Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

Analisi di raggruppamento dei giocatori di tennis professionisti per caratteristiche fisiche e statistiche di performance

Relatore Prof. Francesco Lisi

Dipartimento di Scienze Statistiche

Laureando: Giovanni Luzzatto

Matricola n. 2007339

Anno Accademico 2022/2023

A chi ha sempre creduto in me.

Indice dei contenuti

Introduzione

Uno dei contesti in cui la statistica sta assumendo un ruolo preponderante è sicuramente quello sportivo: sempre più atleti, allenatori, dirigenti e talent scout stanno infatti capendo quanto lo studio e l'analisi della grande mole di dati raccolti durante allenamenti e competizioni possano portare ad un notevole miglioramento della prestazione e ad un vantaggio competitivo nei confronti degli avversari.

L'elaborazione di queste informazioni può essere utile ai singoli atleti per identificare in modo oggettivo i propri punti di forza e di debolezza sia fisici che tecnici, per capire come sfruttare nel modo migliore le proprie abilità e per impostare allenamenti specifici per lavorare sulle proprie carenze.

Grande rilevanza dello studio di dati e prestazioni viene poi assunta negli sport tattici, ovvero quelle discipline in cui giocatori e allenatori devono affrontare la gara con una precisa pianificazione strategica, che molto spesso deve essere poi modificata nel corso della sfida per adattarsi alle situazioni che si vengono a creare e al comportamento assunto dall'avversario. In questo caso l'analisi statistica permette di individuare gli schemi più comunemente usati dai rivali e di pianificare quindi gli adattamenti tattici e strategici più appropriati per le possibili circostanze di gioco.

L'analisi dei dati può inoltre supportare i talent scout nell'individuazione di giocatori promettenti in una determinata disciplina, permettendo quindi un'analisi basata non soltanto sull'osservazione sul campo, ma anche su informazioni oggettive quali forza, velocità, resistenza e altre statistiche di performance.

Un ultimo importante obiettivo è la previsione dei risultati delle competizioni e il calcolo delle quote nelle scommesse: attraverso modelli statistici e algoritmi di previsione, infatti, è possibile valutare le probabilità di vittoria e di sconfitta di un atleta o di una squadra sulla base di variabili quali la classifica o il ranking, la forma fisica, lo storico degli scontri diretti e delle prestazioni recenti, le condizioni ambientali e le tattiche di gioco predilette.

Viste proprio le potenzialità sopra descritte, in tutti i principali sport (ed in particolare in quelli di strategia) il numero di analisti statistici è in crescente ascesa: tutte le principali società professionistiche di calcio, basket, baseball, football e pallavolo sono infatti dotate di un team di match analysis e svolgono pressoché quotidianamente sedute di video-analisi tecnica e tattica.
Restringendo il focus solo sugli sport individuali la situazione risulta essere abbastanza differente, in quanto generalmente solo gli atleti d'élite dispongono delle risorse finanziare necessarie per essere seguiti costantemente da un intero team di professionisti e la figura del match analyst riveste solitamente un ruolo marginale rispetto a quella del coach, del fisioterapista o dello psicologo.
L'interesse di questa relazione si rivolge proprio ad uno sport individuale, il tennis, in cui l'analisi delle caratteristiche dell'avversario e la loro comparazione con le proprie risultano essere uno dei principali strumenti atti alla pianificazione della strategia da utilizzare durante la partita: poiché ad alto livello le differenze tecniche e fisiche tra i giocatori non sono così significative, a primeggiare sono gli atleti che riescono a sfruttare nel modo migliore i propri punti di forza contro quelli di debolezza dell'avversario.
Lo scopo dell'elaborato è quello di effettuare un'analisi di raggruppamento dei giocatori di tennis professionisti per caratteristiche fisiche e di performance, cercando anche di individuare particolari combinazioni di variabili che portino ad una classificazione facilmente interpretabile e che permettano di cogliere le caratteristiche discriminanti dei vari cluster creati.
Il lavoro è strutturato nel seguente modo:

- nel Capitolo 1, dopo una concisa introduzione storica sull'applicazione della statistica nello sport, saranno descritti brevemente i più rilevanti studi presenti in letteratura riguardanti la cluster analysis di tennisti professionisti e l'individuazione delle statistiche di performance più significative per il successo in questa disciplina.
- nel Capitolo 2 verrà presentata una trattazione teorica dell'analisi di raggruppamento, in cui saranno illustrati i princìpi su cui si basa, le sue criticità e i principali metodi e algoritmi utilizzabili, con particolare

attenzione a quelli che verranno effettivamente impiegati nella successiva analisi empirica.

- il Capitolo 3 sarà invece dedicato all'illustrazione dei dati raccolti e alla discussione delle analisi preliminari, volte alla pulizia del dataset, e delle analisi esplorative delle variabili più significative.
- nel Capitolo 4 verranno infine proposte le analisi di raggruppamento effettuate, con particolare attenzione all'interpretazione dei diversi risultati ottenuti.

Capitolo 1

La statistica nello sport

1.1 Introduzione delle analisi statistiche nello sport e nel tennis

La prima applicazione di un metodo scientifico nel mondo dello sport si deve all'ex giocatore di baseball Billy Beane, che una volta divenuto Direttore Sportivo della squadra professionistica della Major League Baseball degli Oakland Athletics, ha iniziato a adottare la Sabermetrica come unico criterio per l'analisi delle prestazioni e per il processo di scouting di nuovi giocatori.

La Sabermetrica, introdotta dallo scrittore e statistico Bill James a partire dal 1977, è l'analisi empirica delle statistiche del baseball che misurano l'attività degli atleti in partita e può essere utilizzata soprattutto per la valutazione delle prestazioni passate di un giocatore e per la previsione di quelle future.

Oltre all'introduzione di questo metodo, uno dei contributi quantitativi più importanti di James è stata sicuramente la scoperta di un "Teorema Pitagorico" per la previsione delle vittorie nelle partite di baseball: egli ha dimostrato che una semplice funzione del quadrato dei punti (run) segnati e di quelli concessi da una squadra in una stagione fornisce una stima molto accurata della percentuale di vittorie nel corso dell'annata (James, 1981). Questo risultato ha ottenuto l'etichetta di "Pitagorico" a causa dei termini elevati al quadrato per i run di una squadra e della rivale, anche se successivamente è stato suggerito un esponente ottimale compreso tra 1.8 e 1.9 (Davenport & Walner, 1999; Braunstein, 2010). Nel tempo il teorema è diventato uno degli strumenti più importanti per gli analisti e gli allenatori di baseball: oltre alla previsione delle vittorie stagionali, il modello è stato anche utilizzato per ottenere diverse misure di coerenza delle prestazioni (Braunstein, 2010) e per determinare squadre con un rendimento insolitamente al di sopra o al di sotto delle aspettative (Vollmayr-Lee, 2002; Cha et al., 2007).

Vista la riconosciuta utilità di tale formula, molti autori hanno cercato di estenderla con vario successo ad altre discipline (Hamilton, 2011; Caro and Machtmes, 2013): diffusa è stata la sua applicazione agli sport di squadra (basket, hockey e football americano in particolare), mentre più raro è stato il suo adattamento agli sport individuali.

Focalizzandosi ora sul tennis, sport oggetto di questo elaborato, si può sicuramente evidenziare come, al pari delle altre discipline non di squadra, l'analisi dei principali indicatori di performance e l'elaborazione di algoritmi di analisi e previsione abbiano avuto uno sviluppo molto più lento, con un aumento del numero di studi che è avvenuto soltanto negli ultimi 10-15 anni.

Un contributo rilevante per il progresso della ricerca nel tennis è certamente dovuto a Jeff Sackman, autore e sviluppatore di software, che nel 2015 ha parlato di analisi del tennis alla Sloan Sport Analytics Conference e successivamente ha iniziato la pubblicazione di un ampio database contenente i risultati di tutte le partite di tennis professionistiche dal 1968, le classifiche mondiali ed una serie di indicatori sui singoli giocatori. Questo suo contributo, assieme alla successiva pubblicazione di "The Match Charting Project", database in continuo aggiornamento contenente i dati dettagliati punto per punto delle partite di tennis professionistiche, ha permesso agli analisti di poter disporre con facilità delle informazioni necessarie per lo studio di questa disciplina sportiva.

1.2 Letteratura

Di seguito è proposto un breve resoconto di 4 studi significativi presenti in letteratura riguardo le statistiche di performance e l'analisi di raggruppamento nel tennis, esposti in ordine cronologico di divulgazione.

Una delle più importanti pubblicazioni sulla performance nel tennis ha cercato di applicare il "Teorema Pitagorico" di James a questo sport (Kovalchik, 2016): dopo aver raccolto i dati presenti sul sito dell'ATP relativi alle partite giocate tra il 2004 e il 2014 dai giocatori classificati nelle prime 100 posizioni del ranking all'inizio di ogni anno e aver eliminato la collinearità tra le variabili, sono state considerate 10 misure di performance (3 relative al servizio e 7 relative alla risposta) tra loro sufficientemente incorrelate. Sono poi stati adattati diversi modelli sulla base di

quello di James: alcuni presentavano un'unica variabile esplicativa, altri avevano diverse combinazioni di esse. Tra i modelli aventi una sola variabile indipendente solo 4 hanno raggiunto una proporzione di varianza spiegata del numero di vittorie attese superiore al 50%: il miglior valore esplicativo lo ha mostrato quello costruito con la percentuale di break point vinti ($R^2 = 0.85$), seguito da quelli aventi rispettivamente il numero di opportunità di break per partita ($R^2 = 0.70$), la percentuale di punti totali vinti al servizio ($R^2 = 0.73$) e la percentuale di punti vinti sulla propria prima di servizio ($R^2 = 0.51$). L'adattamento del modello avente come unica esplicativa la percentuale di break point convertiti si è rivelato molto simile a quello del modello di James per il baseball: l'esponente migliore è risultato 1.83, il coefficiente di determinazione R^2 0.85 e il margine di errore di vittorie attese molto ridotto (circa 2 partite su una media di 50 a stagione).
Sempre per quanto riguarda i fattori critici del successo nel tennis, un'analisi condotta sui primi 100 giocatori del ranking ATP nel 2007 ha cercato di esaminare le relazioni tra la classifica mondiale e 14 statistiche di performance, con l'obiettivo di individuare gli indicatori maggiormente legati alla vittoria (Reid et al., 2017). Tramite una regressione lineare condotta con approccio backward gli autori sono giunti ad un modello di previsione ($R^2 = 0.52$) contente soltanto due variabili (ovvero le percentuali di punti vinti servendo una seconda battuta e rispondendo ad una seconda battuta dell'avversario) e con la seguente equazione di previsione:

$$classifica\ prevista = 548.5 - 666.6 * p1 - 319.9 * p2$$

dove $p1$ è la percentuale di punti vinti servendo la seconda di servizio e $p2$ è la percentuale di punti vinti rispondendo alla seconda di servizio dell'avversario.
Successiva a quest'ultimo studio è la prima pubblicazione avente come obiettivo il raggruppamento dei tennisti professionisti per caratteristiche fisiche e di gioco e l'individuazione dei fattori di performance discriminanti tra i vari gruppi (Cui et al., 2019). La cluster analysis (basata su peso, altezza, mano dominate, tipologia di rovescio e anni di esperienza) di 189 atleti ha portato all'individuazione di 4 gruppi: Giocatori destrimani di grande corporatura con rovescio bimane,

Giocatori destrimani di media corporatura con rovescio monomane, Giocatori destrimani di piccola corporatura con rovescio bimane e Giocatori mancini con rovescio bimane e la variabile più rilevante nella suddivisione dei giocatori è stata la tipologia di rovescio, seguita da mano dominate, altezza e peso.

La successiva applicazione dell'analisi multivariata della varianza (MANOVA) delle variabili di performance relative ai tornei del Grand Slam nel periodo 2015-2017 nei cluster ottenuti ha portato a numerosi risultati rilevanti: gli indicatori relativi al servizio ed alla risposta si sono rivelati essere i più importanti per la differenziazione dei tennisti nei gruppi in tutti i tornei del Grand Slam e i dati più significativi per la prestazione sono stati la velocità massima della battuta, la percentuale di ace nel punteggio di "deuce" e di "advantage" e la percentuale di punti conclusi con un vincente. I gruppi dei Giocatori mancini con rovescio bimane e dei Giocatori destrimani di piccola corporatura con rovescio bimane si sono dimostrati quelli più omogenei nelle prestazioni nei diversi tornei mentre i Giocatori destrimani di grande corporatura con rovescio bimane e i Giocatori destrimani di media corporatura con rovescio monomane hanno mostrato più differenze nel rendimento sulle differenti superfici, risultando però nel complesso migliori degli altri, soprattutto nella velocità del servizio e nel numero di ace, di punti a rete e di colpi vincenti per partita. Come già dimostrato in studi precedenti (Reid & Morris, 2013; Cui et al., 2017) l'esperienza, pur avendo una scarsa influenza nella determinazione dei cluster, si è rivelata invece un fattore rilevante per la performance.

Due sono i limiti dello studio riconosciuti dagli stessi autori: la mancata considerazione del livello dell'avversario nelle partite, che può influire notevolmente sulle prestazioni e quindi sulle statistiche raccolte, e la dimensione del campione (proprio a causa dello scarso numero di giocatori mancini con rovescio ad una mano, ad esempio, l'algoritmo ha unito questi giocatori a quelli destrimani con rovescio monomane).

Altri risultati interessanti sono infine emersi da un'analisi di raggruppamento su tennisti professionisti uomini che è stata effettuata utilizzando il metodo di partizionamento delle k-means (illustrato nel dettaglio nel Paragrafo 2.5) a partire

dai dati sugli incontri professionistici dal 2011 pubblicati da Sackman (Austin, 2021).

Anche in questo caso sono stati individuati 4 cluster, nonostante l'autore affermi che il secondo gruppo potesse essere ulteriormente suddiviso in modo più dettagliato a seconda del livello di gioco. Il primo gruppo contiene gli atleti di statura maggiore, caratterizzati dalla più alta percentuale di ace e di punti vinti sulla prima di servizio e dall'aver giocato un maggior numero di incontri su erba e cemento rispetto ai loro colleghi. Il secondo gruppo risulta il più numeroso e presenta giocatori definiti da Austin "mediocri": hanno disputato in media il minor numero di partite tra quelli considerati, probabilmente a causa di un'alternanza tra i tornei Atp e quelli minori del Challenger Tour, hanno vinto il minor numero di punti e partite e giocato in modo uniforme su tutte le superfici. Vi sono poi gli "all-courter", ovvero i migliori giocatori individuali, caratterizzati dalla più alta percentuale di punti vinti in totale e sulla propria seconda di servizio e dal minor numero di opportunità di break concesse per partita. L'ultimo cluster raggruppa invece i giocatori esperti della terra battuta, ovvero i tennisti che si specializzano in questa superficie e vi giocano un numero di partite annuali superiore a quello dei rivali: sono caratterizzati dalle più basse percentuali di prime di servizio in campo e di punti vinti sempre sulla prima battuta, ma si garantiscono il maggior numero di opportunità di break per incontro.

Emergono quindi i diversi approcci ed obiettivi di questi ultimi due studi: nel primo caso è stato proposto un raggruppamento basato sulle caratteristiche fisiche e di gioco ed è stata poi applicata un'analisi multivariata della varianza per identificare le differenze nelle statistiche di performance tra i vari gruppi, mentre nel secondo caso la formazione dei cluster si è fondata direttamente su quest'ultime.

La presente relazione cerca di inserirsi in questo contesto, tanto che nella prima parte dell'analisi empirica si proverà a replicare separatamente entrambe le proposte, per poi sperimentare una cluster analysis sulle componenti principali degli indicatori di performance.

L'ultimo obiettivo sarà invece quello di cogliere se e come le statistiche medie dei gruppi creati cambino in modo statisticamente significativo al variare della superficie di gioco: si potrà analizzare, ad esempio, se le prestazioni dei più forti

giocatori individuali si mantengono costanti in ogni tipologia di campo di gioco o se i giocatori abili al servizio perdono di competitività su una superficie più lenta come la terra battuta.

Capitolo 2

L’analisi di raggruppamento

2.1 Definizione

L’analisi di raggruppamento (o cluster analysis) è un processo volto a suddividere un insieme di unità statistiche o di variabili in gruppi (o cluster), in modo che vi sia un’alta similarità all’interno dei gruppi e una bassa similarità tra i vari gruppi.
Mentre la procedura di classificazione riguarda un numero noto di gruppi preesistenti e si propone di assegnare le nuove osservazioni ad uno di essi, nell’analisi di raggruppamento non si fanno ipotesi a priori relative al numero o alla struttura dei gruppi ma la suddivisione viene effettuata sulla base della distanza (o dissimilarità) tra le osservazioni o le variabili considerate: per questo motivo nell’applicazione pratica di tale tecnica una buona conoscenza del fenomeno in questione aiuta l’analista a distinguere i raggruppamenti buoni da quelli meno ragionevoli.
A livello empirico raramente possono essere esaminate tutte le possibili partizioni di *n* unità in *k* gruppi, visto che il loro numero cresce molto rapidamente all’aumentare della quantità di unità considerate secondo la seguente formula legata ai numeri di Stirling di seconda specie (dove $S(n,k)$ è il numero di possibili partizioni di *n* oggetti in *k* gruppi):

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$$

Per ovviare a questo problema, sono stati elaborati molti algoritmi di clustering che portano all’individuazione di gruppi ragionevoli senza avere la necessità di esaminare tutti i possibili raggruppamenti.

2.2 Classificazione dei metodi di analisi di raggruppamento

I metodi di clustering vengono classificati in due tipologie:

- *clustering esclusivo*: i gruppi risultanti non possono avere elementi in comune in quanto ogni elemento può essere assegnato ad uno e un solo gruppo.
- *clustering non esclusivo (o fuzzy clustering)*: un elemento può appartenere a più di un cluster. Tale applicazione si basa sull'idea che nella realtà gli oggetti siano distribuiti in modo da rendere difficile la loro attribuzione ad un gruppo piuttosto che ad un altro, e permette di associare un oggetto ai gruppi con un certo grado di appartenenza.

A loro volta i metodi esclusivi si suddividono nelle seguenti categorie:

- *metodi gerarchici*: il numero di gruppi non viene stabilito a priori e l'algoritmo, che può essere agglomerativo o divisivo, procede ad effettuare una serie di partizioni nidificate.
- *metodi di partizionamento*: creano un raggruppamento dei dati in *k* cluster con l'obiettivo di ottimizzare un criterio di partizionamento oggettivo (solitamente basato sulla similarità), in modo che ciascun gruppo contenga almeno un oggetto e ciascun oggetto appartenga ad uno ed un solo gruppo.
- *metodi basati sulla densità*: sono algoritmi che considerano i cluster come regioni nello spazio dei dati dense di oggetti e separate tra loro da regioni a bassa densità; permettono di individuare cluster di forma arbitraria e facilitano il filtraggio del "rumore", rappresentato dalle regioni a bassa densità.
- *metodi basati sulla griglia*: prevedono la discretizzazione di uno spazio multidimensionale in una struttura a griglia con un numero finito di celle nelle quali si collocano i vari oggetti, garantendo una velocità di calcolo elevata e dipendente solo dal numero di celle create e non da quello degli oggetti presenti.

- *metodi basati sul modello*: sono algoritmi che cercano di ottimizzare la corrispondenza tra un modello ipotizzato dall'utente e i dati, che vengono assunti come generati da una composizione di distribuzioni di probabilità.

È bene sottolineare che alcuni algoritmi esclusivi integrano le idee di vari metodi e per questo motivo può risultare difficile classificarli in modo rigoroso in una delle categorie sopra descritte. In questo elaborato verranno applicati i metodi gerarchici ed un metodo di partizionamento, illustrati in dettaglio rispettivamente nei paragrafi 2.4 e 2.5.

2.3 Distanza e dissimilarità

Per poter ottenere un raggruppamento partendo da un insieme di dati complesso è necessario calcolare una misura di "vicinanza" o "similarità": anche se la scelta della misura da adottare può essere molto soggettiva, è comunque fondamentale considerare la tipologia delle variabili rilevate (discrete, continue, dicotomiche) e la scala di misurazione adottata. Solitamente le variabili vengono raggruppate sulla base dei coefficienti di correlazione o di misure di associazione simili mentre la similarità tra gli oggetti viene indicata da una distanza.

Poiché questo elaborato si focalizza sul raggruppamento di unità statistiche, risulta utile richiamare la definizione di distanza tra due osservazioni e illustrarne le principali tipologie esistenti, evidenziando poi come il calcolo della dissimilarità sia influenzato dalla tipologia di variabili rilevate.

Dal punto di vista matematico, date due osservazioni p-dimensionali $\boldsymbol{x}' = [x_1,\ x_2,\ \dots,\ x_p]$ e $\boldsymbol{y}' = [y_1,\ y_2,\ \dots,\ y_p]$ una funzione reale $d(\boldsymbol{x}, \boldsymbol{y})$ è una distanza se soddisfa le seguenti proprietà:

- $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ (simmetria)
- $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ (non negatività)
- $d(\boldsymbol{x}, \boldsymbol{x}) = 0$ (identità)

Inoltre, una distanza è anche una metrica se soddisfa due proprietà aggiuntive:

- $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ d se e solo se $\boldsymbol{x} = \boldsymbol{y}$

- data una terza osservazione $\boldsymbol{z}' = [z_1,\ z_2,\ \dots,\ z_p]$, $d(\boldsymbol{x},\boldsymbol{y}) \leq d(\boldsymbol{x},\boldsymbol{z}) + d(\boldsymbol{y},\boldsymbol{z})$ (disuguaglianza triangolare)

La definizione più generale possibile di distanza tra due oggetti caratterizzati da vettori p-dimensionali contenenti caratteri esclusivamente quantitativi è quella di Minkowski:

$$d(\boldsymbol{x},\boldsymbol{y}) = [|x_1 - y_1|^m + |x_2 - y_2|^m + \cdots + |x_p - y_p|^m]^{\frac{1}{m}}$$

Da tale espressione si possono poi individuare i seguenti casi particolari:

- se *m*=1 si ottiene la distanza della città a blocchi (detta anche di Manhattan), secondo cui la distanza tra due oggetti in uno spazio p-dimensionale è data dalla somma dei valori assoluti delle differenze tra le loro componenti:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sum_{h=1}^{p} |x_h - y_h|$$

- se *m*=2 si ottiene la distanza Euclidea, che soddisfa anche le proprietà necessarie per essere definita una metrica ed è quella comunemente più utilizzata. È definita come la lunghezza del segmento che congiunge due oggetti in uno spazio p-dimensionale:

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

- se $m \rightarrow \infty$ si ottiene la distanza di Lagrange, secondo cui la distanza tra due oggetti in uno spazio p-dimensionale è la massima differenza tra le componenti dei vettori che li descrivono:

$$d(\boldsymbol{x},\boldsymbol{y}) = max_{h\in\{1,2,\dots,p\}}|x_h - y_h|$$

Per calcolare la distanza tra unità statistiche su cui sono state rilevate soltanto variabili di carattere quantitativo si può quindi ricorrere ad una delle definizioni

riportate, tenendo presente che quella Euclidea è la formulazione più completa e solitamente più utilizzata.

In Figura 2.1 sono rappresentati, con colori diversi, l'insieme dei punti nel piano cartesiano che hanno distanza unitaria dall'origine: si può notare come la loro disposizione nello spazio bidimensionale cambi a seconda della tipologia di distanza utilizzata.

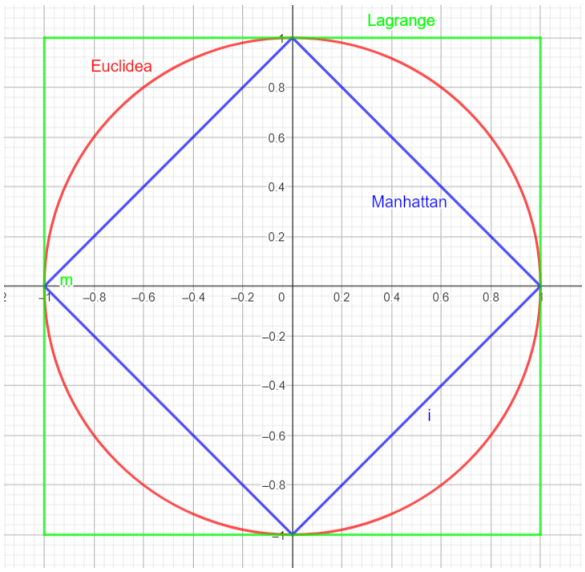


Figura 2.1: Insieme dei punti del piano cartesiano aventi distanza unitaria dall'origine, secondo la tipologia di distanza utilizzata (rosso per la distanza Euclidea, verde per quella di Lagrange e blu per quella di Manhattan).

Meno intuitiva è sicuramente la definizione di dissimilarità tra oggetti su cui sono state rilevate soltanto variabili qualitative o dicotomiche: in tal caso risulta difficile calcolare una distanza tra le osservazioni e vi è quindi la necessità di ricorrere a determinati indici o coefficienti di similarità, atti a fornire una misura del grado di associazione tra le osservazioni.

Un indice di similarità $S(u_i, u_j)$ è definito come una funzione che associa alla coppia di unità statistiche (u_i, u_j) un numero reale compreso tra 0 ed 1, dove 0 indica massima dissimilarità e 1 massima similarità, e soddisfa 3 proprietà:

- $S(u_i, u_j) \geq 0$ (non negatività)
- $S(u_i, u_j) = S(u_j, u_i)$ (simmetria)
- $S(u_i, u_j)$=1 se $u_i = u_j$

Nel caso di variabili binarie l'aspetto più rilevante per determinare la similarità tra due unità è il numero di "positive matches", ovvero la frequenza di item presenti contemporaneamente nei due soggetti. In questa situazione una prima e importante distinzione tra i coefficienti disponibili è quella che divide i coefficienti simmetrici da quelli asimmetrici: nel caso in cui i valori nulli rilevati sulle unità rappresentino un dato certo la scelta dovrà cadere su un coefficiente simmetrico, mentre se lo zero indica l'assenza di informazione bisogna utilizzare coefficienti asimmetrici, in modo da evitare che la simultanea assenza di un dato in due soggetti non generi un'elevata similarità tra essi.

Per ciascuna coppia di unità che si desidera confrontare avendo a disposizione due vettori *p*-dimensionali contenenti le stesse variabili dicotomiche è innanzitutto necessario indicare con *a* il numero di items (ovvero caratteristiche, elementi) comuni tra i due oggetti, con *b* e *c* il numero di items presenti esclusivamente nell'uno o nell'altro vettore e con *d* il numero di items nulli in entrambi (ovviamente la somma dei quattro valori appena citati deve essere *p*, ovvero il numero di variabili binarie di ciascun vettore), per poi procedere alla costruzione dei possibili coefficienti.

Tra i principali indici simmetrici che si possono utilizzare dopo aver individuato il numero di concordanze e discordanze vi sono:

- Indice di concordanza semplice (Sokal & Michener, 1958), che rappresenta il rapporto tra il numero di elementi concordanti tra le due unità ed il numero totale di caratteristiche rilevate. Esso non distingue fra i casi di concordanza su valori 0 e su valori 1 e quindi non risente del criterio utilizzato per la codifica binaria dell'informazione:

$$S_{j,k} = \frac{a + d}{p}$$

- Il coefficiente proposto da Rogers & Tanimoto (1960), che si differenzia da quello di concordanza semplice poiché attribuisce un peso doppio alle discordanze:

$$S_{j,k} = \frac{a + d}{a + 2b + 2c + d}$$

- Indice di Sokal & Sneath (1963), che attribuisce un peso doppio alle concordanze:

$$S_{j,k} = \frac{2a + 2d}{2a + b + c + 2d}$$

Tra gli indici asimmetrici troviamo invece:

- Indice di Jaccard (1900,1901,1908), che per escludere le co-assenze viene calcolato come rapporto fra le concordanze e il numero di osservazioni non nulle dei vettori:

$$S_{j,k} = \frac{a}{a + b + c}$$

- Indice di Sorensen (1948), che si differenzia da quello di Jaccard per il peso doppio attribuito alle concordanze:

$$S_{j,k} = \frac{2a}{2a + b + c}$$

- Versione asimmetrica del coefficiente di Rogers & Tanimoto, proposta da Sokal & Sneath nel 1963 per escludere le osservazioni nulle:

$$S_{j,k} = \frac{a}{a + 2b + 2c}$$

Se su due unità statistiche u_i e u_j sono rilevate esclusivamente variabili qualitative su scala sconnessa si può utilizzare un indice di corrispondenza semplice, ottenuto come proporzione di variabili in cui le due unità hanno la stessa modalità:

$$S_{j,k} = \frac{\sum_{j=1}^{p} I\{x_{ij} = x_{jp}\}}{p}$$

Se invece due unità statistiche presentano variabili qualitative su scala ordinale, una possibile soluzione per evitare di perdere l'ordinamento tra le *m* modalità è quella di trasformarle in numeri interi da 1 a *m* e, dopo aver normalizzato il risultato, calcolare la distanza tra le unità.

Nel caso in cui vi sia la rilevazione di variabili sia quantitative sia qualitative, una prima idea potrebbe essere trasformare i caratteri qualitativi in quantitativi e

ricondurre il tutto ad una stessa scala, ma questa operazione potrebbe portare a diversi risultati ed alla perdita di informazione. Una soluzione efficace è quella data dall'Indice di Gower (1971), costruito in modo da trattare ciascuna variabile di un vettore p-dimensionale in maniera ottimale in rapporto alla sua natura. Tale coefficiente, mediante una variabile indicatrice che assume valore unitario nel caso in cui le unità siano confrontabili rispetto ad un carattere e valore nullo in caso contrario (ovvero se una delle due ha un valore mancante o si ha una co-assenza per un carattere binario asimmetrico), permette di calcolare la media delle similarità individuali per ogni descrittore sulla coppia di osservazioni secondo la seguente formula

$$S_{j,k} = \frac{\sum_{i=1}^{p} w_i s_i}{\sum_{i=1}^{p} w_i}$$

dove s_i è la similarità tra le due osservazioni relativa all'i-esima variabile e w_i è la variabile indicatrice.

La valutazione delle similarità tra le singole variabili può essere effettuata con svariati metodi, ma in origine Gower proponeva di indicare $s_i = \frac{1-|x_{ij}-x_{ik}|}{R_i}$ (dove x_{ij} e x_{ik} sono i valori dell'i-esima variabile nelle osservazioni *j* e *k* e R_i è l'intervallo di variazione dell'i-esima variabile nelle osservazioni disponibili) per i caratteri qualitativi ordinali o quantitativi e $s_i = 1$ o $s_i = 0$ rispettivamente nelle situazioni di concordanza e discordanza per caratteri binari o qualitativi su scala sconnessa, trattando il caso della concordanza da doppio zero come valore nullo o mancanza di informazione.

2.4 Cluster gerarchico

Come già indicato in precedenza i metodi gerarchici procedono ad effettuare una serie di partizioni nidificate sulla base delle distanze tra le osservazioni e i possibili algoritmi si dividono in agglomerativi e divisivi.

Nel caso agglomerativo viene applicata una strategia bottom-up in quanto l'algoritmo parte dalle singole unità statistiche (pertanto inizialmente il numero di

cluster è pari a quello degli oggetti considerati) per poi raggrupparle man mano in gruppi sempre più numerosi in base alla loro similarità: l'ultima iterazione, quindi, porta alla fusione di tutti gli oggetti in un singolo gruppo.
I metodi divisivi operano invece in modo inverso, attuando una strategia top-down: l'unico gruppo iniziale contenente tutte le unità viene diviso in sottogruppi sempre più piccoli fino a quando ciascun oggetto forma un cluster singolo oppure fino a quando non vengono soddisfatte le eventuali condizioni di terminazione determinate a priori e legate solitamente al numero di cluster o alla loro distanza.
Per rappresentare la successione delle partizioni generate da un clustering gerarchico viene utilizzato un dendrogramma (Figura 2.2), ovvero un albero che permette di visualizzare come gli oggetti vengono raggruppati passo dopo passo.

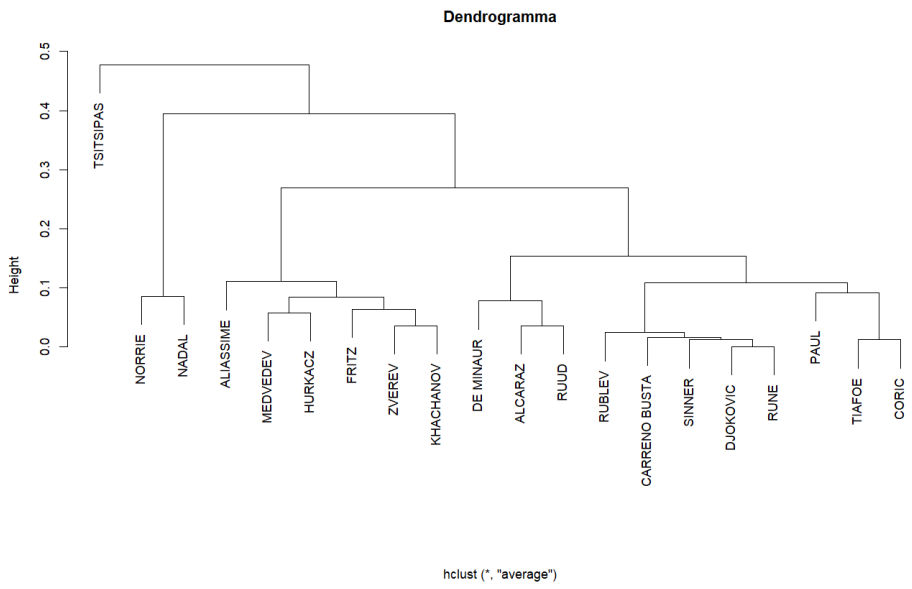


Figura 2.2: Esempio di dendrogramma utilizzato per rappresentare la successione delle partizioni generate da un algoritmo di clustering gerarchico.

In questo elaborato verrà utilizzato esclusivamente il metodo agglomerativo e la distanza tra i gruppi che si andranno a formare sarà calcolata in quattro modi diversi (legame singolo, legame completo, legame medio, metodo di Ward) per poter analizzare le differenze nei risultati ed individuare la miglior interpretazione possibile.
Utilizzando il legame singolo ("single linkage") la distanza tra due gruppi è definita come la minima distanza tra due dei loro rispettivi elementi, ossia la distanza tra

i loro membri più vicini. Come si vedrà nella successiva analisi empirica, questo metodo tende a formare cluster allungati, aggiungendo gli elementi non ancora classificati in gruppi già esistenti piuttosto che formarne di nuovi (effetto di "concatenazione"), e risulta essere più sensibile degli altri alla presenza di osservazioni anomale (privilegiando più la differenza tra i gruppi piuttosto che l'omogeneità al loro interno). Tale metodologia può essere utile quando si vuole creare un assorbimento dei cluster minori in quelli maggiori oppure quando i gruppi "veri" in cui sono divise le unità presentano una forma allungata.

Il legame completo ("complete linkage") identifica come distanza tra due gruppi la massima distanza tra due dei loro rispettivi elementi, ossia la distanza tra i loro membri più lontani.

Il legame medio ("average linkage") definisce la distanza tra due cluster come la media aritmetica delle distanze tra tutte le possibili coppie di elementi dei due gruppi. Essendo basato sulla media delle distanze i risultati sono considerati solitamente più attendibili e meno influenzati dalla presenza di outlier. La configurazione prodotta dal legame medio risulta nella maggior parte dei casi abbastanza simile a quella del legame completo, ma la fusione tra i gruppi avviene ad un livello diverso poiché le distanze sono definite in modo differente.

Il metodo di Ward si differenzia invece dai tre precedenti in quanto non si basa sulla distanza tra gli elementi ma sulla minimizzazione della perdita di informazione (definita come la somma delle devianze interne ai cluster) generata dalla fusione di due gruppi: ad ogni passo vengono calcolate le devianze associate a tutti i possibili raggruppamenti e viene poi fatta l'aggregazione che porta al gruppo con devianza minima. Questo metodo trae quindi origine dalla differenza tra la varianza tra i gruppi (between) e quella entro i gruppi (within), concetto che permette di confrontare la dispersione dei dati tra i vari gruppi e all'interno degli stessi. Il metodo di Ward tende a creare cluster di forma ellittica e per via del suo approccio può essere considerato precursore dei metodi di partizionamento non gerarchici, in cui la scelta del numero ottimale di gruppi si basa proprio sulla proporzione di varianza totale spiegata dal raggruppamento.

Dall'analisi dei metodi di clustering gerarchico si riscontrano due importanti criticità di questi algoritmi: la sensibilità agli outlier e l'impossibilità di ricollocare

nel corso del raggruppamento un elemento inserito in un gruppo errato in una iterazione precedente. Un'altra problematica è data poi dalla mancata scalabilità dell'algoritmo, che necessita il ricalcolo della matrice di distanza tra i cluster ad ogni nuova iterazione. Infine, valori comuni nella matrice di dissimilarità o distanza possono generare soluzioni multiple, specialmente considerando un numero di gruppi elevato: tali soluzioni non sono necessariamente negative, ma l'analista deve esserne a conoscenza per poter interpretarle nel modo corretto e confrontare i diversi raggruppamenti formati.

Per valutare a livello empirico la bontà di un raggruppamento di tipo gerarchico su un insieme di dati non vi sono test statistici specifici ma possono essere utilizzate varie tecniche, la cui analisi è però legata sempre all'interpretazione del singolo: in primis può essere utile confrontare i risultati ottenuti applicando i diversi metodi e, per ognuno di essi, utilizzare diverse distanze. Inoltre può essere valutata la stabilità dell'algoritmo di classificazione prima e dopo aver aggiunto delle "perturbazioni", ad esempio degli outlier: se i gruppi risultano tra loro ben distinti, i raggruppamenti prima e dopo la perturbazione non dovrebbero presentare differenze rilevanti. Infine, attraverso l'analisi della Silhouette (quantità che assume valore negativo se un'unità risulta più vicina ad un altro gruppo rispetto che al proprio, ed è quindi erroneamente classificata, e positivo in caso contrario) è possibile individuare e quantificare gli oggetti inseriti in un cluster sbagliato.

2.5 Metodi di partizionamento

I metodi di raggruppamento non gerarchici, o di partizionamento, sono algoritmi di tipo esclusivo ideati per suddividere gli elementi (e non variabili) in un numero predefinito di gruppi compatti e ben separati attraverso l'ottimizzazione di un criterio di partizionamento oggettivo.

Il metodo più popolare e di semplice applicazione è senza dubbio quello delle k-means elaborato da MacQueen (1967), in cui ogni cluster è rappresentato dal proprio centroide. Esistono due diversi approcci di inizializzazione di tale metodo: nel primo viene effettuata una partizione degli oggetti in k gruppi mentre nel secondo vengono individuati k punti come centroidi di partenza ed in entrambi i

casi le selezioni avvengono in modo del tutto casuale per evitare l'insorgenza di distorsione. A questo punto l'algoritmo procede in maniera iterativa assegnando ciascun oggetto al cluster del centroide più vicino e ricalcolando quindi i centroidi dei cluster così formati, per terminare quando il criterio predeterminato converge oppure quando non vi sono quindi più spostamenti degli elementi da un gruppo all'altro. Tale algoritmo risulta abbastanza scalabile ed efficiente per processare un gran numero di dati perché la sua complessità computazionale è $O(nkt)$, dove n è il numero totale di oggetti, k quello di cluster richiesti e t il numero di iterazioni necessarie per raggiungere la convergenza (solitamente $k \ll n$ e $t \ll n$).
La decisione sul numero k ottimale di gruppi non è facile: dal punto di vista pratico si è portati ad effettuare i raggruppamenti per diversi valori di k e a confrontarli poi tra loro sulla base di un determinato criterio. Una possibilità consiste nell'analizzare la proporzione di varianza spiegata da un numero diverso di raggruppamenti (Figura 2.3), con l'obiettivo di individuare quando tale quantità risulta buona (in genere è auspicabile almeno il 60/70% di quella totale) e quando l'incremento marginale che si ottiene creando un ulteriore cluster è irrilevante.

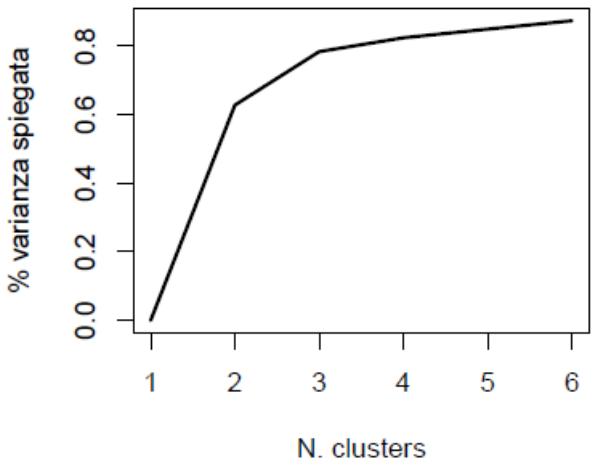


Figura 2.3: Grafico raffigurante la variazione della percentuale di varianza totale spiegata in funzione del numero di gruppi creati.

Anche l'utilizzo di questo metodo necessita di alcuni accorgimenti per evitare di ottenere risultati errati o poco affidabili e interpretabili. L'algoritmo può essere infatti applicato solo in presenza di variabili numeriche (vista la necessità di

calcolare ad ogni passo la media di ciascun gruppo), non è adatto per dati che presentano un raggruppamento “naturale” in cluster non convessi o con dimensioni molto differenti tra loro ed è sensibile alla presenza di outlier, che possono influenzare sensibilmente i centrodi. Risulta inoltre utile eseguire più volte l’algoritmo con una diversa situazione di partenza per verificare l’attendibilità degli esiti ottenuti, in quanto è stato verificato a livello empirico che l’assegnazione finale degli elementi ai cluster può non essere unica e risulta in qualche modo dipendente dalla partizione iniziale o dalla selezione iniziale dei centroidi.

2.6 Criticità e possibili soluzioni

Una prima problematica che si deve spesso affrontare quando si effettua un’analisi di raggruppamento di elementi su cui sono stati rilevati esclusivamente caratteri quantitativi riguarda l’unità di misura delle variabili: se queste sono state rilevate su scale tra loro molto diverse, la distanza tra le osservazioni sarà inevitabilmente condizionata dai caratteri aventi scala maggiore. Per risolvere questo problema si può ricorrere alla standardizzazione del dataset, facendo così in modo che tutte le variabili abbiano pari peso nella determinazione delle distanze.

Una seconda criticità, detta “Maledizione della dimensionalità” (“Curse of dimensionality”), può essere riscontrata analizzando un insieme di unità statistiche descritte da un numero elevato di variabili: all’aumentare del numero di caratteri considerati gli oggetti diventano sempre più sparsi nello spazio occupato e le definizioni di distanza necessarie per gli algoritmi di clustering diventano meno significative. In situazioni altamente dimensionali è stato dimostrato come, per alcune distribuzioni, la differenza tra la distanza di un punto dal suo punto più lontano e quella dal suo punto più vicino tenda a 0 al crescere del numero delle dimensioni considerate.

In questi casi quindi la riduzione della dimensionalità dei dati è un passaggio necessario per diminuire la complessità computazionale dell’algoritmo e permettere al contempo una più semplice interpretazione dei risultati: una delle tecniche più utilizzate in statistica è l’Analisi delle componenti principali (Pca), alla

base della quale vi è l'assunzione che la matrice di covarianza delle variabili originarie possa essere ugualmente spiegata da una serie di trasformazioni lineari delle variabili stesse.

Dato quindi il vettore p-dimensionale $\mathbf{X}' = [X_1,\ X_2, \dots,\ X_p]$ si costruiscono p nuove variabili $Y_1,\ Y_2, \dots, Y_p$ con varianza massima e tra loro incorrelate attraverso le seguenti combinazioni lineari

$$Y_1 = e_1' X$$
$$Y_2 = e_2' X$$
$$\dots..$$
$$Y_p = e_p' X$$

dove $e_1, e_2, \dots, e_p$ sono gli autovettori della matrice di varianza-covarianza di X associati ai corrispondenti autovalori $\lambda_1, \lambda_2, \dots, \lambda_p$, ordinati in modo decrescente. Delle nuove p variabili costruite, si deve decidere il numero $k \ll p$ di componenti da considerare per ridurre la dimensionalità, basandosi su criteri quali l'entità degli autovalori, la proporzione di varianza spiegata dalla singola componente, la proporzione totale di varianza spiegata dalle componenti che si vogliono considerare e la loro interpretazione. Come per la determinazione del numero di cluster da considerare, anche in questa situazione non vi è un unico criterio oggettivo ma la scelta dipende molto dalla sensibilità dell'analista. Una volta stabilito il numero di componenti rilevanti, l'applicazione di un algoritmo di clustering su di esse permette di evitare i problemi causati da un'elevata dimensionalità e ottenere risultati facilmente interpretabili.

Capitolo 3

I dati

3.1 Raccolta dei dati e analisi preliminari

Per effettuare l'analisi prefissata sono stati utilizzati i dati ufficiali relativi alle caratteristiche fisiche e di gioco e alle statistiche di performance presenti sul sito dell'*Atp*, integrandoli con quelli relativi alla velocità media del servizio disponibili su *UltimateTennisStatistics*. Poiché in entrambi i casi le informazioni riguardanti ciascun tennista sono presentate sotto forma tabellare in un'apposita pagina dedicata al singolo atleta, per ottenere un unico database si è resa necessaria un'attività di *scraping*, ovvero una tecnica informatica di estrazione di dati da un sito web mediante l'utilizzo di programmi software. Attraverso le librerie "rvest" e "tidyverse" disponibili in R, programma utilizzato anche per tutte le successive analisi, è stato possibile estrarre i dati necessari e ottenere 4 dataset grezzi, contenenti le variabili di interesse relative all'intera carriera dei primi 300 giocatori del ranking Atp del 20/03/2023, divise rispettivamente per "tutte le superfici", terra, erba e cemento.

Le analisi preliminari sui dataset costruiti hanno evidenziato la necessità di effettuare alcune operazioni sia sul campione sia sulle variabili. Innanzitutto, a causa della grande variabilità del numero di partite disputate da ogni tennista, si è reso opportuno standardizzare le grandezze espresse in valore assoluto per poter permettere il confronto tra i giocatori: il numero di ace, di doppi falli, di break point concessi e di opportunità di break sono stati normalizzati per i game giocati, il numero di tiebreak invece per gli incontri disputati.

Per quanto riguarda il campione analizzato, invece, sono state eliminate prima le unità statistiche aventi uno o più valori mancanti e successivamente anche quelle con un numero di incontri disputati inferiore a 5, in modo da prevenire la distorsione causata da un numero ridotto di informazioni (per la scelta del numero minimo di incontri ci si è basati su quanto proposto da Kovalchik, 2016). In seguito

alle prime procedure di raggruppamento è emersa anche la necessità di eliminare l'atleta Hong, che avendo una percentuale di break point convertiti e di game di risposta vinti pari a 0 costituisce un outlier con un impatto rilevante nella formazione dei gruppi.

I 4 dataset finali contengono così 182 atleti su cui sono state osservate le 40 variabili di seguito codificate e descritte:

- *Ranking*: posizione nella classifica Atp del 20/03/2023
- *Age*: età
- *Hand*: mano utilizzata per giocare
- *Backhand*: tipologia di rovescio
- *Weight*: peso
- *Height*: altezza
- *First_av_speed*: velocità media della prima di servizio
- *Second_av_speed*: velocità media della seconda di servizio
- *First_sv*: percentuale di prime di servizio in campo
- *First_sv_pw*: percentuale di punti vinti sulla prima di servizio
- *Second_sv_pw*: percentuale di punti vinti sulla seconda di servizio
- *Aces_sgp*: numero di ace per turno di servizio
- *Df_sgp*: numero di doppi falli per turno di servizio
- *Aces_df*: rapporto tra numero di ace e di doppi falli
- *Bpf_sgp*: break point affrontati per turno di servizio
- *Bps*: percentuale di break point salvati
- *Sgw*: percentuale di game al servizio vinti
- *Tspw*: percentuale di punti totali vinti al servizio
- *First_ret_pw*: percentuale di punti vinti in risposta alla prima di servizio
- *Second_ret_pw*: percentuale di punti vinti in risposta alla seconda di servizio
- *Bpo_rgp*: opportunità di break per turno di risposta
- *Bpc*: percentuale di break point convertiti
- *Rgw*: percentuale di game in risposta vinti
- *Rpw*: percentuale di punti totali vinti in risposta
- *Match_tot*: partite disputate in carriera

- *Perc_clay_match*: percentuale di partite disputate sulla terra
- *Perc_grass_match*: percentuale di partite disputate sull'erba
- *Perc_hard_match:* percentuale di partite disputate sul cemento
- *Perc_win*: percentuale di partite vinte
- *Perc_lose*: percentuale di partite perse
- *Perc_win_clay*: percentuale di partite vinte sulla terra
- *Perc_lose_clay*: percentuale di partite perse sulla terra
- *Perc_win_grass*: percentuale di partite vinte sull'erba
- *Perc_lose_grass*: percentuale di partite perse sull'erba
- *Perc_win_hard*: percentuale di partite vinte sul cemento
- *Perc_lose_hard*: percentuale di partite perse sul cemento
- *T_match*: numero di tiebreak giocati per partita
- *Perc_t_win*: percentuale di tiebreak vinti
- *Perc_t_lose*: percentuale di tiebreak persi

3.2 Analisi esplorative

Prima di iniziare con le procedure di raggruppamento, un'analisi esplorativa dei dati raccolti senza distinzione per superficie può permettere di cogliere le principali caratteristiche del campione analizzato e di individuare le relazioni tra le variabili più rilevanti, che possono poi essere sfruttate nella fase di interpretazione dei gruppi ottenuti.

Si può innanzitutto osservare che il rovescio preponderante nel tennis moderno è quello bimane, praticato dal 90% dei giocatori considerati, mentre la percentuale di mancini risulta del 15.4%. Appare inoltre evidente la presenza di una rilevante eterogeneità fisica tra i giocatori professionisti di alto livello: il campo di variazione di peso e altezza è rispettivamente di 44 kg e 41 cm e, sebbene si noti una forte correlazione lineare positiva tra le due grandezze (0.76), vi sono atleti con corporature molto differenti (si pensi che per una stessa altezza il peso può variare anche di 20 kg).

Come osservabile dagli scatterplot riportati in Figura 3.1, si registra poi una correlazione abbastanza forte tra le caratteristiche fisiche e alcuni indicatori di performance: peso e altezza presentano un'associazione lineare positiva con le

prestazioni al servizio (in particolare velocità media, percentuale di break point salvati, numero di ace e rapporto tra questi ultimi ed i doppi falli, diminuzione delle occasioni di break concesse) e negativa con le performance in risposta (riduzione del numero di opportunità di break e delle percentuali di punti e di game vinti).

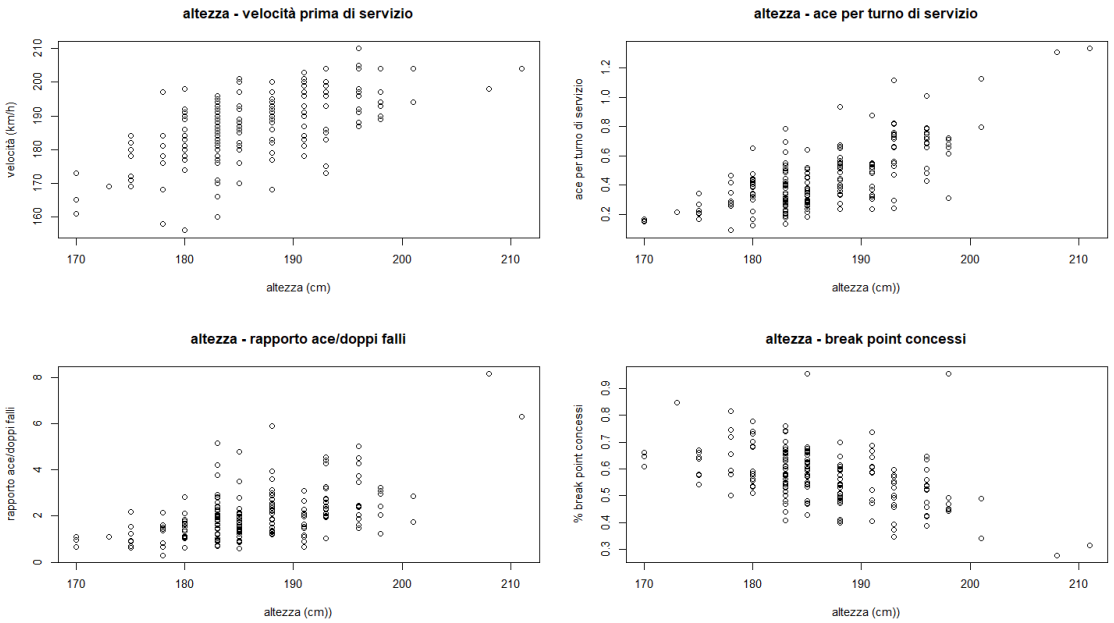


Figura 3.1: Scatterplot raffiguranti le statistiche di performance maggiormente correlate con l'altezza degli atleti: i coefficienti di correlazione lineare risultano rispettivamente di 0.60 per la velocità della prima di servizio, 0.70 per il numero medio di ace per turno di battuta, 0.52 per il rapporto tra ace e doppi falli e -0.48 per il numero di break point concessi.

Sebbene la correlazione indichi soltanto una relazione sistematica tra due variabili (ovvero la tendenza di queste a variare insieme) e non includa il concetto di causa-effetto, la contestualizzazione dei legami riscontrati nello sport del tennis si collega all'idea secondo cui atleti più alti o pesanti (ovvero con più massa muscolare) abbiano un vantaggio al servizio grazie ad un angolo di impatto più favorevole e/o una maggior forza, ma tendano ad essere più svantaggiati in risposta in quanto, generalmente, difettano di rapidità e possono avere difficoltà nello scambio da fondo campo quando non sono in una situazione di attacco.
Focalizzando invece l'attenzione sulle correlazioni all'interno delle sole statistiche di performance emergono relazioni ancor più importanti di quelle segnalate in

precedenza, che possono fornire indicazioni molto utili sugli aspetti chiave che determinano le vittorie a livello ATP. Ad essere fortemente correlate con la percentuale di game vinti al servizio non sono tanto la velocità della battuta, il numero di doppi falli o la percentuale di prime in campo, quanto più la proporzione di punti vinti sulla prima di servizio (e, ma meno rilevante, anche sulla seconda), il numero di ace e il loro rapporto con i doppi falli, il numero di break point concessi e il numero di quelli salvati (grafici rappresentati in Figura 3.2). Vi è inoltre una correlazione lineare pressoché perfetta (0.97) tra la percentuale di punti totali vinti al servizio e quella dei game.

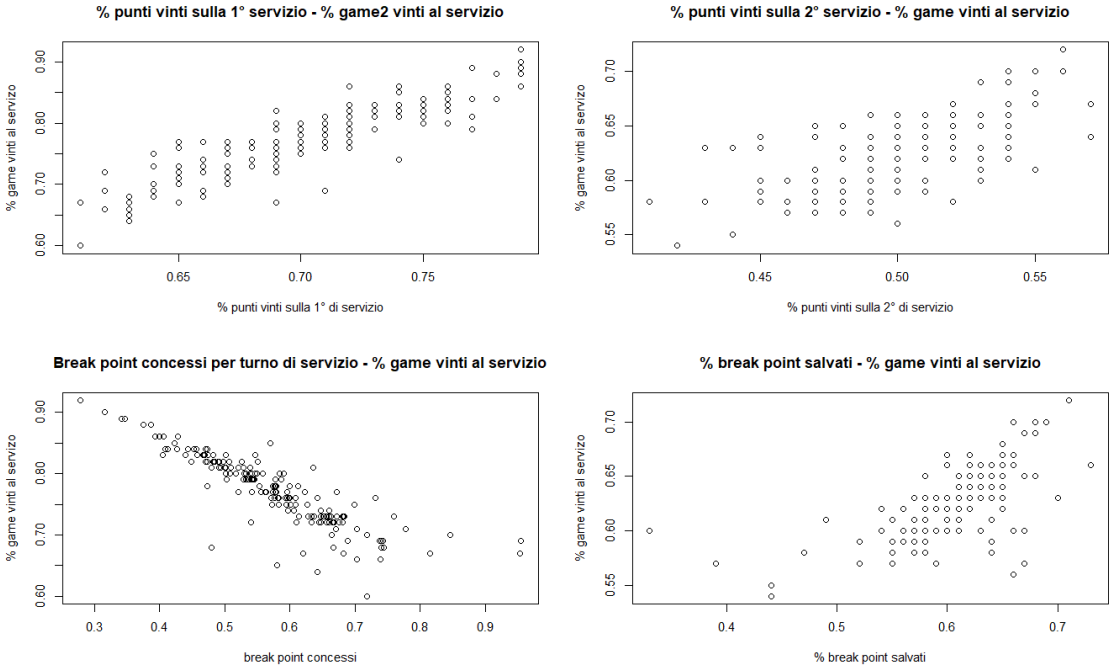


Figura 3.2: Scatterplot raffiguranti la variazione della % di game vinti al servizio in funzione delle statistiche di performance ad essa maggiormente correlate.

Ad essere fortemente correlate con la proporzione di game vinti in risposta sono invece il numero di break point avuti e i punti vinti sulla prima di servizio dell'avversario (Figura 3.3).

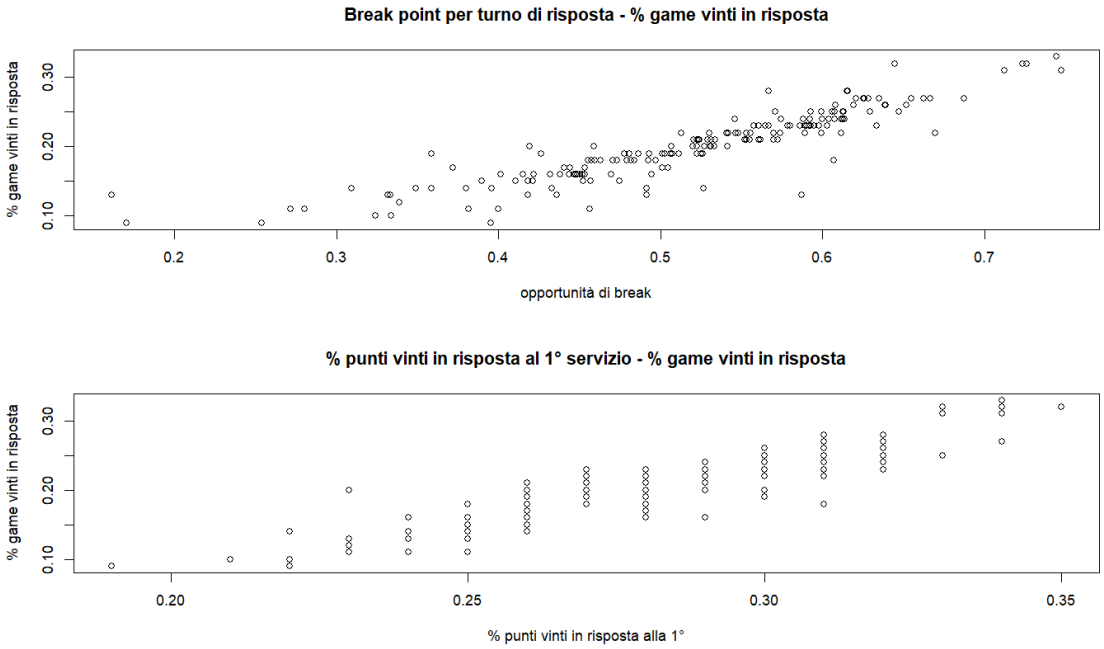


Figura 3.3: Scatterplot raffiguranti la variazione della % di game vinti in risposta in funzione delle statistiche di performance ad essa maggiormente correlate.

Emergono infine due ultimi aspetti, forse scontati ma che vale la pena segnalare perché caratteristici di uno sport come il tennis. Il numero di incontri giocati in carriera, oltre ad essere ovviamente legato all'età, è abbastanza correlato alla percentuale di vittorie ottenute (coefficiente di 0.56): questa disciplina, infatti, non prevede un calendario di partite/gare da giocare a priori, come succede per molti sport di squadra e per altri sport individuali (nuoto, ciclismo, golf), ma poiché si può partecipare ad un solo torneo a settimana, il numero di partite dipende fortemente dal numero di vittorie e dunque dai turni disputati in ciascun tabellone. Il secondo fatto è che la percentuale di vittorie è fortemente correlata positivamente (coefficiente di 0.85) con la percentuale di punti totali vinti: tale dato evidenzia quindi l'importanza di vincere più punti dell'avversario, nonostante non di rado a trionfare siano giocatori che hanno perso la maggioranza dei punti disputati in una singola partita (Lisi et al., 2019).

Capitolo 4

Analisi empirica

4.1 Analisi di raggruppamento per caratteristiche fisiche e di gioco

La prima analisi di raggruppamento proposta si basa esclusivamente sulle caratteristiche fisiche e di gioco: essendoci quindi due caratteri quantitativi (peso e altezza) e due dicotomici (mano dominante e tipologia di rovescio) un approccio sensato da utilizzare è un clustering gerarchico applicato alla matrice di distanze calcolata mediante l'indice di Gower, in modo che la dissimilarità tra due unità risulti la media pesata di quella relativa alle singole variabili, ciascuna considerata secondo la propria tipologia. Tale metodo fornisce infatti esiti più coerenti e facilmente interpretabili rispetto a quelli che si otterrebbero assegnando i valori 1 e 0 alle modalità di ciascuna delle due variabili dicotomiche ed effettuando il raggruppamento con il metodo delle k-means considerando tutti i caratteri come quantitativi.

I risultati ottenuti dal clustering gerarchico con distanza euclidea mostrano delle differenze nel processo aggregativo delle unità in gruppi da parte dei diversi legami utilizzati: per il legame completo ed il legame medio la caratteristica rilevante per una prima divisione in due gruppi risulta essere la tipologia di rovescio, mentre negli altri casi si ottengono delle bipartizioni particolari, con il legame singolo e il metodo di Ward che separano rispettivamente i due mancini con rovescio monomane e i mancini con rovescio bimane dagli altri giocatori.

Focalizzandosi poi su una divisione in quattro gruppi si possono individuare altre interessanti analogie e diversità tra i metodi utilizzati: con il legame medio e il legame singolo si giunge ad una ripartizione per ciascuna combinazione mano dominante-tipologia di rovescio e la discriminazione per caratteristiche fisiche inizia a influenzare il processo di raggruppamento soltanto a partire da un numero di gruppi pari a 5; il legame completo e il metodo di Ward producono invece la

stessa suddivisione presentata nello studio analizzato nel paragrafo 2.2 (Cui et al., 2019), evidenziando il gruppo dei giocatori con rovescio monomane (18 componenti), il gruppo dei giocatori mancini con rovescio bimane (26 componenti), il gruppo dei giocatori destrimani di piccola-media corporatura con rovescio bimane (84 componenti) e il gruppo dei giocatori destrimani di grande corporatura con rovescio bimane (54 componenti), con piccole differenze tra loro nella suddivisione dei tennisti destrimani con rovescio bimane a seconda della conformazione fisica (in Figura 4.1 sono osservabili, rispettivamente da sinistra verso destra, i raggruppamenti sopracitati prodotti dal legame completo).

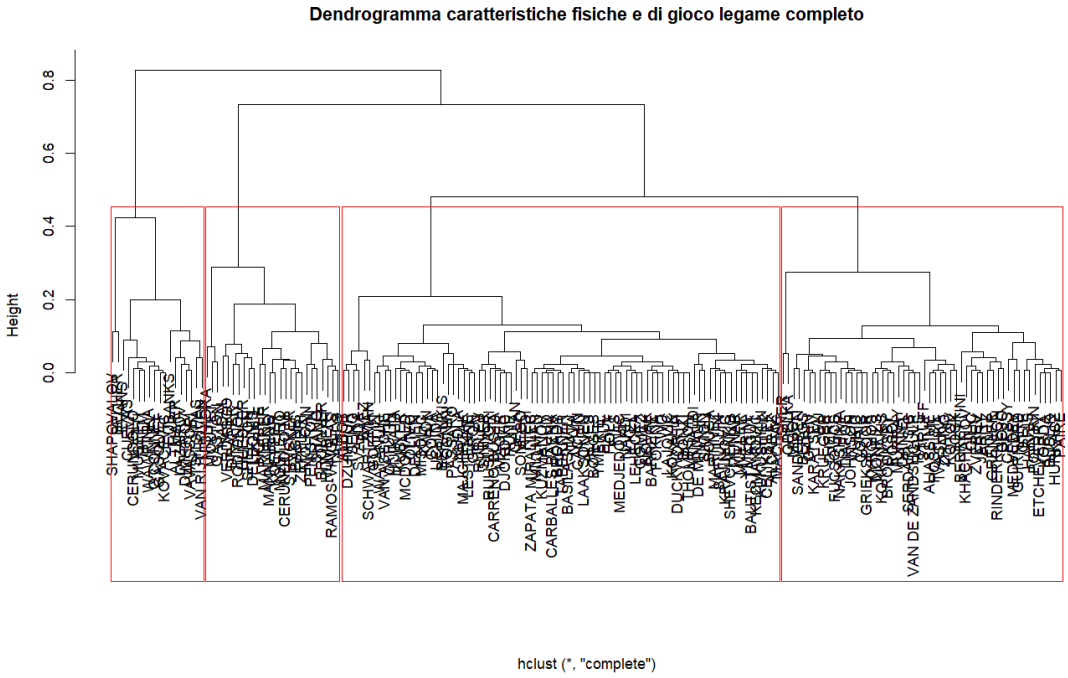


Figura 4.1: Dendrogramma raffigurante il raggruppamento per caratteristiche fisiche e di gioco prodotto dal legame completo: da sinistra verso destra vi sono i Giocatori con rovescio monomane, i Giocatori mancini con rovescio bimane, i Giocatori destrimani di piccola-media corporatura con rovescio bimane e i Giocatori destrimani di grande corporatura con rovescio bimane.

Per verificare che la differenza nella corporatura media nei due gruppi di giocatori destrimani con rovescio bimane sia statisticamente significativa è necessario ricorrere al test non parametrico di Wilcoxon, visto il rifiuto dell'ipotesi di normalità per la distribuzione di peso e altezza nei due cluster da parte dei test di Jarque-

Bera, D'Agostino e Shapiro-Wilk: con un p-value pressoché nullo sia per il test sull'altezza sia per quello sul peso, si conclude che la differenza di corporatura media nei due gruppi è statisticamente significativa.

4.2 Analisi di raggruppamento per statistiche di performance

La seconda analisi di raggruppamento considera tutte le statistiche di performance raccolte, ad eccezione dei dati relativi ai tiebreak e alle percentuali di vittorie e sconfitte assolute e divise per superficie. Essendoci solo variabili quantitative, opportunamente standardizzate in modo tale che le uniche due espresse su una scala notevolmente maggiore rispetto alle altre (ovvero velocità media della prima e della seconda di servizio) non influenzino in modo preponderante la distanza tra le osservazioni, possono essere utilizzati sia il clustering gerarchico sia il metodo di partizionamento delle k-means.

Uno dei passaggi più delicati nell'applicazione dell'algoritmo delle k-means consiste nella scelta del numero ottimale di gruppi in cui suddividere le unità: per cercare di individuare a priori tale quantità possono essere svolte due analisi di tipo quantitativo e grafico simili tra di loro e basate sul presupposto che, al crescere del numero di gruppi, aumenta anche la similarità all'interno di essi e dunque la proporzione di varianza spiegata.

In Figura 4.2 viene infatti riportata la percentuale di varianza totale spiegata da un diverso numero di raggruppamenti: si può notare come la proporzione risulti non molto elevata (circa il 60% del totale) anche considerando una divisione in 8 cluster e come l'incremento marginale sia meno rilevante oltre i 4 o 5 gruppi creati.

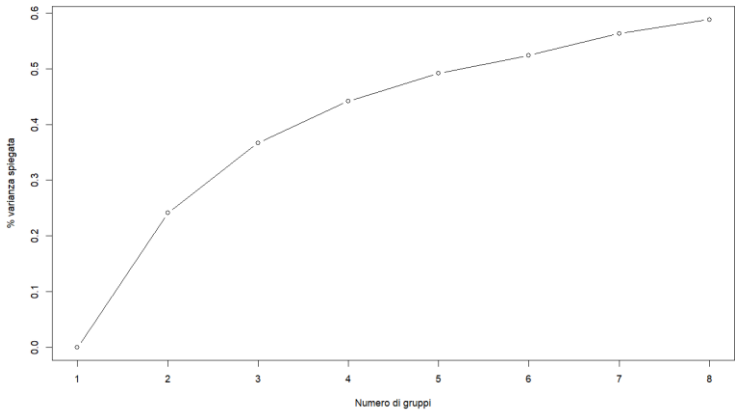


Figura 4.2: Variazione della percentuale di varianza totale spiegata in funzione del numero di raggruppamenti prodotti dal metodo delle k-means applicato alle statistiche di performance.

In figura 4.3 viene invece rappresentato l'output grafico del secondo approccio che può essere utilizzato, che si focalizza sul valore della varianza entro i gruppi: anche in questa situazione si registra una diminuzione marginale della variabilità interna a partire da un numero di gruppi pari a 4 o 5.

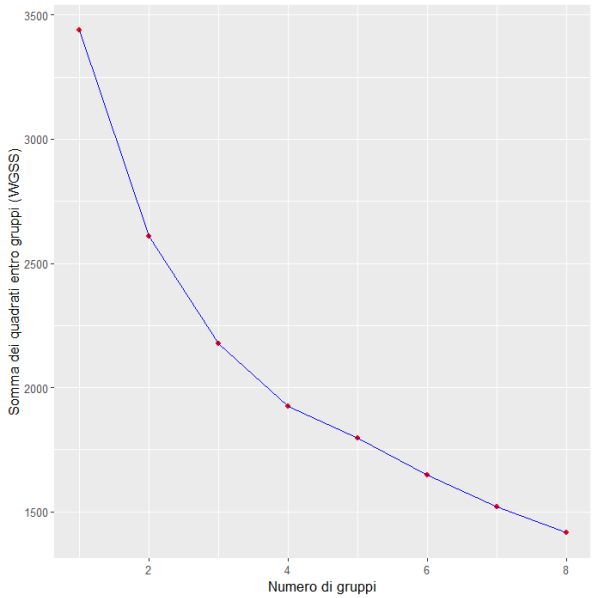


Figura 4.3: Variazione della varianza entro i gruppi in funzione del loro numero in seguito all'applicazione del metodo delle k-means alle statistiche di performance.

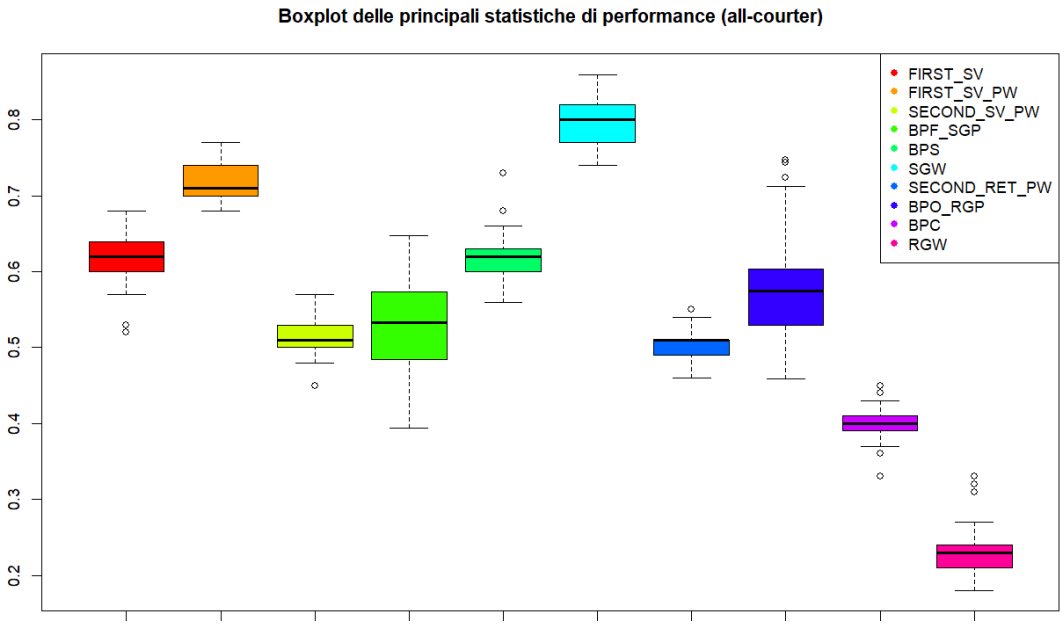
Da quanto appena constatato sembra dunque ragionevole effettuare un raggruppamento in 4 cluster, ma è sempre utile provare anche a diminuire e ad aumentare tale numerosità per cogliere le variabili discriminanti per la suddivisone delle unità e per cercare possibili miglioramenti nell'interpretazione dei risultati.

Iniziando quindi con una semplice bipartizione si può notare come l'algoritmo suddivida i 182 atleti in base alle differenze nelle performance complessive al servizio ed alla risposta, creando un primo gruppo di tennisti più abili alla battuta e un secondo gruppo di tennisti più forti in ribattuta. Nel complesso i giocatori che mostrano prestazioni migliori al servizio prediligono le superfici veloci (erba e cemento) e presentano una percentuale media di vittorie statisticamente più alta (pvalue molto vicino a 0) rispetto al gruppo dei ribattitori, che preferiscono invece disputare un numero maggiore di tornei sulla terra battuta.

Considerando poi il raggruppamento in 4 cluster, grazie alla buona conoscenza del fenomeno in questione e all'analisi sia dei centrodi sia delle variabili non considerate per il calcolo della matrice delle distanze, i risultati prodotti dall'algoritmo risultano facilmente comprensibili, vista anche la loro analogia con gli esiti dello studio di Austin. Il gruppo più numeroso, composto da 61 tennisti, è quello degli all-courter, ovvero i migliori giocatori individuali, ed è caratterizzato dal ranking medio nettamente più alto (posizione 72) e dalle migliori percentuali di incontri vinti in carriera e di punti vinti con la seconda di servizio, in risposta ad una seconda di servizio dell'avversario e in totale. Seguendo lo spunto di Austin, risulta sensato interpretare il secondo cluster come quello dei giocatori "mediocri", in quanto caratterizzati dal più basso rapporto tra incontri vinti e disputati, dalla quasi totalità delle statistiche di performance peggiori in assoluto e da uno scarso livello di esperienza, dovuto principalmente ad un'età media piuttosto giovane e ad un numero di partite giocate a livello Atp di gran lunga inferiore rispetto a quello dei rivali. Il terzo cluster raggruppa invece quei tennisti che possono essere definiti "grandi battitori", caratterizzati dalle migliori performance complessive al servizio (fatta eccezione per proporzione di prime in campo e di punti vinti con la seconda) e da quelle in assoluto meno buone in risposta. I componenti del quarto cluster presentano infine le peggiori prestazioni

assolute alla battuta (fatta eccezione per il numero di prime di servizio in campo), a cui compensano però con i migliori dati in risposta: tali tennisti possono essere identificati come “specialisti della terra battuta”, visto l’elevato numero di incontri disputati su tale superficie (il 44% circa del totale, notevolmente maggiore rispetto a quello di tutti gli altri gruppi) e alle caratteristiche (bassa velocità del servizio ma buon numero di prime battute in campo, abilità in risposta) che ben si adattano ai campi più lenti.

Alla luce di quanto affermato da Austin nella sua analisi e dell’eterogeneità presente nei vari gruppi, osservabile anche nei boxplot riportati in Figura 4.4 (che rappresentano, come esempio, i principali indicatori al servizio e alla risposta per gli all-courter e i giocatori mediocri), è possibile ipotizzare che aumentando il numero di cluster si registri una più dettagliata stratificazione degli atleti secondo le differenze di performance: il gruppo degli all-courter potrebbe essere sezionato in un insieme formato da coloro che occupano posizioni di rilievo nel ranking (o che oggi, a causa dell’età o degli infortuni, sono indietro in classifica ma che in passato hanno occupato le prime posizioni) e coloro che invece presentano un livello inferiore, così come ci si aspetta una partizione più approfondita per livello di gioco tra i giocatori mediocri e nel gruppo dei battitori.



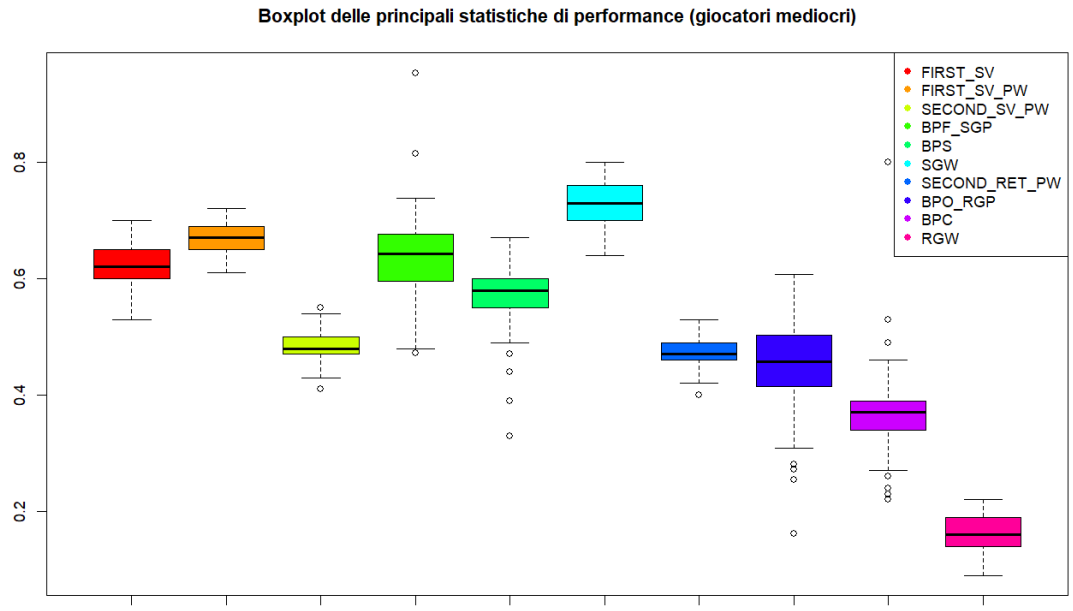


Figura 4.4: Boxplot delle principali statistiche di performance per i giocatori "all-courter" e "mediocri".

Proprio come ipotizzato, aumentando a 6 il numero di cluster non si registra la creazione di raggruppamenti con nuove caratteristiche (ad esempio un gruppo di specialisti dell'erba o delle superfici veloci) ma una più dettagliata segmentazione degli atleti per livello di prestazione, che non sembra però portare ad un miglioramento nella caratterizzazione e interpretabilità dei gruppi stessi.

Aumenta il livello di performance necessario per entrare sia tra gli all-courter, che perdono quasi la metà dei componenti rispetto al caso precedente e migliorano la posizione media nel ranking (ora 47) e la percentuale di vittorie in carriera (che passa dal 56% al 62%, risultando statisticamente più alta della precedente anche a un livello di significatività dell'1%), sia tra i grandi battitori, il cui numero di elementi scende a 7 e le cui statistiche medie al servizio crescono in modo considerevole. Troviamo poi gli specialisti della terra battuta, un gruppo con buoni dati relativi al servizio (contenente tutti i giocatori che non fanno più parte dei grandi battitori e alcuni provenienti dagli all-courter del caso precedente), un cluster non molto numeroso contenente gli atleti con le prestazioni peggiori e infine i tennisti che presentano una distribuzione delle performance tra battuta e risposta simile agli all-courter ma che non raggiungono il livello prestazionale di

questi ultimi. Il confronto della Silhouette riportata in Figura 4.5 porta a ritenere la qualità dei due raggruppamenti molto simile: il valor medio e il numero di elementi non correttamente classificati risultano rispettivamente di 0.17 e 22 per la divisione in 4 cluster e 0.16 e 19 per quella in 6 cluster.

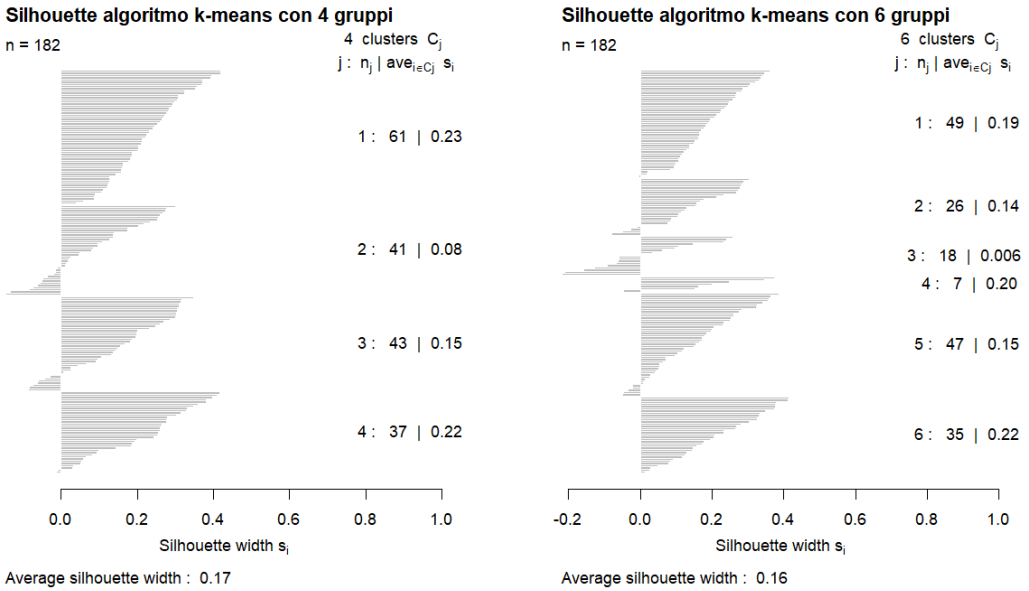


Figura 4.5: Silhouette relative ai risultati prodotti dal metodo di partizionamento per un numero di gruppi rispettivamente pari a 4 e 6.

L'applicazione agli stessi indicatori di performance di un clustering gerarchico basato sulla distanza euclidea porta subito ad escludere l'utilizzo del legame singolo e del legame medio, che si rivelano inadatti in questa situazione in quanto, come visibile nel dendrogramma riportato in Figura 4.6, portano alla formazione di cluster molto allungati, aggiungendo gli elementi ai gruppi già esistenti invece di crearne di nuovi.

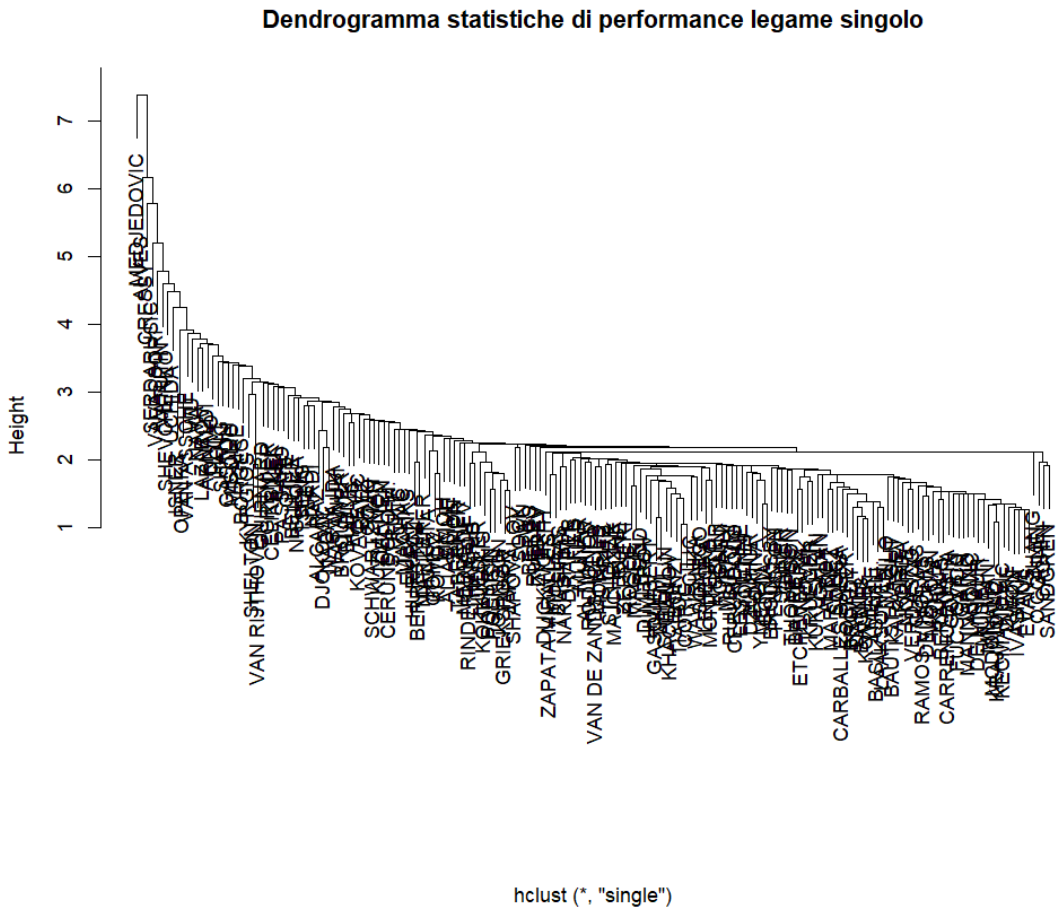


Figura 4.6: Dendrogramma raffigurante la successione delle partizioni prodotte dal legame singolo applicato alle statistiche di performance.

Per una suddivisione in 4 gruppi il legame completo si mostra incapace di raggruppare i giocatori in modo rappresentativo, nonostante la Silhouette evidenzi un valor medio più elevato (0.22 contro 0.16) e un numero di unità erroneamente classificate molto minore (13 contro 34) rispetto al metodo di Ward, che fornisce invece una caratterizzazione dei cluster addirittura analoga a quella fornita dal metodo di partizionamento precedente. In questo caso le principali differenze consistono nella diminuzione del numero di all-courter e di specialisti della terra battuta (in favore dell'aumento dei giocatori mediocri e dei battitori) e nell'assegnazione di alcuni tennisti considerati completi e aventi una classifica molto buona (come Tsitsipas, Tiafoe, Zverev, Khachanov, Coric, Cilic e Dimitrov) proprio al gruppo dei battitori piuttosto che a quello degli all-courter in cui sono stati inseriti dall'algoritmo delle k-means.

4.3 Analisi di raggruppamento per componenti principali delle statistiche di performance

Nonostante il numero di variabili considerate nelle precedenti analisi non sia particolarmente elevato, può risultare interessante individuare ed interpretare le componenti principali delle statistiche di performance, per poi verificare se i raggruppamenti prodotti applicando il metodo di partizionamento delle k-means a tali componenti siano quanto meno simili a quelli ottenuti nel paragrafo antecedente.

Dall'osservazione dell'entità degli autovalori associati a ciascuna componente (nel caso di dati standardizzati si è soliti considerare rilevanti quelli maggiori di 1) e dell'andamento dello screeplot in Figura 4.7 sembra ragionevole considerare 4 componenti principali, che insieme spiegano l'82% circa della varianza complessiva.

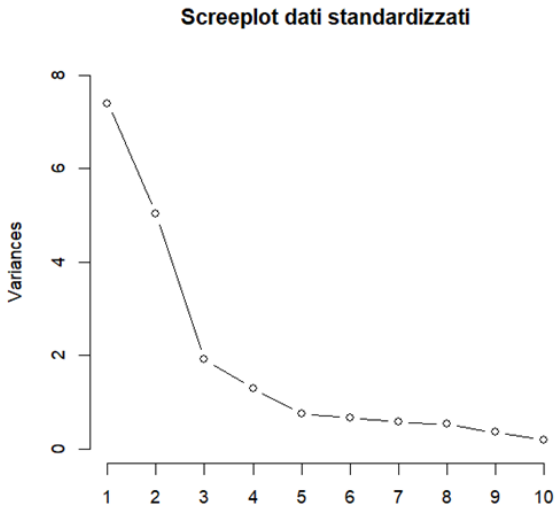


Figura 4.7: Screeplot dei dati standardizzati, che rappresenta la varianza spiegata da ciascuna componente principale: si può notare il "gomito" (ovvero una notevole diminuzione della varianza marginale) in corrispondenza della quarta o quinta componente.

L'interpretazione delle componenti principali ottenute alla luce dei loadings (pesi) associati alle singole variabili (riportati in Tabella 4.1) non risulta però immediata: la prima componente può rappresentare le prestazioni massime del giocatore al servizio, dati i pesi positivi di quasi tutte le statistiche relative alla battuta e

negativi di quelle relative alla risposta; la seconda componente può essere ritenuta un indice complessivo delle cattive performance dell'atleta, poiché presenta pesi negativi per la quasi totalità delle variabili (giocatori di alto livello avranno quindi punteggi negativi e grandi in valore assoluto rispetto a questa componente); la terza componente può essere letta come un indicatore di "solidità" alla battuta, visti gli elevati pesi positivi per le percentuali di prime di servizio in campo e di punti vinti con la seconda e per la variabile indicante il rapporto tra aces e doppi falli; la quarta componente può essere associata all'abilità di vincere i punti chiave alla risposta, dato il peso positivo molto alto della percentuale di break point convertiti.

VARIABILI	PC1	PC2	PC3	PC4
FIRST_AV_SPEED	0.277	0.046	-0.211	0.192
SECOND_AV_SPEED	0.256	0.057	-0.219	0.253
FIRST_SV	-0.058	-0.045	0.489	0.284
FIRST_SV_PW	0.318	-0.153	-0.167	0.023
SECOND_SV_PW	0.141	-0.255	0.271	-0.152
ACES_SGP	0.330	-0.022	-0.124	0.078
DF_SGP	0.0366	0.105	-0.619	0.065
ACES_DF	0.279	-0.063	0.280	0.039
BPF_SGP	-0.286	0.182	-0.118	-0.175
BPS	0.205	0.184	-0.114	-0.367
SGW	0.307	-0.227	0.029	-0.068
TSPW	0.309	-0.225	0.056	0.028
FIRST_RET_PW	-0.205	-0.315	-0.102	0.162
SECOND_RET_PW	-0.191	-0.304	-0.068	-0.149
BPO_RGP	-0.196	-0.336	-0.134	-0.163
BPC	-0.111	-0.065	-0.013	0.712
RGW	-0.216	-0.333	-0.129	0.151
RPW	-0.217	-0.347	-0.091	0.055
TPW	0.091	-0.423	-0.034	0.061

Tabella 4.1: Loadings (pesi) delle statistiche di performance per ognuna delle componenti principali considerate.

In Figura 4.8 sono rappresentate le variabili originarie e le unità statistiche sul piano delle prime due componenti principali: in accordo con quanto sostenuto in precedenza si osserva che ad avere i valori più alti rispetto alla prima componente sono i tennisti con le migliori performance assolute al servizio (Isner (46), Opelka (138) e Cressy (37)), mentre focalizzandosi sulla seconda componente principale si osservano valori positivi elevati per i giocatori meno competitivi e valori negativi per coloro che occupano le prime posizioni del ranking.

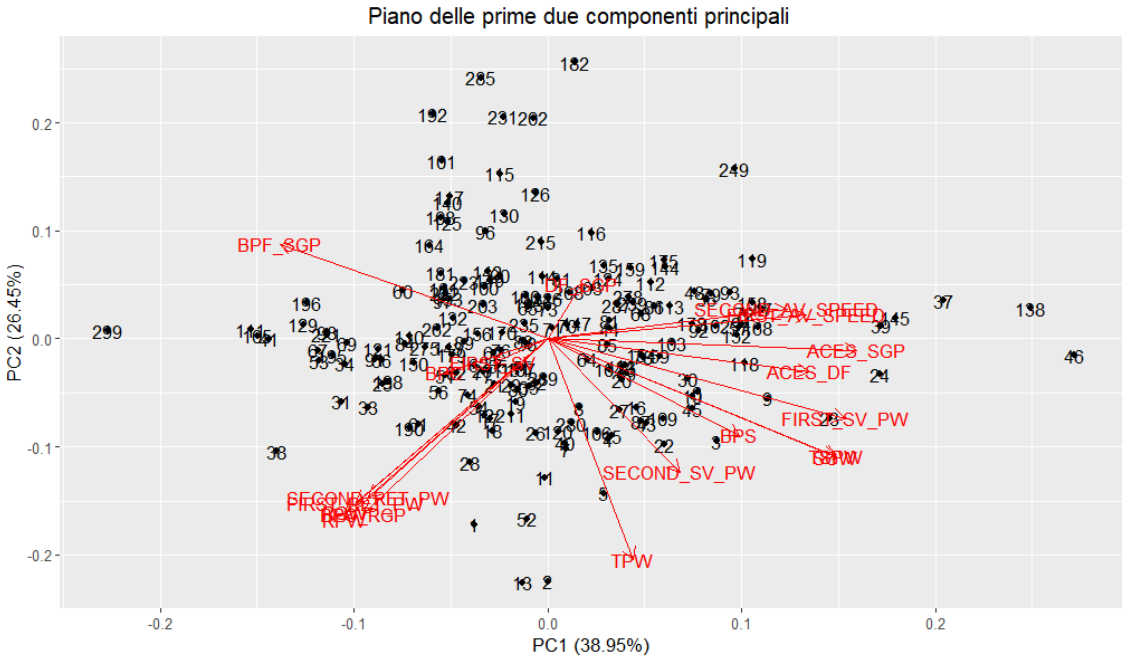


Figura 4.8: Rappresentazione delle variabili originarie e delle unità statistiche (identificate dal ranking del giocatore) sul piano delle prime due componenti principali.

In seguito, all'assegnazione ad ogni giocatore di un punteggio rispetto a ciascuna componente principale mediante la combinazione lineare dei pesi delle variabili è quindi possibile effettuare un'analisi di raggruppamento basata sulle quattro nuove grandezze ottenute, tutte di carattere quantitativo. I grafici utili ad individuare il numero ottimale di gruppi riportati in Figura 4.9 mostrano un andamento molto simile a quelli ottenuti considerando tutte le singole statistiche di performance e suggeriscono quindi una suddivisione in 4 cluster.

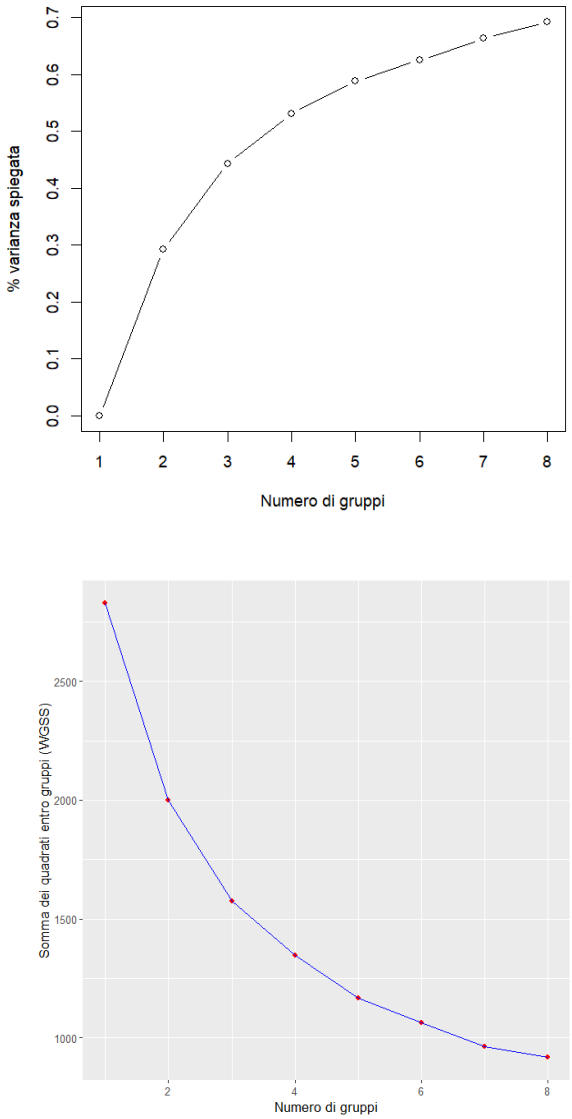


Figura 4.9: Variazione della percentuale di varianza totale spiegata (grafico in alto) e della varianza entro i gruppi (grafico in basso) in funzione del numero di raggruppamenti prodotti dal metodo delle k-means applicato alle componenti principali delle statistiche di performance.

I risultati ottenuti sono molto rilevanti: i raggruppamenti formati dal metodo di partizionamento delle k-means applicato alle componenti principali presentano non solo la stessa caratterizzazione di quelli prodotti considerando tutte le statistiche di performance, ma anche gli stessi elementi all'interno di ogni cluster, fatta eccezione per due soli tennisti (Tsitsipas e Sock) che vengono inseriti tra i battitori invece che tra gli all-courter. Anche provando a restringere ulteriormente

la dimensionalità dei dati i raggruppamenti restano pressoché immutati: utilizzando 3 componenti principali l'unico ulteriore cambiamento consiste nello spostamento di Uchida dal gruppo dei giocatori mediocri a quello dei battitori, mentre considerandone solo 2 lo stesso Uchida torna nei giocatori mediocri e si registra il solo spostamento di Berankis, che viene inserito tra i giocatori mediocri invece che tra gli all-courter (in cui si trovava nei casi di 3 e 4 componenti e di tutte le statistiche di performance).

L'analisi delle componenti principali si dimostra così una tecnica statistica estremamente efficace, in quanto permette di cogliere in modo sintetico gli aspetti fondamentali che caratterizzano le prestazioni dei tennisti professionisti senza alterare in alcun modo la caratterizzazione e la composizione dei raggruppamenti prodotti basandosi su un maggior numero di informazioni. Inoltre, come riportato in Figura 4.10, una riduzione del numero di componenti adottate porta ad un aumento della Silhouette media e ad una diminuzione del numero di unità erroneamente classificate.

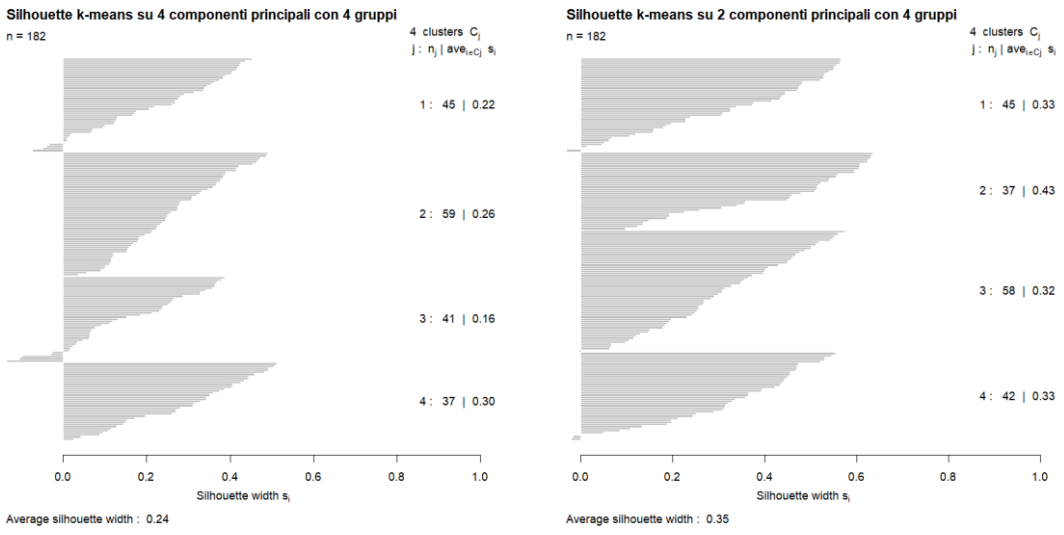


Figura 4.10: Silhouette per il raggruppamento in 4 cluster prodotto dall'algoritmo delle k-means rispettivamente sulle prime 4 e sulle prime 2 componenti principali.

4.4 Confronto delle performance dei gruppi nelle diverse superfici

I dataset contenenti le variabili suddivise per superficie di gioco possono infine permettere il confronto delle statistiche di performance medie di uno stesso gruppo nelle diverse superfici e dei vari gruppi a parità di superficie.

Come prevedibile, le prestazioni su erba e cemento dei giocatori forti al servizio si rivelano migliori rispetto a quelle sulla terra, dove tali atleti riescono però a vincere un numero maggiore di game alla risposta; al contrario, gli esperti della terra battuta evidenziano un peggioramento rilevante delle performance su erba rispetto a quelle sulle altre due superfici, in cui mostrano invece un comportamento molto simile. Per i giocatori mediocri si riscontra una maggior competitività sul cemento mentre gli all-courter si dimostrano i giocatori più versatili e continui, con prestazioni comparabili su tutte i diversi campi di gioco.

Soffermandosi ora sulle singole superfici, i test statistici permettono di verificare se le prestazioni dei gruppi possono essere ritenute equivalenti oppure significativamente differenti tra loro: nei pochi casi in cui si accetta l'ipotesi di normalità delle variabili viene eseguito il t-test a due campioni, mentre negli altri casi viene adoperato, come già fatto in precedenza, il test non parametrico di Wilcoxon per il confronto delle mediane (in tutti i casi il livello di significatività viene fissato al 5%).

Sull'erba risaltano subito valori medi molto simili tra gli all-courter e i battitori: dal punto di vista statistico i due gruppi si rivelano equivalenti per il valor medio di molte variabili relative alla battuta (percentuale di prime di servizio in campo, di break point salvati e concessi e di punti totali vinti), mentre gli atleti appartenenti al primo dei due cluster presentano valori significativamente superiori nei punti e nei game vinti alla risposta.

Sulla terra battuta si registrano prestazioni medie alla risposta statisticamente uguali tra gli specialisti di questa superficie e gli all-courter, ma questi ultimi sovraperformano i rivali nei più importanti indicatori relativi alla battuta (percentuali di punti vinti con la prima di servizio, di break point salvati e di punti e game vinti al servizio).

Focalizzandosi, infine, sul cemento si notano due coppie di cluster con livelli prestazionali medi molto simili: all-courter e battitori da un lato, giocatori mediocri e specialisti della terra battuta dall'altro. Dal confronto tra all-courter e battitori emergono performance significativamente superiori rispettivamente per i primi nelle statistiche relative alla risposta e per i secondi nelle percentuali di punti vinti con la prima di servizio e di game vinti alla battuta, mentre i due gruppi si equivalgono per numero di prime di servizio in campo e di punti vinti servendo la seconda. I giocatori mediocri e gli specialisti della terra battuta presentano invece valori medi non significativamente differenti per le percentuali di prime di servizio in campo, di punti vinti servendo una seconda battuta, di break point salvati e di punti e game vinti al servizio, mentre in risposta i secondi risultano superiori.
Queste comparazioni incrociate delineano in modo netto la supremazia generale degli all-courter, che presentano almeno una parte delle statistiche di performance migliori in ogni superficie, e la buona competitività dei battitori, che risultano sicuramente agevolati dal fatto che la maggior parte dei tornei viene disputata sui campi veloci (erba e cemento), ma che nel complesso si rivelano leggermente inferiori ai migliori giocatori individuali soltanto in risposta.

Conclusioni

Con questa relazione si sono volute presentare diverse tipologie di analisi di raggruppamento relative ai tennisti professionisti, partendo da quanto presente in letteratura e cercando poi di sviluppare lo studio ad un livello successivo.
Dal punto di vista statistico non vi sono state particolari difficoltà nell'applicazione dei metodi conosciuti; gli aspetti più delicati sono stati l'identificazione degli outlier che avrebbero influenzato negativamente il processo di raggruppamento, l'individuazione degli algoritmi più adatti nelle diverse situazioni e l'interpretazione dei risultati, vista soprattutto la mancanza di una classificazione "naturale" o già esistente dei soggetti con cui confrontare quanto ottenuto.
In particolare, con la divisione in gruppi per caratteristiche fisiche e di gioco si è voluto soltanto individuare quale variabile, tra quelle considerate, fosse maggiormente influente per la formazione dei cluster e non si sono volute esaminare le differenze di prestazioni medie tra i vari gruppi, in quanto non ritenuto un elemento così rilevante vista l'eterogeneità presente all'interno dei cluster così formati.
L'attenzione del lavoro si è quindi rivolta all'analisi delle statistiche di performance, sulle quali sono stati ottenuti i risultati più rilevanti. In primo luogo è stato mostrato come un aumento del numero di cluster porti ad una segmentazione sempre più dettagliata dei gruppi preesistenti per livello di prestazione, mentre in seguito all'analisi delle componenti principali si è capito che tutte le informazioni relative alle prestazioni dei tennisti possono essere riassunte in modo efficace con due soli indicatori, uno relativo alle performance al servizio e uno relativo alle performance complessive, ottenibili mediante opportune combinazioni lineari delle variabili raccolte.
Un ultimo importante risultato evidenzia il ruolo del servizio nel tennis moderno: su tutte le superfici, infatti, i grandi battitori sovraperformano i giocatori più abili in risposta e presentano addirittura un livello complessivo molto vicino a quello degli all-courter, mostrandosi inferiori soltanto in ribattuta.

Un ulteriore interessante sviluppo dell'analisi potrebbe essere costituito dal raggruppamento dei tennisti per stile di gioco: disponendo di informazioni aggiuntive (quali il numero di vincenti, di errori non forzati e di discese a rete, il rapporto tra dritti e rovesci giocatati, le rotazioni date alla pallina, la distanza media dalla linea di fondocampo in risposta e durante lo scambio, la distanza media percorsa in un punto e in un incontro) si potrebbero dividere in modo oggettivo gli atleti a seconda della tipologia di gioco in gruppi come attaccante da fondo campo, contrattaccante da fondo campo, giocatore d'attacco, giocatore completo a tutto campo, giocatore serve&volley. Infine, dal confronto incrociato tra questo raggruppamento e quello per statistiche di performance sarebbe interessante individuare gli stili di gioco al giorno d'oggi più efficaci su ogni superficie e quelli prevalenti tra i più forti giocatori individuali.

Appendice A

Lo scopo di questa appendice è quello di riportare il raggruppamento dei tennisti in 4 cluster prodotto dal metodo di partizione delle k-means (a sinistra del cognome di ogni giocatore è indicata la posizione nel ranking Atp del 20/03/2023.

ALL-COURTER		BATTITORI		SPECIALISTI TERRA BATTUTA		GIOCATORI MEDIOCRI	
1	ALCARAZ	6	ALIASSIME	25	FOKINA	63	CACHIN
2	DJOKOVIC	9	HURKACZ	31	CERUNDOLO	65	BARRERE
3	TSITSIPAS	10	FRITZ	33	BAEZ	73	ETCHEVERRY
4	RUUD	23	BERRETTINI	34	NISHIOKA	90	MMOH
5	MEDVEDEV	24	KYRGIOS	38	SCHWARTZMAN	96	GOMEZ
7	RUBLEV	30	SHAPOVALOV	41	ZAPATA-MIRALLES	97	DANIEL
8	RUNE	36	GRIEKSPOOR	51	RAMOS-VINOLAS	100	BAGNIS
11	SINNER	37	CRESSY	53	YMER M.	101	SHEVCHENKO
12	NORRIE	39	SHELTON	56	MOLCAN	108	VAN ASSCHE
13	NADAL	44	LEHECKA	60	LESTIENNE	114	ALTMAIER
14	TIAFOE	45	NAKASHIMA	61	BROOKSBY	115	PASSARO
15	ZVEREV	46	ISNER	66	MUNAR	116	ARNALDI
16	KHACHANOV	47	HUESLER	67	CORIA	117	KOTOV
17	CARRENO BUSTA	48	BUBLIK	69	MOUTET	122	KUDLA
18	DE MINAUR	57	JARRY	82	GARIN	124	ZEPPIERI
19	PAUL	59	SONEGO	84	CARBALLES BAENA	125	HIJIKATA
20	CORIC	68	BORGER	89	GALAN	126	RIEDI
21	MUSETTI	72	RINDERKNECH	91	FOGNINI	130	NARDI
22	CILIC	78	HUMBERT	98	GASTON	131	RODIONOV
26	KORDA	79	HALYS	105	CERUNDOLO M.	135	BERGS
27	DIMITROV	81	MONTEIRO	110	ALBOT	140	TSENG
28	BAUTISTA AGUT	86	O'CONNEL	111	DELLIEN	142	NOVAK
29	EVANS	92	ZHANG	121	MARTINEZ	146	GRENIER
32	VAN DE ZANDSHCULP	93	POPYRIN	129	MULLER	149	YMER E.
35	KECMANOVIC	94	KOKKINAKIS	132	BASILASHVILI	153	KOVALIK
40	GASQUET	102	KOVACEVIC	143	MACHAC	159	UCHIDA
42	GOFFIN	103	STRUFF	150	MISOLIC	160	WU
43	DRAPER	112	DUCKWORTH	156	SOUSA	161	BROADY
49	BONZI	113	GOJO	179	DELBONIS	164	ALVES
52	MURRAY	118	STRICKER	190	SIMON	181	SHANG
54	RUUSUVORI	119	EUBANKS	195	DZUMHUR	182	NAVA
55	MCDONALD	138	OPELKA	196	OLIVO	192	MEDJEDOVIC
58	DJERE	144	MARTERER	198	KUZMANOV	193	HOLT
62	MANNARINO	145	VAN RIJTHOVEN	221	GUINARD	202	VAVASSORI
64	WU	152	JOHNSON	262	KOEPFER	203	LAAKSONEN
70	KUBLER	158	POSPISIL	275	KUKUSHKIN	208	CELIKBILEK
71	GIRON	168	VUKIC	299	VATUTIN	215	KRUEGER
74	FUCSOVICS	175	BELLIER			223	SVAJDA
75	KWON	178	HARRIS			231	ILKEL
76	LAJOVIC	239	SANDGREN			232	TABERNER
77	KRAJINOVIC	249	SERDARUSIC			285	LAZAROV
80	IVASHKA	278	ROSOL				
83	THOMPSON	287	MAGER				
85	OTTE						
87	WAWRINKA						
104	SAFIULLIN						
106	THIEM						
107	KARATSEV						
109	FILS						
120	BROUWER						
123	WATANUKI						
147	MAJCHRZAK						
151	PENISTON						
154	SOCK						
162	TABILO						
170	PAIRE						
201	MILLMAN						
211	VERDASCO						
235	BERANKIS						
280	MONFILS						
289	CUEVAS						

Bibliografia e sitografia

Austin D., 2021. Atp tennis cluster analysis. *Towards Data Science.*

Biondi M., 1987. Osservazioni comparative sul comportamento di tre indici di similarità per dati binari*. Biogeographia, The Journal of Integrative Biogeography, 11(1).*

Braunstein A., 2010. Consistency and Pythagoras. *Journal of Quantitative Analysis in Sports 6(1): 1-16.*

Caro C. & Machtmes R., 2013. Testing the Utility of the Pythagorean Expectation Formula on Division One College Football: An Examination and Comparison to the Morey Model. *Journal of Business & Economics Research 11(12): 537-542.*

Cha D.U., Glatt D.P., Sommers P.M., 2007. An empirical Test of Bill James' Pythagorean Formula. *Journal of Recreational Mathematics 35(2): 117-130.*

Cui Y., Gómez M. A., Gonçalves B., Sampaio J., 2019. Clustering tennis players' anthropometric and individual features helps to reveal performance fingerprints. *European Journal of Sport Science.*

Davenport C. & Woolner K., 1999. Revisiting the Pythagorean Theorem: Putting Bill James' Pythagorean Theorem to the Test. *The Baseball Prospectus.*

Gower J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics, 27: 857-871.*

Hamilton H.H., 2011. An extension of the Pythagorean Expectation for Association Footbal. *Journal of Quantitative Analysis in Sports 7(2): 1-18.*

Jaccard P., 1900. Contribution au problème de l'immigration postglaciaire de la flore alpine. *Bull. Soc. vaudoise Sci. nat., 36: 87- 130.*

Jaccard P., 1901. Etude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. vaudoise Sci. nat., 37: 547-579.*

Jaccard P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. vaudoise Sci. nat., 44: 223-270.*

James B, 1981. Baseball Abstract. *Self-published, Lawrence, KS, 1981.*

Johnson R. & Wichern D. (2014). Applied Multivariate Statistical Analysis.

Kovalchik S. A., 2016. Is there a Pythagorean theorem for winning in tennis? *Journal of Quantitative Analysis in Sports*.

La Rocca A., 2018. Fuzzy Clustering: la logica, i metodi*. Istat.*

Lewis M., 2003. Moneyball: The art of winning an unfair game. *W.W. Norton & Company.*

Lisi F., Grigoletto M., Canesso T., 2019. Winning tennis matches with fewer points or games than the opponent. *Journal of Sports Analytics.*

Machar R., McMurtrie D., Crespo M., 2017. The relationship between match statistics and top 100 ranking in professional men's tennis. *Tandfonline.*

MacQueen, J. B., 1967. Some methods for classification and analysis of multivariate observations. *In L. M. Le Cam & J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California: University of California Press.*

Paderi F., 2017. Basic of Clustering.

Pegoraro E., 2019. Statistica per Data Science con R (V. 03).

Rogers D.J. & Tanimoto T.T., 1960. A computer program for classifying plants. *Science (Wash. D.C.), 132: 1115-1118.*

Sokal R.R. & Michener C.D., 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull., 38: 1409-1438.*

Sokal R.R. & Sneath P.H.A., 1963. Principles of numerical taxonomy. *W.H. Freeman, San Francisco, xvi+359 pp.*

Sørensen T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biol. Skr., 5: 1-34.*

Vollmayr-Lee B., 2002. More Then You Probably ever Wanted to Know about the “Pythagorean” Method.

https://www.atptour.com

https://github.com/JeffSackmann

https://www.ultimatetennisstatistics.com