TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Creació del primer cens de tots els acadèmics especialitzats en estadística esportiva i descripció dels seus perfils.

Autor: Martí Oliver i Artacho

Director: Martí Casals Toquero i Daniel Fernández Martínez

Departament: Estadística i Investigació Operativa

Convocatòria: Juny 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM I PARAULES CLAU

L'estadística esportiva és un àmbit que ha agafat gran importància els últims anys amb la disponibilitat de diferents recursos com articles científics i llibres en revistes especialitzades, la divulgació a partir de conferències, *podcast* i pel·lícules de l'àmbit que es poden trobar per exemple a la secció d'estadística esportiva de l'*American Statistics Association (ASA)*. Tot i que hi ha estadístics o acadèmics que contribueixen en aquest àmbit, no es disposa ni es coneix encara d'un cens d'aquests professionals. És per això que l'objectiu d'aquest TFG ha estat la creació d'un cens de tots els investigadors i acadèmics especialitzats en estadística esportiva. Aquest primer cens consta de 191 investigadors de 26 països d'arreu del món dels quals se'ls hi ha fet arribar una enquesta que ha permès conèixer i ampliar el perfil d'aquests. La taxa de resposta de l'enquesta ha estat del 50,49% (n = 103). Del total dels 103 investigadors, se n'han seleccionat 85 i s'ha pogut descriure el perfil d'aquests (tres *clusters*) a partir de les tècniques de *clustering k-means i KAMILA.* El primer *cluster* està format pels pioners o els que han impulsat per primera vegada l'estadística esportiva en el seu propi país (ex: USA, UK, Itàlia) i ja són acadèmics molt sèniors, en el segon *cluster* hi ha un grup majoritari d'edats entre 18-29 i 30-44 anys i són els que han continuat el que van crear els del primer *cluster* i a més hi ha la nova fornada d'investigadors més joves que combinen tant l'acadèmia com la indústria de l'esport amb experiència com analista en alguna organització o equip professional, mentre que al tercer *cluster* hi ha estadístics australians reconeguts sobretot en el camp de *Sports Science* i *Exercise Medicine*. El cens realitzat ha ajudat a crear un primer mapa mundial i veure on es troben distribuïts tots aquests investigadors. Sembla confirmar-se que Estats Units és el país líder en aquest àmbit amb un 24,71% dels investigadors i el segueix Itàlia amb un 17,65% dels investigadors.

Paraules claus: cens, *sports analytics,* estadística esportiva, enquesta, *clustering*, *k-means, KAMILA*.

ABSTRACT AND KEYWORDS

Sports statistics is a field that has gained significant importance in recent years with the availability of various resources such as scientific articles and books in specialized journals, dissemination through conferences, podcasts, and movies in the field, which can be found, for example, in the sports statistics section of the American Statistics Association (ASA). Although there are statisticians and academics contributing to this area, there is still no census or comprehensive knowledge of these professionals. Therefore, the objective of this bachelor's thesis has been to create a census of all researchers and academics specialized in sports statistics. This initial census consists of 191 researchers from 26 countries around the world, to whom a survey has been sent, allowing us to learn and expand their profiles. The survey response rate has been 50.49% (n = 103). Out of the total 103 researchers, 85 have been selected, and their profiles have been described (in three clusters) using k-means and KAMILA clustering techniques. The first cluster consists of pioneers or those who first promoted sports statistics in their own country (e.g., USA, UK, Italy) and are already senior academics. The second cluster includes a majority group aged between 18-29 and 30-44, who have continued what the first cluster initiated. Additionally, there is a new generation of younger researchers who combine academia and the sports industry, with experience as analysts in organizations or professional teams. The third cluster consists of recognized Australian statisticians, particularly in the field of Sports Science and Exercise Medicine. The conducted census has helped create an initial global map and identify the distribution of these researchers. It seems to confirm that the United States is the leading country in this field, with 24.71% of the researchers, followed by Italy with 17.65% of the researchers.

Keywords: census, sports analytics, sports statistics, survey, clustering, k-means, KAMILA.

CLASSIFICACIÓ AMS

62H30 Classification and discrimination; cluster analysis

AGRAÏMENTS

A en Martí Casals i en Daniel Fernández, tutors del treball, per proposar-me el tema, per la seva confiança dipositada en mi, les moltes hores de dedicació amb el seu constant suport i sobretot, per tot l'aprenentatge al llarg de tot el Treball de Final de Grau.

Als meus companys, i a la vegada amics, d'universitat, pel seu interès i acompanyament.

A la meva família, pels seus ànims i paciència.

ÍNDEX DE CONTINGUTS

1. INTRODUCCIÓ 5

1.1. Estat de l'art de l'estadística esportiva 5

1.2. Objectiu del treball 7

2. METODOLOGIA 8

2.1. Procés de creació del cens 8

2.1.1. Identificació de les fonts d'informació 8

2.1.2. Selecció de les variables d'estudi 8

2.1.3. Procediment d'extracció de dades 10

2.1.4. Enquesta com a eina d'obtenció d'informació 11

2.1.5. Tractament de les dades 12

2.1.6. Diagrama de flux 13

2.1.7. Selecció de les variables de la base de dades per a l'anàlisi estadístic 14

2.2. Mètode per analitzar la base de dades: clustering 15

2.2.1. Clustering jeràrquic 16

2.2.2. Clustering particional 18

2.2.3. Tècnica de clustering alternativa: KAMILA 22

2.2.3.1. Descripció 22

2.2.3.2. Formulació 22

2.2.4. Estadístic de Hopkins 28

2.3. EDA 29

3. RESULTATS 31

3.1. EDA 31

3.1.1. EDA del cens definitiu (n = 191) 31

3.1.1.1. Anàlisi univariant 31

3.1.1.2. Anàlisi bivariant 32

3.1.2. EDA de la base de dades d'anàlisi (n = 85) 35

3.1.2.1. Anàlisi univariant 36

3.1.2.2. Anàlisi bivariant 38
3.2. K-MEANS 41
3.2.1. Estadístic de Hopkins 41
3.2.2. Inicialització 41
3.2.3. Execució de l'algoritme 42
3.3. KAMILA 43
3.3.1. Inicialització 43
3.3.2. Execució de l'algoritme 43
3.3.3. Caracterització dels clusters 46
3.4. World map 50
4. DISCUSSIONS I CONCLUSIONS 52
4.1. Discussions dels resultats 52
4.2. Limitacions 54
4.3. Aplicacions pràctiques i accions futures 55
4.4. Conclusions 56
5. WEBGRAFIA 57
6. ANNEX 62
6.1. Enquesta 62
6.1.1. Correu electrònic 62
6.1.2. Model de l'enquesta 63
6.2. Codi R 65
6.3. Enllaços directes per la recerca d'investigadors 65

ÍNDEX DE FIGURES

Figura 2.1 PRISMA: Diagrama de flux de l'obtenció del cens definitiu i la base de dades d'anàlisi. 14

Figura 2.2 Possibles diferents clusters segons quina mesura utilitzem 15

Figura 2.3 Imatge d'un dendrograma i del seu tipus de clustering jeràrquic aglomerat o divisiu. 17

Figura 2.4 Formació dels clusters amb l'algoritme k-means. 19

Figura 2.5 Exemple del mètode del colze (Elbow method). 20

Figura 2.6 *Exemple del silhouette method.* 21

Figura 2.7 Exemple del gap statistic method. 21

Figura 2.8 Algoritme KAMILA 24

Figura 2.9 Exemple de clusters definits aplicant el prediction strength. 27

Figura 2.10 Exemple per determinar el nombre de clusters gràficament tenint en compte el prediction strength. 28

Figura 3.1 Matriu de correlacions entre les variables numèriques. 34

Figura 3.2 Gràfic de les variàncies segons cada component 35

Figura 3.3 Gràfic PCA de les variables numèriques. 35

Figura 3.4 Matriu de correlacions entre les variables numèriques de la base de dades de l'anàlisi estadístic. 40

Figura 3.5 Gràfic de les proporcions de les variàncies de la base de l'anàlisi estadístic segons cada component 40

Figura 3.6 Gràfic PCA de les variables numèriques de la base de l'anàlisi estadístic. 41

Figura 3.7 Gràfic de l'elbow method per determinar la k òptima. 42

Figura 3.8 Gràfic dels clusters després d'aplicar l'algoritme k-means. 42

Figura 3.9 Gràfic dels valor prediction strength segons el nombre de clusters. 44

Figura 3.10 Gràfic dels clusters després d'aplicar l'algoritme KAMILA. 44

Figura 3.11 Gràfic dels valor prediction strength després d'executar de nou l'algoritme. 45

Figura 3.12 Gràfic dels clusters després d'aplicar de nou l'algoritme KAMILA. 45

Figura 3.13 Boxplots de les variables numèriques estratificats per cada cluster. 48

Figura 3.14 Mapa mundial dels països dels investigadors del cens. 51

ÍNDEX DE TAULES

Taula 2.1 Taula descriptiva de les variables del cens 9

Taula 2.2 Taula descriptiva de les noves variables de la base de dades per a l'anàlisi estadístic..... 15

Taula 2.3 Taula de conceptes importants per l'algoritme KAMILA..... 24

Taula 3.1 Taula resum de l'anàlisi univariant de les variables més rellevants del cens definitiu. 31

Taula 3.2 Taula resum de l'anàlisi bivariant de totes les variables més rellevants del cens definitiu. . 32

Taula 3.3 Taula bivariant entre la variable Country categorized i Survey answered?. 33

Taula 3.4 Taula bivariant entre la variable Research group segons i Survey answered?..... 34

Taula 3.5 Taula resum de l'anàlisi univariant de les variables més rellevants de la base de dades de l'anàlisi..... 36

Taula 3.6 Taula bivariant i dels p-valors entre la variable Country categorized i years worked in sports statistics in academia. 39

Taula 3.7 Taula bivariant i dels p-valors entre la variable Years worked in sports statistics in acadèmia i Highest level of education. 39

Taula 3.8 Valors prediction strength segons cada valor de k..... 43

Taula 3.9 Taula amb Informació de totes les variables sobre tenint en compte que la variable resposta és el cluster..... 46

Taula 3.10 Taula de les variables significatives, segon cada cluster. 47

Taula 3.11 Resum amb les característiques principals de cada cluster 49

Taula 3.12 Noms i freqüències de tots els països del cens. 50

1. INTRODUCCIÓ

1.1. Estat de l'art de l'estadística esportiva

En els últims anys, la disciplina de l'estadística ha experimentat una gran importància i evolució en diferents àmbits com l'educació, la investigació i la transferència de coneixement a la indústria (Aerts et al., 2021; Alamar, 2013a; Gelman & Vehtari, 2021). Tant és aquest creixement que el passat octubre del 2012 la revista *Harvard Business Review* va publicar que la feina de científic de dades és considerada la professió més sexy del segle XXI, i amb més sortides professionals (Davenport & Patil, 2022). És conegut que els professionals de l'estadística poden ser bioestadístics, bioinformàtics, econometristes, especialistes en estadística oficial o ecòlegs, entre moltes altres especialitats, de forma que desenvolupen les seves tècniques estadístiques i sobretot el seu pensament estadístic en funció de les seves necessitats específiques per a la resolució de problemes que hi ha al món.

L'evolució i els canvis de les tècniques estadístiques que han anat apareixent al llarg dels últims anys, ha comportat un canvi en la presa de decisions informades en diferents àmbits, de manera que els estudis de les dades han pres una gran importància en diverses especialitats (Gelman & Vehtari, 2021).

Una de les especialitzacions que ha crescut més és l'estadística esportiva que, en altres països, especialment als Estats Units, és més coneguda com a *Sports Analytics* (Alamar, 2013b). La funció d'una persona especialista en estadística esportiva és recopilar, analitzar i interpretar les dades dels esdeveniments esportius de forma que s'obté informació valuosa, per exemple, pels jugadors, equips, lligues i organitzacions de la indústria de l'esport. Amb aquesta informació també es pretén per exemple que es millorin les estratègies del joc i la presa de decisions.

L'estadística esportiva és recolzada per l'*American Statistical Association* (ASA), on durant les *Joint Statistical Meetings* celebrades l'any 1992 a la ciutat de Boston, es va crear la *Statistics in Sport Section* (SiS), i que tenia com a objectiu fomentar el desenvolupament de l'estadística i les seves aplicacions a l'esport. A més, al llarg dels anys, han anat apareixent diverses contribucions que han provocat que s'anés guanyant importància, com per exemple, la creació de pel·lícules (Moneyball (Miller B., 2011) o Draft Day (Reitman I., 2014)), i llibres ((Albert et al., 2016) o bé (Oliver, 2004)). Recentment s'han creat consultories dedicades especialment a l'esport, com per exemple *Zelus Analytics* (Bornn L. & Fearing D., 2019). Aquesta és una consultoria externa que treballa en diversos esports (beisbol, bàsquet, futbol...) per donar suport a les operacions internes de cada client, amb la característica que

només operen amb un nombre limitat de clients, de forma que així no tots els equips tenen la mateixa informació. A més a més, també han anat apareixent diverses revistes científiques sobre estadística esportiva (Swartz, 2020a), més enfocades a *sports analytics,* com per exemple *Journal of Quantitative Analysis in Sport* (https://www.degruyter.com/journal/key/jqas/html) i *Journal of Sports Analytics* (https://www.iospress.com/catalog/journals/journal-of-sports-analytics), entre d'altres ja més aplicades a la computació, *sports sciences* i *sports medicine*. També hi ha organitzacions que es dediquen a realitzar conferències anuals, amb un clar enfoc d'estadística esportiva com són per exemple la *New England Symposium on Statistics in Sports (NESSIS)* o bé la *MathSport International* on cada any s'organitza a un país i universitat diferent, de forma que es facilita el diàleg i la col·laboració entre professionals i experts de l'estadística i l'esport de diferents parts del món. Cal afegir que, per compte propi, molts investigadors han començat a publicar articles sobre esport i estadística, tal i com comenta Tim B. Swartz: "*With the increasing fascination of sport in society and the increasing availability of sport related data, there are great opportunities to carry out sports analytics research"* (Swartz, 2020b).

La tendència creixent de publicacions en l'àmbit de l'estadística esportiva i alguns *special issues* recordant la seva importància en algunes revistes és cada vegada més freqüent (Groll & Liebl, 2023; Sainani et al., 2021 i alguns volums complets de *special issues* com els de la revista *Italian Journal of Applied Statistics* (https://www.sa-ijas.org/ojs/index.php/sa-ijas/issue/archive). Entre aquests articles n'hi ha de grans investigadors i acadèmics estadístics esportius ja reconeguts mundialment, com el recent d'en Ben Baumer, d'en Michael Lopez o d'en Luke Bornn (Baumer et al., 2023; Fernández et al., 2021; Lopez & Matthews, 2015). Un dels llibres més populars d'aquest camp és però, el ja comentat dels investigadors Jim Albert, Mark E. Glickman, Tim B. Swartz i Ruud H. Koning, coneguts tots mundialment (Albert et al., 2016). Això també ha provocat que a les universitats i institucions de tot hagin aparegut diferents departaments de *sports analytics/statistics* (ex: Bringham Young, Brescia, Simon Fraser, Harvard, Victoria, entre moltes altres). A més, també hi ha societats i instituts d'estadística que han ajudat al creixement sobre esport i estadística (ex: *l'International Statistical Institute o l'Institute of Mathematical Statistics*), i aquest 2023, la *National Science Foundation* ha creat una xarxa nacional als Estats Units de desenvolupament i difusió de continguts esportius per a la divulgació, investigació i educació en ciència de dades. Un dels objectius principals és crear un marc educatiu basat en l'aprenentatge de problemes i aplicacions al món real, especialment centrat en els esports (2023 SCORE Network).

Tot i que hi ha un creixement palès a nivell acadèmic de l'estadística esportiva, actualment encara no existeix un cens exhaustiu que intenti recollir tots els investigadors i acadèmics

especialitzats en aquesta àrea a nivell mundial. Aquesta manca d'informació pot dificultar la coordinació entre investigadors i la identificació de les línies de recerca més prometedores en l'àmbit de l'estadística esportiva.

1.2. Objectiu del treball

És per tot això explicat anteriorment que l'objectiu principal d'aquest treball és crear un cens exhaustiu de tots els investigadors i acadèmics especialitzats en estadística esportiva, tant si treballen en grups de *sports analytics/statistics,* com si ho fan de manera autònoma.

Les finalitats d'aquest cens són millorar la visibilitat dels investigadors en aquest camp, fomentar la creació de xarxes de col·laboració i promoure el desenvolupament de noves investigacions. Amb aquesta iniciativa, es pretén obtenir una base de dades única a nivell mundial que permeti una millor comprensió i avanç en l'àmbit de l'estadística esportiva i a la vegada, que serveixi per tenir-la com a base perquè aquesta es pugui anar retro alimentant de forma dinàmica al llarg del temps, de forma que cada cop sigui el més completa possible. Addicionalment, gràcies a la creació del cens, s'intentarà plasmar la informació d'aquest en un mapa mundial, per tal d'obtenir una imatge ràpida i clara del cens.

2. METODOLOGIA

2.1. Procés de creació del cens

2.1.1. Identificació de les fonts d'informació

Malgrat no haver-hi cap referència científica ni informació disponible a internet d'on poder començar a crear una base de dades per obtenir el cens, els meus tutors em van facilitar una llista de noms d'alguns professionals i grups de recerca coneguts mundialment en aquest camp. Aquesta llista pròpia va ser el punt de partida i l'ajuda que necessitava per iniciar la recerca. A partir d'aquí es van utilitzar diferents estratègies per intentar tenir la base de dades el màxim completa. A banda de la llista ja esmentada, es van cercar acadèmics i investigadors utilitzant paraules clau com *sports statistics* o *sports analytics* tant al cercador de *Google* com a diferents pàgines web on s'hi publiquen articles científics, com són *Google Scholar* o *ResearchGate* (veure annex 6.3). A més, la cerca no es va limitar només a través de les paraules claus comentades anteriorment, sinó que també es va analitzar a fons el perfil individual de cada investigador.

Així doncs, quan s'identificava un possible investigador per afegir a la base de dades, s'analitzava tant el seu perfil de *Google Scholar* com de *ResearchGate*, mirant els seus articles que tenien relació amb l'esport, ja que la gran majoria de vegades, l'article no l'havia escrit sol, i per tant, possiblement l'havia escrit amb algú amb un perfil similar, que podria ser inclòs a la base de dades. A més, també s'analitzava a quina universitat pertanyia, i s'intentava identificar si hi havia algun grup de recerca relacionat amb l'esport, ja que, en cas afirmatiu, el més probable és que hi haguessin més investigadors.

La recerca d'informació va començar el dia 6 de febrer i aquesta va durar aproximadament un mes i mig, fins al dia 18 de març, en que es va decidir no seguir buscant més ja que, les persones que trobava o bé no complien amb els criteris, que s'explicaran a continuació, o bé ja estaven a la base de dades.

2.1.2. Selecció de les variables d'estudi

Inicialment, abans de realitzar la recerca, es van definir la informació i primeres variables que voldríem que tingués cada registre del cens per tal que, un cop identificat un possible

investigador, es seguís un mateix criteri per a tots, i així totes les variables estarien completes i tindrien la mateixa informació. Aquestes variables es poden veure a la taula 2.1.

Taula 2.1 Taula descriptiva de les variables del cens

VARIABLE	**DESCRIPCIÓ**	**TIPUS**
Name	Nom i cognom de l'investigador/a.	Qualitativa
Gender	Gènere de l'investigador/a.	Qualitativa
Email	Correu electrònic de l'investigador/a.	Qualitativa
City	Ciutat on es troba la universitat de l'investigador/a.	Qualitativa
Country*	País on es troba la universitat de l'investigador/a.	Qualitativa
Center	Universitat o centre al que es troba l'investigador/a.	Qualitativa
Type	Indica si la universitat o centre al que es troba l'investigador/a és públic o privat.	Qualitativa dicotòmica
Research group of Sports analytics?	Indica si la universitat o centre al que es troba l'investigador/a té un grup de *sports analytics/statistics* o no.	Qualitativa dicotòmica
Link of center or research group	Enllaç del grup de *sports analytics/statistics* de la universitat o centre al que es troba l'investigador/a, sempre i quan aquesta en tingui.	Caràcter especial
Number of membres of research group	Nombre de persones que formen el grup de *sports analytics/statistics* de la universitat o centre al que es troba l'investigador/a, sempre i quan aquesta en tingui.	Numèrica discreta
Research group name	Nom del grup de *sports analytics/statistics* de la universitat o centre al que es troba l'investigador/a, sempre i quan aquesta en tingui.	Qualitativa
Total number of citations in Google Scholar	Número total de cites de tots els articles publicats per l'investigador/a que es troben a *Google Scholar*.	Numèrica discreta
Total number of citations in Google Scholar since 2018	Número total de cites de tots els articles publicats per l'investigador/a des del 2018 que es troben a *Google Scholar*.	Numèrica discreta
Total number of citations acording to H-index	Número total de cites de tots els articles publicats per l'investigador/a que es troben a *Google Scholar*, tenint en compte l'índex H.	Numèrica discreta
Total number of citations acording to H-index since 2018	Número total de cites de tots els articles publicats per l'investigador/a des del 2018 que es troben a *Google Scholar*, tenint en compte l'índex H.	Numèrica discreta
Number of citations of the article with most citations	Número total de cites de l'article amb més cites publicat per l'investigador/a a *Google Scholar*	Numèrica discreta
Link of Google Scholar	Enllaç al perfil de *Google Scholar* de l'investigador/a.	Caràcter especial

Country*: Per dur a terme l'anàlisi estadístic, s'han agrupat els països per tal de no tenir tants nivells.

L'índex H és un sistema de mesura que va ser proposat l'any 2005 pel professor Jorge Hirsch (Hirsch, 2005), de la Universitat de Califòrnia, que s'utilitza per mesurar la qualitat dels articles publicats. Per calcular-lo, s'ordenen els articles de més a menys número de cites, i s'enumeren del primer fins a l'últim. Quan un article que conté H cites coincideix amb la seva enumeració, aquell serà l'índex H. Per exemple, si el desè article amb més cites conté 10 cites, en aquest cas H valdrà 10. És a dir, la variable de l'índex H ens indica quants articles, ja siguin totals, o des del 2018, han tingut més de H cites, segons cada investigador.

De forma que, completant totes les variables esmentades, es podia investigar a fons tant les publicacions de l'investigador (a través de *Google Scholar*) com el seu entorn (a través de la universitat i/o grup de recerca), i com a conseqüència, era més fàcil poder trobar nous investigadors.

2.1.3. Procediment d'extracció de dades

Per poder dur a terme la creació completa de la base de dades, s'han acabat seguint uns criteris d'inclusió comentats seguidament que volien apropar-se a l'objectiu del cens i on es volia englobar bàsicament les persones que es dediquen a l'acadèmica del camp de l'estadística esportiva. Els criteris aplicats per saber si un investigador era afegit a la base de dades o no, han estat els següents:

- La persona està actualment lligada a l'acadèmia. Si anteriorment s'ha dedicat a la docència i/o recerca en una institució universitària, però ara mateix només treballa a la indústria, aquesta persona no s'inclourà. Tot i així, hi ha persones que han estat grans pioneres en l'acadèmia i l'estadística esportiva però recentment han passat a treballar en la indústria esportiva. Aquests són per exemple els experimentats en estadística esportiva com en Luke Bornn i en Michael Lopez. També trobem casos similars en joves que han acabat recentment el seu PhD i que han tingut una producció rellevant en aquest àmbit com són per exemple els casos d'en Javier Fernández, en Paolo Cintia i en Bart Spencer. És per aquest motiu que, excepcionalment, aquestes cinc seran incloses al cens.
- La persona ha de tenir els estudis finalitzats. Si la persona actualment ja publica articles sobre *sports analytics/statistics* però no els ha finalitzat (ja sigui grau, màster o doctorat) aquesta no s'inclourà.
- La persona ha de tenir estudis lligats a l'àmbit de l'estadística. No és una condició necessària per excloure alguna persona, però si no en té, s'haurà d'analitzar molt bé el seu perfil i veure tots els articles sobre esport publicats. És a dir, si una persona té

uns estudis que no estan relacionats amb l'estadística però al llarg dels anys s'ha anat especialitzant en *sports analytics/statistics*, també s'inclourà.

- Hi ha moltes persones que es defineixin com a *sports scientist* que fan aplicació de l'estadística en diferents disciplines del cos humà, com per exemple, la fisiologia de l'exercici, la biomecànica o l'exercici i la psicologia de l'esport. És a dir, molt lleugerament toquen l'esport. Tot i així, aquest no és el perfil que busquem i per tant com a norma general no els inclourem, però si després d'analitzar el seu perfil ens adonem que realment una de les seves principals línia de recerca és l'estadística esportiva, finalment sí que serà acceptat al cens.

Una eina per trobar nous investigadors/es ha estat *Google Scholar.* Al perfil de cada persona, cadascuna es pot posar el seus àmbits d'investigació, de forma que és més fàcil trobar-los. Així doncs, s'ha filtrat per les paraules *sports analytics* i *sports statistics*. Amb aquests filtres, apareix molta gent (més de 200), ordenats de més a menys cites. Totes aquelles que tenen menys de 10 cites, no s'han inclòs a la base de dades, perquè quan s'analitzaven els articles de les que tenien poques cites (però més de 10), tots s'acabaven descartant perquè o bé eren articles que no tenien relació amb *sports analytics/statistics* o bé eren articles publicats només a conferències. Per tant, si aquests que ja tenen poques cites no eren acceptats, té poc sentit perdre el temps mirant si incloure o no els que tenien 10 o menys cites.

Simultàniament a aqueta cerca, se'n va realitzar una de literatura gris a través de les eines de cerca de *Google*. És a dir, es va filtrar a través de *Google* per tal d'intentar trobar noves persones. Dos exemples de la cerca més utilitzada van ser a través dels termes *sports analytics research group* i *sports statistics research group*.

Durant el mes i mig que va durar la recerca, inicialment els cens contava amb 257 investigadors. Tot i així, un cop es va donar per tancada aquesta, es va tornar a revisar un per un cada investigador per tal que complissin els criteris definits, de forma que 53 persones van quedar descartades, i el cens definitiu inicial era de 204 persones. D'aquestes 204, 187 teníem clar que complien el perfil que estàvem buscant, i n'hi havia 17 que crèiem que el complien però no n'estàvem segurs, i per tant, es va decidir deixar-los momentàniament al cens, tal i com s'indica a la figura 2.1.

2.1.4. Enquesta com a eina d'obtenció d'informació

A partir de les variables anteriorment mencionades (taula 2.1) on s'identificaven inicialment els investigadors, es va decidir crear una enquesta per ampliar la informació i així tenir un

cens definitiu. Aquesta va ser enviada exclusivament a les 204 persones que teníem al cens definitiu inicial.

Per a la creació de la *survey* s'ha utilitzat l'eina *Qualtrics*. Es tracta d'una plataforma que permet crear enquestes, tant gratuïtament com pagant, i a més permet exportar els resultats en molts formats diferents (CSV, Excel...) de forma que després puguin ser exportats a altres programes molt fàcilment. A més, també té l'opció de crear un enllaç únic per a cada persona, per tal que un cop respon, es pugui saber qui ha respost. Aquesta opció ha estat l'escollida de la forma d'enviament de l'enquesta, ja que era necessari saber què havia respost cadascú per tal de poden confeccionar el cens definitiu.

L'enquesta, enviada personalment a cada investigador mitjançant un correu electrònic on es posava en context sobre qui eren les persones encarregades de la realització del treball, i s'informava que la resposta no seria anònima, però que només seria utilitzada exclusivament per a la realització d'aquest, té un total de set preguntes, les quals cinc són tancades amb una sola opció, una és dicotòmica que passa a ser oberta o tancada segons la resposta, i finalment l'altra és d'opció múltiple amb una opció oberta.

El primer enviament de l'enquesta va ser els dies 30 i 31 de març (es pot veure a l'annex 6.1 què es va enviar). Es va deixar un període de temps perquè tothom pogués respondre fins al punt que el dia 17 d'abril, l'havien respost un total de 83 persones. Aquest mateix dia, es va realitzar un recordatori a les 121 persones que no havien respost, amb la intenció d'obtenir una taxa de respostes més elevada. Així doncs, després del recordatori realitzat el dia 17 d'abril, la van respondre 20 més, de forma que en total **l'enquesta va ser resposta per 103 persones de 204 disponibles, el que suposa una taxa de resposta del 50,49%.**

2.1.5. Tractament de les dades

Gràcies a les 103 respostes de l'enquesta, ens ha permès modificar aquest cens i ajustar-lo encara una mica més a la realitat. De les 17 persones que estaven en dubte, n'han respost 11. 10 han respost que es consideren un perfil acadèmic i 1 no es considera un perfil acadèmic. Aquest investigador i els 6 que estaven en dubte i no han respost, han quedat descartats del cens. Les altres 91 respostes corresponen a investigadors que havíem definit que sí complien el perfil, però tot i així, n'hi ha hagut 5 que han dit que no es consideren de perfil acadèmic i 1 no es dedica a *sports analytics/statistics*, de forma que aquestes 6 persones també han quedat descartades. Els altre 85 restants, han dit que es consideren un perfil acadèmic. L'enquesta ens ha servit per descartar un total de 13 persones. Hi ha 101 persones no han contestat i per tant, és possible que algú el qual creiem que compleix els

criteris, realment no és d'un perfil acadèmic, tal i com ens ha passat amb 6 persones, però com que no han respost l'enquesta, no ho podrem arribar a saber. Per tant, del cens inicial de 204 persones, hem passat a tenir un cens definitiu de 191 persones.

A través de les respostes de l'enquesta, crearem una altra base de dades addicional, ja que en aquesta tindrem més informació que al cens inicial. D'aquestes 103 respostes, hem vist que n'hi ha 6 que han dit que no es consideren acadèmics i 1 que no es dedica a *sports analytics/statistics.* A més, hi ha 11 investigadors que han respost però no tenen perfil de *Google Scholar* i per tant, tampoc els inclourem a aquesta base de dades ja que sinó tindríem varies variables amb valors *missings*. Per tant, la base de dades que serà analitzada tindrà un total de 85 registres.

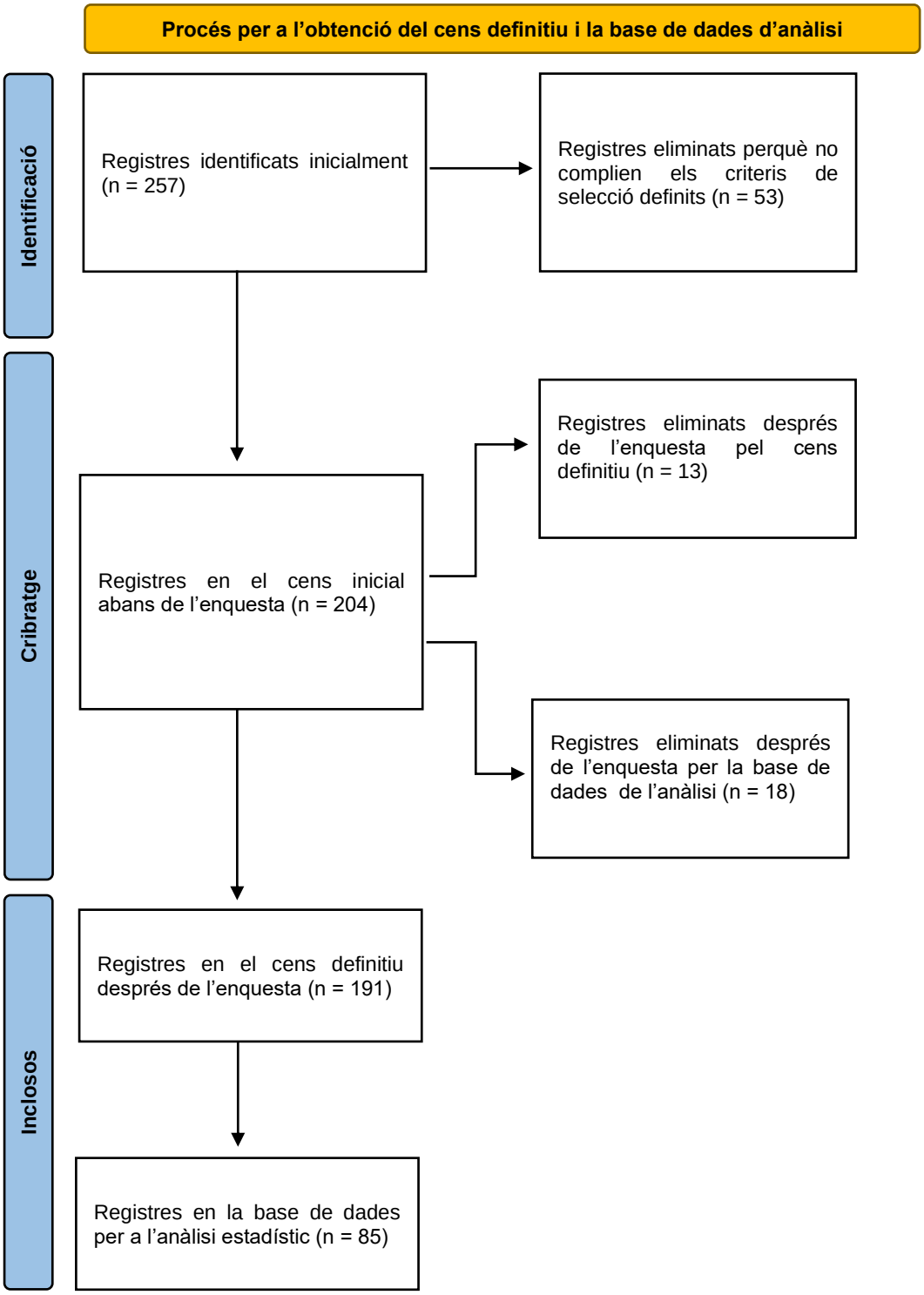
2.1.6. Diagrama de flux

Per identificar bé tot el procés dut a terme, s'ha seguit la guia *PRISMA*. La declaració *PRISMA* (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*), va ser publicada per primera vegada l'any 2009 i es va dissenyar per ajudar als autors de revisions sistemàtiques a documentar la seva revisió (Tricco et al., 2018). Des de el perquè d'aquesta, com s'ha fet i com s'ha trobat la informació. Al llarg d'aquests anys s'han produït avenços en la metodologia i la terminologia de les revisions sistemàtiques, fet que va provocar que l'any 2020 es creés una nova declaració *PRISMA* (Page et al., 2021) que té com a referència la del 2009, però que conté algunes millores. Aquesta nova declaració consta de 27 ítems que s'han de seguir per tal que el diagrama de flux sigui correcte.

Una revisió sistemàtica és un tipus d'estudi científic en el que es recopila tota la informació generada per investigacions d'un tema o pregunta determinat. El seu objectiu és proporcionar una síntesi completa i imparcial en un sol document, utilitzant mètodes rigorosos i transparents, aportant resultats el màxim fiables a partir dels quals es puguin extreure conclusions i prendre decisions.

Així doncs, encara que ja s'ha explicat via escrita quin ha estat el procés per identificar tots els investigadors, a la figura 2.1 es pot visualitzar un diagrama de flux *PRISMA* de tot el procés dut a terme. Des de l'inici fins al final de la creació del cens.

Figura 2.1 PRISMA: Diagrama de flux de l'obtenció del cens definitiu i la base de dades d'anàlisi.



2.1.7. Selecció de les variables de la base de dades per a l'anàlisi estadístic

A través de la figura 2.1 hem pogut veure com s'han escollit les persones que formen part de la base de dades per a l'anàlisi estadístic. Tot i així, les variables d'aqueta base de dades no

són exactament les mateixes que les del cens definitiu. S'ha decidit mantenir només les variables *Gender, Type, Research group* i totes les relacionades amb les cites de *Google Scholar*. A més, se n'han afegit algunes gràcies a la informació de l'enquesta. Les noves variables es poden trobar a la taula 2.2.

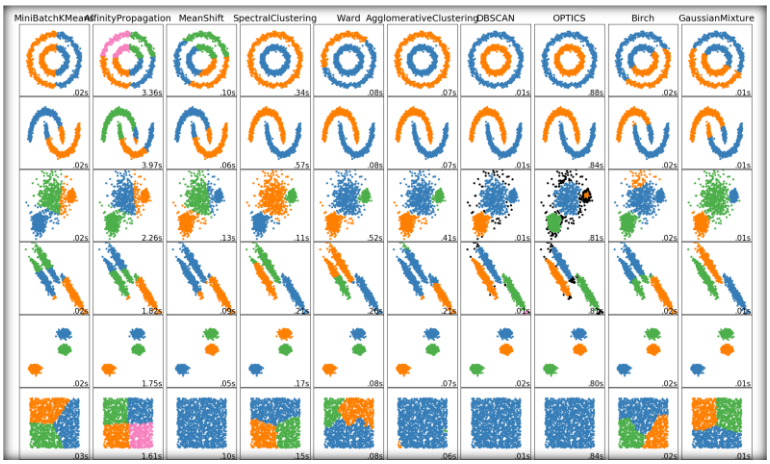
Taula 2.2 Taula descriptiva de les noves variables de la base de dades per a l'anàlisi estadístic

VARIABLE	**DESCRIPCIÓ**	**TIPUS**
Age group	Grup d'edat de l'investigador/a.	Qualitativa
Country categoized	País on es troba la universitat de l'investigador/a. La variable es troba agrupada per tenir menys categories.	Qualitativa
Highest level of education	Nivell d'estudis de l'investigador/a.	Qualitativa
Years worked in sports statistics in academia	Anys treballats a l'acadèmia sobre temes de *sports statistics.*	Numèrica discretitzada
Profile	Perfil acadèmic de l'investigador/a.	Qualitativa
Number of sports studied	Número d'esports amb els que treballa l'investigador/a.	Qualitativa

2.2. *Mètode per analitzar la base de dades: clustering*

Els mètodes *clustering* són una tècnica de *machine learning* i d'aprenentatge no supervisat bastats en agrupar o identificar *clusters* dins d'un conjunt de dades, tenint en compte una determinada mesura de similitud entre les observacions, podent obtenir diferents *clusters* segons la mesura utilitzada (figura 2.2). Un mètode d'aprenentatge no supervisat és aquell on no es necessita tenir una variable resposta. La finalitat és doncs, obtenir diferents grups de forma que les observacions dins de cada grup siguin el més similars possibles entre elles, i el més diferents possibles respecte els altres grups.

Figura 2.2 Possibles diferents clusters segons quina mesura utilitzem



Hi ha molts mètodes de *clustering* diferents, els quals es poden agrupar en dos subgrups, segons quina és la informació per assignar una observació a un *cluster*. El primer subgrup és el *hard clustering* i aquest es basa en distàncies matemàtiques, de forma que assigna a cadascun dels elements del conjunt de dades a un únic *cluster*. Dins aquest subgrup hi trobem dos mètodes principals: el *clustering* jeràrquic i el *clustering* particional. L'altre subgrup és el *soft clustering* i aquest es basa en probabilitats. Aquest algoritme assigna un vector amb la probabilitat de que cadascun dels elements del conjunt de dades pertanyin a cadascun dels possibles *clusters*. En aquest cas els diferents *clusters* no estan connectats, sinó que tots els registres pertanyen a tots els *clusters* amb un grau de pertinença proporcional a la seva probabilitat. Dins d'aquest subgrup hi trobem dos mètodes principals: el *clustering* probabilístic i el *fuzzy clustering*.

En general, el *clustering* és una tècnica que s'utilitza amb només variables numèriques, de forma que aquestes no poden contenir valors *missings*. També és important que, independentment del mètode que s'esculli, les dades estiguin normalitzades.

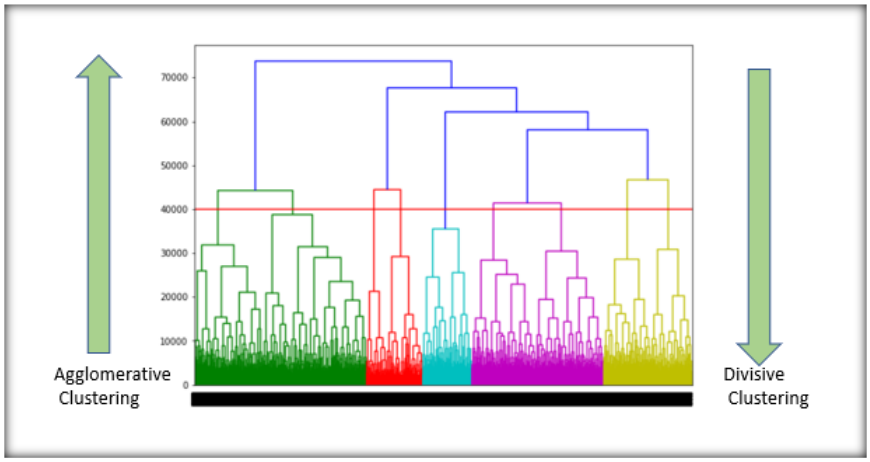
Tenint en compte aquesta explicació, es pot veure que hi ha molts mètodes diferents, tot i així, a continuació s'explicaran els dos més comuns i que a la pràctica s'utilitzen més: el *clustering* jeràrquic i el *clustering* particional.

2.2.1. Clustering jeràrquic

El *clustering* jeràrquic és un mètode que permet obtenir representacions de les observacions en forma d'arbre, anomenat dendrograma. L'objectiu és crear *clusters* mitjançant un agrupament jeràrquic entre les diferents observacions del conjunt de dades. Hi ha dos tipus d'agrupaments (figura 2.3). El primer tipus, el *clustering* jeràrquic aglomerat té un ordre ascendent. Cada observació comença sent un únic grup i aquests es van agrupant mentre es puja la jerarquia, de forma que finalment s'obté un únic grup. L'altre tipus, el *clustering* jeràrquic divisiu és exactament el contrari. Es comença amb un únic grup i té un ordre descendent. Es van desagrupant les observacions mentre es baixa la jerarquia, de forma que finalment s'obté tants grups com observacions tenim.

Les unions que es troben més a baix del dendrograma corresponen al grup d'observacions més similars entre sí mentre que les que s'uneixen més amunt del dendrograma són les més diferents entre sí. Quan més llarga és la línia vertical, més gran és la distància entre els grups.

Figura 2.3 Imatge d'un dendrograma i del seu tipus de clustering jeràrquic aglomerat o divisiu.



El més important en aquest tipus de *clustering* és l'elecció de la mesura de similitud. Segons quina escollim, el dendrograma pot tenir una forma o una altra. Hi ha dos tipus de mesures diferents: la similitud entre dues observacions i la similitud entre dos grups d'observacions.

Per calcular la similitud entre dues observacions tenim:

- Distància Euclidiana:

És la distància entre dos punts mesurada en línia recta en un espai euclidià. La seva fórmula es dedueix mitjançant el teorema de Pitàgores. La seva fórmula és (1).

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

- Distància Manhattan:

S'assembla a la distància Euclidiana però aquesta és la suma de les diferències de les seves coordenades, en valor absolut. La seva fórmula és (2).

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

Existeixen altres mesures com poden ser la distància de correlació de *Pearson*, la distància de correlació de *Spearman* o la distància de correlació de *Kendall*, tot i que s'utilitzen molt menys.

Per mesurar la similitud entre dos grups d'observacions tenim:

- Enllaç complet:

Calcula la distància entre els dos punts més llunyans de cada *cluster*. La seva fórmula és (3).

$$\max\{d(a,b): a \in A, b \in B\} \tag{3}$$

- Enllaç simple:

És el contrari a l'enllaç complet ja que calcula la distància entre els dos punts més propers de cada *cluster*. La seva fórmula és (4).

$$\min\{d(a,b): a \in A, b \in B\} \tag{4}$$

- Distància entre mitjanes:

Calcula la distància entre les mitjanes de la distància de cada *cluster*. La seva fórmula és (5).

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b) \tag{5}$$

2.2.2. Clustering particional

A diferència del *clustering* jeràrquic, on primer aplicàvem l'algoritme i llavors decidíem el nombre de *clusters*, el *clustering* particional requereix saber quants *clusters* tenim, és a dir, per dur a terme l'algoritme, inicialment hem d'indicar el nombre *k* de *clusters*. L'algoritme més comú d'aquest tipus de *clustering* és el *k-means*.

Aquest algoritme, creat per MacQueen el 1967, classifica les observacions de forma que siguin el més semblants possibles entre les del mateix *cluster* (similitud dins de classe alta) i el més diferents possibles entre les de diferents *clusters* (similitud entre classes baixa). A més, cada *cluster* està representat per un centroide, que és la mitjana dels punts assignats al *cluster* corresponent.

Matemàticament, la variació interna dels grups es pot mesurar tal i com s'indica a (6),

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{6}$$

on x_i és un punt que pertany al *cluster* C_k i μ_k és el valor mitjà dels punts assignats al *cluster* C_k. La suma de la variació interna dels grups es defineix com la variació entre els diferents grups. Matemàticament, aquesta és (7).

$$Variació\ entre\ classes = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k}\sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{7}$$

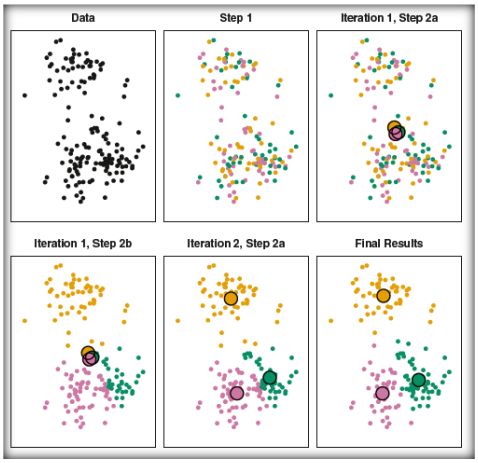
L'algoritme consta d'una sèrie de passos senzills que, a part de visualitzar-los a la figura 2.4, són els següents.

1- S'escullen *k* grups, de forma que serà el nombre de *clusters* que tindrem. Aquests *k* s'assignen a cada observació de forma aleatòria.

2- Es va iterant fins que l'assignació de cada *cluster* deixa de canviar:
2a. Per cada un dels *k clusters*, es calcula el seu centroide.

2b. S'assigna a cada observació al *cluster* on el centroide és el més proper. A més, a cada iteració es minimitza la variància dins de classes i es maximitza la variància entre classes.

Figura 2.4 Formació dels clusters amb l'algoritme k-means.



Com hem comentat, per executar l'algoritme es necessita assignar un nombre *k* de *clusters*. Si no es coneix aquest nombre, hi ha diferents mètodes per seleccionar quin és el nombre òptim de *clusters*. Els més populars són els següents:

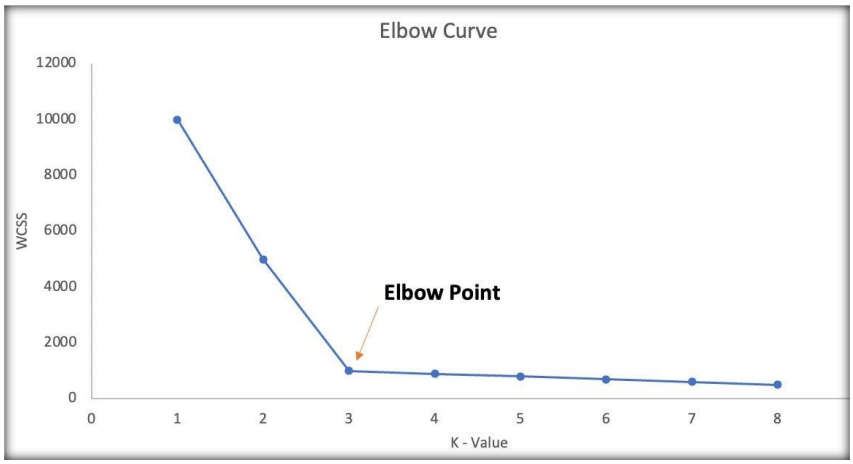
- Elbow method (mètode del colze):

El mètode del colze, inventat el 1953 per Robert L. Thorndike, consta de quatre passos diferents. Primer es calcula l'algoritme *k-means* per a diferents valors de *k,* on *k* = 1, 2, ..., 10. Seguidament, per a cada valor de *k* es calcula la suma quadràtica total entre els grups d'observacions (*wss*), tal i com s'indica en (8).

$$wss = \sum_{k=1}^{k} W(C_k) \tag{8}$$

Després, es dibuixa la corba segons el valor de cada *wss* i cada valor de *k*. Finalment, aquell punt on la corba faci una forma de colze, serà el número de cluster òptim (figura 2.5).

Figura 2.5 Exemple del mètode del colze (Elbow method).



- *Silhouette method*

El mètode *Silhouette*, inventat per Peter Rousseeuw el 1990, és una mesura que compara la similitud d'un punt dins un *cluster*, en comparació els altres *clusters*, pels diferents valors de *k*. Aquesta mesura s'anomena *average silhouette*, i el nombre òptim de *clusters k* és el que maximitza aquest valor, dins un rang de valors possibles de *k*. Per calcular l'*average silhouette*, primer necessitem calcular el *silhouette coeficient*. Aquest, per un punt particular es calcula a partir de la fórmula (9),

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

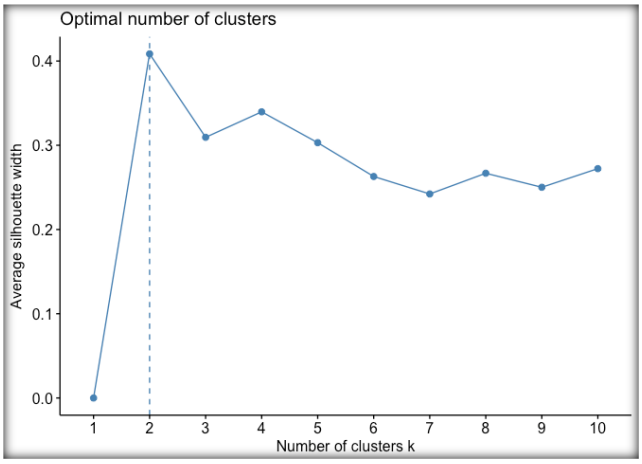
on $a(i)$ és la distància mitjana entre i i els altres punts del *cluster* on es troba i, i $b(i)$ és la distància mitjana entre i i tots els *clusters* als que no pertany i. Un cop tenim aquest valor, ja podem calcular l'*average silhouette*, que es calcula a partir de la fórmula (10),

$$\hat{S}(i) = \frac{\sum_{i=1}^{n} S(i)}{n} \tag{10}$$

on n és el nombre total de punts del *cluster k.*

Un cop tenim calculat aquest valor segon cada valor de *k*, podem representar-ho gràficament, de forma que ens permetrà veure fàcilment quin és el *cluster* amb un major valor i així identificar el valor de *k* (figura 2.6).

Figura 2.6 *Exemple del silhouette method.*



- *Gap statistic method*

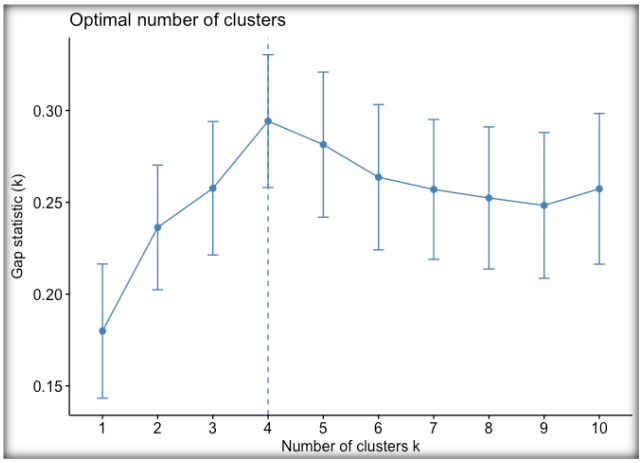
El mètode *Gap statistic*, inventat per Tibshirani, Wlther i Hastie el 2001, el que fa és comparar la variació total interna dels *clusters* per a diferents valors de *k*, amb els seus valors esperats sota una distribució de referència nul·la de les dades. El valor per aquest estadístic es calcula tal i com indica (11),

$$Gap_n(k) = E_n^* \log(W_k) - \log(W_k) \tag{11}$$

on E_n^* s'obté mitjançant *bootstrapping* generant B còpies de la base de dades de referència. Aquesta base de dades s'ha generat inicialment amb les simulacions de Monte Carlo en un procés de mostreig. Pel que fa a W_k, com s'ha vist anteriorment, s'obté amb la variació interna dels grups.

Igual que en els altres dos mètodes, el valor d'aquest estadístic es pot representar gràficament, i el valor òptim de *k* és aquell on el valor del *gap estadístic* és màxim (figura 2.7).

Figura 2.7 Exemple del gap statistic method.



2.2.3. Tècnica de clustering alternativa: KAMILA

2.2.3.1. Descripció

Malgrat que el que hem vist fins ara són les tècniques de *clustering* més conegudes i més utilitzades, la característica principal és que en totes s'utilitzen només variables numèriques, ja que aquestes es basen en distàncies matemàtiques. Pel que fa a la base de dades del treball, creada a través del cens i l'enquesta realitzada, és cert que té alguna variable numèrica, però en general, hi predominen les variables categòriques. És per aquest motiu que el més eficient no és utilitzar un mètode de *clustering* només per a variables numèriques, sinó que s'ha optat per una tècnica que sigui capaç de modelar variables numèriques i categòriques alhora, de forma que utilitzarem el paquet *KAMILA*.

L'algoritme *KAMILA* (*KAy-means for MIxed LArge data*) (Foss et al., 2016) és una tècnica de *clustering* que permet combinar equitativament variables quantitatives i qualitatives. Aquest combina dos dels algoritmes més populars, per les variables numèriques utilitza el ja esmentat *k-means*, mentre que per les variables categòriques utilitza els models de mixtura gaussiana-multi-nominal (Peel, D., & MacLahlan, G. (2000); StataCorp, L. L. C. (2017)). Igual que l'algoritme *k-means*, el *KAMILA* no fa suposicions paramètriques fortes en relació a les variables quantitatives. Tot i així, aconsegueix evitar el tractament de les dades desbalancejades entre variables quantitatives i qualitatives, basant-se en la distància Euclidiana.

El model de mixtura gaussià-multi-nominal és un algoritme basat en models en els que cada distribució representa un *cluster*, i aquesta distribució és una normal. Cada observació se l'assigna a la distribució que té la probabilitat més alta de pertànyer a una de les distribucions. De manera similar, l'algoritme *KAMILA* és capaç d'equilibrar les variables qualitatives i quantitatives sense la necessitat de seleccionar pesos, es basa amb un estimador de densitat adequat calculat a partir de les pròpies dades, cosa que fa reduir les suposicions més estrictes de la distribució gaussiana.

En resum, l'algoritme *KAMILA* és òptim per treballar amb dades mixtes, ja que a més també s'eviten suposicions paramètriques restrictives i no fa falta indicar ponderacions úniques per les variables.

2.2.3.2. Formulació

Per formular l'algoritme *KAMILA*, prèviament es necessita conèixer alguna definició i notació bàsica. Definirem el següent:

- V_1, ..., V_i, ..., V_{85} és un vector 6 x 1 de variables numèriques independent i idènticament distribuïda (i.i.d.) amb densitat *h* on $V_i = (V_{i1}, ..., V_{i6}, ..., V_{85,6})^T$, amb V_i ~

f_v (v) = $\sum_{g=1}^{G} \pi_g h(v; \mu_g)$ on G correspon al nombre de *clusters*, μ_g és el P x 1 centroide del *cluster* gth de V_i i π_g és la probabilitat a priori d'observar la població gth. A més $\sum_{g=1}^{G} \pi_g = 1$ i $\pi_g \in [0,1]$.

- W_1, ..., W_i, ..., W_{85} és un vector 16 x 1 de variables categòriques i.i.d. on cada element és una barreja de variables aleatòries on $W_i = (W_{i1}, ..., W_{iq}, ..., W_{85,16})^T$ amb $W_{iq} \in \{1,..., l, ..., L_q\}$ i $W_i \sim f_w(w) = \sum_{g=1}^{G} \pi_g \prod_{q=1}^{Q} m(w_q; \theta_{gq})$ on $m(w; \theta) = \prod_{l=1}^{L_{16}} \theta_l^{I\{w=l\}}$ és la funció de probabilitat multi nominal, I{·} és l'indicador de la funció i θ_{gq} és el vector L_q x 1 de paràmetres de la funció multi nominal corresponent a la variable aleatòria qth del *cluster* gth.
- A la iteració t, $\hat{\mu}_g^{(t)}$és l'estimador del centroide de la població *g* i $\hat{\theta}_{gq}^{(t)}$és l'estimador dels paràmetres de la distribució multi nominal corresponent a la variable aleatòria discreta qth de la població g.

En general, quan les variables categòriques no són independents, les podem modelitzar per una nova variable categòrica amb un nivell categòric per cada combinació de nivells de les variables dependents.

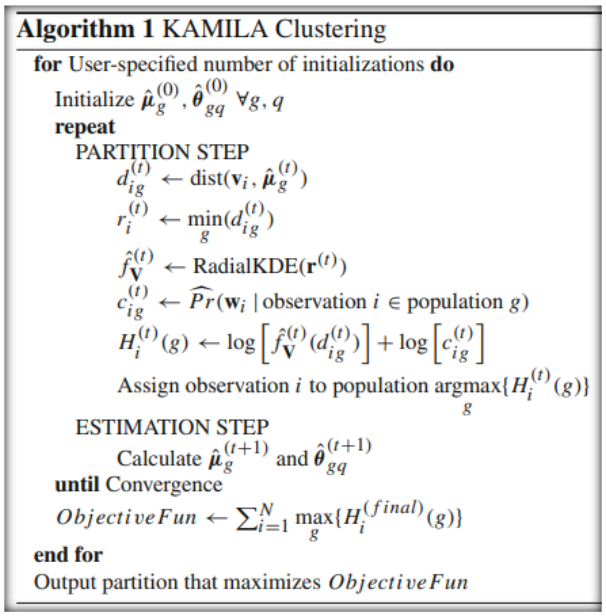
A més, el *KAMILA* també ens permet identificar agrupacions esfèriques o el·líptiques que s'han d'especificar abans d'executar l'algoritme. Una agrupació esfèrica o el·líptica és aquella funció que es defineix mitjançant equacions cúbiques.

El que volem és calcular d'una manera eficient les densitats conjuntes de distribucions esfèriques multivariades utilitzant estimacions de densitat *kernel*. Una estimació de densitat *kernel* (Cao et al., 1994) és un mètode que s'utilitza per estimar densitat de dades que no tenen comportaments estadístics paramètrics, permetent passar d'un problema difícil de classificació no lineal a un problema senzill de classificació lineal.

Aquestes estimacions de densitat *kernel* pateixen problemes de dimensionalitat i sobreajustament quan hi ha moltes dades, que provoquen que hi hagi estimacions de densitat molt altes en alguns punts i estimacions de densitat siguin molt baixes en altres punts. La solució perquè això no passi és primer fer agrupacions esfèriques i després agrupacions el·líptiques. Utilitzat les propietats d'aquestes distribucions ens permet obtenir estimacions de densitat *kernel* més precises i ràpides de calcular, de forma que obtenim densitats simètriques radialment al voltant d'un vector mitjà, és a dir, densitat que només depenen de la distància entre la mostra i el centre de la distribució.

Així doncs, amb aquesta petita base, ja ens serveix definir el pseudocodi de l'algoritme *KAMILA* (figura 2.8).

Figura 2.8 Algoritme KAMILA



Algorithm 1 KAMILA Clustering

for User-specified number of initializations **do**

Initialize $\hat{\boldsymbol{\mu}}_g^{(0)}, \hat{\boldsymbol{\theta}}_{gq}^{(0)}\ \forall g, q$

repeat

PARTITION STEP

$d_{ig}^{(t)} \leftarrow \text{dist}(\mathbf{v}_i, \hat{\boldsymbol{\mu}}_g^{(t)})$

$r_i^{(t)} \leftarrow \min_g(d_{ig}^{(t)})$

$\hat{f}_{\mathbf{V}}^{(t)} \leftarrow \text{RadialKDE}(\mathbf{r}^{(t)})$

$c_{ig}^{(t)} \leftarrow \widehat{Pr}(\mathbf{w}_i \mid \text{observation } i \in \text{population } g)$

$H_i^{(t)}(g) \leftarrow \log\left[\hat{f}_{\mathbf{V}}^{(t)}(d_{ig}^{(t)})\right] + \log\left[c_{ig}^{(t)}\right]$

Assign observation i to population $\underset{g}{\text{argmax}}\{H_i^{(t)}(g)\}$

ESTIMATION STEP

Calculate $\hat{\boldsymbol{\mu}}_g^{(t+1)}$ and $\hat{\boldsymbol{\theta}}_{gq}^{(t+1)}$

until Convergence

$ObjectiveFun \leftarrow \sum_{i=1}^{N} \max_g\{H_i^{(final)}(g)\}$

end for

Output partition that maximizes $ObjectiveFun$

Taula 2.3 Taula de conceptes importants per l'algoritme KAMILA

Conceptes	**Descripció**
Dist (V_i, $\hat{\mu}_g^{(t)}$)	Distància Euclidiana entre la variable numèrica V_i i el valor $\hat{\mu}_g^{(t)}$ calculat.
$\underbrace{min}_{g}(d_{ig}^{(t)})$	Valor mínim de les distàncies Euclidianes calculades.
RadialKDE($r^{(t)}$)	Funció de densitat Kernel del valor mínim calculat.
$\widehat{Pr}\ (w_i \mid observation\ i \in population\ g)$	Probabilitat de la variable categòrica w_i condicionada a que l'observació *i* pertanyi al *cluster g*.

Els passos de l'algoritme són els següents:

1) Per començar primer té en compte les variables numèriques i s'inicialitza amb $\hat{\mu}_g^{(0)}$que té un valor aleatori d'una distribució uniforme amb límits iguals al mínim i al màxim de la variable p^{th}, mentre que $\hat{\theta}_{gq}^{(0)}$ s'obté d'una extracció d'una distribució de Dirichlet (Di Nardo et al., 2021) on tots els paràmetres de forma són 1.

2) A continuació, l'algoritme s'inicia més d'un cop i s'executa fins que s'arriba al número màxim d'iteracions especificat inicialment o bé fins que la població no canvia respecte la iteració anterior. Cada iteració consta de dos passos generals, un de partició i un d'estimació.

3) A la iteració t^{th}, quan ja es té un valor per $\hat{\mu}_g^{(t)}$i $\hat{\theta}_{gq}^{(t)}$, al pas de partició s'assignen N observacions en els G grups. Llavors es calcula la distància Euclidiana entre la observació *i* i el valor $\hat{\mu}_g^{(t)}$. Es calcula tal i com indica (12),

$$d_{ig}^{(t)} = \sqrt{\sum_{p=1}^{P} [\xi_p (v_{ip} - \hat{\mu}_{gp}^{(t)})]^2}, \tag{12}$$

on ξp és una ponderació opcional corresponent a la variable *p*.

4) A continuació, es calcula la distància mínima distància per a la observació i[th] com $r_i^{(t)} = \underbrace{min}_{g} (d_{ig}^{(t)})$.

5) Després es calcula la densitat *kernel*. Utilitzant el paràmetre calculat de la distància mínima i la fórmula (13),

$$\hat{f}_R^{(t)}(r) = \frac{1}{N h^{(t)}} \sum_{\ell=1}^{N} k\left(\frac{r - r_\ell^{(t)}}{h^{(t)}}\right), \tag{13}$$

on $h^{(t)}$ és el paràmetre d'amplada corresponent a la iteració *t*, s'obté el valor necessari per aplicar la funció de densitat *kernel*, que és (14),

$$f_{\mathbf{V}}(\mathbf{v}) = \frac{f_R(r)\, \Gamma(\frac{p}{2} + 1)}{p r^{p-1} \pi^{p/2}}, \tag{14}$$

on Γ correspon a la funció Gamma que es calcula de la forma, $\Gamma(n) = (n-1)!$

A més, en aquesta funció π no és una probabilitat, tal i com s'havia definit anteriorment, sinó que fa referència al número irracional 3,14.

Aquesta és la part que es treballa amb les diferents variables numèriques. Cal destacar que, per un millor funcionalment de l'algoritme, és important estandarditzar-les. Un cop fet això, l'algoritme treballa amb les categòriques.

6) Aquest suposa independència entre totes les variables categòriques Q dins d'un *cluster g*, i calcula la probabilitat logarítmica d'observar el vector categòric i[th] que pertany a la població com a $log\left(c_{ig}^{(t)}\right) = \sum_{q=1}^{Q} \xi_q * \log(m(w_{iq}; \hat{\theta}_{gq}^{(t)}))$ on m(·;·) és la probabilitat de la funció multi nominal i ξ_q és una ponderació opcional corresponent a la variable *q*. Si no s'especifica, la ponderació serà per defecte 1.

7) Un cop calculat això, es realitza l'assignació del *cluster* calculant el logaritme de la densitat *kernel* multiplicat per la distància Euclidiana i sumant la probabilitat logarítmica. La fórmula és (15).

$$H_i^{(t)}(g) = \log\left[\hat{f}_{\mathbf{V}}^{(t)}(d_{ig}^{(t)})\right] + \log\left[c_{ig}^{(t)}\right], \tag{15}$$

L'observació *i* s'assignarà al grup *g* que maximitzi el valor $H_i^{(t)}(g)$.

8) Quan l'observació s'ha assignat a un grup, s'acaba el pas de partició i comença el d'estimació. Aquest consisteix en calcular els valors $\hat{\mu}_g^{(t+1)}$ i $\hat{\theta}_{gq}^{(t+1)}$ per tots els valors de g, p, q. Sigui $\Omega_g^{(t)}$el conjunt d'índexs de les observacions assignades a la població *g* en la iteració *t*, els dos paràmetres es calculen tal i com indica (16).

$$\hat{\boldsymbol{\mu}}_g^{(t+1)} = \frac{1}{\left|\Omega_g^{(t)}\right|} \sum_{i \in \Omega_g^{(t)}} \mathbf{v}_i \tag{16}$$

$$\hat{\theta}_{gq\ell}^{(t+1)} = \frac{1}{\left|\Omega_g^{(t)}\right|} \sum_{i \in \Omega_g^{(t)}} I\{w_{iq} = \ell\} \tag{17}$$

9) Els passos de partició i estimació es repeteixen fins que s'obté una solució o s'arriba al màxim número de repeticions. És a dir, l'objectiu final és que per a cada inicialització, calculem la funció objectiu (18).

$$\sum_{i=1}^{N} \max_g \{H_i^{(final)}(g)\}. \tag{18}$$

Per determinar el nombre òptim de *clusters* tenim diferents formes de fer-ho. Amb el software estadístic R, un cop s'ha executat l'algoritme, hi ha l'opció de determinar automàticament quin és aquest número, mitjançant una simple línia de codi. També hi ha l'opció de determinar-lo gràficament a través de la mesura *prediction strength* o força de predicció (Tibshirani & Walther, 2005), que el que fa és estimar la força de predicció de l'algoritme *KAMILA* per un nombre de *clusters* determinats. El *cluster* que tingui una major *prediction strength,* serà el nombre de *cluster* òptim. Aquest valor es calcula de la següent manera:

1) Es divideix la base de dades en dues (les anomenem *training* i *test*).
2) S'aplica l'algoritme, en el nostre cas, el *KAMILA*, a les dues bases de dades per un valor *k* en concret.
3) Es crea una matriu D[C(X_{tr},k),X_{te}] de mida n_{te} x n_{te} on n_{te} correspon al nombre d'observacions de la base de dades *test* i i C(X_{tr},k) correspon al nombre de *clusters* obtingut després d'executar l'algoritme amb la base *training.*

4) A l'element ii'th de la matriu descrita al pas anterior se li assigna 1 si tant l'element i com el i' de la base de dades *test* pertanyen al mateix *cluster*, i se li assigna 0 en cas contrari.
5) Ara hem de determinar si els centroides prediuen correctament les dades. Per cada parell d'observacions (ii') de la matriu *test* que hem vist que pertanyen al mateix *cluster* (i per tant, si li ha assignat valor 1), es determina si el mateix parell d'observacions ii' de la matriu *training* també es troben al mateix *cluster* i per tant, tenen el mateix centroide.
6) El valor *prediction strength* es calcula amb la fórmula (19),

$$\mathrm{ps}(k) = \min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{i \neq i' \in A_{kj}} D[C(X_{\mathrm{tr}}, k), X_{\mathrm{te}}]_{ii'}. \qquad (19)$$

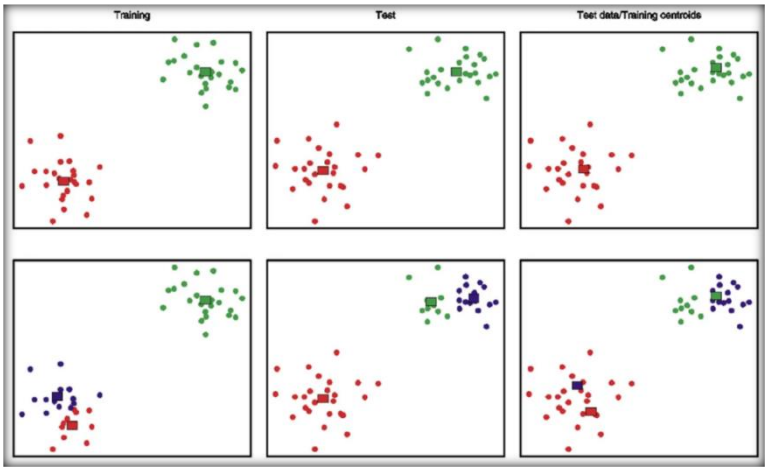
on n_{kj} és el nombre d'observacions del *cluster* jth.

7) Finalment, per cada *cluster* de la base de dades *test*, calculem la proporció del parell d'observacions que també s'han assignat al mateix *cluster* però utilitzant la base de dades *training*. El *prediction strength* és el mínim d'aquesta quantitat sobre els *k clusters*.

Realitzem els passos 2)-7) per a cada *cluster* i a continuació, seleccionem el nombre de *cluster* amb un *prediction strength* més alt, sempre que superi el llindar marcat. Es recomana que aquest llindar es trobi entre 0,8 i 0,9.

Un exemple visual dels passos descrits es pot visualitzar a la figura 2.9.

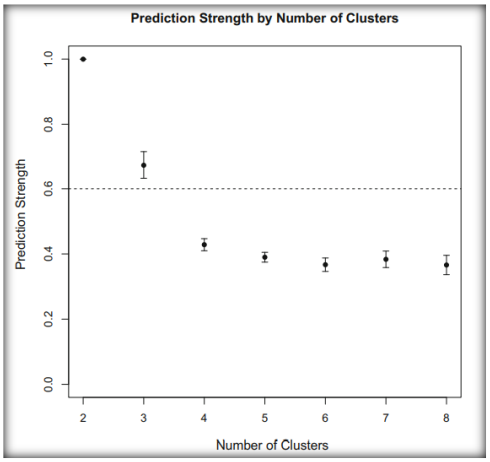
Figura 2.9 Exemple de clusters definits aplicant el prediction strength.



La primera columna pertany a només la base de dades *training*, la segona a la base de dades *test* i la tercera als dos alhora. Si ens fixem amb la primera fila, amb *k = 2* tant una com l'altra base defineixen el mateix centroide i per tant, com que coincideix, tindrà un *prediction*

strength elevat. En canvi, per la segona fila, amb *k = 3* els centroides no són els mateixos i un cop es combinen, la predicció no és gaire bona, cosa que provocarà que el *prediction strength* sigui més baix. Un exemple del gràfic que es crea és la figura 2.10.

Figura 2.10 Exemple per determinar el nombre de clusters gràficament tenint en compte el prediction strength.



Tal i com es pot veure, només amb dos i tres *clusters* s'obté una valor de *prediction strength* que supera el llindar establert. Tot i així, com que el primer és 1 i per tant supera el 0,9, que és el valor màxim que es recomana establir, escollirem que el nombre òptim de *clusters* és 3.

2.2.4. Estadístic de Hopkins

L'estadístic de Hopkins (HOPKINS & SKELLAM, 1954) és un estadístic que serveix per mesurar la tendència del *cluster* en el conjunt de dades. La hipòtesi nul·la és que les dades es distribueixen de forma uniforme mentre que la hipòtesi alternativa és que les dades no es distribueixen de forma uniforme. Si s'obté un valor proper a 1 significa que les dades poden ser fàcilment clusteritzades. Un valor proper a 0,5 significa que les dades són totalment aleatòries, i un valor propera 0 significa que les dades estan distribuïdes uniformement. Per tant, per tal de poder realitzar el *clustering* correctament, ens interessarà tenir un valor de l'estadístic de Hopkins proper a 1. Aquest es calcula tal i com indica (19),

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} \tag{19}$$

on *d* és la dimensió de les dades, u_i és la distància mínima entre $y_i \in Y$ i la observació més propera de X, i w_i és la distància mínima de $x_i \in X$ i la observació més propera $x_j \in X$, on $x_i \neq x_j$.

2.3. EDA

Un *EDA* (*Exploratory Data Analysis*), tal i com es pot intuir del seu nom, serveix per examinar, estudiar, descriure i resumir les característiques d'un conjunt de dades, intentant maximitzar la seva comprensió.

Encara que un *EDA* pot ser molt extens i englobar diferents conceptes d'anàlisi estadístic, n'hi ha dos que són bàsics per poder-lo dur a terme. Aquest són els següents:

- Mesurament i descripció: Es duu a terme a través de l'estadística descriptiva univariant. Per les variables categòriques es tracta de conèixer les freqüències de cada categoria mentre que per les variables numèriques es tracta de conèixer les mesures estadístiques més importants: tant mesures de tendència central (mitjana, mediana,...) com mesures de dispersió (variància, rang,...).
- Comparació: Es duu a terme a través de l'estadística descriptiva bivariant. Per les variables categòriques es tracta de crear taules de contingència per veure possibles relacions. La hipòtesi nul·la és que les dues categories són iguals mentre que la hipòtesi alternativa és que les dues categories són diferents. Per saber si aquestes comparacions ho són estadísticament, tenim dues opcions.
  Si les freqüències són de 6 o més, es realitza una prova chi-quadrat (Dzhaparidze & Nikulin, 1995). L'estadístic a calcular és (20),

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}. \tag{20}$$

  on O_i són les freqüències observades i E_i són les freqüències esperades. Aquest segueix una distribució chi-quadrat amb (número de línies - 1) * (número de columnes - 1) graus de llibertat.

  Si les freqüències són de 5 o menys, es realitza una prova exacte de Fisher (Fisher, 1922). Aquesta prova es basa en la distribució hipergeomètrica i el p-valor es calcula tal i com indica (21).

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!} \tag{21}$$

  Per comparar dues variables numèriques, primer necessitem saber si segueixen una distribució normal o no. Per saber-ho, hem de fer una prova de Shapiro-Wilks (SHAPIRO & WILK, 1965). La hipòtesi nul·la és que els dades segueixen una distribució

normal, mentre que la hipòtesi alternativa és que les dades no segueixen una distribució normal. L'estadístic a calcular és (22).

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{22}$$

En cas que segueixin una distribució normal, es realitzarà una prova *t-student* (David & Gunnink, 1997) on la hipòtesi nul·la és que no hi ha diferència entre els les dues variables i la hipòtesi alternativa és que hi ha diferència entre les dues variables. L'estadístic a calcular és (23),

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{23}$$

on

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}}.$$

Aquest estadístic segueix una *t-student* amb n_1+n_2-2 graus de llibertat.

Si les dades no segueixen una distribució normal, la prova a realitzar és la U de Mann-Whitney (Fay & Proschan, 2010). L'estadístic de la prova és el valor mínim de (24),

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \\ U_2 &= n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \end{aligned} \tag{24}$$

On R_1 i R_2 corresponen a la suma dels rangs. Com que és una prova no paramètrica, l'estadístic no segueix cap distribució.

A banda de totes aquestes proves, amb les variables numèriques també es poden estudiar les seves correlacions, ja sigui amb una matriu bàsica de correlacions o bé amb un anàlisi més complexa com pot ser un anàlisi multivariant.

3. RESULTATS

3.1. EDA

3.1.1. EDA del cens definitiu (n = 191)

3.1.1.1. Anàlisi univariant

A partir dels registres del cens definitiu després de l'enquesta (**n = 191)** (figura 2.1) es descriuen cadascuna de les característiques dels acadèmics esportius (taula 3.1)

Taula 3.1 Taula resum de l'anàlisi univariant de les variables més rellevants del cens definitiu.

Variables categòriques (n = 191)	Categoria				n (%)
Survey answered?	*Yes*				96 (50,26%)
	No				95 (49,74%)
Gender	*Male*				166 (86,91%)
	Female				25 (13,09%)
Country categorized	*United States*				54 (28,27%)
	Italy				34 (17,80%)
	Canada				19 (9,95%)
	Australia				13 (6,81%)
	Spain				12 (6,28%)
	*Other: Rest of Europe**				47 (24,61%)
	*Other: Rest of the world***				12 (6,28%)
Type	*Public*				150 (78,53%)
	Private				41 (21,47%)
Currently belong to a sports statistics research group?	*No*				97 (50,79%)
	Yes				94 (43,21%)
Variables numèriques (N = 191)	Mínim	Màxim	Mediana	Mitjana	Desviació típica
Number of members	2	60	33,16	31,5	26,788
Total citations in GS	14	58650	942	3851,7	8610,173
Total citations in GS since 2018	11	56598	595	1999,7	5350,345
Total citations in GS acording H-index	2	108	14	21,13	18,029
Total citations in GS acording H-index since 2018	2	67	12	15,80	11,661
Number of citations of the article with most citations	6	34771	168	766	2904,971

Other: Rest of Europe*: Belgium, Czech Republic, Denmark, France, Germany, Greece, Iceland, Ireland, Luxemburg, Netherlands, Norway, Serbia, Sweden, United Kingdom.

Other: Rest of the World**: Argentina, China , India, Iran, Japan, South Africa, South Korea.

Podem destacar que el 50,26% han respost l'enquesta mentre que el 49,74% no l'han respost, hi ha una gran presència d'homes (més del 85%) respecte les dones i en relació als països, Estats Units destaca sobre la resta amb 54 persones i el segueix Itàlia amb 34. Pel que fa al tipus d'universitat, la gran majoria són públiques (gairebé el 80%) i la resta són privades, de les quals un 73,17% pertanyen a Estats Units. En referència als grups de recerca, el 50,79% pertanyen a un grup.

En quant a les variables numèriques, cal dir que per calcular els diferents paràmetres no s'han tingut en compte aquelles persones que no pertanyen a cap grup de recerca i/o no tenen perfil de *Google Scholar*. Exceptuant la variable del nombre de membres del grup de recerca, podem pensar que les variables en relació a les cites tenen un comportament semblant, ja que hi ha molta diferència entre el valor mínim i màxim, i la desviació típica és molt gran.

3.1.1.2. Anàlisi bivariant

Com s'ha vist a la taula 3.1, a partir de les 191 persones s'ha identificat aquelles segon si han contestat o no l'enquesta. Aquesta serà la nostra variable resposta per veure el seu comportament respecte les altres variables esmentades i per fer-ho, es realitzarà amb el paquet *compareGroups()* (Isaac Subirana, Hector Sanz, Joan Vila (2014)) del programari estadístic R. Per defecte, aquesta funció ja detecta les proves que ha de realitzar però, es pot comprovar fàcilment quines s'han de dur a terme. Pel que fa a les variables categòriques, amb la taula 3.1 de l'anàlisi univariant es pot veure que totes les categories tenen una freqüència superior a 5 i per tant, la prova a realitzar és la de chi-quadrat. En relació a les variables numèriques, després de realitzar un test de Shapiro-Wilk, s'obté un p-valor inferior a 0,05 per a totes les variables numèriques i per tant, aquestes no segueixen una distribució normal i per dur a terme la comparació es realitzarà un test U de Mann-Whitney.

Els resultats de la comparació entre la variable resposta i la resta de variables es troben a la taula 3.2.

Taula 3.2 Taula resum de l'anàlisi bivariant de totes les variables més rellevants del cens definitiu.

Variable	**N**	**p-valor**	**Mètode**
Gender	191	0,978	Categòric. Test chi-quadrat
Country categorized	191	0,022	Categòric. Test chi-quadrat
Type	191	0,308	Categòric. Test chi-quadrat
Research group?	191	0,096	Categòric. Test chi-quadrat
Number of members	191	0,060	Continu no normal
Total citations in GS	191	0,825	Continu no normal

Total citations in GS since 2018	191	0,751	Continu no normal
Total citations in GS acording H-index	191	0,991	Continu no normal
Total citations in GS acording H-index since 2018	191	0,863	Continu no normal
Number of citations of the article with most citations	191	0,842	Continu no normal

Si ens fixem amb la columna del p-valor, veiem que amb un nivell de significació del 0,05 només és significativa la variable *Country categorized*. I amb un nivell de significació del 0,1 també ho són les variables *Research group?* i *Number of members*. La resta, no són significatives. És a dir, no hi ha diferències entre el grup que ha respost l'enquesta i el que sí l'ha respost. El fet que aquestes tres variables siguin significatives provoca que els resultats puguin afectar la validesa externa, és a dir, no es pot aplicar aquests resultats d'una mostra concreta, a tota la població. Malgrat tot, es pot analitzar quines són les principals diferències, sempre tenint en compte que les conclusions extretes només seran aplicables a la mostra.

Pel que fa a la de *Country categorized*, aquesta significació, que és de 0,022, és deguda a que el 100% de les persones d'Espanya han respost l'enquesta i no n'hi ha cap que no l'hagi respost. El fet que aquests investigadors espanyols coneguin qui els hi ha passat l'enquesta, ha influenciat en que Espanya tingui una taxa de resposta del 100%. Un altre país on hi ha diferències importants entre un grup i altre és Estats Units, on les persones que no han respost l'enquesta representen el 33,7% mentre que els que sí l'han respost representen només el 22,9%. La resta de països tenen un percentatge similar tots dos grups. Aquesta informació la trobem a la taula 3.3.

Taula 3.3 Taula bivariant entre la variable Country categorized i Survey answered?.

País	**NO (N = 95)**	**SÍ (N = 96)**
Austràlia	5 (5,26%)	8 (8,33%)
Canada	10 (10,5%)	9 (9,38%)
Itàlia	17 (17,9%)	17 (17,7%)
Altres: Resta d'Europa	25 (26,3%)	22 (22,9%)
Altres: Resta del món	6 (6,32%)	6 (6,25%)
Espanya	0 (0,00%)	12 (12,5%)
Estats Units	32 (33,7%)	22 (22,9%)

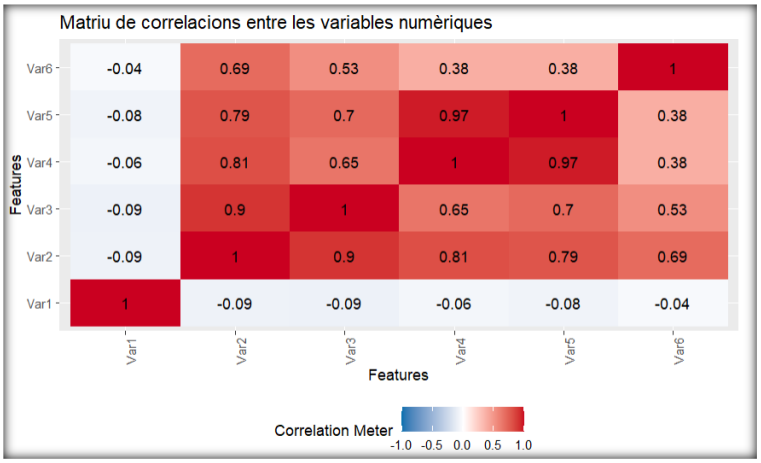
L'altra variable categòrica significativa era la de si l'investigador pertany a un grup de recerca o no. Amb la taula 3.4 podem veure que la diferència entre un grup i altre és deguda a que, els que no han respost l'enquesta hi predomina la gent que pertany a un grup de recerca, mentre que els que han respost l'enquesta hi predominen les persones que no pertanyen a un grup de recerca.

Taula 3.4 Taula bivariant entre la variable Research group segons i Survey answered?.

Grup de recerca?	NO (N = 95)	SÍ (N = 96)
NO	42 (44,2%)	55 (57,3%)
SÍ	53 (55,8%)	41 (42,7%)

En relació a les variables numèriques, encara que només una sigui significativa, realitzem una matriu de correlacions de totes elles i, degut a que el seu nom és llarg i no es visualitzaria bé la matriu, les anomenarem *Var1-6* on l'ordre de les variables és el mateix amb el que apareixen a la taula 3.1. Amb el gràfic de correlacions (figura 3.1) podem veure que la variable 1 no té gairebé relació amb cap altra i la poca que té és negativa. En canvi, la correlació entre la variable 4 i la 5 és la més alta de totes. Molt semblant és la correlació entre les variables 2 i 3. Finalment, pel que fa a la variable 6, té unes correlacions més baixes que la resta, exceptuant la variable 1.

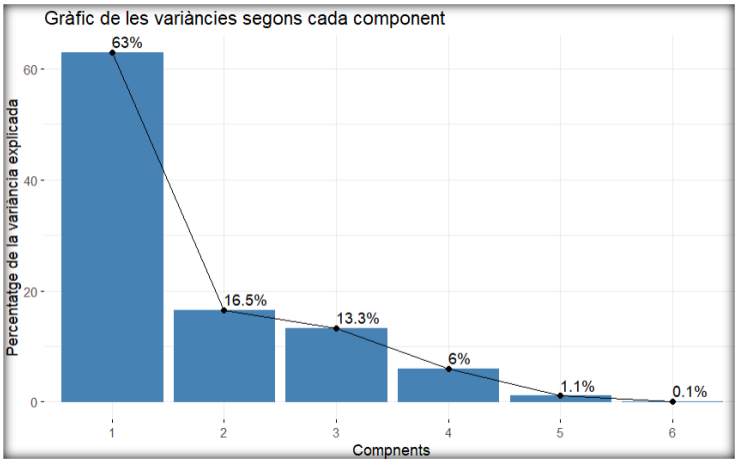
Figura 3.1 Matriu de correlacions entre les variables numèriques



Una altra forma de representar la relació entre les diferents variables numèriques és a través de la tècnica multivariant de l'Anàlisi de Components Principals (PCA). Aquesta s'utilitza per descriure un conjunt de dades a través de components, que s'ordenen de més a menys inèrcia i aquesta equival a la proporció de la variabilitat de les dades. Quan la inèrcia o proporció acumulada és major del 80%, es pot dir que gairebé tenim tota la informació. Així doncs, l'objectiu principal del PCA és poder representar totes les variables numèriques amb el mínim nombre de components.

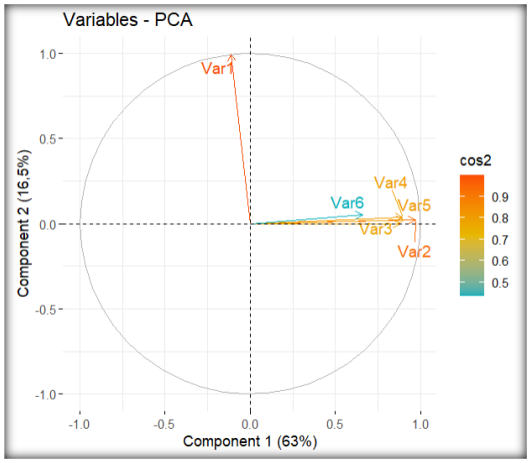
Amb l'ajuda dels paquets *FactoMineR (*Sebastien Le, Julie Josse, Francois Husson (2008)) i *factoextra* (Kassambara A, Mundt F (2020)) podem representar gràficament la inèrcia de cada variable, representades a la figura 3.2. Amb les dues primeres components tenim el 79,5% de la variància explicada i per tant, com que és molt proper al 80%, podem afirmar que amb només les dues primeres components tenim la informació necessària per a la correlació entre totes les variables numèriques.

Figura 3.2 Gràfic de les variàncies segons cada component



També podem representar les variables segons aquestes dues components (figura 3.3). La primera component només és explicada per la primera variable que correspon al nombre de persones en el grup de recerca. En canvi, la segona és explicada per les altres cinc variables numèriques.

Figura 3.3 Gràfic PCA de les variables numèriques.



Com a conclusió final del PCA, es pot dir que tots els resultats obtinguts són lògics ja que les cinc variables que estan més correlacionades entre elles (des de la *Var2* fins la *Var6)*, al final la informació és la mateixa (nombre de cites) i només varia en alguna aspecte.

3.1.2. EDA de la base de dades d'anàlisi (n = 85)

Amb l'*EDA* del cens definitiu (n = 191) hem arribat a la conclusió que no tenim gaires variables per diferenciar les persones que han respost l'enquesta de les que no l'han respost, ja que, exceptuant alguna variable, no hi ha diferències significatives entre els dos grups. Com que tenim una altra base de dades, on s'obté nova informació gràcies a l'enquesta realitzada (n = 85), s'intentarà veure si realment hi ha dos (o més) perfils diferents.

3.1.2.1. Anàlisi univariant

A continuació trobem la descripció univariant de la base de dades de l'anàlisi (n = 85) (taula 3.5).

Taula 3.5 Taula resum de l'anàlisi univariant de les variables més rellevants de la base de dades de l'anàlisi.

Variables categòriques (N = 85)	Categoria	n (%)
Gender	*Male*	74 (87,06%)
	Female	10 (11,76%)
	Prefer not to say	1 (1,18%)
Age group	30-44	40 (47,06%)
	45-59	26 (30,59%)
	60 +	14 (16,47%)
	18-29	5 (5,88%)
Country categorized	*United States*	21 (24,71%)
	Italy	15 (17,65%)
	Spain	8 (9,41%)
	Australia	7 (8,24%)
	*Other**	34 (40%)
Type	*Public*	62 (72,94%)
	Private	23 (27,06%)
Highest level of education	*Doctorate or PhD*	81 (95,29%)
	Master's degree	4 (4,71%)
Years worked in sports statistics	*5 years or more*	66 (77,65%)
	3-4 years	15 (17,65%)
	1-2 years	2 (2,35%)
	Less than 1 year	2 (2,35%)
Profile	*Full-time academic of sports statistics/analytics field.*	74 (87,06%)
	Part-time sports statistician in academia and part-time in the sports industry.	6 (7,06%)
	Academic sports statistician with years of expertise and contributions, but currently working in the sports industry.	5 (5,88%)
Currently belong to a sports statistics research group?	*No*	48 (56,47%)
	Yes	37 (43,53%)
Sport	*Soccer*	15 (17,65%)
	Basketball	6 (7,06%)
	Basketball, Soccer	4 (4,71%)

	Baseball	4 (4,71%)
	Running	4 (4,71%)
	Soccer, Tennis	4 (4,71%)
	American Football	3 (3,53%)
	Basketball, Baseball, American Football	3 (3,53%)
	Cricket	3 (3,53%)
	Basketball, Soccer, American Football	2 (2,35%)
	Tennis	2 (2,35%)
	*Other***	35 (41,18%)

Variables numèriques (N = 85)	Mínim	Màxim	Mediana	Mitjana	Desviació típica
Total citations in GS	58	50435	897	327	8190,987
Total citations in GS since 2018	58	24031	536	1672	3628,936
Total citations in GS acording H-index	3	108	14	19,92	17,462
Total citations in GS acording H-index since 2018	3	67	12	15,31	11,340
Number of the article with most cit.	20	34771	162	837,7	3861,164
Number of sports studied	1	9	2	2,165	1,519

Other*: Belgium, Canada, Czech Republic, Denmark, France, Germany, Greece, India, Ireland, Japan, Luxemburg, Netherlands, Norway, Sweden, United Kingdom.

Other**: Combinacions de freqüència 1 entre molts altres esports.

En aquesta base de dades segueixen predominant els homes. A més, la principal diferència d'aquesta variable és que tenim una categoria nova, ja que algú prefereix no indicar quin és el seu gènere. El rang de l'edat de les persones que predomina és el dels que tenen entre 30 i 44 anys, amb un 47,06%. Estats Units és el país amb més investigadors (un 24,71%), seguit d'Itàlia (un 17,65%), i els investigadors que treballen a una universitat pública segueixen sent majoria. Pel que fa al nivell d'estudis gairebé el 100% és el de doctorat o PhD, casi un 80% porta més de cinc anys treballant a la docència i aproximadament el 90% s'hi dedica plenament. A més, predominen les persones que no pertanyen a cap grup de recerca. En referència als esports estudiats, el *soccer* (més conegut per nosaltres com a futbol), estudiat de forma individual, és l'esport que predomina. El segueixen el bàsquet, el beisbol, el *running* i la combinació *soccer* i bàsquet. Cal remarcar que la variable era multi resposta i per tant, realment hi ha esports que s'estudien molt més, però al ser una combinació única, aquesta no apareix a la taula. També s'ha mirat individualment i el *soccer* és l'esport que predomina (24,76%) i el segueix el bàsquet (16,67%), el tenis (7,62%) i el futbol americà (6,67%).

Pel que fa a les variables numèriques, totes les que tenen relació amb el nombre de cites de *Google Scholar* segueixen la mateixa tendència que el cens definitiu, on hi ha una gran diferència entre el mínim i el màxim i la desviació típica és molt gran. A més, s'incorpora una nova variable en referència al nombre d'esports.

3.1.2.2. Anàlisi bivariant

Per l'anàlisi bivariant, continuarem utilitzant el paquet *compareGroups()* però ara la gran diferència és que no tenim una variable resposta definida. Així doncs, per trobar possibles relacions entre variables, les combinarem totes amb totes.

Per les comparacions de les variables categòriques, a la taula 3.5 podem veure que hi ha categories amb una freqüència de 5 o menys. Aquestes són les variables *Gender, Age group, Highest level of education, Years worked in sports statistics* i *Profile*. De forma que un cop comparem aquestes haurem de fer la prova exacte de Fisher. En canvi, per les variables *Country categorized, Type* i *Currently belong to a sports statistics research group?* realitzarem la prova chi-quadrat.

En quant a les variables numèriques, després de fer el test de Shapiro-Wilks s'ha obtingut un p-valor inferior a 0,05 en totes elles i per tant, les dades no segueixen una distribució normal i haurem de fer la prova U de Mann-Whitney.

Malgrat que s'han comparat totes les variables amb totes i s'han analitzat detalladament, per no fer-se repetitiu, exposarem només els resultats d'aquelles que tenen rellevància.

Variable resposta *Country categorized*

Quan la variable resposta és *Country categorized* tenim diverses variables significatives. Una d'aquestes és la variable *Years worked in sports statistics in academia*. El seu p-valor és de 0,011 i per tant, significa que, segons els anys treballats a l'estadística esportiva, hi ha diferències amb almenys un país. Tot i així, ens interessa saber entre quins hi ha aquestes diferències. Amb la taula 3.6 podem veure les freqüències segons cada país i els anys que porta l'investigador treballats a l'estadística esportiva. Tots els països tenen un percentatge molt elevat en la categoria de 5 anys o més treballats excepte a Espanya, on aquesta és molt més baixa. Per tant, és possible que aquest país sigui el que provoca que la variable sigui significativa. Un cop feta la comparació dos a dos per a cada país, els p-valors obtinguts, que també es troben a la taula 3.6, confirma que els investigadors espanyols segons els anys treballats són diferents, ja que les dues úniques comparacions significatives són les del grup *Other vs Spain* i *Spain vs US.* Per tant, la conclusió és que la nacionalitat pot estar lligada als anys treballats.

Taula 3.6 Taula bivariant i dels p-valors entre la variable Country categorized i years worked in sports statistics in academia.

Anys treballats a l'acadèmia	**Austràlia (N = 7)**	**Itàlia (N = 15)**	**Altres (N = 34)**	**Espanya (N = 8)**	**Estats Units (N = 21)**
1-2 anys	0 (0,00%)	1 (6,67%)	0 (0,00%)	1 (12,5%)	0 (0,00%)
3-4 anys	1 (14,3%)	3 (20,0%)	5 (14,7%)	5 (62,5%)	1 (4,76%)
5 anys o més	6 (85,7%)	11 (73,3%)	28 (82,4%)	2 (25,0%)	19 (90,5%)
Menys d'1 any	0 (0,00%)	0 (0,00%)	1 (2,94%)	0 (0,00%)	1 (4,76%)
Comparació segons p-valor					
Austràlia	-	1,000	1,000	0,167	0,743
Itàlia	1,000	-	0,677	0,166	0,450
Altres	1,000	0,677	-	0,018	0,677
Espanya	0,167	0,166	0,018	-	0,007
Estats Units	0,743	0,450	0,677	0,007	-

Variable resposta *Years worked in sports statistics in academia*

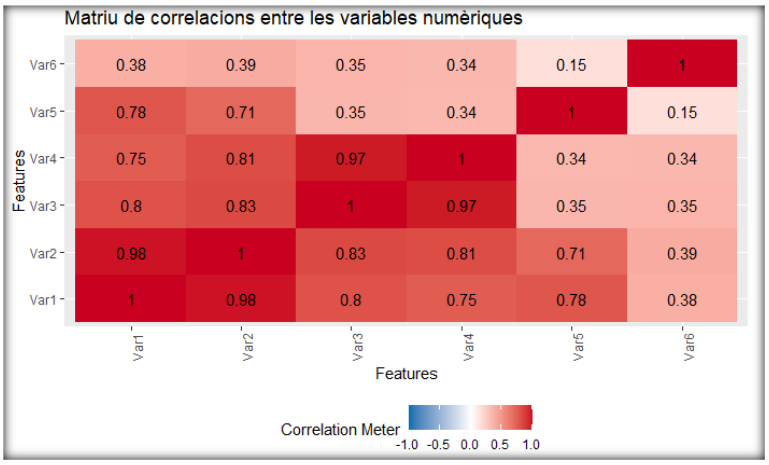
Quan la variable resposta és *Years worked in sports statistics in academia* també tenim algunes variables significatives. Una d'aquestes és la variable *Highest level of education*. El seu p-valor és de 0,040 i per tant, significa que, segons els anys treballats a l'estadística esportiva, hi ha associació amb el nivell d'educació. Amb la taula 3.7 podem veure les freqüències segons els anys que porta l'investigador treballats a l'estadística esportiva i el seu nivell d'educació. A més també podem veure els p-valors quan es fa la comparació dos a dos. Si suposem un nivell de significació de 0,1, l'única categoria significativa és la comparació entre el grup que porta 3-4 anys estudiant i el que en porta 5 o més, la resta de comparacions tenen un p-valor molt gran. Per tant, la conclusió és que els anys treballats a l'estadística esportiva poden influir amb el nivell d'educació.

Taula 3.7 Taula bivariant i dels p-valors entre la variable Years worked in sports statistics in acadèmia i Highest level of education.

Nivell d'educació	**1-2 anys (N = 2)**	**3-4 anys (N = 15)**	**5 anys o més (N = 66)**	**Menys d'1 any (N = 2)**
Doctorar o PhD	2 (100%)	12 (80,0%)	65 (98,5%)	2 (100%)
Màster	0 (0,00%)	3 (20,0%)	1 (1,52%)	0 (0,00%)
Comparació segons p-valor				
1-2 anys	-	1,000	1,000	.
3-4 anys	1,000	-	0,094	1,000
5 anys o més	1,000	0,094	-	1,000
Menys d'1 any	.	1,000	1,000	-

Per acabar amb l'anàlisi bivariant, comparem la relació de les variables numèriques entre totes elles. De nou, creem una matriu de correlacions (figura 3.4) on var1-6 correspon a la variables numèriques de la taula 3.5.

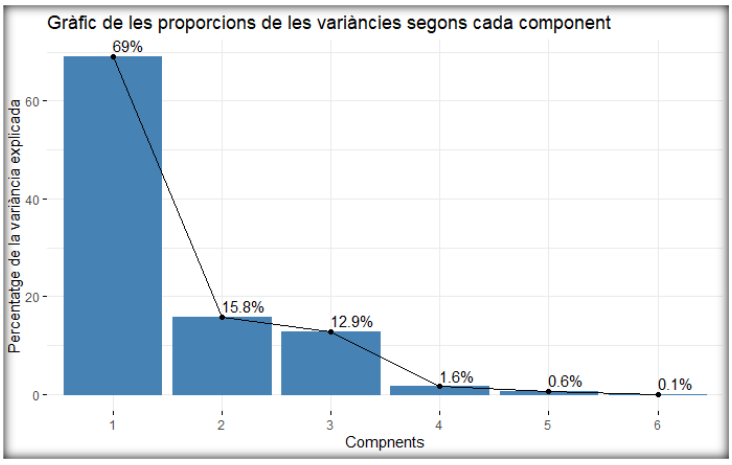
Figura 3.4 Matriu de correlacions entre les variables numèriques de la base de dades de l'anàlisi estadístic.



La primera variable gairebé té una correlació perfecta amb la segona. Una correlació molt semblant és la de la tercera variable i la quarta. Pel que fa a la cinquena variable, aquesta té força correlació amb la primera i segona variable i finalment la sisena no té gaire correlació amb cap variable.

Amb la figura 3.5 podem veure el percentatge de variància explicada amb cada component. Amb només les dues primeres ja tenim més del 80%, concretament el 84,8%. Així doncs, amb només dues components ja en tenim prou per explicar tota la informació de les variables numèriques.

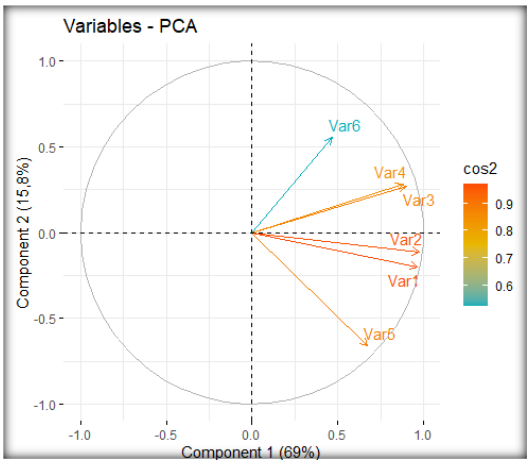
Figura 3.5 Gràfic de les proporcions de les variàncies de la base de l'anàlisi estadístic segons cada component



Per acabar, representem les variables a través d'un gràfic de PCA (figura 3.6). Les quatre primeres variables, corresponents al nombre de cites, són les que millor representen la segona component. Pel que fa a la primera component, podem dir que tant la cinquena com

la sisena variable la representen de la mateixa manera, amb la diferència que aquesta última té una correlació més baixa.

Figura 3.6 Gràfic PCA de les variables numèriques de la base de l'anàlisi estadístic.



Amb l'anàlisi bivariant de la base de dades d'anàlisi hem pogut veure que quan creuem totes les variables amb totes, n'hi ha més d'una que ens indica que hi ha alguna diferència entre alguna categoria. Per aquest motiu, el més probable és que mínim hi hagi dos grups diferenciats. La millor tècnica estadística per diferenciar grups és el *clustering* i, un dels mètodes més utilitzat i recurrent és el *k-means*. Així doncs, el següent pas serà utilitzar aquest algoritme per intentar esbrinar quins són els perfils dels diferents grups d'investigadors que tenim a la base de dades.

3.2. K-MEANS

3.2.1. Estadístic de Hopkins

Abans de començar amb l'execució de l'algoritme *k-means*, calculem l'estadístic de Hopkins, amb l'ajuda de la funció *get_clust_tendency()* (Hopkins & Skellam, 1954). Recordem que ens interessa que aquest tingui un valor proper a 1, ja que ens indicarà que les dades són clusteritzables.

Després de calcular-lo, obtenim que té un valor de 0,979 i per tant, podem concloure que és adient utilitzar un mètode *clustering* per analitzar les dades.

3.2.2. Inicialització

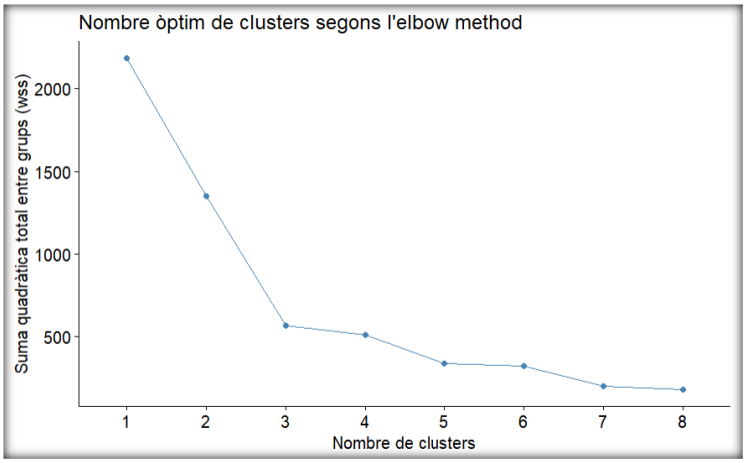
Per poder executar l'algoritme, es necessiten una sèrie de passos previs. Un cop carregada la base de dades, seleccionem només les variables numèriques ja que, tal com s'ha explicat a la

metodologia, aquest algoritme només es basa en distàncies. A més, també estandarditzem les dades perquè aquestes estiguin en la mateixa escala.

Recordem que per executar l'algoritme *k-means* necessitem conèixer el nombre de *clusters*. Així doncs, el següent pas abans d'executar-lo és determinar la *k*.

Per fer-ho, utilitzem el mètode del colze. Els resultats es poden veure a la figura 3.7 i aquell punt on hi ha la forma de colze, i per tant, és l'òptim, és quan $k = 3$.

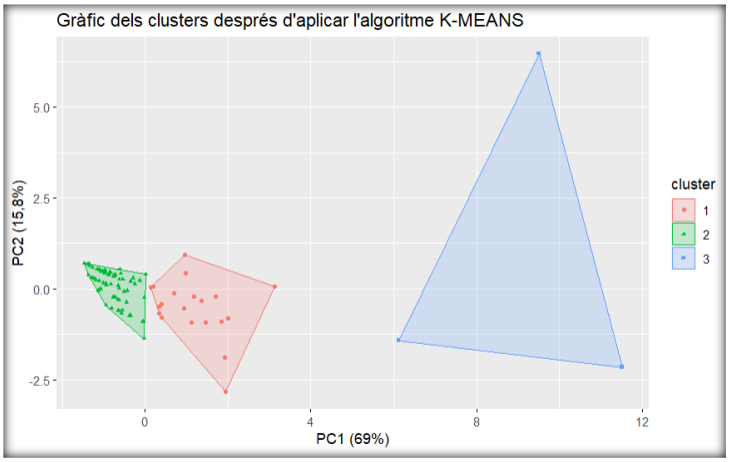
Figura 3.7 Gràfic de l'elbow method per determinar la k òptima.



3.2.3. *Execució de l'algoritme*

Un cop s'ha determinat que 3 són el nombre òptim de *clusters*, ara ja estem a total disposició d'executar l'algoritme. Després de fer-ho, els resultats gràfics obtinguts es poden visualitzar a la figura 3.8.

Figura 3.8 Gràfic dels clusters després d'aplicar l'algoritme k-means.



Al primer *cluster* només hi ha tres persones. Amb l'ajuda d'R, es pot saber que aquests tres punts corresponen a les observacions 28, 35 i 72. Després de visualitzar els valors de totes les variables d'aquestes tres observacions, ens n'adonem que són tres investigadors on el

nombre de cites és exageradament alt. Concretament a la variable *Total Citations in GS* tenen 50435, 50030 i 27292 cites respectivament. En canvi, la quarta persona amb més cites en té 11707. És conegut que, degut a que és un mètode basat en distàncies, l'algoritme *k-means* és molt sensible als valors atípics. Després d'estar pensat com ho podríem fer perquè aquestes observacions no influïssin, una primera opció va ser eliminar-les. Tot i així, la base de dades d'anàlisi també està formada per diverses variables categòriques que no es tenen en compte en cap moment, ja que només s'utilitzen les numèriques. És per aquest motiu que es va acabar descartant l'opció d'eliminar les observacions atípiques i es va voler donar importància a les variables categòriques. Aquí és quan va aparèixer el *KAMILA*. Un algoritme de mixtura que té en compte tant les variables numèriques com les categòriques.

3.3. KAMILA

3.3.1. Inicialització

Perquè l'algoritme sigui executat sense problemes necessitem fer una sèrie de passos previs. Primer de tot, creem una *dataframe* amb totes les variables numèriques i seguidament, les estandarditzem perquè no hi hagi problemes amb la interpretació dels resultats. A continuació creem un altra *dataframe* amb totes les variables categòriques i transformem totes les categories a factor. Un cop realitzat això, ja estem a disposició d'executar l'algoritme.

3.3.2. Execució de l'algoritme

L'algoritme s'inicialitzarà 10 vegades i el nombre màxim d'iteracions serà 100. Aquest s'executarà per *clusters* amb valors des de $k = 2$ fins a $k = 10$. A més, el mètode per escollir el nombre de *clusters* serà el del *prediction strength* i assumirem que aquest té un valor inicial de 0,65. Per tant, un cop definits els paràmetres inicials, executem l'algoritme amb la funció *kamila,* del paquet d'R *kamila* (Foss AH, Markatou M (2018)). Els valors *prediction strength* obtinguts segons cada *cluster* es troben a la taula 3.8.

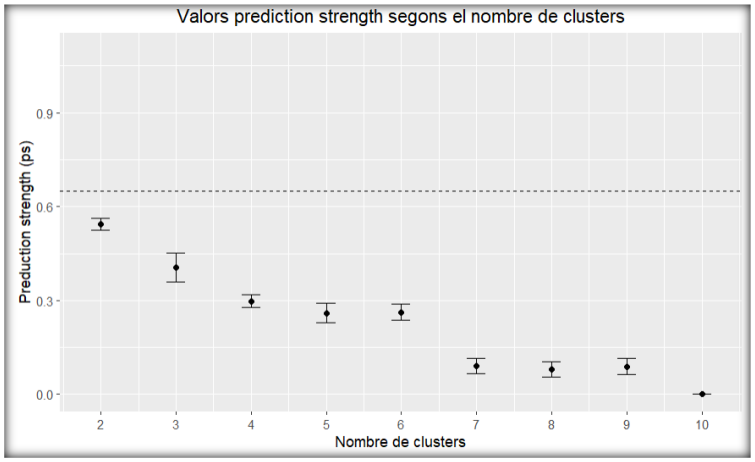
Taula 3.8 Valors prediction strength segons cada valor de k.

k	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Valor ps	0,615	0,389	0,257	0,189	0,222	0,165	0,087	0,175	0,08

Tal i com s'ha vist a la metodologia, es recomana que el nombre de *clusters* que s'escull tingui un valor *prediction strength* entre 0,8 i 0,9. Encara que no hem obtingut cap *cluster*

amb aquest valor, amb *k* = 2 s'obté un valor proper i per tant aquest serà el nombre de *clusters* que escollirem. Gràficament aquests valors es veuen reflectits a la figura 3.9.

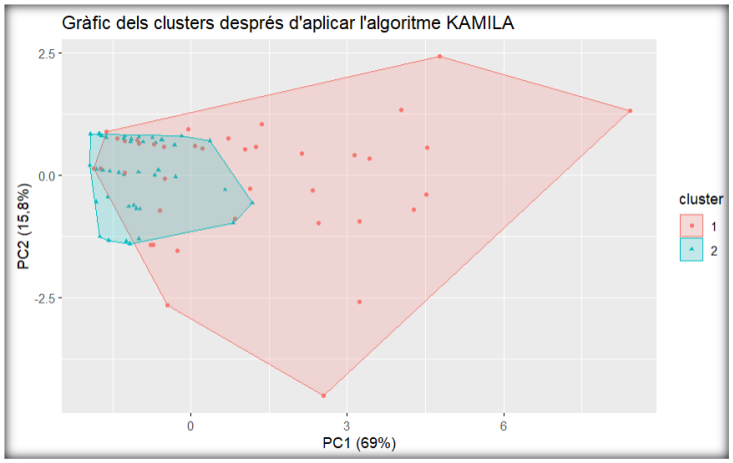
Figura 3.9 Gràfic dels valor prediction strength segons el nombre de clusters.



Amb la figura 3.9 veiem que no hi ha cap valor *prediction strength* que sobrepassi la línia discontínua (valor marcat inicialment com a 0,65). Tot i així, quan *k = 2* el valor és molt proper. Finalment, amb l'ajuda d'una de les opcions que ens proporciona R, aquesta també ens indica que el nombre òptim de *clusters* és 2 i per tant, concloem que aquest serà el nostre nombre de *clusters*.

A partir d'ara, creem una nova variable anomenada *cluster*, que només tindrà valors 1 o 2 i segons a quin *cluster* pertanyi l'observació, tindrà un valor o altra. Després de fer-ho, sabem que hi ha 3 observacions (el 3,53%) que pertanyen al primer *cluster* i 82 (el 96,47%) que pertanyen al segon *cluster*. El gràfic on es pot visualitzar els dos *clusters* és la figura 3.10.

Figura 3.10 Gràfic dels clusters després d'aplicar l'algoritme KAMILA.

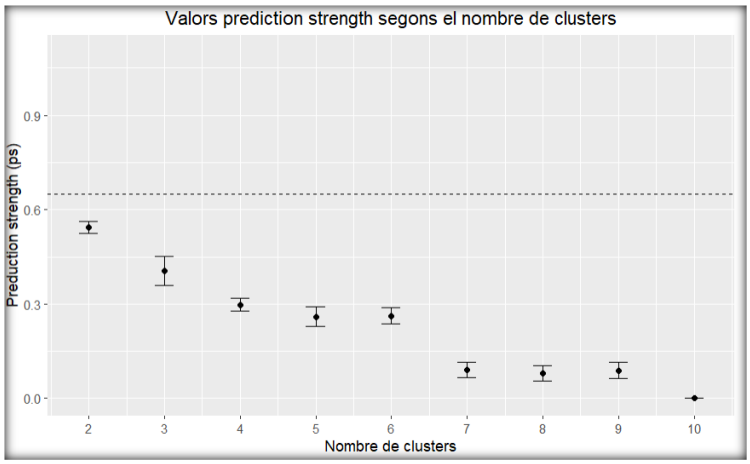


Ens n'adonem que tenim exactament el mateix problema que amb el *k-means*, i és que hi ha un *cluster* amb només tres observacions, que pertanyen a les tres observacions tres d'aquells investigadors que tenen moltes cites a *Google Scholar*. Com que ara ja hem utilitzat

tant un algoritme de només variables numèriques i un de mixtura, l'últim recurs per tal que aquestes tres observacions no influeixin és eliminar-les. Per tant, el que ara fem serà assumir que aquestes tres observacions ja formem un *cluster* elles soles i executarem tot l'algoritme sense tenir-les en compte.

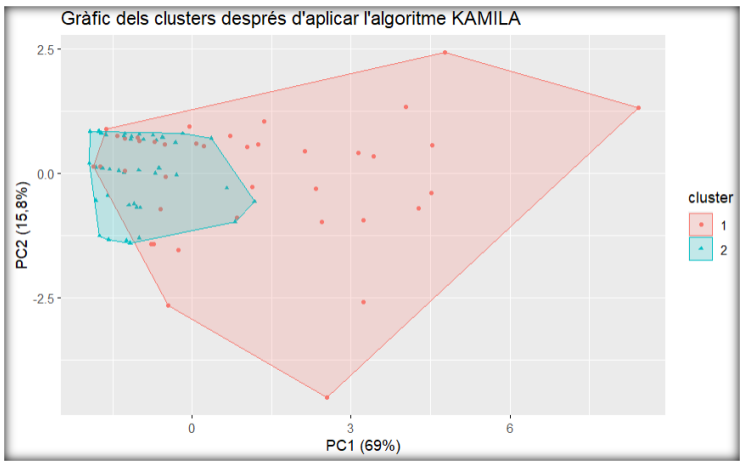
A la figura 3.11 podem veure els nous valors *prediction strength* segons cada *cluster* un cop s'ha executat l'algoritme. No n'hi ha cap que superi la línia discontínua però tot i així, quan *k* = 2 és quan s'obté un major valor. A més, amb l'opció automàtica d'R també ens indica que el nombre òptim és 2.

Figura 3.11 Gràfic dels valor prediction strength després d'executar de nou l'algoritme.



Ara modifiquem la variable anomenada *cluster*, que havíem creat anteriorment, que també prendrà valors 1 o 2. Després de fer-ho, sabem que hi ha 39 observacions (el 47,56%) que pertanyen al primer *cluster* i 43 (el 52,44%) que pertanyen al segon *cluster*. El gràfic on es pot visualitzar els dos nous *clusters* és la figura 3.12.

Figura 3.12 Gràfic dels clusters després d'aplicar de nou l'algoritme KAMILA.



Així doncs, arribem a la conclusió que tenim 3 *clusters* diferents. Un d'aquests és el que pertany a aquells tres investigadors amb moltíssimes cites a *Google Scholar,* mentre que els altres dos són els de la figura 3.12. Tot i així, no tenim informació d'aquests dos *clusters* així

que ara l'objectiu serà intentar conèixer una mica millor les característiques dels investigadors d'aquests.

3.3.3. Caracterització dels clusters

Per conèixer quines són les variables que són diferents respecte els *clusters*, tal i com ho hem fet amb l'*EDA*, utilitzarem la funció *compareGroups()*, on la variable resposta serà el nombre de *cluster*, que podrà prendre valors 1 o 2. Després de creuar totes les variables amb aquesta, a la taula 3.9 tenim informació dels resultats obtinguts.

Taula 3.9 Taula amb Informació de totes les variables sobre tenint en compte que la variable resposta és el cluster.

Variable	**N**	**p-valor**	**Mètode**
Gender	82	0,399	Categòric. Test chi-quadrat
Age group	82	<0,001	Categòric. Test chi-quadrat
Country categorized	82	<0,001	Categòric. Test exacte de Fisher
Type	82	0,002	Categòric. Test exacte de Fisher
Highest level of education	82	0,119	Categòric. Test chi-quadrat
Years worked in sports statistics in academia	82	<0,001	Categòric. Test chi-quadrat
Profile	82	0,429	Categòric. Test chi-quadrat
Currently belong to a sports statistics research group?	82	0,051	Categòric. Test exacte de Fisher
Number of members	82	< 0,001	Continu no normal
Total citations in GS	82	< 0,001	Continu no normal
Total citations in GS since 2018	82	< 0,001	Continu no normal
Total citations in GS acording H-index	82	< 0,001	Continu no normal
Total citations in GS acording H-index since 2018	82	< 0,001	Continu no normal
Number of citations of the article with most citations	82	0,001	Continu no normal
Number of sports studied	82	0,604	Continu no normal

Pel que fa a les variables categòriques, les variables *Age group, country categorized, Type, Years worked in sports statistics in academia* tenen un p-valor inferior a 0,05 i per tant són significatives i hi ha diferències entre els dos *clusters*. Pel que fa a les variables numèriques, totes són significatives excepte la del nombre d'esports. Així doncs, el que farem serà entrar en més detall amb aquestes variables. A la taula 3.1o tenim informació sobre totes elles.

Si ens fixem en la variable *Age group*, la principal diferència en que al primer *cluster* hi ha la gran majoria d'investigadors que pertanyen al grup d'edat de 45-59 anys i 60+ anys. En canvi, al segon *cluster* hi ha gairebé totes les persones més joves, que pertanyen al grup d'edat de 18-29 anys i de 30-44 anys. La conclusió és que al segon *cluster* hi ha gent més jove que al primer.

Pel que fa a la variable *country categorized*, la principal diferència és que en el primer *cluster* hi ha les persones que són de països amb poca freqüència (per saber quins són aquests països, es pot saber mirant la taula 3.5) i també els que són d'Itàlia. En canvi, al segon *cluster* hi ha tota la gent Espanya i la gran majoria d'Estats Units. La conclusió és que al primer *cluster* hi ha els investigadors dels països on la seva freqüència és petita i d'Itàlia mentre que al segon *cluster* hi ha la gent d'Espanya i Estats Units.

En quant a la variable *Type*, al primer *cluster* està format per la majoria d'investigadors que pertanyen a una universitat pública, mentre que al segon *cluster* els que pertanyen a una privada tenen molta més presència. La conclusió és que al primer *cluster* hi ha els investigadors d'una universitat pública mentre que al segon *cluster*, encara que també n'hi ha algun que és d'una universitat pública, la majoria són d'una privada.

Per la variable *Years worked in sports statistics in academia* es pot veure que totes les persones que formen el primer *cluster* porten 5 o més anys treballats, mentre que al segon *cluster*, a banda d'haver-ni alguna que també porta 5 o més anys treballats, hi ha tots els altres que porten menys de 5 anys. La conclusió és que al primer *cluster* hi ha la gent amb més experiència laboral en l'àmbit de l'estadística esportiva mentre que al segon *cluster* hi ha els investigadors amb menys experiència.

Finalment, pel que fa a les variables numèriques, el primer valor que apareix són el nombre de cites mitjanes, mentre que el valor entre parèntesis és la desviació típica del *cluster*. Es pot veure clarament que al primer *cluster* hi trobem aquelles persones que tenen moltes cites a *Google Scholar.* A més, la desviació també és major. La conclusió és que al primer *cluster* hi trobem els investigadors amb moltes cites i al segon hi trobem els que en tenen menys.

Taula 3.10 Taula de les variables significatives, segon cada cluster.

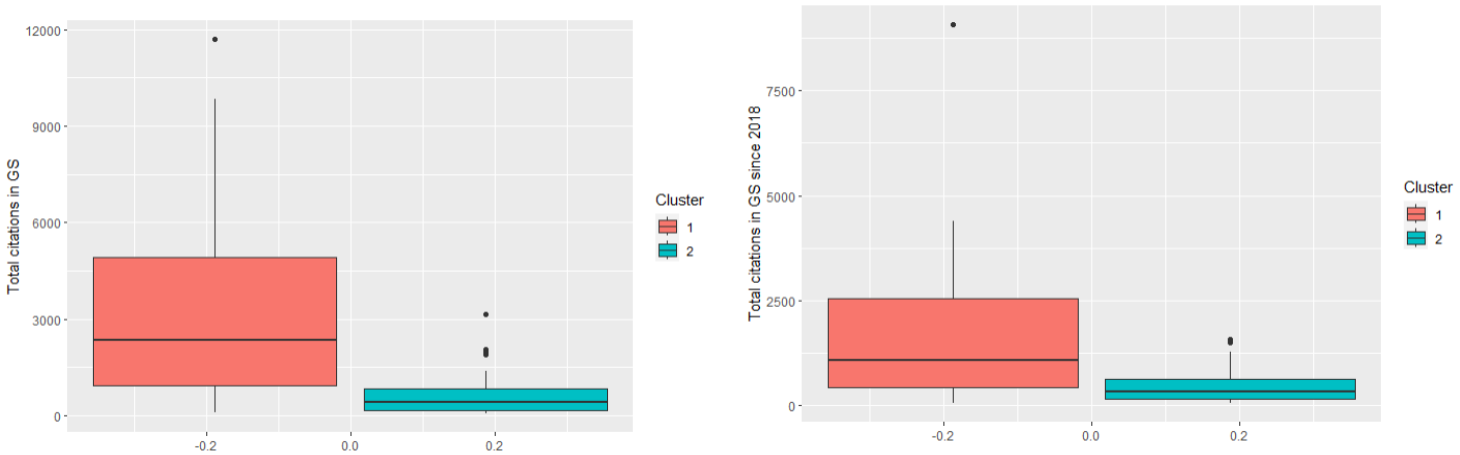
Variable		***Cluster* 1 (N = 39)**	***Cluster* 2 (N = 43)**	**P-valor**
Age group	Categories			< 0,001
	18-29	0 (0,00%)	5 (11,6%)	0,004
	30-44	12 (30,8%)	28 (65,1%)	0,002
	45-59	17 (43,6%)	9 (20,9%)	0,268
	60+	10 (25,6%)	1 (2,33%)	Ref.
Country categorized	Categories			<0,001
	Australia	1 (2,56%)	4 (9,30%)	0,105
	Italy	10 (25,6%)	5 (11,6%)	1,000
	Other	24 (61,5%)	10 (23,3%)	Ref.
	Spain	0 (0,00%)	8 (18,6%)	0,005
	United States	4 (10,3%)	16 (37,2%)	0,005
Type	Categories			0,002
	Private	4 (10,3%)	19 (44,2%)	Ref.

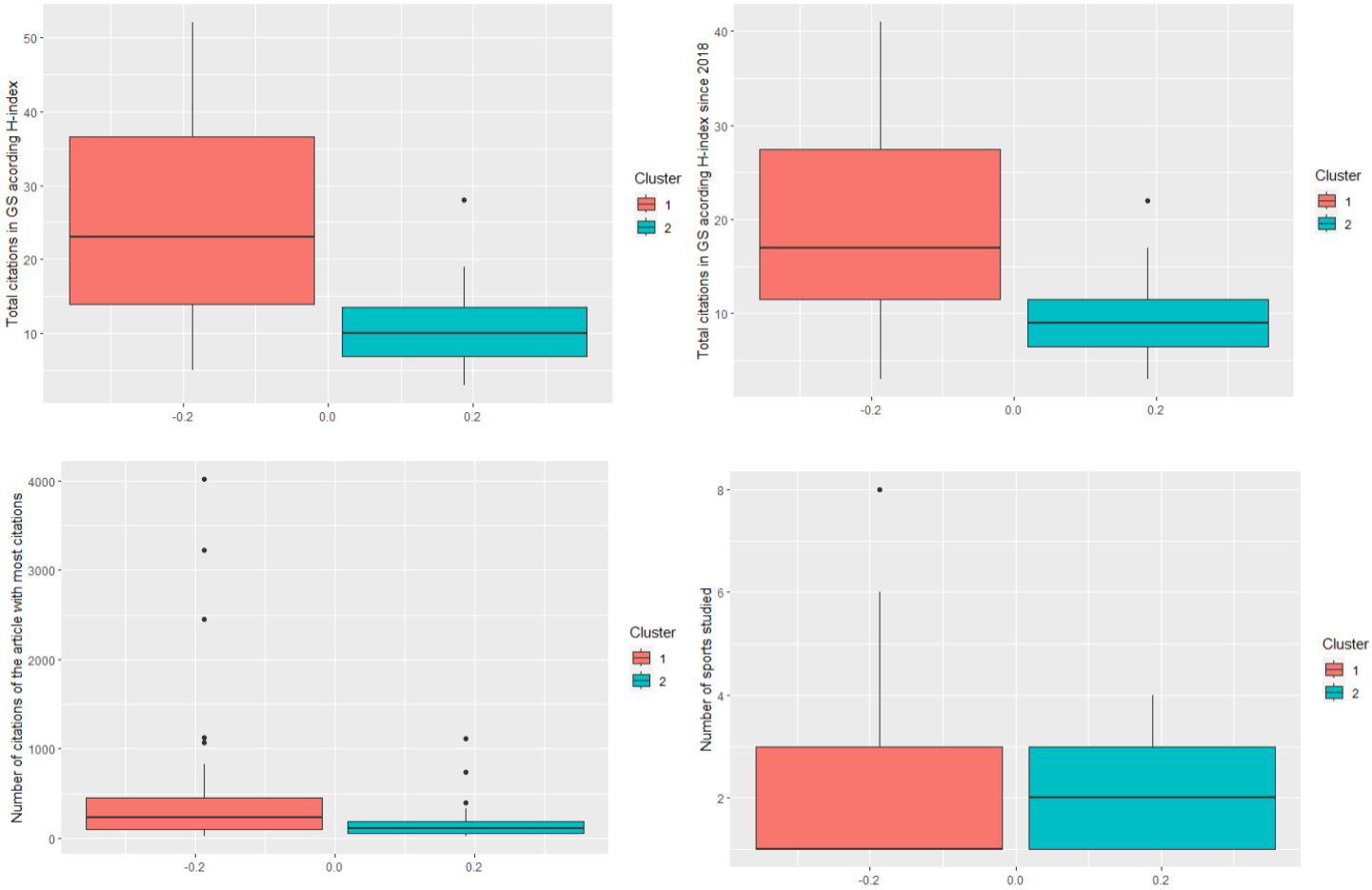
	Public	35 (89,7%)	24 (55,8%)	0,002
Years worked in sports statistics in academia	Categories			<0,001
	1-2 years	0 (0,00%)	2 (4,65%)	0,156
	3-4 years	0 (0,00%)	15 (34,9%)	0,156
	5 years or more	39 (100%)	24 (55,8%)	Ref.
	Less than 1 year	0 (0,00%)	2 (4,65%)	0,156
Total citations in GS		3144 (2958)	642 (644)	< 0,001
Total citations in GS since 2018		1705 (1758)	472 (421)	< 0,001
Total citations in GS acording H-index		25 (13,2)	11 (5,56)	< 0,001
Total citations in GS acording H-index since 2018		18,9 (9,74)	9,49 (4,12)	< 0,001
Number of citations of the article with most citations		528 (853)	162 (198)	0,012

Ref. : És la categoria referència i la que es compara amb la resta. Exemple: Per la variable *Age group* es compara 60+ vs 18-29, 60+ vs 35-44 i 60+ vs 45-59.

Amb la prova U de Mann-Whitney realitzada hem pogut confirmar la diferència entre totes les variable numèriques excepte la del nombre d'esports estudiats. Tot i així, un altra forma de compara-ho és a través de *boxplots*. A la figura 3.13 apareix un *boxplot* per cada variable estratificada per *cluster*. El de color vermell correspon al primer *cluster* i el de color blau correspon al segon *cluster*. Els cinc primers gràfics són els de les cinc variables significatives, i es pot veure molt fàcilment el que s'ha vist amb les taules, on pel primer *cluster* hi ha moltes més cites i la desviació típica és molt més gran. A més, com que no es solapen, vol dir que les variables són significatives. Finalment, el sisè gràfic, que correspon a la variable del nombre d'esports estudiats, veiem que són gairebé idèntics i es solapen gairebé totalment, cosa que significa que la variable no és significativa.

Figura 3.13 Boxplots de les variables numèriques estratificats per cada cluster.

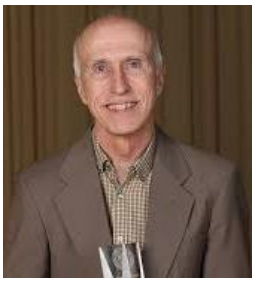




A partir d'aquests resultats comentats per cada *cluster* i variables significatives, podem fer un resum de cada *cluster*. Ho podem visualitzar a la taula 3.11 i a més a més, veiem la imatge de dos investigadors/es que pertany a cada un dels *clusters*.

Taula 3.11 Resum amb les característiques principals de cada cluster

Característiques	***Cluster* 1**	***Cluster* 2**	***Cluster* 3**
Rang d'edat	45-59 anys i 60+ anys	18-29 anys i 30-44 anys	60 + anys
Països	Països de baixa freqüència i Itàlia	Espanya i Estats Units.	Austràlia.
Tipus d'universitat	Pública.	Privada.	Pública.
Anys treballats a l'estadística esportiva	5 anys o més.	Menys de 5 anys.	5 anys o més.
Cites a *Google Scholar*	Nombre elevat de cites (mitjana de 3144 cites totals).	Nombre baix de cites (mitjana de 642 cites totals).	Nombre exageradament alt de cites (mitjana de 42582.

Número d'investigadors	39	43	3
Exemple	Jim Albert:  Marica Manisera: 	Michael Lopez:  Benjamin Baumer: 	William Hopkins:  Caroline Finch: 

*Imatges extretes a través dels perfils públics i personals de cada investigador.

3.4. World map

Un dels objectius d'aquest treball era fer un mapa mundial de tots els investigadors que es dediquen a l'acadèmia i l'estadística esportiva. Un cop acabat el treball, després de moltes hores de treball, s'ha aconseguit un cens amb un total de 26 països diferents. Els diferents països i les seves freqüències es poden visualitzar a la taula 3.12.

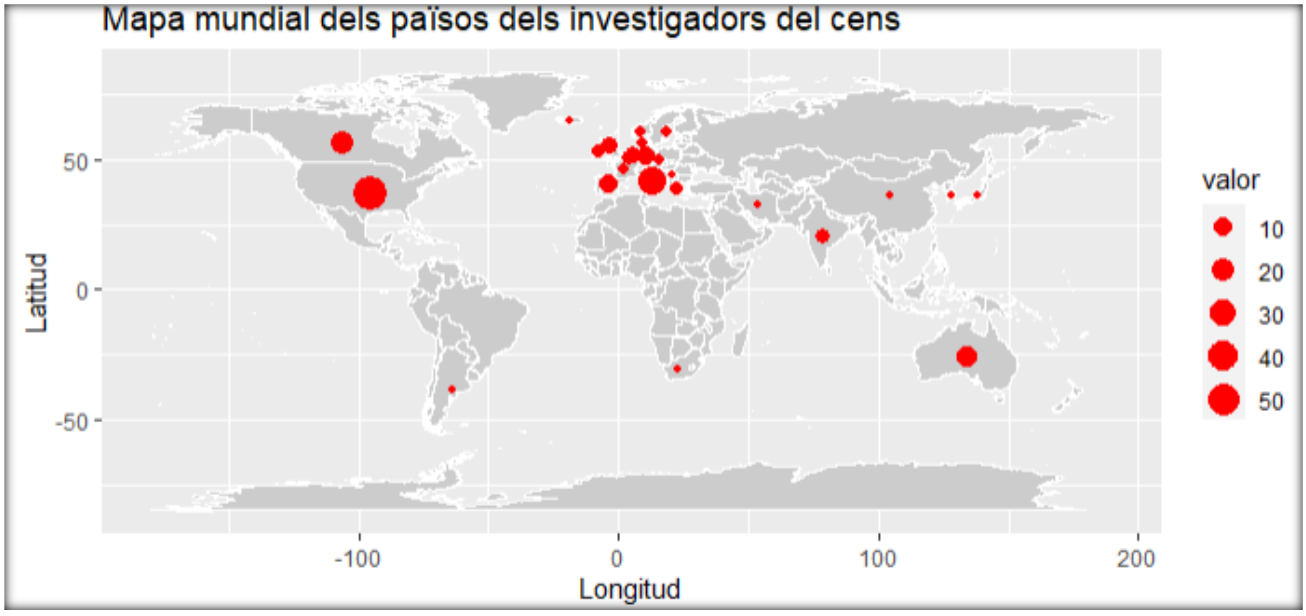
Taula 3.12 Noms i freqüències de tots els països del cens.

País	Freqüència absoluta	Freqüència relativa
Estats Units	54	24,71%
Itàlia	34	17,65%
Canadà	19	9,41%
Austràlia	13	8,24%
Espanya	12	8,24%
Alemanya	10	5,88%
Regne Unit	8	4,71%
Índia	6	3,53%
Bèlgica	5	2,35%
Països Baixos	5	2,35%
Grècia	3	1,18%

Irlanda	3	1,18%
República Txeca	2	1,18%
Dinamarca	2	1,18%
França	2	1,18%
Noruega	2	1,18%
Suècia	2	1,18%
Argentina	1	1,18%
Xina	1	1,18%
Islàndia	1	1,18%
Iran	1	1,18%
Japó	1	1,18%
Luxemburg	1	1,18%
Sèrbia	1	1,18%
Sud-Àfrica	1	1,18%
Corea del sud	1	1,18%

A més a més, per dur a terme el mapa necessitem saber la longitud i la latitud de cada país. Després de buscar-ho pels 26 països, ja estem en total disposició per realitzar un mapa mundial. El mapa resultant és el de la figura 3.14.

Figura 3.14 Mapa mundial dels països dels investigadors del cens.



4. DISCUSSIONS I CONCLUSIONS

4.1. Discussions dels resultats

Aquest treball evidencia clarament l'enorme evolució que l'àmbit de l'estadística esportiva ha experimentat en els últims anys (Alamar, 2013; Gelman & Vehtari, 2021). Les possibilitats i producció dels diferents treballs acadèmics d'estadística esportiva a partir de les diferents revistes científiques disponibles ha anat també en augment (Swartz, 2020a). Entre aquestes destaquen també articles enfocats a l'evolució d'*sports analytics* tan a nivell de pensament estadístic com a nivell computacional (Baumer et al., 2023; Casals et al., 2023). Des del nostre coneixement i després d'una recerca bibliogràfica, no s'ha creat mai cap cens per conèixer tots els investigadors que es dediquen a l'estadística esportiva. Aquesta mancança ha estat el motiu principal que ha impulsat la realització d'aquest Treball de Fi de Grau.

En aquest treball s'ha obtingut un cens exhaustiu que inclou 191 professionals dedicats a l'estudi de l'estadística esportiva. Atès que no existeix cap referència prèvia en aquest àmbit, no teníem una idea clara de la mida esperada d'aquest cens. Per tant, no podem concloure definitivament si hem inclòs la major part dels estadístics esportius de tot el món. No obstant això, creiem que el nostre cens recull o intenta recollir tots els professionals, tenint en compte les limitacions conegudes i amb l'objectiu que en una futura recerca sigui un cens dinàmic i amb opcions de créixer.

La taxa de resposta de l'enquesta que es va enviar als 204 investigadors inicials cercats, segons els que enteníem que complien els criteris establerts, va ser prou bona (50,49%) si ho comparem amb altres estudis similars on s'envien enquestes a acadèmics. En general, es considera que una enquesta té un bon índex de resposta quan aquest oscil·la entre el 5% i el 30%, a més, quan aquesta és del 50% o més, es pot considerar excel·lent (Chung, L., 2022).

A partir dels 103 que van respondre l'enquesta, es va crear una base amb 85 investigadors, després de revisar i eliminar aquells que realment no complien el perfil. Es va intentar conèixer més a fons el perfil d'aquests acadèmics d'arreu del món emprant les tècniques de *clustering*: el *k-means* i el *KAMILA*.

El cens ha revelat una notable disparitat de gènere per aquells professionals acadèmics que es dediquen en l'àmbit de l'estadística esportiva, amb una presència masculina molt superior a la femenina (86,91% vs 13,09%). Això posa de manifest la necessitat de promoure i donar visibilitat en aquest àmbit per tal de reduir aquestes desigualtats percentuals. És important assenyalar que als Estats Units, aquest àmbit s'ha desenvolupat des de fa més temps, com

ho demostra el nombre significatiu de professionals registrats en aquest país, que representen un 24,71% del total. Itàlia, amb el 17,65%, és el segon país amb més investigadors. Aquests resultats poden causar sorpresa, però s'entenen millor sabent que en aquest país hi ha un grup de recerca d'estadística esportiva format per 60 persones. A la resta del món hi ha altres grups que es dediquen a l'estadística esportiva, però tan sols el 50,79% de les persones del cens en pertany a un. Això significa que no n'hi ha els suficients com perquè aquests puguin fer-ne una àmplia difusió.

Amb les tècniques *clustering* emprades per conèixer millor els perfils d'aquest cens s'ha arribat a la conclusió que tenim tres perfils diferenciats:

El primer *cluster* està format per les persones de major edat, concretament les del grup de 45-59 i 60 o més anys. La majoria són d'Itàlia o de països amb poca freqüència, com pot ser el Regne Unit, Alemanya o l'índia (per veure la llista de països complets, veure taula 3.12). Tot i així, també n'apareix algun d'Estats Units. Com a conseqüència de tenir una certa edat, el *cluster* està format per aquells investigadors que porten 5 o més anys treballats a l'estadística esportiva. A més, molts són d'universitats públiques. Finalment, com que ja tenen una llarga trajectòria acadèmica, això provoca que el nombre de cites a *Google Scholar* sigui elevat. Dos exemples d'investigadors són en Jim Albert, i la Marica Manisera (veure foto a la taula 3.11). En Jim Albert va ser un dels grans impulsors i dels primers a dedicar-se a l'estadística esportiva als Estats Units. La Marica Manisera no té un perfil tant sènior però tot i així, pertany al grup d'edat d'entre 45-59 anys. Ella és una de les principals impulsores de l'estadística esportiva a Itàlia, sent una de les dues coordinadores d'un dels grups de recerca més nombrós a tot el món. A més a més, en aquest *cluster* també s'hi troben investigadors com en Tony Myers (impulsor de l'estadística esportiva al Regne Unit), en Tim Swartz (impulsor de l'estadística esportiva a Canadà), l'Andreas Groll (impulsor de l'estadística esportiva a Alemanya) o en Rajitha Silva (impulsor de l'estadística esportiva a l'Índia). Curiosament tots aquests estadístics també són bayesians. És per tot això explicat que el primer *cluster* es pot considerar que hi formen part les persones que van posar la primera pedra.

El segon *cluster*, a diferència del primer, està format per persones més joves. Hi predominen les del rang entre 18-29 i 30-44 anys. Aquí els països que predominants són Espanya i Estats Units i, degut a la seva poca edat, és més probable trobar-hi investigadors amb menys de 5 anys treballats en aquest àmbit. Com a conseqüència de trobar-hi molts estatunidencs, hi predominen les universitats privades. La poca experiència, sobretot degut a l'edat, fa que els investigadors d'aquest *cluster* tinguin poques cites a *Google Scholar.* Dos exemples d'investigadors són en Michael Lopez i en Benjamin Baumer (veure foto a la taula 3.11). Aquests dos són els últims que van ser nombrats investigadors contribuents per la *Statistics in Sports Section (ASA)* els anys 2020 i 2019, respectivament

(https://community.amstat.org/sis/aboutus/honorees). Curiosament, 16 anys abans, el 2003, en Jim Albert també n'havia estat nombrat. Malgrat que en Michael Lopez i en Benjamin Baumer són joves, aquests dos investigadors ja tenen una trajectòria important en l'acadèmia de l'estadística esportiva i és per això que, encara que s'ha vist als resultats que no hi ha diferències significatives com perquè hi hagi un altre grup. Podríem pensar que aquest segon *cluster* es podria dividir en dos, entre els joves que ja tenen una certa experiència i inclús més joves (recent post-docs) que fa molt poc que han entrat en aquest món. És per aquests motius que el segon *cluster* es pot considerar que hi formen part les persones que han continuat el que van iniciar les del primer i la nova fornada d'investigadors del futur.

Finalment el tercer *cluster* està format només per tres investigadors on la característica principal és que les seves cites a *Google Scholar* són extremadament altes. Com és lògic, porten 5 o més anys treballats, en universitats públiques i són australians. Dos d'aquests investigadors són en William Hopkins i la Caroline Finch (veure foto a la taula 3.11). En William Hopkins és un estadístic molt important que és molt conegut en l'àmbit de l'*sport science*. Pel que fa a la Caroline Finch, és una de les bioestadístiques més importants de tot el món. Tot i així, encara que el seu àmbit principal sigui el de la bioestadística, ha contribuït notablement amb l'estadística esportiva. Per aquest motiu, en el tercer *cluster* hi ha els grans estadístics australians en que, des del seu àmbit, han contribuït amb l'estadística esportiva.

Com a conclusió final del *clustering*, una acció divertida, però amb les circumstàncies actuals, impossible de realitzar, podria ser presentar aquests tres perfils a tots els investigadors del cens, i fer-los-hi dues preguntes: la primera, amb quin dels tres *clusters* s'identificarien, i la segona, a quin dels tres *clusters* els hi agradaria formar part. Seria interessant veure els resultats obtinguts i si coincideixen amb els del treball.

El cens també ens ha premés crear un mapa mundial i veure visualment on es troben distribuïts tots aquests investigadors.

4.2. Limitacions

Recordant els països dels quals pertanyen els investigadors, la gran majoria són de Nord-Amèrica o d'Europa. Dels altres continents n'hi ha algun, però molts menys. Això no significa que no n'hi hagi cap, sinó que simplement o bé la informació no es troba disponible a través d'internet, o bé com que s'utilitza una altra llengua, com per exemple, els països africans o asiàtics, es fa tot molt més complicat.

El fet que els tutors treballin a Espanya i coneguin a diverses persones dins d'aquest àmbit, ha provocat que persones, sobretot espanyoles, que potser no s'haguessin trobat si s'haguessin buscat només per internet, apareguin al cens. És a dir, amb gairebé total certesa es pot dir que en el cens apareixen tots els investigadors d'estadística esportiva de l'estat espanyol però per exemple, a França, que és el país del costat d'Espanya, només n'hi apareixen dues. No significa que només n'hi hagi dues en tot el país, sinó que simplement pel fet de no tenir ningú coneixedor d'aquest àmbit en aquest país, ha provocat que no s'hagin identificat més persones. Aquest és el biaix que té el cens.

Una altra gran limitació ha estat el temps. Tot i començar amb molta antelació, per poder dur a terme una cerca que tingui en compte tots els investigadors de tot el món, requereix d'una dedicació total i d'un període de temps molt més llarg que, amb les condicions actuals, no es podia assumir. Una opció podria haver sigut investigar un per un els prop de 200 països que hi ha a tot el món, però el cost era extremadament gran i aquesta acció no es va dur a terme.

Finalment, una darrera limitació són els criteris definits. Aquests van ser creats per nosaltres, segons el que s'entenia quin era el perfil que s'estava buscant. Els criteris van descartar alguna persona que potser, vist amb des d'una altra perspectiva, realment hauria d'haver estat inclosa. Així doncs, en cas de tornar-se a realitzar el treball, potser seria oportú definir i contrastar els criteris amb professionals de l'estadística esportiva que no tinguessin relació amb el treball.

4.3. Aplicacions pràctiques i accions futures

La realització i finalitat dels cens no ha estat només per poder realitzar el Treball de Fi de Grau. A banda d'això, aquest ha sigut una primer pedra d'un cens únic al món. La intenció és que aquest cens no quedi només aquí, i és que es pugui millorar tot el que sigui possible. Encara no hi ha una ruta clara i definida de com poder-lo ampliar, però una primera opció que segurament es durà a terme serà la publicació i l'explicació d'aquest a una revista científica, per tal que pugui arribar al màxim de persones. Així, quantes més coneguin la seva existència, més possible serà trobar nous investigadors que es dediquin a *sports analytics/statistics*.

També, per a una futura recerca podria ser interessant intentar obtenir més informació dels investigadors, ja que l'enquesta que es va realitzar, contenia poques preguntes perquè pogués ser resposta ràpidament i fàcilment. Així doncs, amb més informació, podria ser possible que s'obtinguessin perfils molt més detallats o inclús nous perfils que, amb el treball realitzat, no hem trobat.

4.4. Conclusions

L'àmbit de *sports analytics/statistics* és desconegut per moltes persones. Tot i així, ja fa diversos anys que està en constant creixement, ja sigui a través de publicacions de revistes, llibres o la creació de grups de recerca. La realització d'aquest treball és un reflex d'aquest creixement, el qual pretén ajudar a seguir aquesta línia, amb la creació d'un cens de totes les persones que treballen en aquest àmbit que actualment no existeix.

Gràcies al cens realitzat, s'ha pogut identificar un nombre important d'investigadors influents en l'estadística esportiva i es poden treure conclusions important com el país de procedència d'aquests o quins grups de recerca hi ha a tot el món. Malgrat que té molt marge de millora, aquest primer cens serveix per tenir identificats una bona part dels professionals de l'estadística esportiva.

A partir a les respostes de l'enquesta realitzada per tots aquests investigadors, les tècniques de *clustering* ens han servit per conèixer i estudiar comportaments i perfils d'aquests.

Per acabar, estic molt satisfet per la realització d'aquest Treball de Fi de Grau en el que haurà servit per aportar el meu gra de sorra en aquest àmbit tant nou, però alhora tant consolidat i amb tant de futur com és el de l'estadística esportiva.

5. WEBGRAFIA

Admin. (2019, July 12). *Clustering Jerárquico - Agrupar elementos con minería de datos*. ESTRATEGIAS DE TRADING. https://estrategiastrading.com/clustering-jerarquico/

Aerts, M., Molenberghs, G., & Thas, O. (2021). Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What. *Annual Review of Statistics and Its Application*, *8*(1), 25-39. https://doi.org/10.1146/annurev-statistics-040620-032820

Alamar, B. C. (2013). *Sports Analytics*. Columbia University Press. https://doi.org/10.7312/alam16292

Albanese, N. C. (2022, January 15). Four R packages for Automated Exploratory Data Analysis you might have missed. *Towards Data Science*. https://towardsdatascience.com/four-r-packages-for-automated-exploratory-data-analysis-you-might-have-missed-c38b03d4ee16

Albert, J., Glickman, M. E., Swartz, T. B., & Koning, R. H. (2016). *Handbook of Statistical Methods and Analyses in Sports* (J. Albert, M. E. Glickman, T. B. Swartz, & R. H. Koning, Ed.; 1st Edition). Chapman and Hall/CRC. https://doi.org/10.1201/9781315166070

Arjona, M. (2021, November 23). *Sport Data Analytics y el futuro del deporte de élite*. Canal Gestión Empresarial. https://www.inesem.es/revistadigital/gestion-empresarial/sport-data-analytics/

ASA Community. (n.d.). StatisticsinSportsSection. Retrieved June 27, 2023, from https://community.amstat.org/sis/journals

Baldacci, H. E. (2021). KAMILA clustering for a mixed-type data analysis of Illinois medicare data. *Digital Commons @ Butler University* .

Banerji, A. (2021a, May 18). *K means clustering*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

Banerji, A. (2021b, May 18). *K means clustering*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

Baumer, B. S., Matthews, G. J., & Nguyen, Q. (2023). *Big Ideas in Sports Analytics and Statistical Tools for their Investigation*. 1-26.

Campos. (2021, June 23). *Diagrama de flujo PRISMA 2020*. BiblioGETAFE. https://bibliogetafe.com/2021/06/23/diagrama-de-flujo-prisma-2020/

Cao, R., Cuevas, A., & González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, *17*(2), 153-176. https://doi.org/10.1016/0167-9473(92)00066-Z

Casals, M., Fernández, J., Martínez, V., Lopez, M., Langohr, K., & Cortés, J. (2023). A systematic review of sport-related packages within the R CRAN repository. *International Journal of Sports Science & Coaching*, *18*(2), 621-629. https://doi.org/10.1177/17479541221136238

Chung, L. (2022a, February 17). *¿Cuál es una buena tasa de respuesta en las encuestas a clientes en 2022?* Delighted. https://delighted.com/es/blog/average-survey-response-rate

Chung, L. (2022b, February 17). *¿Cuál es una buena tasa de respuesta en las encuestas a clientes en 2022?* Delighted. https://delighted.com/es/blog/average-survey-response-rate

Cui, B. (2020, December 8). *Introduction to DataExplorer*. https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html

Davenport, T. H., & Patil, D. (2022). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business. Review*, *90*.

David, H. A., & Gunnink, J. L. (1997). The Paired t Test Under Artificial Pairing. *The American Statistician*, *51*(1), 9. https://doi.org/10.2307/2684684

Di Nardo, E., Polito, F., & Scalas, E. (2021). A Fractional Generalization of the Dirichlet Distribution and Related Distributions. *Fractional Calculus and Applied Analysis*, *24*(1), 112-136. https://doi.org/10.1515/fca-2021-0006

Dung, N. C. (n.d.). *Principal component analysis (principal component methods in R)*. Retrieved June 27, 2023, from https://rstudio-pubs-static.s3.amazonaws.com/323416_ab58ad22d9e64ba2831569cf3d14a609.html

Dzhaparidze, K. O., & Nikulin, M. S. (1995). On the computation of chi-square-type statistics. *Journal of Mathematical Sciences*, *75*(5), 1910-1921. https://doi.org/10.1007/BF02365082

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*(none). https://doi.org/10.1214/09-SS051

Fernández, J., Bornn, L., & Cervone, D. (2021). A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, *110*(6), 1389-1427. https://doi.org/10.1007/s10994-021-05989-6

Fisher, R. A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87. https://doi.org/10.2307/2340521

Foss, A., Markatou, M., Ray, B., & Heching, A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, *105*(3), 419-458. https://doi.org/10.1007/s10994-016-5575-7

Garzón, C. (2019, February 1). Métodos de estimación de densidad de Kernel (De ODF a EBSD). *Medium*. https://medium.com/@garzonsergio/m%C3%A9todos-de-estimaci%C3%B3n-de-densidad-de-kernel-de-odf-a-ebsd-b4a143dc9eee

Gelman, A., & Vehtari, A. (2021). What are the Most Important Statistical Ideas of the Past 50 Years? *Journal of the American Statistical Association*, *116*(536), 2087-2097. https://doi.org/10.1080/01621459.2021.1938081

Gil Martínez, C. (2018, June). *RPubs*. Métodos de Clustering. https://rpubs.com/Cristina_Gil/Clustering

Google Scholar. (n.d.). Retrieved June 27, 2023, from https://scholar.google.com/

Groll, A., & Liebl, D. (2023). Editorial special issue: Statistics in sports. *AStA Advances in Statistical Analysis*, *107*(1-2), 1-7. https://doi.org/10.1007/s10182-022-00453-9

Heras, J. M. (2020, January 25). *Clustering (Agrupamiento), K-Means con ejemplos en python - IArtificial.net*. Jose Martinez Heras. https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569-16572. https://doi.org/10.1073/pnas.0507655102

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

Kaushik, S. (2016, November 3). *Clustering*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Lemire, J. (2021, December 12). Zelus is not mythology; it's an analytics firm that's god to pro teams. *Sports Business Journal*.

https://www.sportsbusinessjournal.com/Daily/Issues/2021/10/12/Technology/zelus-is-not-just-greek-mythology-it-is-an-analytics-firm-that-is-god-to-nba-mlb-and-european-soccer-teams.aspx

Lewinson, E. (2020, June 27). Prediction Strength — a simple, yet relatively unknown way to evaluate clustering. *Towards Data Science*. https://towardsdatascience.com/prediction-strength-a-simple-yet-relatively-unknown-way-to-evaluate-clustering-2e5eaf56643

Lopez, M. J., & Matthews, G. J. (2015). Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, *11*(1). https://doi.org/10.1515/jqas-2014-0058

New England Symposium on Statistics in Sports. (n.d.). Retrieved June 27, 2023, from https://www.nessis.org/index.html

ResearchGate. (n.d.). ResearchGate. Retrieved June 27, 2023, from https://www.researchgate.net/

Rodríguez, D. (2022, January 14). Diferencias entre Hard y Soft Clustering. *Analytics Lane*. https://www.analyticslane.com/2022/01/14/diferencias-entre-hard-y-soft-clustering/

Royo, M. (2023). *BiblioGuías: Revisiones sistemáticas: Definición: ¿qué es una revisión sistemática?* BiblioGuías at Biblioteca Universidad de Navarra. https://biblioguias.unav.edu/revisionessistematicas/que-es-una-revision-sistematica

Oliver, D. (2004). *Basketball on Paper: Rules and Tools for Performance Analysis*.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, *74*(9), 790-799. https://doi.org/10.1016/j.recesp.2021.06.016

Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., Knight, E. J., & Bargary, N. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, *55*(2), 118-122. https://doi.org/10.1136/bjsports-2020-102607

Sánchez, L. D. (2023). *Guías de la BUH: Evaluación de la Investigación: Índice H*. Guías de La BUH at Universidad de Huelva. https://guiasbuh.uhu.es/c.php?g=655120&p=4605523

SHAPIRO, S. S., & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Sports, S., & Physiology, E. (2019, May 21). *What is sport science?* Sydney Sports and Exercise Physiology. https://ssep.com.au/what-is-sport-science/

Swartz, T. B. (2020a). Where Should I Publish My Sports Paper? *The American Statistician*, *74*(2), 103-108. https://doi.org/10.1080/00031305.2018.1459842

Swartz, T. B. (2020b). Where Should I Publish My Sports Paper? *The American Statistician*, *74*(2), 103-108. https://doi.org/10.1080/00031305.2018.1459842

Team, T. A. (2021, June 28). *Everything on hierarchical clustering*. Towards AI. https://towardsai.net/p/l/everything-on-hierarchical-clustering

Tibshirani, R., & Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511-528. https://doi.org/10.1198/106186005X59243

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., … Straus, S. E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, *169*(7), 467-473. https://doi.org/10.7326/M18-0850

UC business analytics R programming guide. (2018). *K-means cluster analysis*. https://uc-r.github.io/kmeans_clustering#fn:kauf

Wikimedia, C. de los proyectos. (2023e, May 23). Wikipedia. Wikipedia. https://es.wikipedia.org/wiki/Wikipedia

6. ANNEX

6.1. Enquesta

6.1.1. Correu electrònic

El correu electrònic que es va enviar a totes els investigadors va ser el següent:

Title: Survey for developing a census of researchers involved in sports statistics.

Email: Dear Dr. [Last Name],

I hope this email finds you well. Our team, consisting of Dr. Martí Casals (National Institute of Physical Education of Catalonia (INEFC), University of Barcelona, Spain https://mon.uvic.cat/ceeaf/team/marti-casals/), Dr. Daniel Fernández (Universitat Politècnica de Catalunya (UPC), Spain https://grbio.upc.edu/en/about-us/cv/daniel-fernandez), and undergraduate student in statistics: Martí Oliver (UCP), is conducting a research study titled "Description of the academic profile of sports statistics around the world." This study aims to create a census of the academic researchers involved in sports statistics.

As a respected expert in the field of sports statistics, we would like to invite you to participate in our study by completing a brief survey consisting of 7 questions, which should take no more than 3 minutes of your time. Your participation in this study is crucial in helping us gather data and insights on the current state of sports statistics research worldwide. Your expertise and contribution will be invaluable in helping us achieve our research objectives.

Please note that while the survey is not anonymous, we will ensure that your responses are kept confidential to the extent permitted by law. We understand the sensitivity of the information provided and assure you that it will be used solely for research purposes and share with you the final census.

Please find the link to the survey attached here [LINK], and we kindly request that you complete the survey at your earliest convenience.

Thank you very much for considering our request and taking the time to participate in our study. We appreciate your valuable contributions.

Sincerely,

6.1.2. Model de l'enquesta

L'enquesta que es va enviar a totes els investigadors va ser la següent:

<u>Survey:</u> Survey for developing a census of researchers involved in sports statistics

As a respected expert in the field of sports statistics, we would like to invite you to participate in our study by completing a brief survey consisting of 7 questions, which should take no more than 3 minutes of your time. Your participation in this study is crucial in helping us gather data and insights on the current state of sports statistics research worldwide.

1. What is your gender?

- Male
- Female
- Prefer not to say.
- Other

2. Which age group do you belong to?

- 18-29
- 30-44
- 45-59
- 60+
- I don't answer.

3. What is the highest level of education you have completed?

- High school
- Bachelor's degree
- Master's degree
- Doctorate or PhD

4. How many years have you worked in sports statistics in academia?

- None
- Less than 1 year
- 1-2 years
- 3-4 years
- 5 years or more

5. Which best describes your profile?

- Full-time academic of sports statistics/analytics field
- Part-time sports statistician in academia and part-time in the sports industry
- Academic sports statistician with years of expertise and contributions, but currently working in the sports industry.
- I do not consider myself an academic profile.

6. Do you currently belong to a sports statistics research group?

- Yes, namely: [open-ended response]
- No

7. What sport have you worked on the most on a statistical level? (Select all that apply)

- Basketball
- Baseball
- Soccer
- American Football
- e-Sports
- Cricket
- Tennis
- Ice Hockey
- Running
- Cycling
- Other: [open-ended response]

6.2. Codi R

Els codis i les bases de dades de R que s'han utilitzat per realitzar el treball es poden trobar al següent enllaç de GitHub: https://github.com/martioliver5/codi_TFG.git

6.3. Enllaços directes per la recerca d'investigadors

A continuació es mostres quatre enllaços directe a les pàgines *Google Scholar* i *ResearchGate* amb les paraules claus que es van utilitzar per trobar nous investigadors.

https://scholar.google.com/citations?hl=ca&view_op=search_authors&mauthors=label:sports_analytics

https://scholar.google.com/citations?hl=ca&view_op=search_authors&mauthors=label%3Asports_statistics&btnG=

https://www.researchgate.net/search.Search.html?query=sports+analytics&type=publication

https://www.researchgate.net/search.Search.html?query=sports+statistics&type=publication