusters. By incorporating probability, it provides a more flexible representation of cluster assignments. A recent overview of clustering methods, including recent proposals such as spectral clustering, and the topic of classification of unstructured data is covered in (Aggarwal and Reddy, 2014). In the scientific literature on the field of sports, the book that stands out above all is the Handbook of Statistical Methods and

Analyzes in Sports Albert et al. (2017), which offers the most up-to-date overview of research in this field. This book mentions how various clustering methods are used to analyze some features of different sports such as baseball and basketball Albert et al. (2017). Thus in baseball, the k-means and model-based clustering algorithms are used to classify the different types of throws with different speeds and movements and whether they will be well hit or not. Whereas in basketball, clusters allow us to identify the different types of players by studying the space where they move on the court to decide defensive pairings.

The main objective of this study is to perform a scoping review (from now on, SR) of the application of clustering techniques in the field of sports sciences. The aim is to map the existing literature on this topic, identify research gaps, and provide an overview of the available evidence. We assess the quality of the reporting and the information reported in original articles in the field of sports sciences through PRISMA-SCR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines Tricco et al. (2018). To the best of our knowledge, this would be the first time of such a SR. The plan of the article is as follows: Section 2 describes the Methods, including the study design, the search strategy (i.e., the inclusion and exclusion criteria to select the studies), the data extraction, and the identification of the studies. Results summarizing the main insights about the general characteristics of the selected articles, of the sport, and of the clustering methods are described in Section 3, and we conclude with a discussion in Section 4.

2 Methods

2.1 Study design

A scoping review of the performance of clustering methods in the articles of the field of sports sciences was carried out following the recommendations of the PRISMA-SCR guidelines Tricco et al. (2018).

Search strategy

The search for the scoping review was performed on May 7th, 2023, using the following databases: Web of Science (WoS), PubMed, SportDiscus, and CINAHL (Cumulative Index to Nursing and Allied Health Literature). All databases were searched using the same combination of keywords and Boolean operators: sport* and ("clustering" or "cluster analysis").

Selection of studies

As study eligibility criteria, the following inclusion and exclusion criteria in the scoping review were defined:

- *Inclusion criteria*: The articles included had to be original articles, written in English, applying clustering techniques. Additionally, those studies have to be pub-

lished by sports science journals included in the list of journals published by Journal Citation Reports (JCR) in 2020. We also included in the scoping review two more journals, which did not appear in the JCR list: the Journal of Quantitative Analysis in Sports (JQAS; `https://www.degruyter.com/journal/key/jqas/html`) and the Journal of Sports Analytics (JSA; `https://journalofsportsanalytics.com/`) which can be found and also cited in the ASA sports section Swartz (2020).

- *Exclusion criteria*: Dissertations, conference proceedings, and non-original articles such as editorials, opinion or commentary letters, and reviews were not included in the SR.

The following two-step procedure was applied as the selection process. First, the articles were selected based on their title and abstract. Second, if the title and/or abstract were not clear enough, a full-text revision of the article was undertaken to assess whether they met the eligibility criteria.

2.2 Data extraction

The information collected in the selected studies was grouped into three categories: a) general characteristics of the selected articles; b) characteristics of the particular sport studied; and c) characteristics of the clustering techniques. Thus, the first category (Table 1) allows us to have a general and basic idea of the article in which information such as the authors' names, the country where the data are obtained, the year and the journal of publication (which includes the Impact Factor (IF) and the quartile ranking), the number of participants and their age range, and the purpose of the study. The other two categories are focused on the purpose of the SR (Table 2 and Table 3). Among the general characteristics of sport, variables such as the type of sport stud-

Table 1: Description of the general characteristics of the articles included in the SR.

Variable	**Description**
Author (citation)	APA format citation of the authors
Country	Country of the data of the article
Publication Year	Year of publication of the article
Journal Name	Name of the Journal where is published the article
IF (Quartile)	Impact Factor and Quartile of the Journal
Longitudinal study	If it treats or no of a longitudinal study
N (participants)	Number of participants of the study
Age (participants)	Average or age interval of the participants
Principal Aim	Description of the principal aim of the study

ied in the article, the gender of the studied participants (i.e., male, female, or both), the competition category (i.e., professional, amateur, or both), the data source (i.e., league, association, organization, federation), and the sports sciences field and focus of the study (i.e., sports performance analysis, sports technology, movement Integration, health and Sports non-governmental organization (NGOs) (Table 2). The last category is related to the description of the outcomes, the type of clustering method performed, the clustering technique applied, the number and size of the resulting clusters, the criterion to choose the optimal number of clusters (e.g., elbow, silhouette, or gap statistic, among others), the software and package used to run the cluster analysis, and if the article itself shares the code or data used (Table 3). Data were col-

Table 2: Description of the characteristics of the particular sport studied.

Variable	**Description**
Author (citation)	APA format citation of the authors
Sport	Sport or sports in which the article focuses
Gender	Sex of the participants in which the article focuses *(Male, Female, Both)*
Category participants	Category of the participants (*Professional, Amateur, Both*)
Name of source data	Name of the source of the data of the study
	(League, Association, Organization, Federation, etc.)
Category classification	Five categories:
	1) Sports Performance Analysis;
	2) Sports technology;
	3) Movement integration;
	4) Health;
	5) Sports NGOs

lected, stored in a database, and then checked for discrepancies between the three authors MP, DF, and MC. Discrepancies were resolved by consensus after re-reviewing conflicting articles. The data is available on a publicly accessible GitHub repository (`https://github.com/marticasals/Clustering_SR_Rerorting_SportsSC`).

2.3 Identification of studies

Figure 1 depicts the Preferred Reporting Items for Scoping reviews and Meta-Analyses (PRISMA-SCR) flow chart to summarize all stages of the selection process. In the first stage, the search engines (PubMed, WoS, SportDiscus and CINAHL) are already filtered by language and by type of scientific literature (WoS). Thus, the set of collected items broken down by search engine was: 1450 (PubMed), 2120 (WoS), 812 (SportDiscus), and 336 (CINAHL). After reviewing whether all of these articles were part of any of the fields of sport sciences, those that were not published in journals included in the Journal

Table 3: Description of the characteristics of the clustering techniques.

Variable	**Description**
Author (citation)	APA format citation of the authors
Description of Outcomes	Description of the outcomes as it is written in the article
Type of clustering method	Type of clustering method used in the article: (*Hierarchical clustering*, *Partitioning methods*, *Model-based Clustering*, *Fuzzy Clustering*, *Density-based clustering*, etc.)
Name of the clustering method used	Name of clustering method used in the article: PCA, *k-means*, *k-medoids*, *Kamila*, *hclust*, *Gaussian Mixture Models*, etc.
Number of clusters	Number of clusters
Cluster size per cluster	Size of each cluster
Method to decide the number of clusters	Method to decide the optimal number of clusters: *Elbow*, *Gap*, *Silhouette*, *Dendrogram*, etc.
Software used	Software used
Package used	Package used
Data shared	If the article shares or not the data that uses in the study.
Code shared	If the article shares or not the code of the software.
Repository of Data or Code shared	If the article shares a repository with the data or the code used.
Practical implication results	Summary of the practical implication of the results of the study.

Citation Reports (JCR), Journal of Quantitative Analysis in Sports (JQAS), and Journal of Sports Analytics (JSA) were excluded, resulting in 1330 articles in PubMed, 1739 in WoS, 648 in SportDiscus, and 235 in CINAHL. Of these 766 articles, 247 were excluded for being duplicates, resulting in a total of 519 articles. After inspection of the title and abstract, we excluded 24 non-original articles, one article written in French and one in Dutch, which were not filtered in the first stage, and 88 more articles with a non-sports theme. From the remaining 405 articles, we proceeded to the second phase of the SR process in which the full text of the articles was reviewed to determine whether clustering methods were applied. Thus, of the 405 articles that remained after the first phase of the review, it was observed that in 127 articles the word "clustering" was only used to refer to the grouping of data in its initial structure, and not to describe a clustering method in the statistical analysis of the data. Therefore, there were only 278 articles left to be reviewed carefully. Figure 1 summarizes the number of articles identified and the reasons for exclusion in each PRISMA-SCR stage.

3 Results

3.1 General characteristics of the selected articles

Table 4 shows that out of the 278 selected articles, 73 (26.2%) did not specify the country of the studied dataset. Additionally, 27 articles (9.7%) reported that the data has been collected from multiple countries. Of the remaining articles, 16 (5.8%), 35 (12.6%), and 16 (5.8%) reported data from Australia, the United States, and the United Kingdom, respectively. The distribution of articles by region reveals that 63.0% reported data from various European countries, 13.5% from American countries, 18.0% from Asia, 2.7% from Oceanic countries, and 0.9% from African countries, with 0.4% not specifying the data source. Table 4 also shows illustrates the concentration of original articles on clustering techniques in sports science over the past 14 years. The number of original articles per year ranged from 9 to 31, with 2021 recording the highest number at 31 articles, representing 11.1% of the total. In other years, the number of articles ranged from one to fourteen, constituting between 0.4% and 2.9% of the total. Furthermore, we can observe that 262 out of 278 selected articles are not longitudinal studies, accounting for 94.24% of the total. This predominance is expected in this field, as many popular clustering methods are tailored for cross-sectional frameworks.

Figure 2 depicts the publication trend of original articles on the use of clustering techniques in sports sciences by year. The first article in this domain was published in 1996, and since then, the number of articles published in journals of JCR, JQAS, and JSA has steadily increased. The trend demonstrates an exponential growth, indicating a positive trajectory in publications. It's important to note that the drop in the number of articles in 2023 is due to the completion of our scoping review in May 2023.

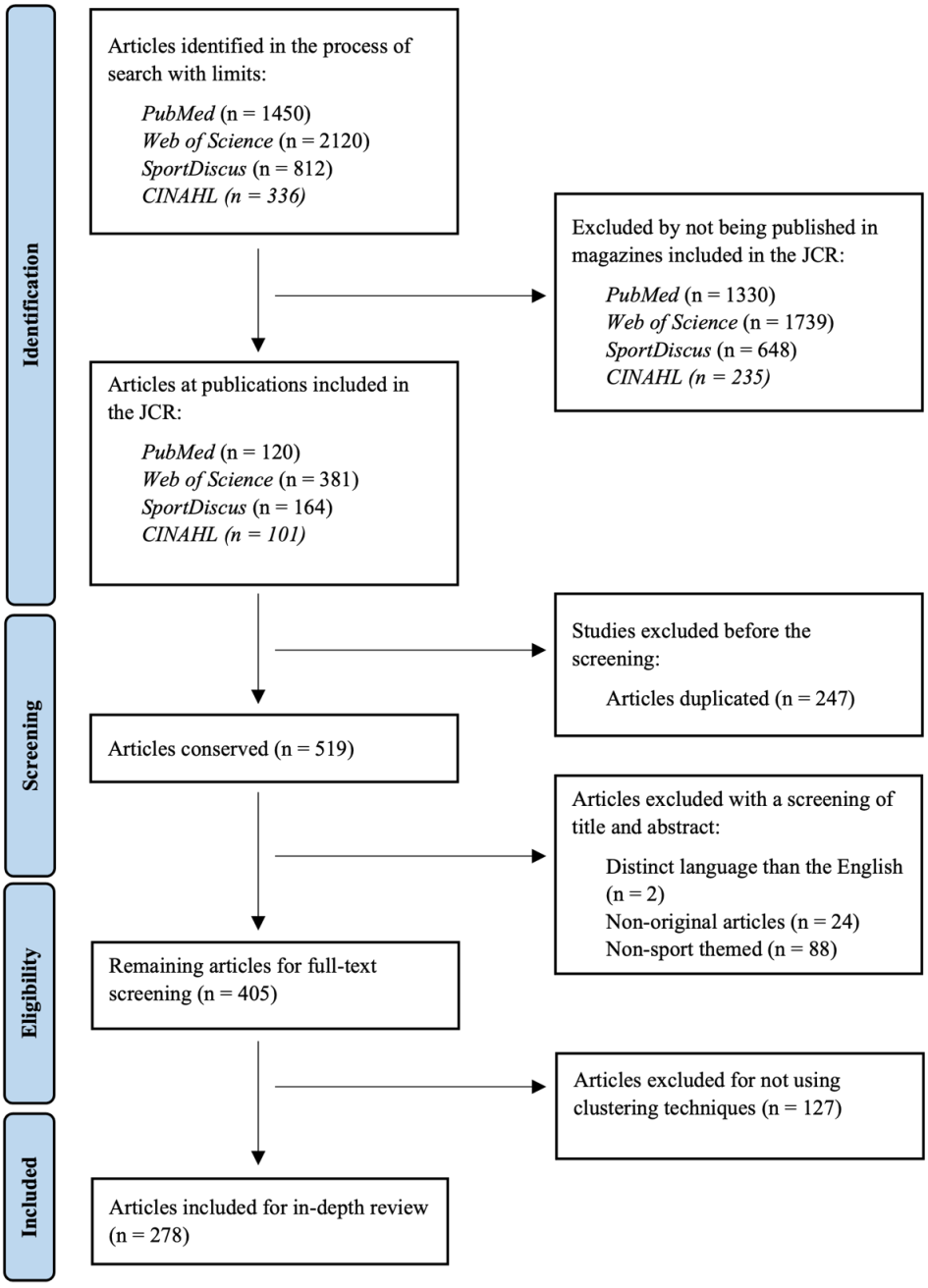

Figure 1: PRISMA-SCR: Scoping review flow chart about the application of clustering techniques in original articles in the field of sports sciences.

Table 4: Frequency and percentage table of the general characteristics of the articles included in the SR.

Variable (N = 54)	**Category**	**n (%)**
Country	*International*	27 (9.7%)
	Australia	16 (5.8%)
	USA	35 (12.6%)
	UK	16 (5.8%)
	No reported	73 (26.2%)
	*Others**	111 (39.9%)
Publication Year	*2023*	7 (2.5%)
	2022	26 (9.3%)
	2021	31 (11.1%)
	2020	14 (5.0%)
	2019	22 (7.9%)
	2018	23 (8.3%)
	2017	9 (3.2%)
	2016	19 (6.8%)
	2015	21 (7.7 %)
	2014	11 (4.0%)
	2013	14 (5.0%)
	2012	9 (3.2%)
	2011	14 (5.0%)
	2010	8 (2.9%)
	2009	13 (4.7%)
	2008	5 (1.8%)
	2007	8 (2.9%)
	2006	3 (1.1%)
	2005	1 (0.4%)
	2004	2 (0.7%)
	2003	5 (1.8%)
	2002	1 (0.4%)
	2001	3 (1.1%)
	2000	4 (1.4%)
	1999	1 (0.4%)
	1998	2 (0.7%)
	1996	2 (0.7%)
Longitudinal study	*No*	262 (94.2%)
	Yes	16 (5.8%)

* This category grouped the countries with less frequency: Belgium, Brazil, Canada, Chile, China, Czech Republic, Finland, France, Germany, Greece, Hungary, India, Ireland, Italy, Japan, Korea, Malaysia, New Zealand, Norway, Poland, Portugal, Russia, Singapore, Slovenia, Spain, Sweden, Switzerland, Thailand, Tunisia and not available.

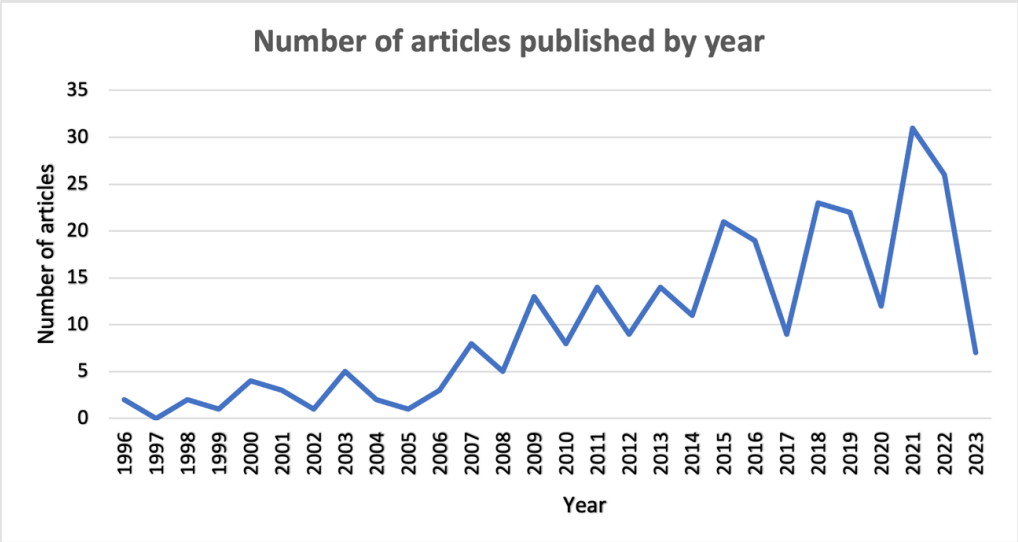

Figure 2: Publication trend of original articles on the use of clustering techniques in the field of sports sciences by year. The reason for the drop in the number of articles in 2023 is that our scoping review ended in May 2023.

3.2 General characteristics of the sport

Table 5 presents a summary of the attributes of the sports studied in the 278 selected articles. Among these articles, 14.4% did not focus on a single sport but rather addressed athletes from various sports disciplines, categorizing them as multidisciplinary. Additionally, 7.9% of the articles utilized clustering methods to study the physical activity of participants without considering their specific sport, while 10.8% did not report the type of sport studied. In terms of specific sports, football (soccer) and basketball were the most prominent, with 12.2% and 7.5% of articles focusing on these sports, respectively. Tennis emerged as the predominant racket sport, featured in 2.2% of articles. Other popular sports such as rugby, volleyball, cricket, and aquatic sports like water polo or swimming were grouped under the 'others' category, representing 42.4% of the total.

he majority of articles predominantly focus on male participants (81.6%), with 37.7% exclusively focusing on males and 43.9% including both genders. Additionally, 47.8% of articles involve professional sports participants, while 19.4% involve amateurs. Information on participant status was not reported in 23.8% of articles, and 8.6% examined both professional and amateur categories, with data unavailable in 0.4% of articles. Concerning to type of sports category, the articles were categorized into five types: Sports performance analysis, Sports technology, Movement Integration, and Health. Notably, no articles were classified under the Sports NGOs category. During the systematic review process, the majority of articles were classified as Sports performance analysis (66.2%), followed by Health (28.1%), Movement Integration (3.24%), and Sports Technology (1.4%). Additionally, a small percentage of articles (1.89%) were allocated to

Table 5: Frequency and percentage table of the general characteristics of the sport.

Variable (N = 54)	**Category**	**n (%)**
Sport	*Multidisciplinary*	40 (14.4%)
	Physical Activity	22 (7.9%)
	Soccer	34 (12.2%)
	Tennis	6 (2.2%)
	Basketball	21 (7.5%)
	Golf	6 (2.2%)
	No reported	30 (10.8%)
	Others[*]	118 (42.4%)
	Not available	1 (0.4%)
Gender	*Male*	105 (37.7%)
	Female	20 (7.2%)
	Both	122 (43.9%)
	No reported	30 (10.8%)
	Not available	1 (0.4%)
Category participants	*Professional*	133 (47.8%)
	Amateur	54 (19.4%)
	Both	24 (8.6%)
	No reported	66 (23.8%)
	Not available	1 (0.4%)
Category classification	*Sports Performance Analysis*	184 (66.2%)
	Health	78 (28.1%)
	Movement Integration	9 (3.2%)
	Sports Technology	4 (1.4%)
	Sports NGOs	2 (0.7%)
	Not available	1 (0.4%)

[*] This category grouped the sports modalities with less frequency: Aerobic fitness, American Football, Archery, Athletics, Australian Rules Football, Badminton, Ballet Dancer, Baseball, Bocce Ball, Classical ballet, Collegiate wrestling, Crawl swimming, Cricket, Cross-Country Sit Skiing, Cycling, Dance, Decathlon, Field Hockey, Futsal, Gymnastics, Handball, Hockey, Ice Hockey, Judo, Jujitsu, Long Jump, Netball, Race Running, Rowing, Rugby, Running, Ski, Squash, Surf, Swimming, Table-tennis, Volleyball, Water-polo, Wheelchair Basketball, Wheelchair Racers, Wheelchair Rugby.

both the Sports Technology and Sports NGOs categories.

3.3 Characteristics of clustering techniques

Table 6 presents the main characteristics of the clustering techniques employed in the selected articles. Notably, hierarchical clustering emerged as the most popular technique, utilized in 31.6% of the articles, followed by partitional clustering (26.3%) and two-step clustering (9.3%). Less common clustering methods included model-based clustering (1.4%), density-based spatial clustering (1.8%), and high-dimensional clustering (0.7%). Additionally, 14.4% of articles employed two or more clustering techniques simultaneously, categorized as 'more than one category type'. However, other clustering methods were much less frequent, and 11.2% of articles did not specify the clustering technique used. To provide more specific insights, the frequency of specific clustering methods used in the selected articles is reported. The k-means algorithm emerges as the most popular method, appearing in at least 29.5% of the 278 articles. However, this count represents a minimum, as k-means is often used in combination with other methods, categorized under the "More than one category type" level. Following k-means, Ward's linkage within hierarchical clustering is the second most popular algorithm, utilized in 15.4% of the articles. Other methods include the Average linkage method (1.8%), Mixed data method (1.4%), Gaussian mixture models (1.1%), and Spectral clustering (1.1%). Additionally, 26.62% of articles do not specify the clustering method used. Figure 3 depicts the distribution of clustering techniques used in the articles, categorized by the name of the method. It is noteworthy that partitional clustering is primarily composed of the k-means algorithm. However, the graph does not include techniques that were not reported. Table 7 presents the software and data sharing characteristics of the methods used. The predominant software for clustering in the selected articles were SPSS and R, constituting 24.8% and 9.0%, respectively. MATLAB, Statistica, and Python were less commonly used, representing 4.3%, 3.6% and 2.5%, respectively. Additionally, combinations of these software, along with others such as SAS and Minitab, were found in 3 (1.1%) articles, categorized under "more than one type" (3.6%). However, software usage was not reported in 121 articles (43.5%).Furthermore, while the software used was mentioned, details regarding the specific analysis methods, such as the R packages utilized (e.g., mclust, hclust), were absent in 244 articles (87.7%). Lastly, none of the articles included in this scoping review shared their data, code, or repository.

4 Discussion

The scoping review of 278 selected articles unveils a notable historical insight: the inaugural publication utilizing clustering techniques in sports science emerged in 1981, followed by a 26-year hiatus before a resurgence in 2007. Since then, the field has experienced a consistent and growing trend in articles employing clustering techniques. However, this growing trend does not align with the apparent interest in clustering methods observed on Google Trends (Figure 4), where interest decreased in 2006 and subsequently depicts stability with low interest.

Table 6: Characteristics of clustering techniques.

Variable (N = 54)	**Category**	**n (%)**
Type of clustering method	*Hierarchical clustering*	88 (31.6%)
	Partition clustering	73 (26.3%)
	Model-based clustering	4 (1.4%)
	Density-based Spatial clustering	5 (1.8%)
	High-dimensional clustering	2 (0.7%)
	Two-step clustering	26 (9.3%)
	More than one type	40 (14.4%)
	Others[a]	8 (2.9%)
	No reported	31 (11.2%)
	Not available	1 (0.4%)
Name of the clustering method used	*k-means*	82 (29.5%)
	Ward's method	43 (15.4%)
	Schwarz's Bayesian Criterion	11 (4.0%)
	Average linkage method	5 (1.8%)
	Mixed Data method	4 (1.4%)
	Gaussian mixture models	3 (1.1%)
	Spectral clustering	3 (1.1%)
	More than one type	35 (12.6%)
	Others[b]	17 (6.1%)
	No reported	74 (26.6%)
	Not available	1 (0.4%)
Method to decide the number of clusters	*Dendrogram*	53 (19.1%)
	Silhouette	15 (5.4%)
	BIC	7 (2.5%)
	Elbow	4 (1.4%)
	Calinski-Harabasz index	4 (1.4%)
	Gap	3 (1.1%)
	More than one type	29 (10.5%)
	Others[c]	9 (3.2%)
	No reported	153 (55.0%)
	Not available	1 (0.4%)

[a] This category grouped the type of clustering methods with less frequency: Trajectory-based Longitudinal Clustering, Latent Profile Analysis, Thematic Clustering, Unsupervised Clustering, Cluster phase method, Thematic Clustering and Probabilistic Curve-clustering method. "More than one type" accounts for the combination of two clustering techniques.

[b] This category grouped the clustering methods with less frequency: PCA, Complete Linkage, Cluster Phase Analysis (CPA), k-modes, Linkage Criteria, Log-likelihood method, Random Forest method, Single linkage algorithm, Super-paramagnetic clustering (SPC), Weigthed pair-group method, Latent class analysis hclust. kmlShape. Louvain method. "More than one type" accounts for the combination of two or more clustering methods.

[c] This category grouped the number of cluster selection criteria with less frequency: Majoritary Rule, Agglomeration schedule, Fréchet mean, Group Concept Mapping, Kappa Coeficient, Random Forest and R-square. "More than one type" accounts for the combination of two clustering techniques.

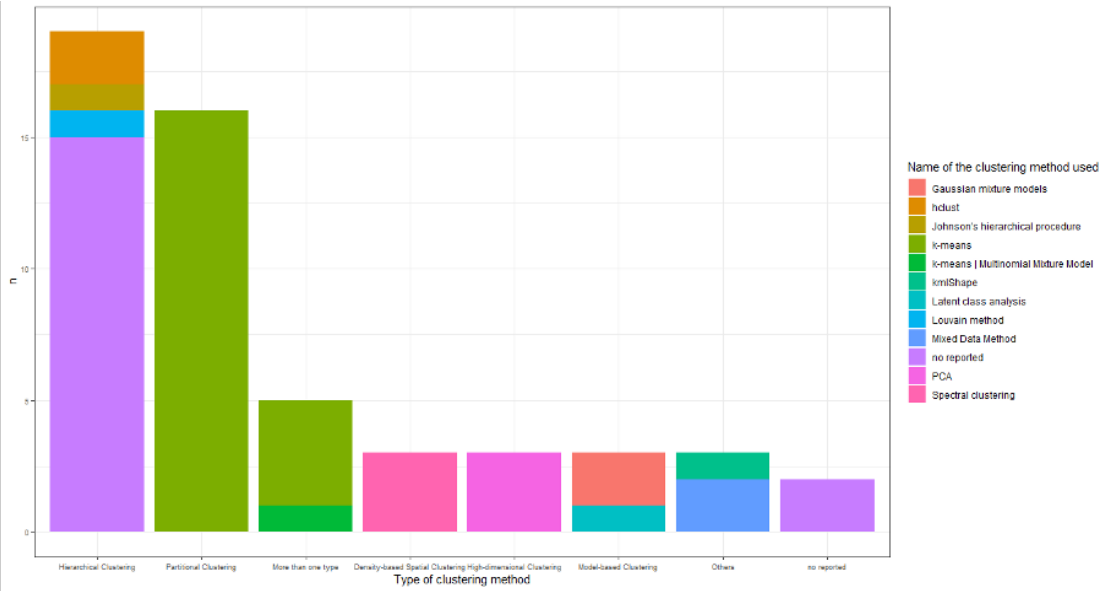

Figure 3: Percentage of articles according to the type of clustering technique and the name of the clustering method used.

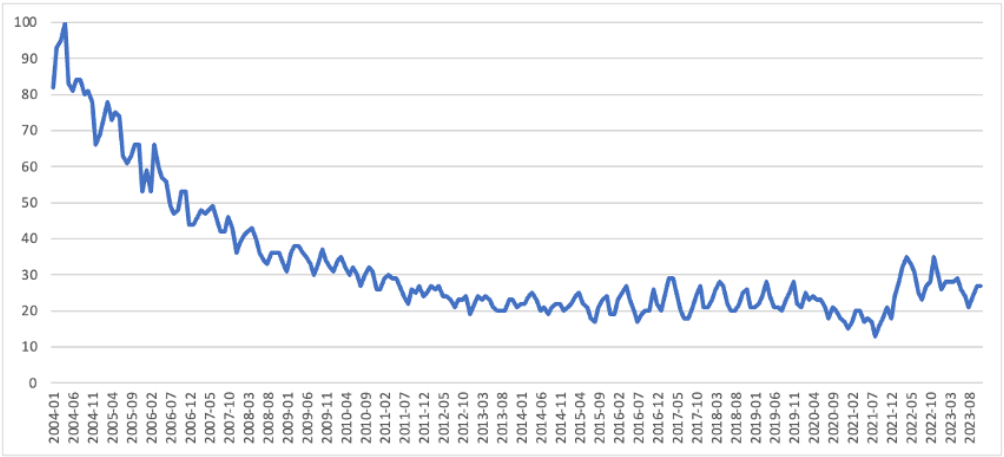

Figure 4: Google Trends data provides a visualization of the interest in the keyword "clustering" from January 2004 to November 2023. This data is derived from a representative sample of Google search queries. In the figure, the blue line represents the normalized relative search volume, which indicates the proportion of searches for the specific term compared to the total number of searches conducted over the specified period. Normalization is performed on a scale from 0 to 100, where 100 denotes the peak popularity of the term within the selected time frame. This metric helps to gauge the level of interest in clustering over time, offering insights into its fluctuation and trends.

Table 7: Software and data sharing characteristics reported.

Variable (N = 54)	**Category**	**n (%)**
Software used	*SPSS*	69 (24.8%)
	R	25 (9.0%)
	MATLAB	12 (4.3%)
	Statistica	10 (3.6%)
	Python	7 (2.5%)
	Minitab	3 (1.1%)
	SAS	3 (1.1%)
	More than one type	17 (6.1%)
	Others[d]	10 (3.6%)
	No reported	121 (43.5%)
	Not available	1 (0.4%)
Package reported	*No*	244 (87.7%)
	Yes	33 (11.9%)
	Not available	1 (0.4%)
Data shared	*No*	278 (100.0%)
Code shared	*No*	278 (100.0%)
Repository of Data or Code shared	*No*	278 (100.0%)

[d] This category grouped the software with less frequency: Anaconda, Concep Systems groupwisdomTM, GraphPad Prims, HLM, JASP, k-Dynami, NCSS, STATA and TreeView. "More than one type" accounts for the combination of two or more software used.

Half of the published articles related to sports science focus on the sports performance of the participants (i.e., sports performance analysis). A small proportion of articles (3.2%) delve into aspects related to movement (Movement integration), while those linking sport to health sciences are less frequent (28.1%) Albert et al. (2017).The underlined aspects in 2017 contribute to enhancing athletes' performance in sports such as football (soccer), basketball, and tennis, either by describing position patterns on the field or by focusing on injury prevention Siedlik et al. (2016); Anıl Duman et al. (2024). Among all selected articles, the United States, Australia, and the United Kingdom emerge as the most frequent countries. Additionally, the prevalence of professional sports participants surpasses that of amateurs (19.4%), likely due to the greater accessibility of data for conducting studies and the notable influence of the Sports Performance Analysis or Sports Analytics field, which is more developed both academically and in the sports industry Alamar (2013); Glickman (2017). The participants in sports studies are predominantly male in 105 selected articles (37.7%), while both genders are represented in 43.9% of the articles. This fact is noteworthy due to the inherent differences in physical characteristics and playing styles between genders. Therefore, results obtained for one gender may not be fully extrapolated to the other Thibault et al. (2010). Recent literature developments involve exploring performance in disabled athletes, a theme not covered in this work (see, for instance, (van der Linden et al., 2021), for a clustering study). An example of clustering methods' application to identify groups of wheelchair basketball players with similar game-related statistics is reported in Cavedon et al. (2024).

Regarding selected clustering techniques, a majority of studies performed hierarchical clustering (31.6%), while partitional clustering, specifically the k-means algorithm, represented 26.3% of the cases. In contrast, certain articles reported the type of clustering technique without specifying the method used (26.6%) or omitted the method for determining the number of clusters (55.0%). Thus, assessing the adequacy of the approaches used becomes challenging, potentially leading to suboptimal outcomes. It is well-recognized that clustering methods can yield disparate results on the same data, and their suitability depends on the data's nature Singh and Gosain (2013). The observed limitations might stem from the initial selection of journals, which did not include generalist statistical journals.

The most frequently used software, either independently or in combination with others, includes SPSS and R. However, a substantial number of articles do not report the software used. However, a substantial number of articles do not report the software used. Even when the software is reported, the specific packages or libraries employed are rarely detailed. Consequently, this omission complicates the assessment of whether the chosen methods were appropriate and prevents the creation of a comprehensive list of widely used tools in this field, which would be valuable for both researchers and practitioners. Additionally, none of the reviewed articles provide access to their data, software code, or repositories. This lack of data and code sharing poses a significant barrier to reproducing or replicating the analyses and verifying the results Resnik and Shamoo (2017); Schwab et al. (2022).

To address these issues, we recommend that future studies prioritize transparency by

specifying the software and packages used and by sharing data and code through public repositories (e.g., GitHub, Zenodo, among others). Such practices would enhance reproducibility and the overall impact of clustering research in sports sciences. As highlighted by Horton and Stoudt (2024), fostering a culture of transparency and accountability in research is critical for building trust and ensuring the integrity of scientific findings. By adopting these recommendations would not only address the concerns raised by reviewers, but also contribute to a more robust and credible body of research in this field.

Recent Advances in the Literature

Since the completion of the database search on May 7, 2023, several new studies have emerged that demonstrate the continued application and refinement of clustering techniques in sports sciences. For example, Liu et al. (2024) applied k-means clustering to para-alpine sit skiing, identifying athlete clusters based on impairment measures and their impact on performance outcomes. Their results highlight the utility of cluster analysis in supporting evidence-based classification systems for Paralympic sports. Similarly, Akhanli and Hennig (2023) employed advanced clustering validity indexes to analyze mixed-type performance data from football players, illustrating novel methods to optimize cluster selection based on specific goals, such as team composition or player comparison. These approaches provide valuable insights into tailoring clustering methods to meet the nuanced demands of performance analysis in team sports. A final example is the work of Bunker et al. (2024), who introduced a multi-agent statistically discriminative sub-trajectory mining (MA-Stat-DSM) method to analyze tracking data from the NBA. Their hierarchical approach revealed critical movement patterns associated with successful plays, offering an innovative application of clustering for tactical analysis in basketball. These recent contributions underscore the expanding utility of clustering techniques across a broad spectrum of sports and research questions. They also highlight the need for continued methodological refinement and reporting transparency to enhance the reproducibility and impact of future studies.

Limitations

One limitation of this scoping review is its narrow focus on articles from JCR journals, JQAS, and JSA. To broaden the scope of future reviews, incorporating more diverse or open-access sources, such as preprint servers, institutional repositories, and grey literature, would provide a more comprehensive perspective on the field and mitigate potential biases associated with restricted access to certain publications. By excluding machine learning journals and those centered on unsupervised techniques, as well as clustering methods by statisticians or data scientists applicable to sports science using real or simulated data, certain relevant research may have been overlooked. However, this selection was made to ensure a consistent level of rigor and relevance in the included articles. It is worth noting that, to the best of our knowledge, this is the first scoping review to specifically examine articles applying clustering techniques in sports sciences. While this approach may have limitations, it also presents a novel endeavor with the potential to

inform and inspire future research in the field. Moving forward, efforts to expand the scope of such reviews could include considering articles from a broader range of journals and disciplines, thereby enhancing the comprehensiveness and applicability of the findings.

Conclusions

This review identifies a predominant focus on sports performance and athlete health over the past 15 years, with hierarchical or partitional clustering methods commonly employed. However, there is substantial room for improvement across all reviewed articles, particularly in data and code transparency, as well as in the reporting of clustering technique details. To advance the clarity and transparency of clustering method reporting in sports sciences, future studies should prioritize detailed reporting of the methods used, including the choice of algorithms, the criteria for determining the number of clusters, and the rationale behind these choices. The adoption of standardized reporting frameworks, such as extensions of PRISMA tailored to clustering studies, could significantly improve the reproducibility of research. Additionally, the implementation of open data and code-sharing practices is crucial. Depositing datasets and scripts in publicly accessible repositories will enable other researchers to validate findings and build upon them. Including practical examples or links to interactive tutorials demonstrating the application of clustering methods in sports contexts would further facilitate knowledge transfer and encourage broader adoption of best practices. Addressing these gaps is particularly important given the growing trend of publishing articles utilizing clustering methods in sports sciences Bullock et al. (2023). By improving methodological transparency and fostering open science, researchers can create a more robust foundation for applying clustering techniques, benefiting both academic research and practical applications in sports performance and health.

Ethics Statement

We did not seek ethical approval for this work, as all information used and reported is freely available via online sources.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All authors wrote the article, critically read it. All authors have read and approved the final version of the manuscript and agree with the order of presentation of the authors.

Acknowledgement

This work has been supported by the Ministerio de Ciencia e Innovación (Spain) [PID2019-104830RB-I00/ DOI (AEI): 10.13039/501100011033], and by grant 2021 SGR 01421 (GRBIO) administrated by the Departament de Recerca i Universitats de la Generalitat de Catalunya (Spain).

References

Aggarwal, C. C. and Reddy, C. K. (2014). *Data Clustering: Algorithms and Applications*. CRC Press.

Akhanli, S. and Hennig, C. (2023). Clustering of football players based on performance data and aggregated clustering validity indexes. *Journal of Quantitative Analysis in Sports*, 19(2):103–123.

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision makers*. Columbia University Press.

Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H. (2017). *Handbook of statistical methods and analyses in sports*. Crc Press.

Anıl Duman, E., Sennaroğlu, B., and Tuzkaya, G. (2024). 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*, 238(1):55–75.

Anzer, G., Bauer, P., and Brefeld, U. (2021). The origins of goals in the german bundesliga. *Journal of Sports Sciences*, 39(22):2525–2544.

Ball, K. and Best, R. (2007). Different centre of pressure patterns within the golf stroke i: Cluster analysis. *Journal of sports sciences*, 25(7):757–770.

Bianchi, F., Facchinetti, T., and Zuccolotto, P. (2017). Role revolution: towards a new meaning of positions in basketball. *Electronic Journal of Applied Statistical Analysis*, 10(3):712–734.

Bullock, G. S., Ward, P., Impellizzeri, F. M., Kluzek, S., Hughes, T., Hillman, C., Waterman, B. R., Danelson, K., Henry, K., Barr, E., et al. (2023). Up front and open? shrouded in secrecy? or somewhere in between? a meta-research systematic review of open science practices in sport medicine research. *Journal of Orthopaedic & Sports Physical Therapy*, 53(12):735–747.

Bunker, R., Duy, V., Tabei, Y., Takeuchi, I., and Fujii, K. (2024). Multi-agent statistically discriminative sub-trajectory mining and an application to nba basketball. *Journal of Quantitative Analysis in Sports*.

Carpita, M., Pasca, P., Arima, S., and Ciavolino, E. (2023). Clustering of variables methods and measurement models for soccer players' performances. *Annals of Operations Research*, 325(1):37–56.

Cartigny, E., Fletcher, D., Coupland, C., and Bandelow, S. (2021). Typologies of dual

career in sport: A cluster analysis of identity and self-efficacy. *Journal of Sports Sciences*, 39(5):583–590.

Cavedon, V., Zuccolotto, P., Sandri, M., Manisera, M., Bernardi, M., Peluso, I., and Milanese, C. (2024). Optimizing wheelchair basketball lineups: A statistical approach to coaching strategies.

Chandran, A., Brown, D., Nedimyer, A. K., and Kerr, Z. Y. (2019). Statistical methods for handling observation clustering in sports injury surveillance. *Journal of athletic training*, 54(11):1192–1196.

Dalton-Barron, N., Palczewska, A., Weaving, D., Rennie, G., Beggs, C., Roe, G., and Jones, B. (2022). Clustering of match running and performance indicators to assess between-and within-playing position similarity in professional rugby league. *Journal of Sports Sciences*, 40(15):1712–1721.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley, Chichester, UK.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Glickman, M. E. (2017). Discussion of practical problems in sports analytics. *JQAS Invited Session*, 490.

Hautbois, C., Djaballah, M., and Desbordes, M. (2020). The social impact of participative sporting events: a cluster analysis of marathon participants based on perceived benefits. *Sport in Society*, 23(2):335–353.

Hodge, K. and Petlichkoff, L. (2000). Goal profiles in sport motivation: A cluster analysis. *Journal of Sport and Exercise Psychology*, 22(3):256–272.

Horton, N. J. and Stoudt, S. (2024). Editorial: Guidelines and best practices to share deidentified data and code. *Journal of Statistics and Data Science Education*, 32(3):227–231.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer US.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Koutroumbas, K. and Theodoridis, S. (2008). *Pattern recognition*. Academic Press.

Kubat, M. and Kubat, M. (2021). Ambitions and goals of machine learning. *An Introduction to Machine Learning*, pages 1–15.

Landesman, P. (2015). Concussion imdb. `https://shorturl.at/mDEJ4`. Accessed on 2024-04-17.

Liu, K., Ji, L., Ma, H., and Lu, Y. (2024). Cluster analysis of multiple impairment measures in evidence-based classification for para-alpine sit skiers. *Scandinavian Journal of Medicine & Science in Sports*, 34(1):e14514.

Lloyd, S. (1982). Least squares quantization in pcm. In *Least squares quantization in*

PCM, 1982., IEEE Transactions on Information Theory, volume 28, pages 129–137. IEEE.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

McNicholas, P. (2016). *Mixture Model-Based Classification*. Chapman and Hall/CRC.

Miller, B. (2011). Moneyball imdb. `https://www.imdb.com/title/tt1210166/?ref_=fn_al_tt_1`. Accessed on 2024-04-17.

Miller, T. W. (2015). *Sports analytics and data science: winning the game with methods and models*. FT press.

Muniz, M. and Flamand, T. (2022). A weighted network clustering approach in the nba. *Journal of Sports Analytics*, 8(4):251–275.

Murray, N. P. and Hunfalvay, M. (2017). A comparison of visual search strategies of elite and non-elite tennis players through cluster analysis. *Journal of sports sciences*, 35(3):241–246.

Musa, R. M., Taha, Z., Majeed, A. P. A., and Abdullah, M. R. (2018). *Machine learning in sports: identifying potential archers*. Springer.

Núñez, J. L., Mahbubani, L., Huéscar, E., and León, J. (2019). Relationships between cardiorespiratory fitness, inhibition, and math fluency: A cluster analysis. *Journal of Sports Sciences*, 37(23):2660–2666.

Resnik, D. B. and Shamoo, A. E. (2017). Reproducibility and research integrity. *Accountability in research*, 24(2):116–123.

Schwab, S., Janiaud, P., Dayan, M., Amrhein, V., Panczak, R., Palagi, P. M., Hemkens, L. G., Ramon, M., Rothen, N., Senn, S., Furrer, E., and Held, L. (2022). Ten simple rules for good research practice. *PLOS Computational Biology*, 18(6).

Siedlik, J. A., Bergeron, C., Cooper, M., Emmons, R., Moreau, W., Nabhan, D., Gallagher, P., and Vardiman, J. P. (2016). Advanced treatment monitoring for olympic-level athletes using unsupervised modeling techniques. *Journal of Athletic Training*, 51(1):74–81.

Singh, D. and Gosain, A. (2013). A comparative analysis of distributed clustering algorithms: A survey. In *2013 International Symposium on Computational and Business Intelligence*, pages 165–169. IEEE.

Smyth, D. (2022). The job market for sports analysts.

Swartz, T. B. (2020). Where should i publish my sports paper? *The American Statistician*, 74(2):103–108.

Thibault, V., Guillaume, M., Berthelot, G., El Helou, N., Schaal, K., Quinquis, L., Nassif, H., Tafflet, M., Escolano, S., Hermine, O., et al. (2010). Women and men in sport performance: the gender gap has not evolved since 1983. *Journal of sports science & medicine*, 9(2):214.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., Weeks, L., et al. (2018). Prisma extension for

scoping reviews (prisma-scr): checklist and explanation. *Annals of internal medicine*, 169(7):467–473.

van der Linden, M. L., Corrigan, O., Tennant, N., and Verheul, M. H. (2021). Cluster analysis of impairment measures to inform an evidence-based classification structure in racerunning, a new world para athletics event for athletes with hypertonia, ataxia or athetosis. *Journal of Sports Sciences*, 39(sup1):159–166.

Zhang, S., Lorenzo, A., Gómez, M.-A., Mateus, N., Gonçalves, B., and Sampaio, J. (2018). Clustering performances in the nba according to players' anthropometric attributes and playing experience. *Journal of sports sciences*, 36(22):2511–2520.

Zuccolotto, P. and Manisera, M. (2020). *Basketball data science: With applications in R*. CRC Press.