Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Clustering di squadre e giocatori NBA per la previsione del net rating dei quintetti: un approccio basato su modelli mistura e random forest

Relatore: Prof. Antonio Canale
Dipartimento di Scienze Statistiche

Laureando: Nicola Artuso
Matricola: 2005387

Anno Accademico 2022/2023

Indice

Introduzione

"With the first pick in 2023 NBA draft, the San Antonio Spurs select: Victor Wembanyama from Metropolitans92, France". Così Adam Silver, commissioner della NBA, il 22 giugno 2023 ha annunciato l'entrata nella National Basketball Association di uno dei giocatori con più aspettative di sempre. Wembanyama, infatti, è il prototipo di giocatore che qualunque squadra vorrebbe: altissimo (2.22 m), leve sostanzialmente infinite (apertura alare di 2.40 m) e, soprattutto, in possesso di un bagaglio tecnico e di movenze che solitamente appartengono a giocatori molto più piccoli e rapidi di lui. L'entrata in scena di questo giocatore non è che l'ultimo di una serie di passi che stanno cambiando la pallacanestro per come era conosciuta fino a quindici anni fa. Se infatti, fino all'inizio degli anni 2000, le posizioni e i ruoli in campo sono sempre stati piuttosto definiti in base a specifiche caratteristiche fisiche e tecniche (utilizzando per esempio i giocatori più piccoli lontano da canestro e i cosiddetti "lunghi", ovvero i giocatori più alti, vicino a quest'ultimo), nell'ultima decade le cose sono cambiate grazie a una serie di fattori. Tra questi è sicuramente presente la maggior attenzione data da General Manager e Staff ai dati. Sotto questo aspetto il libro «*Basketball On Paper*» di Oliver, 2004, ha segnato l'inizio di una vera e propria rivoluzione analitica della pallacanestro. La successiva applicazione di nuove tecnologie, come quella del tracking data a partire dalla stagione 2013-2014, ha permesso di poter avere a disposizione sempre più dati da analizzare per poter creare un vantaggio competitivo. Si aggiunga poi la ricerca e l'arrivo di giocatori capaci di ricoprire più ruoli ed ecco che la classica suddivisione delle posizioni di playmaker, guardia tiratrice, ala piccola, ala forte e centro ha cominciato a suonare sempre più desueta. Proprio in quest'ottica, nell'articolo «*NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball and modelling lineup efficiency of soft lineup aggregate*» di Kalman e Bosch, 2020 (rispettivamente studenti di Purdue University e Syracuse Univer-

sity) si è cercato di individuare dei nuovi ruoli che descrivessero meglio la pallacanestro contemporanea. A tal fine, gli autori hanno utilizzato un modello di mistura normale per clusterizzare giocatori con caratteristiche simili. L'utilizzo di tale modello è in primo luogo vantaggioso perché il numero di gruppi non è noto e, in secondo luogo, a differenza di altri metodi di classificazione (ad esempio il K-means), anzichè inserire ciascuna osservazione in un determinato gruppo, determina la probabilità di queste di far parte di un certo cluster. Tale assegnazione permette di dare una misura delle caratteristiche del giocatore che, a volte, possono risultare ibride e non perfettamente inseribili in un unico gruppo. Una volta quindi eseguito, per ciascun giocatore, quello che viene chiamato 'soft assignment', mediante Random Forest, Kalman e Bosch hanno cercato di prevedere il net-rating (ovvero il differenziale tra punti segnati e punti subiti su 100 possessi) di una qualsiasi combinazione di 5 giocatori (lineup), basandosi sulla composizione della stessa.

La seguente tesi propone una replicazione del lavoro sopra citato, utilizzando però dati più recenti e inserendo una classificazione delle squadre da aggiungere a quella dei giocatori. Il lavoro di Kalman e Bosch, infatti, utilizza dati che partono dalla stagione NBA 2009-2010 fino alla stagione 2017-2018, e i risultati evidenziano dei cluster formati da alcuni tipi di giocatori che, almeno a livello intuitivo, sembrano non esserci più al giorno d'oggi. Per questo i dati presi qui in considerazione vanno dalla stagione NBA 2015-2016 (primo anno con a disposizione tutti i dati derivanti dal tracking data) fino a quella 2022-2023. La classificazione delle squadre invece viene fatta per valutare se possono essere riscontrati diversi stili di gioco e per capire quale tipo di quintetto (lineup) si adatti maggiormente al modo di giocare della squadra.

Il Capitolo 1 è dedicato alla raccolta, al pre-processing dei dati e alla costruzione dei dataset di giocatori, squadre e lineups, con una spiegazione dettagliata di tutte le variabili prese in considerazione. Sono inoltre presenti alcune analisi esplorative che testimoniano l'evoluzione della pallacanestro in questi otto anni.

Il Capitolo 2 presenta invece i metodi e i modelli utilizzati per il clustering e per la previsione del net-rating. In particolare verrà approfondito il modello di mistura gaussiano, il metodo per scelta delle variabili individuali e di squadra e il modello random forest al fine della previsione.

Il Capitolo 3 mostra l'applicazione del modello mistura ai dati relativi ai giocatori e alle squadre. Viene effettuato il clustering dei giocatori e vengono descritte le caratteristiche

dei giocatori che compongono i gruppi ottenuti. Discorso analogo per quanto riguarda il clustering delle squadre. Vengono poi descritti la composizione delle soft-lineups e il concetto di net rating bayesiano

Nel Capitolo 4 si mostra l'applicazione del Random Forest ai dati, vengono quindi analizzati i risultati e ne vengono tratte le conclusioni.

Capitolo 1

Dati NBA 2015-2023

1.1 Raccolta dati e pre-processing

Per raccogliere i dati necessari ai fini dell'analisi è stato implementato uno script in Python utilizzando alcune delle API disponibili dal sito NBA.com/stats. Tramite R poi, è stato creato il dataset dei giocatori aggregando le statistiche individuali basate sul nome del giocatore e sulla stagione di riferimento. Discorso analogo per quanto riguarda il dataset delle squadre, con le statistiche aggregate sulla base del nome della squadra e della stagione di riferimento. Infine, è stato creato il dataset riguardante le lineups, in cui ciascuna osservazione è rappresentata dalla combinazione dei 5 giocatori che compongono il quintetto.

1.2 Variabili individuali

Nel dataset dei giocatori, ciascuna riga corrisponde alle statistiche di un giocatore in una data stagione. Come per il lavoro di Kalman e Bosch, ciò ha permesso di identificare l'evoluzione di ciascun giocatore nel corso degli anni. Nello specifico, le variabili considerate sono state:

- GP: Partite giocate durante la stagione.
- HEIGHT: Altezza del giocatore misurata in pollici.

- OREB PCT: Percentuale di rimbalzi offensivi catturati mentre il giocatore è in campo.
- DREB PCT: Percentuale di rimbalzi difensivi catturati mentre il giocatore è in campo.
- AST PCT: Percentuale dei canestri dei compagni che il giocatore ha assistito.
- STL PCT: Palle rubate su 100 possessi avversari .
- BLOCK PCT: Stoppate date su 100 tiri avversari tentati.
- TOV PCT: Palle perse su 100 possessi .
- USG PCT: Percentuale di possessi di squadra utilizzati dal giocatore.
- PIE: Stima della percentuale di eventi nella partita in cui il giocatore ha contribuito.
- FT RATE: Tiri liberi segnati per tiro dal campo tentato.
- FT PCT: Percentuale di tiri liberi segnati.
- FGA: Tiri dal campo tentati su 100 possessi.
- TWO FG PCT: Percentuale tiri da due punti.
- THREE FG PCT: Percentuale tiri da tre punti.
- TWO FG AST PCT: Percentuale tiri da 2 punti assistiti.
- THREE FGA PCT: Percentuale di tiri tentati da tre punti.
- CORNER3: Percentuale di tiri tentati da tre punti dall'angolo.
- THREE FG AST PCT: Percentuale di tiri da tre assistiti.
- RA: Percentuale di tiri tentati dalla Restricted Area.
- ITP: Percentuale di tiri tentati nel pitturato.
- MID: Percentuale di tiri tentati dalla media distanza.

Il dataset finale si compone quindi di 4245 osservazioni in 27 variabili, le 22 sopra elencate e le seguenti variabili di identificazione:

- PLAYER ID : codice che identifica il giocatore.
- PLAYER NAME: nome del giocatore.
- TEAM ABBREVIATION: sigla della squadra di cui fa parte il giocatore.
- AGE: età del giocatore.
- SEASON: stagione di riferimento.

1.2.1 Analisi esplorative

Una volta costruito il dataset, andiamo però a valutare quanti dei 4245 giocatori presi in considerazione possono effittivamente essere validi per la clusterizzazione. In particolare, guardiamo la distribuzione delle partite giocate (Figura 1.1) e ammettiamo soltanto i giocatori che hanno disputato almeno 30 partite in una data stagione, che risultano essere 3100. Poichè gran parte delle variabili considerate sono su una scala di 100 possessi, la scelta di un minimo di partite giocate è funzionale a fare in modo che la stima delle abilità dei giocatori sia sufficientemente robusta.

Poichè, storicamente, i ruoli nella pallacanestro sono sempre andati di pari passo con l'altezza dei giocatori, visualizziamo se alcune delle variabili prese in considerazione possano essere collegate con quest'ultima. In prima battuta andiamo a vedere come si distribuiscono le altezze dei giocatori nella Figura 1.2.

Vista la bassa frequenza di alcune delle classi di altezza prese in considerazione, decidiamo di aggregare tutti i giocatori al di sotto dei 74 pollici (1.88 m) in un unico gruppo. Analogamente aggreghiamo i giocatori sopra gli 84 pollici (2.13 m).

Con quest'ultima classificazione dell'altezza, vediamo i boxplot di alcune variabili sulla base di quest'ultima:

- Punti per 100 possessi:

  Al top dei punti segnati per 100 possessi (Figura 1.3) ci sono, senza sorprese, alcuni dei migliori giocatori NBA. Per quanto riguarda l'andamento della media, essa sembra essere sostanzialmente costante al variare dell'altezza.

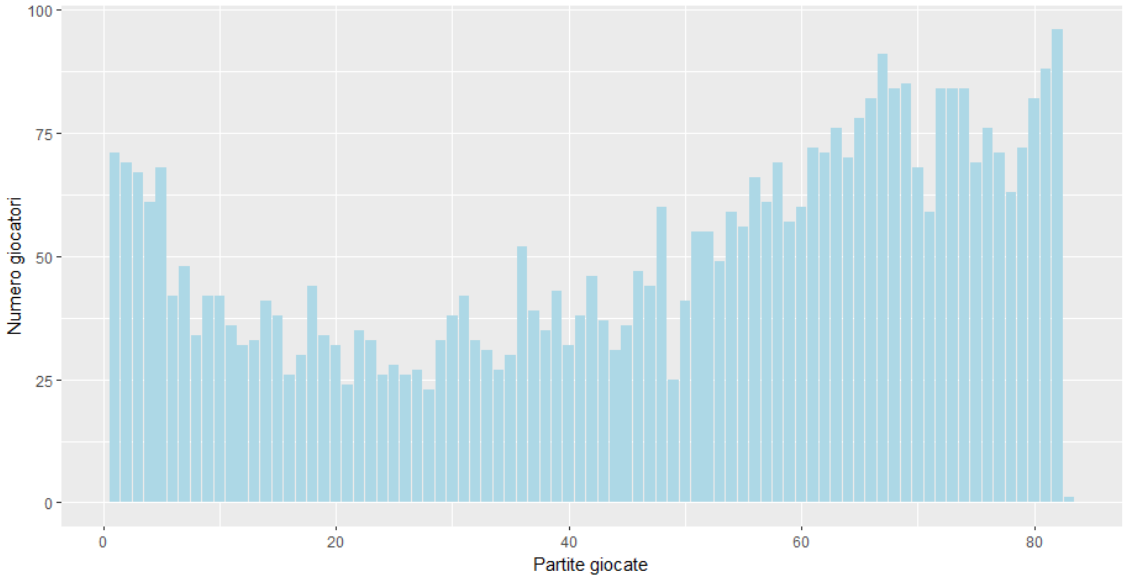

Figura 1.1: istogramma delle partite giocate dai giocatori NBA tra il 2015/2016 e il 2022/2023.

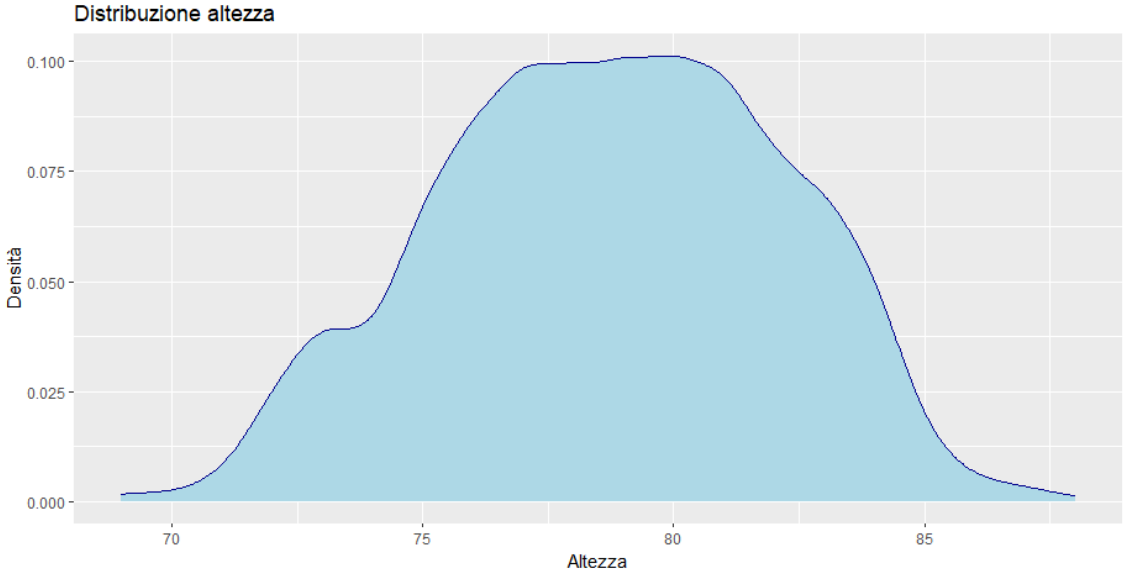

Figura 1.2: Distribuzione dell'altezza dei giocatori NBA tra il 2015/2016 e il 2022/2023.

- Assist percentage:

  Da sempre gli assist sono considerati il pane per i playmaker, che hanno il compito di gestire la squadra e che, nella visione più classica della pallacanestro, sono i giocatori più bassi. Tale visione sembra essere in parte confermata dalla Figura 1.4, con un andamento decrescente della media. Tuttavia i migliori passatori della lega sono tutti sopra il 30 percento di assist percentage, e comprendono giocatori come Nikola Jokic e Domantas Sabonis che invece sono considerati dei 'centri'. Da notare la presenza di LeBron James in due classi di altezza diverse, a testimoniare che le misurazioni dell'altezza, sebbene fatte su giocatori adulti e già formati, alle volte può variare di 1 o 2 pollici tra una stagione e un'altra.

- Rebound percentage:

  I rimbalzi si confermano essere una caratteristica dei giocatori più alti (Figura 1.5): la media della percentuale di rimbalzi catturati (somma della percentuale di rimbalzi offensivi e difensivi) si alza con il crescere dell'altezza. Ciò non stupisce dato che si tratta di un fondamentale (sia difensivo che offensivo) che richiede soprattutto di essere alti per poter avere piu chance di recuperare il pallone dopo un canestro sbagliato.

- Steal percentage:

  Anche le palle rubate sono sempre state una caratteristica dei giocatori più bassi. In generale però, nonostante la media di steal percentage si abbassi al crescere dell'altezza (Figura 1.6), ciò che si nota di più sono i top player in questa categoria, che si distinguono per valori molto sopra la media, come Jose Alvarado e soprattutto Paul Reed.

- Percentuale di tiri tentati da 3 punti:

  I tiri da 3 sono stati per molto tempo un'esclusiva delle 'guardie tiratrici', giocatori più alti dei playmaker ma non esattamente altissimi. In effetti la media della percentuale dei tiri da tre punti effettuati cala a partire dai 78 pollici, ma comunque si rileva un gran numero di giocatori anche molto alti che prende più del 50 percento dei propri tiri da oltre l'arco (Figura 1.7), a conferma ancora una volta della sempre meno netta distinzione dei ruoli.

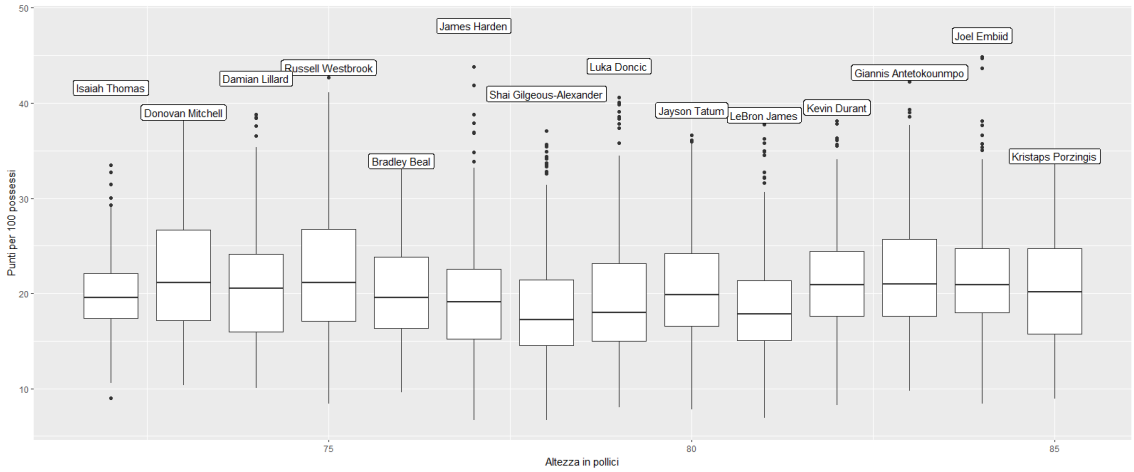

Figura 1.3: Boxplot e migliori giocatori per punti per 100 possessi in base all'altezza.

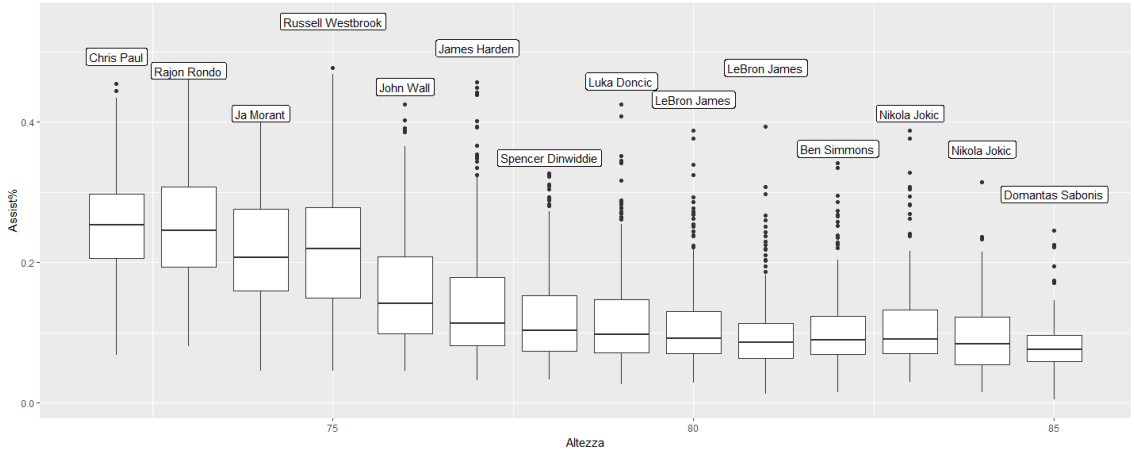

Figura 1.4: Boxplot e migliori giocatori per assist percentage in base all' altezza.

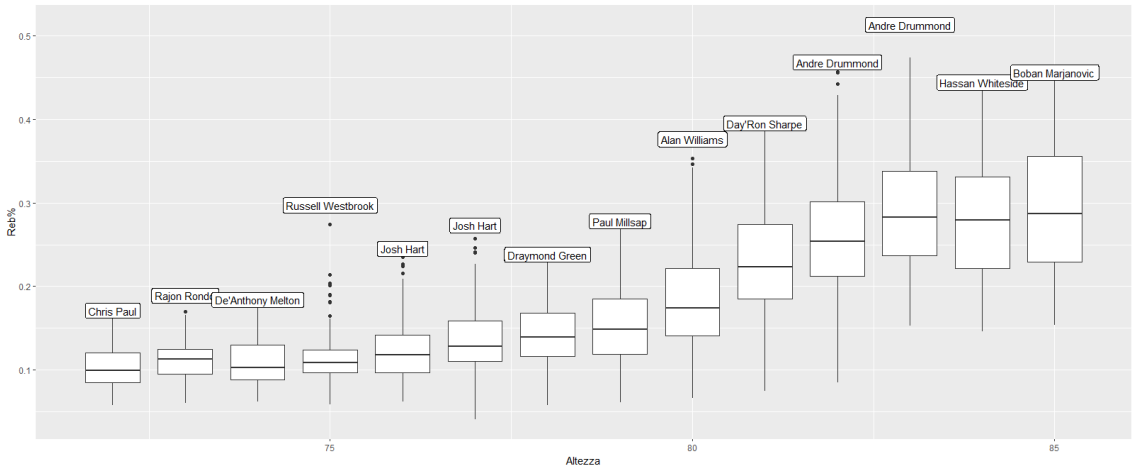

Figura 1.5: Boxplot e migliori giocatori per rebound percentage in base all' altezza.

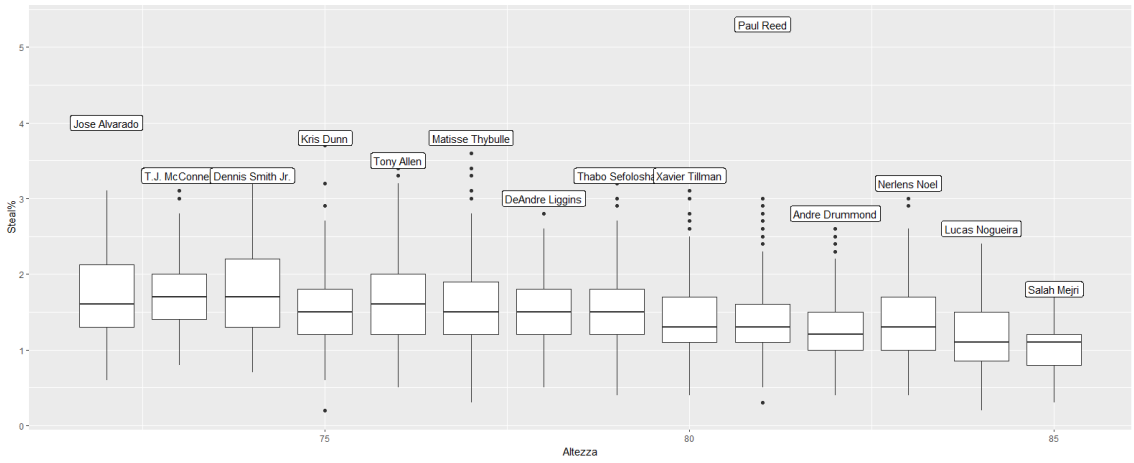

Figura 1.6: Boxplot e migliori giocatori per palle rubate su 100 possessi in base all'altezza.

- Usage percentage e PIE:

  Usage percentage e PIE sono indicatori dell'importanza di un giocatore all'interno della squadra. La media dello usage percentage è più o meno stabile intorno allo 0.2 (come prevedibile dato che essendo 5 i giocatori in campo ci aspettiamo in media un pari contibuto da tutti), ma non ha importanti variazioni sulla base dell'altezza e quindi del ruolo (Figura 1.8). Analogo il discorso per quanto riguarda il PIE (Figura 1.9), dove si registra una leggera crescita per quanto riguarda i giocatori più alti, probabilmente avvantaggiati dal numero di rimbalzi presi, come visto in precedenza.

Le analisi esplorative confermano dunque che i ruoli sono sempre meno chiari basandosi sull'altezza. Nel Capitolo 2 vedremo in base a quali nuove caratteristiche decideremo di classificare i giocatori.

1.3 Variabili di squadra

Per la selezione delle variabili di squadra, è stato scelto di includere, più che variabili relative all'efficienza della squadra (come ad esempio le percentuali realizzative), variabili che indicassero maggiormente il modo di attaccare e di difendere del team. Questa scelta è in linea con l'idea di clusterizzare le squadre per stile di gioco, permettendo di fatto che in uno stesso cluster ci possano essere sia squadre il cui record di vittorie (indice del livello della squadra stessa) è molto alto, sia squadre che invece fanno fatica a raggiungere i playoff. Possiamo dividere le variabili selezionate in alcune categorie: tipi di punti, tiri concessi in difesa, tiri presi in attacco, tipo di azione, statistiche tradizionali, statistiche avanzate.

- Tipi di punti (ogni dato è una media dei punti a partita):
  - POINTS: punti segnati.
  - DRIVE PTS: punti segnati da azioni di 'penetrazione'.
  - CATCH SHOOT PTS: punti segnati da azioni di 'ricevi e tira'.
  - PULL UP PTS: punti segnati da azioni di palleggio arresto e tiro.
  - PAINT TOUCH PTS: punti segnati da azioni 'nel pitturato'.

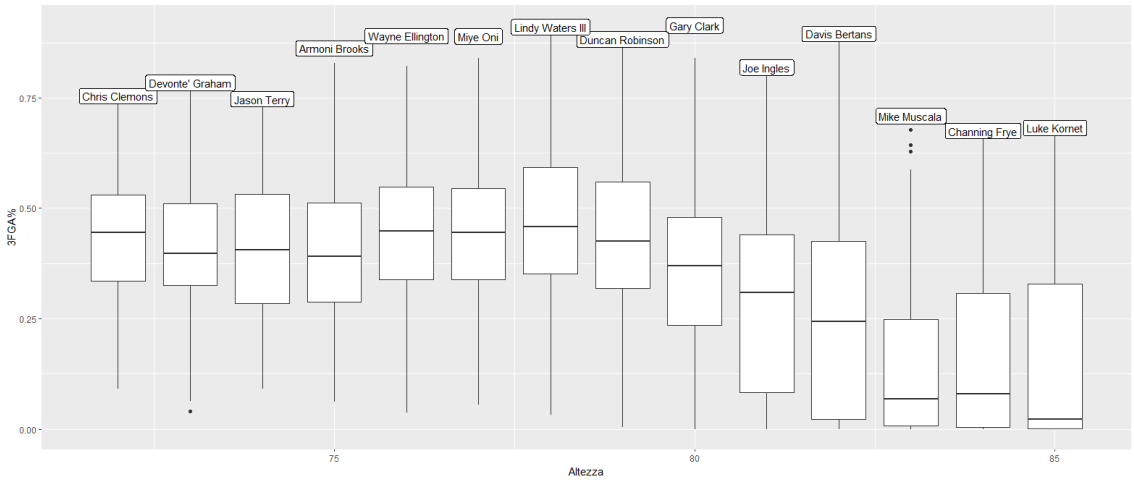

Figura 1.7: Boxplot e migliori giocatori per percentuale di tiri da 3 tentati in base all'altezza.

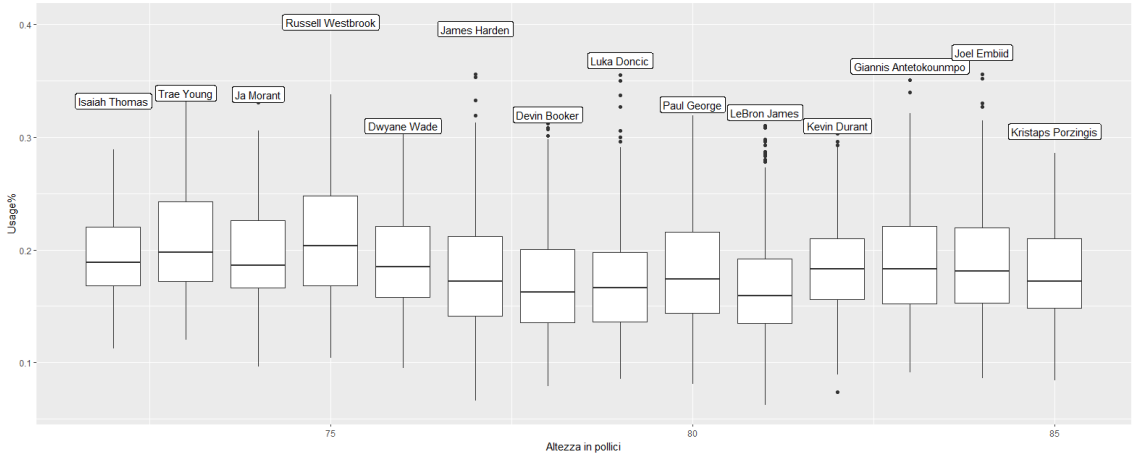

Figura 1.8: Boxplot e migliori giocatori per usage percentage in base all'altezza.

  - POST TOUCH PTS: punti segnati a partire da un azione di attacco 'in post'.

- Tiri presi in attacco (ogni dato è la media a partita del numero di tiri tentati da una certa zona del campo) :

  - Restricted Area: tiri tentati in 'Restricted Area'.
  - In The Paint Non RA: tiri tentati 'nel pitturato', esclusa la zona di 'Restricted Area'.
  - Mid Range: tiri tentati dalla media.
  - Left Corner 3: tiri da 3 tentati dall'angolo sinistro.
  - Right Corner 3: tiri da 3 tentati dall'angolo destro.
  - Above the break 3: tiri da 3 tentati (esclusi gli angoli).
  - Corner 3: tiri da 3 tentati dagli angoli

- Tiri concessi in difesa (ogni dato è la media a partita del numero di tiri tentati dalla squadra avversaria da una certa zona del campo) :

  - OPP RA: tiri concessi in 'Restricted Area'.
  - OPP ITP: tiri concessi 'nel pitturato', esclusa la zona di 'Restricted Area'.
  - OPP Mid: tiri concessi dalla media.
  - OPP LC3: tiri da 3 concessi dall'angolo sinistro.
  - OPP RC3: tiri da 3 concessi dall'angolo destro.
  - OPP AB3: tiri da 3 concessi (esclusi gli angoli).
  - OPP C3: tiri da 3 concessi dagli angoli

- Tipo di azione (ogni dato è proporzione del numero di azioni concluse con un canestro utilizzando un certo tipo di giocata, sul totale delle azioni concluse con un canestro):

  - Isolation: proporzione di azioni di 'isolamento'.
  - Transition: proporzione di azioni in contropiede.
  - Cut: proporzione di azioni 'di taglio'.

  - PRBallHandler: proporzione di azioni di Pick and Roll concluse con un canestro del portatore di palla.
  - PRRollMan: proporzione di azioni di Pick and Roll concluse con un canestro del bloccante.
  - Postup: proporzione di azioni in Post.
  - Spotup: proporzione di azioni di tiro piazzato.
  - Handoff: proporzione di azioni di passaggio consegnato.
  - OffScreen: proporzione di azioni di uscita dai blocchi.
  - OffRebound: proporzione di azioni da rimbalzo offensivo.
  - Misc: proporzione di azioni non elencate tra le precedenti.

- Statistiche tradizionali (ogni dato è inteso su 100 possessi offensivi)
  - FGA: tiri dal campo tentati.
  - FG3A: tiri da 3 tentati.
  - FTA: tiri liberi tentati.
  - OREB: rimbalzi offensivi.
  - DREB: rimbalzi difensivi.
  - REB: rimbalzi totali.
  - AST: assist.
  - TOV: palle perse.
  - STL: palle rubate.
  - BLK: stoppate date.
  - BLKA: stoppate subite.
  - PF: falli commessi.
  - PFD: falli subiti.

- Statistiche Avanzate:

  - NET RATING: differenziale di punti su 100 possessi (nota: questa variabile non verrà utilizzata per la clusterizzazione).
  - AST PCT: proporzione di canestri che sono stati assistiti.
  - OREB PCT: proporzione di rimbalzi offensivi catturati rispetto a quelli disponibili.
  - DREB PCT: proporzione di rimbalzi difensivi catturati rispetto a quelli disponibili.
  - TM TOV PCT: proprozione di azioni conclusasi con una palla persa.
  - PACE: numero di possessi offensivi a partita.

Oltre a queste variabili, sono presenti anche, al fine dell'identificazione delle osservazioni:

- id: valore da 1 a 240 che identifica l'osservazione.
- SEASON: stagione di riferimento.
- TEAM NAME: nome della squadra.
- TEAM ABBREVIATION: sigla della squadra.

Il dataset finale si compone quindi di 240 osservazioni (30 squadre per ciascuna delle 8 stagioni prese in considerazione) e 54 variabili.

Si noti che non tutte le variabili sopracitate verranno usate ai fini della clusterizzazione ma, in base ai metodi descritti nel Capitolo 3, si sceglieranno le variabili maggiormente adatte a descrivere la differenza tra più gruppi.

1.3.1 Analisi esplorative

Come spiegato nel Capitolo 1, la pallacanestro è cambiata molto negli ultimi anni. In particolare, l'aspetto su cui maggiormente ci si è focalizzati è stata l'importanza di tirare con la maggior efficienza possibile. Nella fattispecie si è capito che i tiri maggiormente efficienti sono di tre tipi: i tiri nella restricted area (maggiore la vicinanza al canestro, maggiore la percentuale di realizzazione), i tiri liberi (che, per definizione, sono privi della difesa e, oltre a questo, aumentano il numero di falli avversari) e i tiri da tre punti (valgono il 50 percento in più dei tiri da 2, quindi è chiaro che, a parità di percentuale, i punti per

tiro prodotti sono maggiori). Allo stesso tempo i cosiddetti tiri dal mid-range o i long-two (tiri da due appena dentro la linea da 3 punti), così frequenti negli anni novanta e duemila, sono stati automaticamente considerati brutti tiri perchè non particolarmente efficienti. Le squadre hanno quindi cominciato a tirare poco dal mid-range e molto più spesso da 3 e da vicino a canestro. Altro aspetto evidenziato negli ultimi anni è stato l'aumento del ritmo delle partite, rappresentato dal PACE, ovvero dal numero di possessi in una partita. Più tiri significa più possibilità di fare canestro, ed ecco perchè negli ultimi anni il ritmo sembra essere aumentato. Verifichiamo quindi queste affermazioni attraverso delle analisi grafiche.

A partire dal tiro da tre punti (Figura 1.10), è chiaro come il numero tiri a partita presi da oltre l'arco sia cresciuto ogni anno, per poi stabilizzarsi intorno ai 34/35 tiri a partita. Da notare il gran volume di tiri da tre punti presi dagli Houston Rocket tra il 2016/2017 e il 2019/2020. La presenza di questa squadra non è soprendente: il cosidetto "Morey ball", gioco di parole tra il titolo del film "Moneyball" e il general manager di Houston Daryl Morey, è l'emblema della rivoluzione analitica che si è abbattuta sulla NBA. Non per niente i Rockets del 2017-2018 è stata l prima squadra nella storia della NBA a tentare più tiri da tre che da due (Goldsberry, 2019).

Se guardiamo infatti i tiri dalla media (Figura 1.11), Houston risulta ultima in 6 delle 8 stagioni considerate. Inoltre, coerentemente con quanto detto, la media dei tiri presi da quella distanza è in completa picchiata ed è sostanzialmente dimezzata rispetto a 8 anni fa. Non è un caso infatti che i Chicago Bulls, primi in questa classifica per quanto riguarda le ultime due stagioni (posizione dovuta soprattutto alla presenza di due 'mid-range master' come DeMar DeRozan e Zach LaVine), siano invece ultimi per tiri da 3 tentati nelle medesime stagioni.

Andando a vedere l'andamento delle medie dei tiri tentati dalla restricted-area (Figura 1.12) e dei tiri liberi (Figura 1.13), non sembrano esserci particolari trend di crescita o decrescita. Ciò è dovuto probabilmente al fatto che la 'rivoluzione' di cui abbiamo parlato ha colpito soprattutto il bilanciamento tra tiri da 3 e dal mid-range, piuttosto che tiri di comprovata efficacia come quelli vicino a canestro o i tiri liberi. Si consideri anche il fatto che le difese hanno sempre cercato in primo luogo di limitare le incursioni avversarie nella propria area, impedendo di fatto che la media dei tiri presi da quella zona potesse alzarsi di molto.

Il confronto finale tra le medie indicato in Figura 1.14 evidenzia quanto detto: crescita

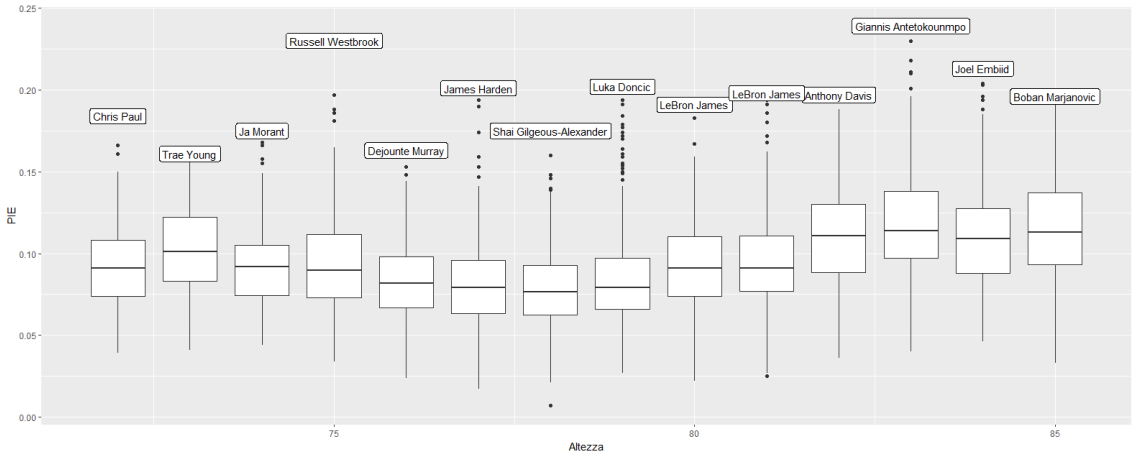

Figura 1.9: Boxplot e migliori giocatori per PIE in base all'altezza.

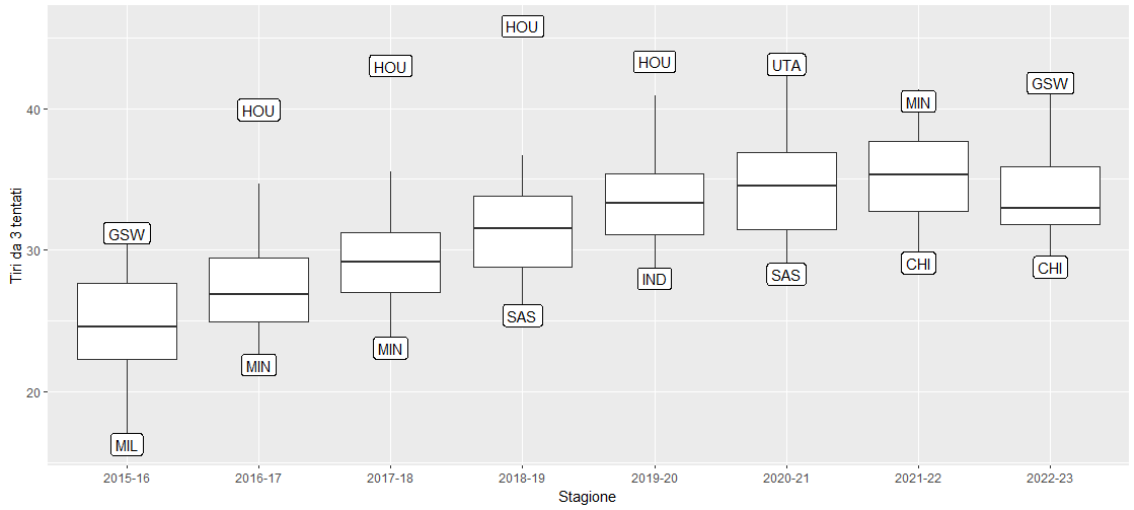

Figura 1.10: Boxplot e squadre con maggiore e minor numero di tiri da tre punti tentati a partita in ogni stagione.

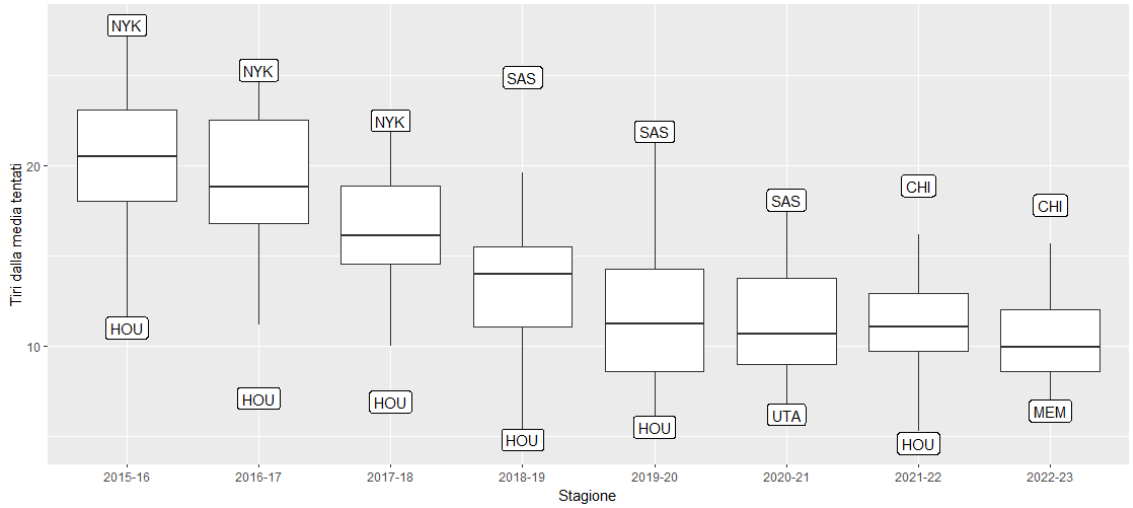

Figura 1.11: Boxplot e squadre con maggiore e minor numero di tiri dalla media tentati a partita in ogni stagione.

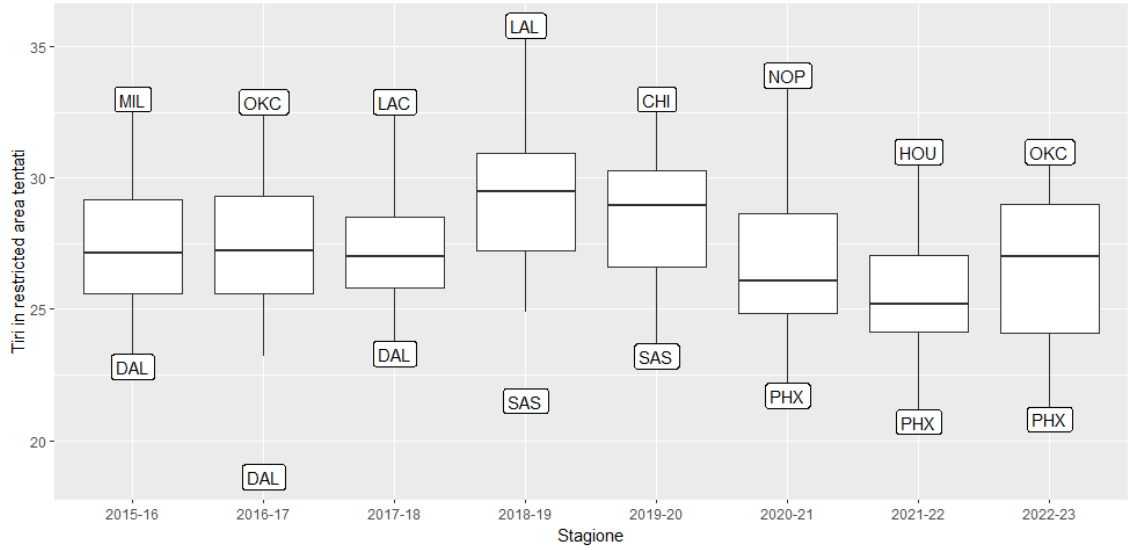

Figura 1.12: Boxplot e squadre con maggiore e minor numero di tiri dalla restricted area tentati a partita in ogni stagione.

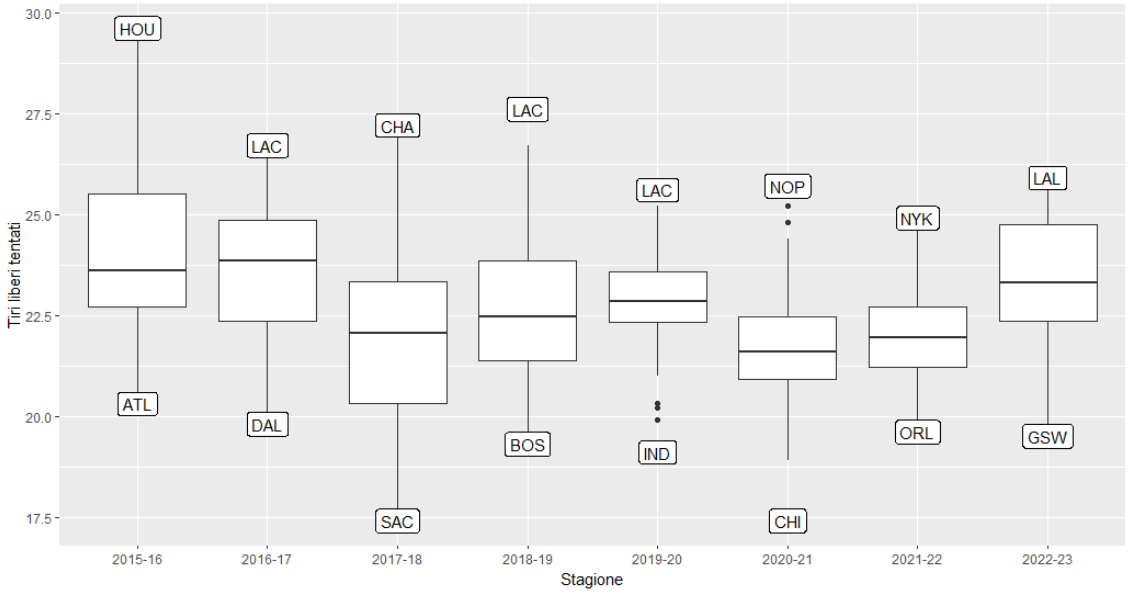

Figura 1.13: Boxplot e squadre con maggiore e minor numero di tiri liberi tentati a partita in ogni stagione.

del tiro da 3 punti, calo dei tiri dal mid-range e una certa stabilità nel tiro dalla restricted area

Guardando invece la media del PACE (Figura 1.15), sembra esserci una generale crescita, anche se soggetta a fluttuazioni e comunque non così evidente.

Per quanto riguarda la distribuzione dei tipi di giocata (Figura 1.16) e dei tipi di punti prodotti (Figura 1.17), non sembrano esserci grandi differenze tra le diverse stagioni, se non un aumento dei punti da 'penetrazione' (Drive) a discapito dei punti dal gomito (Elbow) e dal post. Ciò trova conferma nella diminuzione di giocate dal post e dal dominio di azioni di transizione, spotup e pick and roll.

1.4 Dati sulle lineup

Il dataset delle lineup comprende tutti i quintetti scesi in campo durante le 8 stagioni prese in considerazione. Per ogni osservazione abbiamo:

- GROUP ID: codice formato dai 5 codici dei giocatori del quintetto.
- GROUP NAME: il nome di ciascun giocatore che compone il quintetto.

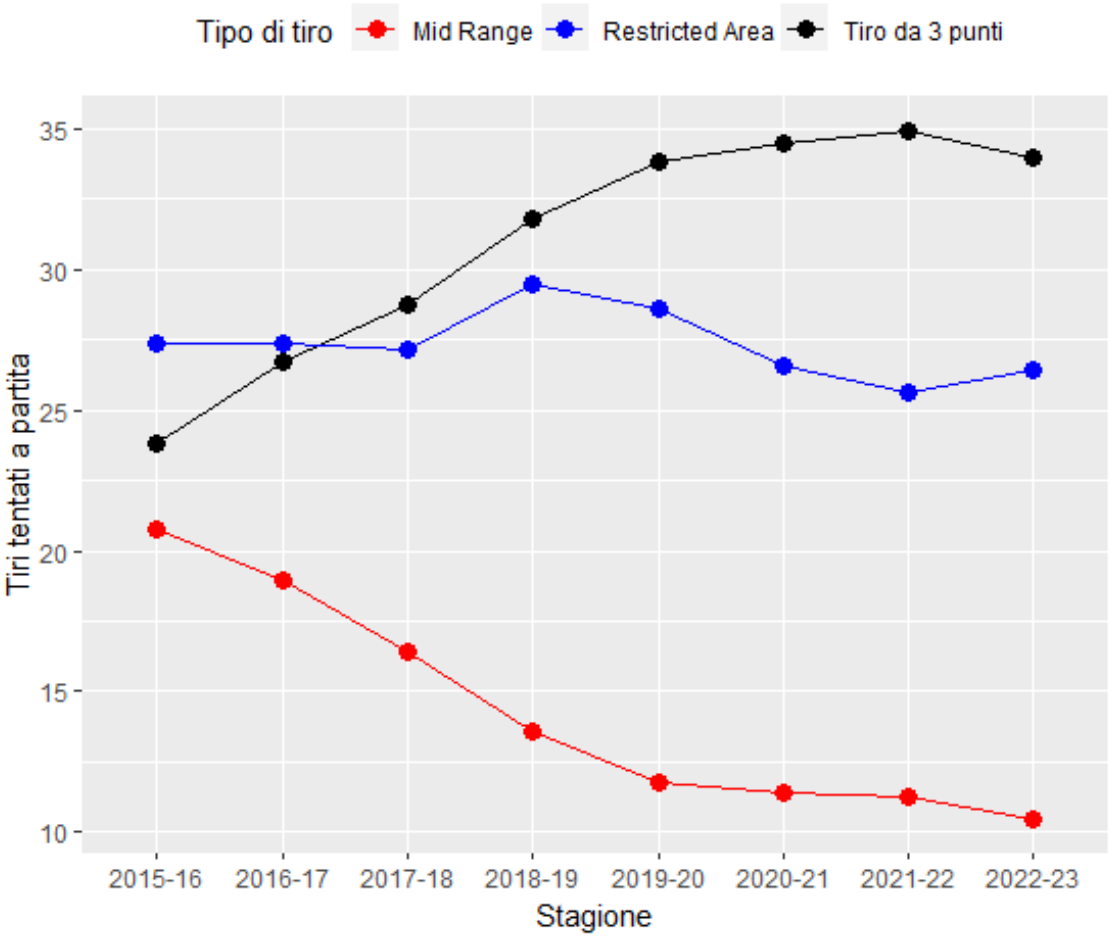

Figura 1.14: Confronto dell'andamento delle medie dei tipi di tiri tentati a stagione.

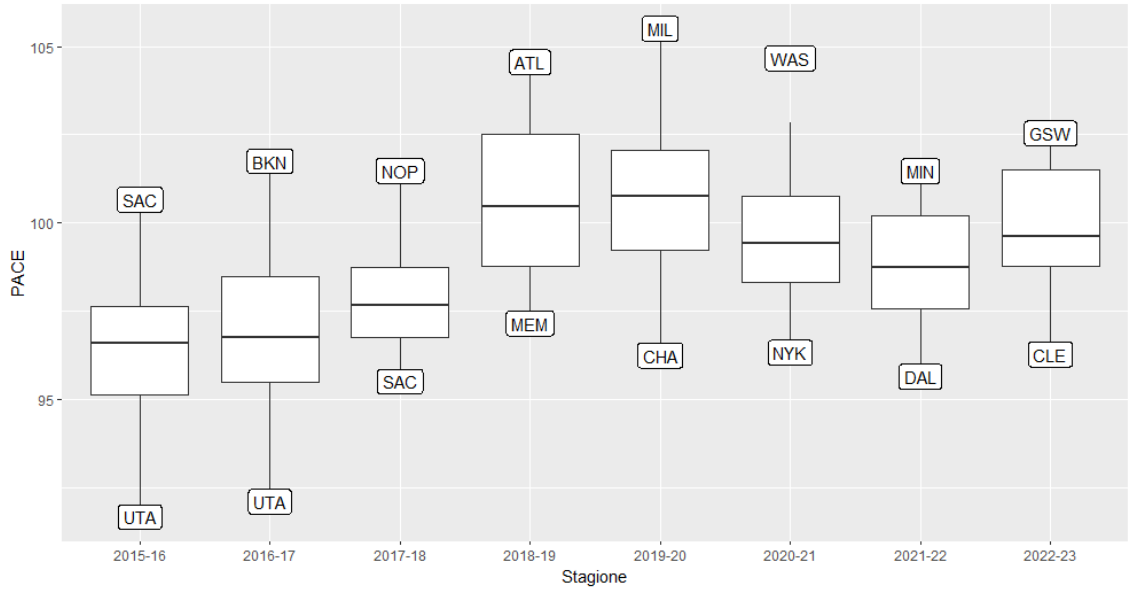

Figura 1.15: Boxplot e squadre con maggiore e minor PACE in ogni stagione.

- TEAM ABBREVIATION: sigla della squadra.
- SEASON: stagione di riferimento.
- NET RATING: differenziale di punti per 100 possessi.
- POSS: numero di possessi.
- id: codice che identifica l'osservazione.

In totale sono quindi presenti 16000 osservazioni in 8 variabili. Come vedremo nel Capitolo 3, il net rating verrà modificato sulla base del numero di possessi giocati dal quintetto. Questi ultimi infatti hanno una forte variabilità, e vanno da un minimo di 19 a un massimo di 2771, rendendo così la misurazione del net rating dei quintetti con meno possessi più distorta e meno veritiera rispetto a lineup con maggior tempo speso assieme.

1.4.1 Analisi esplorative

A conferma di quanto detto sopra, la Figura 1.18 mostra come molti quintetti abbiano giocato meno di 5 log-possessi (circa 150 possessi) e come giocare tanto assieme sia quasi una rarità.

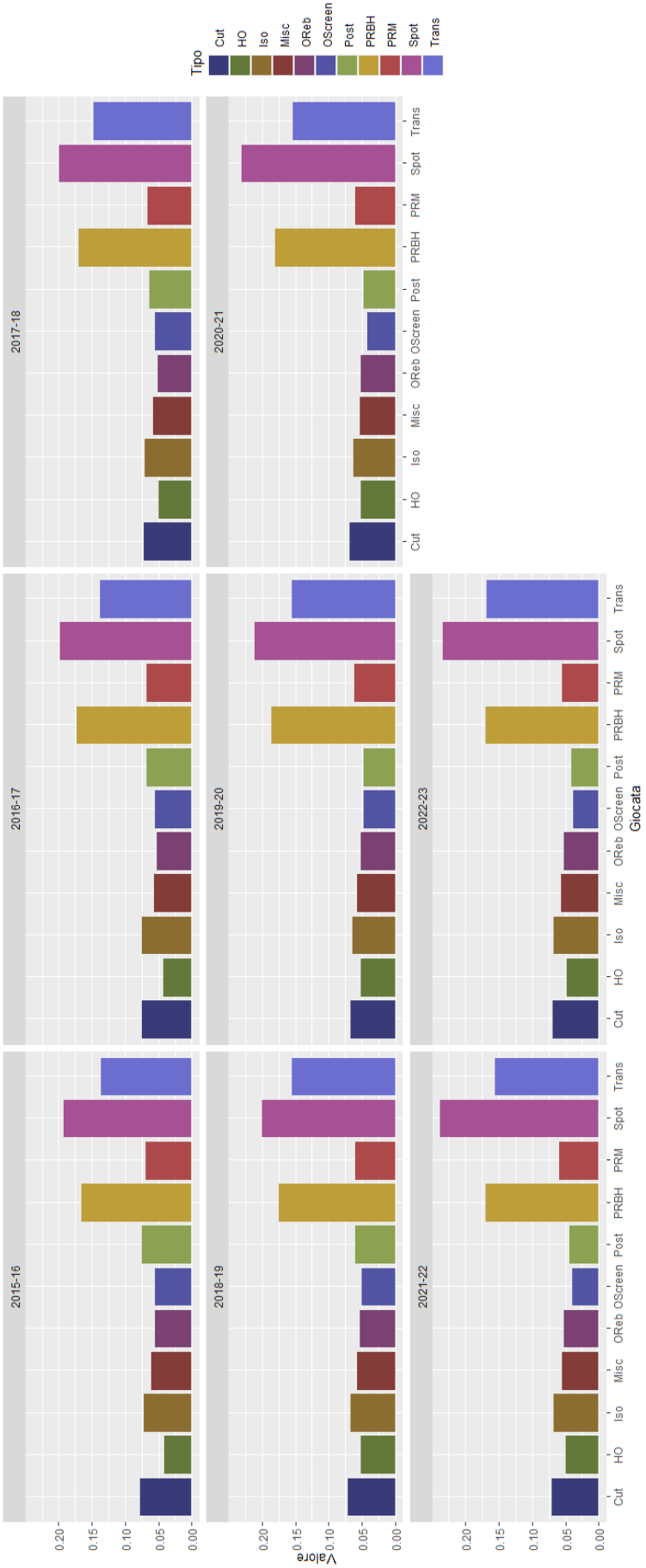

Figura 1.16: Distribuzione dei tipi di giocata per stagione.

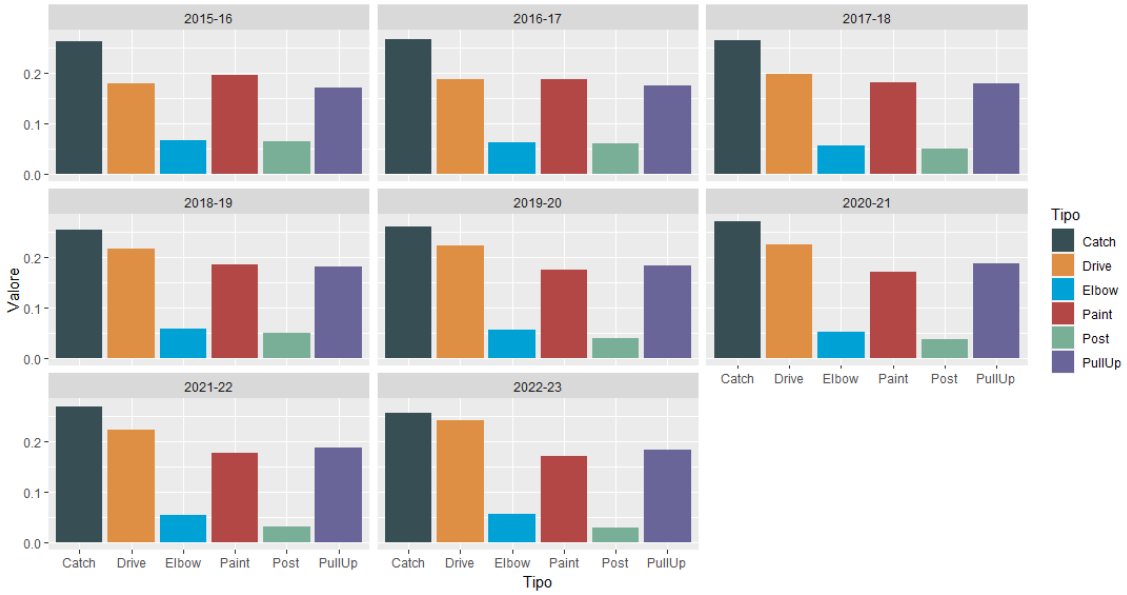

Figura 1.17: Distribuzione dei punti generati per stagione.

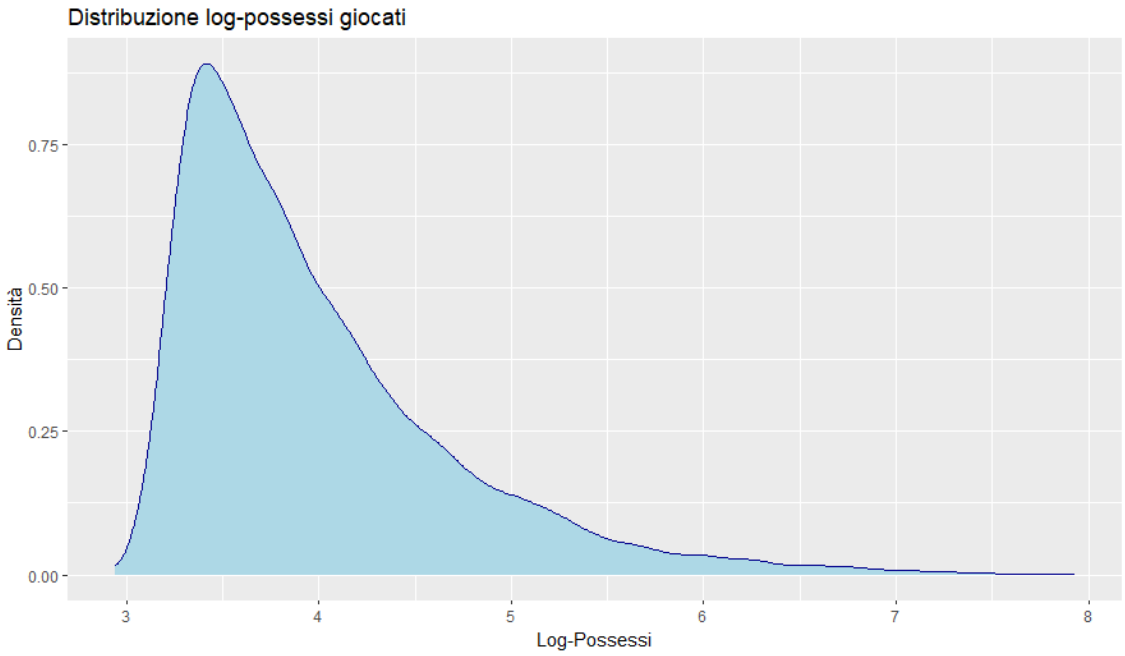

Figura 1.18: Distribuzione numero di log-possessi giocati dai quintetti dalla stagione NBA 2015-2016 alla stagione 2022-2023.

Quintetto	Squadra	Stagione	Possessi
Gortat - Wall - Morris - Beal - Porter Jr.	WAS	16-17	2771
Teague - Gibson - Butler - Wiggins - Towns	MIN	17-18	2287
Howard - Williams - Batum - Walker - Kidd-Gilchrist	CHA	17-18	2273
Westbrook - George - S. Adams - Grant - Ferguson	OKC	18-19	1999
Barnes - Sabonis - Fox - Huerter - Murray	SAC	22-23	1932

Tabella 1.1: I 5 quintetti con il maggior numero di possessi nelle stagioni NBA dall'anno 2015/2016 fino al 2022/2023.

Nella Tabella 1.1 è riportata la top 5 dei quintetti con più possessi giocati assieme. Degno di nota è il quintetto di Sacramento, squadra che, dopo 17 anni di assenza dai playoff, grazie all'offensive rating più alto mai registrato in una stagione NBA, è riuscita ad aggiudicarsi il terzo posto nella Western Conference nella stagione 2022-2023.

Nel Capitolo 3 verrà approfondita la modifica del dataset delle lineup, con l'eliminazione dei quintetti in cui sono presenti giocatori con meno di 30 partite e l'introduzione del bayesian net rating.

Capitolo 2

Metodi e modelli

In questo capitolo vengono introdotti i metodi e i modelli utilizzati per il clustering di squadre e giocatori e per la previsione del net rating. Più precisamente verranno trattati i modelli mistura e l'algoritmo random forest. La teoria riportata in questo capitolo si basa sui libri « *Model-Based Clustering and Classification for Data Science*» di Bouveryon et al., 2019, «*The Elements of Statistical Learning*» di Hastie, Tibshiriani e Friedman, 2009 e «*Data Analysis and Data Mining: An Introduction*» di Azzalini e Scarpa, 2012. Per l'approfondimento dei contenuti qui presentati si invita alla visione dei testi citati.

2.1 Modelli mistura

Si considerti un dataset con n osservazioni $y_1, ..., y_n$ in d variabili, tali che $y_i = (y_{i,1}, ..., y_{i,d})$ per $i = 1, ..., n$. Nei modelli di mistura, la funzione di densità o la funzione di probabilità di un'osservazione y_i è definita dalla media pesata di G funzioni di densità, chiamate componenti della mistura. In particolare, nel modello di mistura gaussiano, tali funzioni di densità provengono tutte da distribuzioni normali:

$$p(y_i) = \sum_{g=1}^{G} \tau_g \phi_g(y_i|\theta_g), \tag{2.1}$$

dove τ_g indica la probabilità che l'osservazione sia generata dall g-esima componente, con $\tau_g \geq 0$ per $g = 1, ..., G$ e $\sum_{g=1}^{G} \tau_g = 1$, mentre $\phi_g(\cdot|\theta_g)$ indica la funzione di densità della g-esima componente, dati i suoi parametri θ_g. Poichè le componenti della mistura

hanno distribuzione normale, si ha $\theta_g = (\mu_g, \Sigma_g)$, dove μ_g indica il vettore d dimensionale delle medie della g-esima componente e Σ_g la corrispondente matrice d x d di varianza-covarianza.

2.1.1 Decomposizione VSO

Una problematica che emerge dall'utilizzo del modello di mistura normale è tuttavia rappresentato dal numero di parametri da utilizzare. Si hanno infatti $G-1$ parametri per gli elementi τ_g, Gd parametri per i vettori delle medie e, infine, $Gd(d+1)/2$ parametri per le matrici di varianza-covarianza. Al crescere del numero di componenti della mistura e del numero di variabili il modello diventa dunque poco parsimonioso, creando possibili problemi di stima. Una soluzione a questo problema è l'utilizzo della decomposizione VSO (Volume-Shape-Orientation) delle matrici di varianza-covarianza delle componenti della mistura:

$$\Sigma_g = \lambda_g D_g A_g D_g^T.$$

In tale decomposizione:

- D_g è la matrice degli autovettori di Σ_g. Essa determina l'orientamento della g-esima componente.
- A_g è la matrice diagonale con valori proporzionali agli autovalori di Σ_g in ordine decrescente. Essa determina la forma della g-esima componente.
- λ_g è la costante di proporzionalità associata. Essa determina il volume della g-esima componente

Il vantaggio dell'utilizzo di questa decomposizione sta nel fatto che, imponendo alcuni vincoli sugli elementi di quest'ultima, nel caso multivariato è possibile definire 14 tipi di modelli più o meno parsimoniosi. Ciascun di questi modelli viene indicato con un identificativo di tre lettere: la prima rappresenta il volume dei cluster, la seconda la forma e la terza l'orientamento. Nello specifico:

- Se la prima lettera è E, allora $\lambda_g = \lambda \ \forall g = 1, ..., G$. Il volume è dunque costante in tutte le componenti. In caso contrario si utlizza la lettera V.

- Se la seconda lettera è E, allora $A_g = A \ \forall g = 1, ..., G$. La forma è dunque la stessa per tutte le componenti. Se, oltre a ciò, $A = I$, con I intesa come matrice di identità, la forma delle componenti risulta sferica e si utlizza la lettera I. Negli altri casi si utlizza la lettera V.

- Se la terza lettera è E, allora $D_g = D \ \forall g = 1, ..., G$. L'orientamento è dunque lo stesso per tutte le componenti. Se, oltre a ciò, $D = I$, con I intesa come matrice di identità, si utlizza la lettera I. Negli altri casi si utlizza la lettera V.

Con tali modelli il numero di parametri per le matrici di varianza-covarianza vanno da un minimo di d (nel caso EII, dove $\Sigma_g = \lambda I \ \forall g = 1, ..., G$) a un massimo di $Gd(d+1)/2$ (nel caso VVV, dove Σ_g è differente per ciascun cluster).

2.1.2 Stima del modello tramite massima verosimiglianza

I modelli descritti sopra possono essere stimati tramite massima verosimiglianza, utilizzando in particolare l'algoritmo Expectation - Maximization, anche detto EM (Dempster, Laird e Rubin, 1977, McLachlan e Krishnan, 1997).

Assumiamo che i dati consistano di n osservazioni (y_i, z_i) per $i = 1, \ldots, n$, dove sono osservati y_i ma non z_i. Siano inoltre (y_i, z_i) indipendenti e identicamente distribuite secondo una distribuzione f con parametri θ. Allora la verosimiglianza dei dati completi risulta essere:

$$L_C(y, z|\theta) = \prod_{i=1}^{n} f(y_i, z_i|\theta), \tag{2.2}$$

dove $y = (y_1, \ldots, y_n)$ e $z = (z_1, \ldots, z_n)$. La verosimiglianza dei dati osservati si ottiene invece integrando 2.2 rispetto a z:

$$L_O(y|\theta) = \int L_C(y, z|\theta) \ dz \tag{2.3}$$

Essa, nel caso gaussiano, può anche essere riscritta come:

$$L_O(y|\theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g \phi_g(y_i|\mu_g, \Sigma_g) \tag{2.4}$$

Nello specifico lo stimatore di massima verosimiglianza di θ basato sui dati osservati massimizza $L_O(y|\theta)$ rispetto a θ.

Come detto sopra, per stimare i parametri del modello verrà utilizzato l'algoritmo EM. Tale algoritmo prevede l'iterazione di due passaggi: l'E-step e l'M-step. Il primo fornisce una stima dei dati non osservati alla luce di quelli osservati e dei parametri stimati. Il secondo invece massimizza la log-verosimiglianza dei dati completi (basata sulla stima dei dati non osservati ottenuta al punto precedente) rispetto ai parametri da stimare. Questi due step vengono ripetuti fino a convergenza, o perlomeno fino al raggiungimento di una certa condizione.

Nel caso del modello di mistura normale, i dati completi sono dati da (y_i, z_i), dove $z_i = (z_{i,1}, \ldots, z_{i,G})$ sono i dati non osservati, con:

$$z_{i,g} = \begin{cases} 1 & \text{se } y_i \text{ appartiene al gruppo } g \\ 0 & \text{altrimenti} \end{cases}$$

Assumiamo che gli z_i siano i.i.d. e provenienti da una distribuzione multinomiale dei G gruppi con probabilità $\tau_1, \ldots, \tau_G$. Assumiamo inoltre che la probabilità di osservare y_i, data z_i, sia $\prod_{g=1}^{G} \phi_g(y_i|\theta_g)^{z_{i,g}}$. La log-verosimiglianza dei dati completi è quindi, in questo caso:

$$l_C(\theta_g, \tau_g, z_{i,g}|y) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{i,g} log[\tau_g \phi_g(y_i|\theta_g)], \tag{2.5}$$

All's-esima iterazione dell'algoritmo EM, vengono quindi calcolate le seguenti quantità

- E-step: viene fornita una stima di $z_{i,g}$, calcolata come:

$$\hat{z}_{i,g}^{(s)} = \frac{\hat{\tau}_g^{(s-1)} \phi_g(y_i|\hat{\theta}_g^{(s-1)})}{\sum_{h=1}^{G} \hat{\tau}_h^{(s-1)} \phi_h(y_i|\hat{\theta}_h^{(s-1)})} \tag{2.6}$$

  dove $\hat{\tau}_g^{(s)}$ è il valore di τ_g dopo s iterazioni. $\hat{z}_{i,g}^{(s)}$ è la stima della probabilità condizionata che l'osservazione i appartenga al gruppo g, date le ossservazioni y_i e i parametri θ_g.

- M-step: viene massimizzata 2.5 rispetto a τ_g e θ_g, con $z_{i,g}$ fissato dall'E-step precedente. Essendo nel caso gaussiano, le stime dei parametri risultano:
  - $\hat{\tau}_g^{(s)} = \frac{\hat{n}_g^{(s-1)}}{n}$
  - $\hat{\mu}_g^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} y_i}{\hat{n}_g^{(s-1)}}$

- $\hat{n}_g^{(s-1)} = \sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)}$
- $\hat{\Sigma}_g^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,g}^{(s-1)} (y_i - \hat{\mu}_g^{(s)})(y_i - \hat{\mu}_g^{(s)})^T}{\hat{n}_g^{(s-1)}}$

L'algortimo viene reiterato, nel caso specifico, fino a quando la differenza tra la log-verosimiglianza al passo s e al passo $s-1$ non è più piccola di una certa soglia (tipicamente 10^{-5}).

La verosimglianza per i modelli mistura non è tuttavia generalmente convessa, perciò è possibile ci siano dei massimi locali (Biernacki, Celeus e Govaert, 2003). A conseguenza di ciò, la stima ottenuta dall'algoritmo EM può dipendere dai valori iniziali che sono stati scelti per i parametri. Una soluzione computazionalmente efficiente, che può essere utilizzata per l'inizializzazione dei parametri dell'algoritmo EM, è quella della classificazione gerarchica basata sui modelli (Banfield e Raftery, 1993). Analogamente ai classici metodi gerarchici agglomerativi, questa tecnica identifica inizialmente ciascuna osservazione in un proprio cluster e, successivamente, ad ogni passo unisce due gruppi in base ad uno specifico criterio. Tale criterio è, in questo caso, la verosimiglianza di classificazione, definita come:

$$L_{CL}(\theta, z|y) = \prod_{i=1}^{n} f_{z_i}(y_i|\theta_{z_i}). \tag{2.7}$$

Tale verosimgilianza viene massimizzata stimando θ e z (quindi parametri e gruppo di appartenenza) contemporaneamente. Si noti che tali stime in generale non sono asintoticamente consistenti (Mariott, 1975) ma, essendo computazionalmente efficienti, risultano utili come stime iniziali per l'algoritmo EM.

2.1.3 Scelta del modello e del numero di cluster

Poichè, di fatto, non conosciamo il numero di gruppi presenti nei nostri dati, la selezione del modello da utilizzare prevede la scelta di due fattori: il modello di clustering (tra i 14 possibili) e il numero di cluster. Nella decisione da prendere ci sarà da valutare il compromesso (tradeoff) tra un modello più semplice (quindi con maggiori restrizioni per quanto riguarda forma, orientamento e volume dei cluster) ma con un maggior numero di gruppi, o un modello più complesso (dunque con maggiore elasticità per quanto riguarda le varie matrici di varianza-covarianza) con meno elementi. Un approccio possibile per tale scelta è la selezione del modello utilizzando la probabilità a posteriori dei modelli (Kass e

Raftery, 1995). Immaginiamo di avere K possibili modelli $M_1, \ldots, M_K$ con probabilità a priori $p(M_k)$, $k = 1, \ldots, K$ (solitamente poste uguali per ogni modello). Siano D i dati a nostra disposizione, allora per il teorema di Bayes abbiamo:

$$p(M_k|D) \propto p(D|M_k)p(M_k) \tag{2.8}$$

La scelta ricade poi sul modello con la maggiore probabilità a posteriori e, poichè $p(M_k)$ è costante per ogni k, allora si guarda la probabilità a priori più grande. Quest'ultima viene ottenuta attraverso il teorema della probabilità totale:

$$p(D|M_k) = \int p(D|\theta_{M_k}, M_k)p(\theta_{M_k}|M_k)\, d\theta_{M_k} \tag{2.9}$$

dove $p(\theta_{M_k}|M_k)$ è la distribuzione a priori di θ_{M_k}, cioè dei parametri del modello M_k. L'integrale di cui sopra è tuttavia difficile da calcolare. Sotto condizioni di regolarità dei modelli però, la 2.9 può essere approssimata dal criterio di informazione di Bayes (BIC):

$$2logp(D|M_k) \approx 2logp(D|\hat{\theta}_{M_k}M_k) - \nu_{M_k}log(n) = BIC_{M_k}, \tag{2.10}$$

dove ν_{M_k} è il numero di parametri indipendenti da stimare nel modello M_k (Haughton, 1988).

Sebbene i modelli di mistura non soddisfino le condizioni di regolarità richieste, l'utilizzo del criterio è appropriato anche nel caso della classificazione basata sui modelli (Leroux, 1992, Keribin, 1998). La scelta ricade dunque sul modello con BIC più elevato.

2.1.4 Selezione delle variabili

Nell'ambito del clustering basato sui modelli, la scelta delle variabili risulta molto importante poichè, nel caso di variabili irrilevanti, i metodi sopracitati possono non risultare particolarmente efficaci a causa del rumore eccessivo generato da tali variabili. Vari approcci sono stati utilizzati in questo ambito, ma quello proposto qui (Maugis, Celeux e Martin-Magniette, 2009) prevede che, tra le variabili rilevate, ci sia:

- Un insieme S di variabili rilevanti ai fini del clustering.
- Un insime $U = S^C$ di variabili irrilevanti.
- Un insieme $R \subseteq S$ di variabiili rilevanti da cui le variabili di U dipendono attraverso un modello di regressione lineare.

Denotando quindi con G il numero di gruppi, m il modello scelto e con r la forma della matrice di regressione Ω (identità, diagonale o generica), la funzione di densità dei dati diventa quindi in questo caso:

$$f(y_i|G, m, r, \mathbf{V}, \theta) = \sum_{g=1}^{G} \tau_g \phi(y_i^S|\mu_g, \Sigma_{g(m)}) \times \phi(y_i^U|a + y_i^R b, \Omega_{(r)}) \tag{2.11}$$

dove $\theta = (\eta, a, b, \Omega)$ è il vettore dei parametri con $\eta = (\tau_1, \ldots, \tau_G, \mu_1, \ldots, \mu_G, \Sigma_1, \ldots, \Sigma_G)$ e $\mathbf{V} = (S, R, U)$.

Indicando ora con BIC_{clust} il valore del BIC per il modello di mistura con le variabili dell'insieme S e con BIC_{reg} il valore del BIC per il modello di regressione delle variabili di U su quelle di R, la scelta delle variabili e del modello si otterrà ottimizzando:

$$CRIT(G, m, r) = BIC_{clust}(y^S|G, m) + BIC_{rec}(y^U|r, y^R). \tag{2.12}$$

2.1.5 Pacchetti R

I metodi descritti saranno implementati utilizzando le librerie di R **MClust** e **clustvarsel**. In particolare. la seconda permetterà di scegliere le variabili più adatte alla clusterizzazione, mentre la prima verrà utilizzata per selezionare il modello e procedere con il clustering.

2.2 Random Forest

Il modello che utilizzeremo per poter fare previsione sul net rating bayesiano (concetto che verrà introdotto nel prossimo capitolo) sarà, come anticipato nei capitoli precedenti, quello del Random Forest. In particolare, tale metodo costruisce un grande numero di alberi di regressione (da qui il termine foresta) de-correlati e, successivamente, calcola la previsione per una nuova osservazione come media delle previsioni di ciascun albero ottenuto. Introduciamo quindi il concetto di albero di regressione e bagging e vediamo nel dettaglio l'algoritmo che permette la previsione della variabile risposta tramite random forest.

2.2.1 Alberi di regressione

Immaginiamo di voler approssimare una generica e ignota funzione $y = f(x)$ con $x \in \mathbb{R}^p$, basandoci su un campione di N osservazioni (x_i, y_i) per $i = 1, \ldots, N$ e utilizzando una funzione a gradini. Per generare una funzione di questo tipo, è necessario dividere lo spazio $\mathbb{R}^p$ in un numero finito J di regioni R_j per $j = 1, \ldots, J$ e assegnare un valore costante c_j a ognuna di tali regioni. Per semplicità, imponiamo che ciascuna suddvisione sia effettuata tramite tagli paralleli agli assi coordinanti. L'approssimazione di $f(x)$ può essere quindi ottenuta come segue:

$$\hat{f}(x) = \sum_{j=1}^{J} c_j I(x \in R_j), \tag{2.13}$$

dove J indica il numero di suddivisioni, c_j sono costanti e $I(x \in R_j)$ è funzione indicatrice della regione R_j generata dalla suddivsione.

Gli alberi di regressione utilizzano un approccio di ottimizzazione passo passo per $\hat{f}(x)$, nel senso che a ogni step dell'algoritmo generano un'approssimazione via via più fine della funzione di regressione, scegliendo una delle p variabili a disposizione e il punto in cui effettuare la suddivisione in modo da minimizzare la devianza totale:

$$D = \sum_{i=1}^{p} [y_i - \hat{f}(x_i)]^2. \tag{2.14}$$

Si noti che questa espressione può anche essere riscritta come somma delle singole devianze delle regioni R_j:

$$D = \sum_{i=1}^{N} [y_i - \hat{f}(x_i)]^2 = \sum_{j=1}^{J} \sum_{x_i \in R_j} (y_i - \hat{c}_j)^2 = \sum_{j=1}^{J} D_j. \tag{2.15}$$

La costante $\hat{c}_j$ che permette tale minimizzazione è data dalla media delle risposte delle osservazioni che appartengono a R_j . L'algoritmo di crescita di un albero di regressione parte inizialmente da $J = 1$, $R_j = \mathbb{R}^p$, $D = \sum_{i=1}^{N} (y_i - \overline{y})^2$ e, ad ogni passo, aumenta a $J + 1$ la dimensione dell'albero.

Per $J = 1$ fino a $J = N$, per ogni $j = 1, ..., J$, dividiamo la regione R_j in due parti R_j^1 e R_j^2 . La nuova devianza della regione R_j risulta quindi essere scomposta nel modo

seguente:

$$D_j^* = \sum_{i \in R_j^1} (y_i - \hat{c}_j^{'})^2 + \sum_{i \in R_j^2} (y_i - \hat{c}_j^{"})^2. \tag{2.16}$$

Poichè l'obiettivo è mimizzare $D = \sum_{j=1}^{J} D_j$, selezioniamo la variabile e il punto di taglio che massimizzano il guadagno in devianza $D_j - D_j^*$. Procedendo fino a $J = N$, si crea un albero con N foglie, che però risulta concettualmente equivalente ad una interpolazione dei dati a nostra disposizione e causa quindi overfitting. E' perciò necessario effettuare una "potatura" dell'albero, abbassando il numero J delle foglie. Si introduce quindi una funzione obiettivo che include una penalizzazione rispetto al grado di complessità J del modello:

$$C_\alpha(J) = \sum_{j=1}^{J} D_j + \alpha J \tag{2.17}$$

dove α è un parametro di penalizzazione non negativo. Fissato α, si può dimostrare che esiste un unico albero più piccolo dell'albero completo che minimizza $C_\alpha(J)$ (Ripley, 1996). Scegliamo dunque tale albero, eliminando sequenzialmente una foglia alla volta, in particolare quella la cui rimozione porta al minor incremento di $\sum_{j=1}^{J} D_j$. Il problema si riduce ora alla scelta del parametro α. Tale scelta può essere effettuata utilizzando una parte del campione come training set per far crescere l'albero e la restante porzione di dati come test set per poter scegliere il parametro di penalizzazione che permette la minimizzazione della devianza. Una volta costruito l'albero, la previsione per una nuova osservazione viene ottenuta facendo cadere quest'ultima dall'albero e osservando su quale foglia si deposita.

2.2.2 Bagging

Uno dei maggiori difetti degli alberi di regressione sta nell'instabilità dei risultati per quanto riguarda piccole perturbazioni o integrazioni dei dati del campione, che possono generare alberi molto diversi fra loro. Una soluzione a questo problema è il bagging. Tale metodo, il cui nome sta per bootstrap aggregation, permette la riduzione della varianza ed è quindi molto utile se applicato agli alberi. Nella pratica, il bagging consiste nel generare B replicazioni bootstrap del training set (dove ciascun bootstrap non è altro che un campionamento casuale con reimmissione delle osservazioni dei dati di addestramento), utilizzare ciascuno dei B campioni ottenuti per creare un modello e ottenere una previsione

$\hat{f}^{*b}(x)$ e infine calcolare la media di tutte le previsioni per ottenere la previsione finale:

$$\hat{f}_{bag}(x) = \frac{1}{B}\hat{f}^{*b}(x) \tag{2.18}$$

Tale procedura sarà utilizzata successivamente all'interno dell'algortimo random forest.

2.2.3 Algoritmo Random Forest

Una volta introdotti i concetti di albero di regressione e bagging, vediamo più nel dettaglio come funziona l'algoritmo di previsione del random forest (Breiman, 2001). Sia N il numero di osservazioni utilizzate come training set, p il numero di variabili esplicative, m un numero minore di p (vedremo nel seguito con quale criterio scegliere tale valore) e sia n_{min} la dimensione minima dei nodi. Applichiamo la procedura del bagging al nostro campione, generando B replicazioni bootstrap di dimensione N a partire dal training set e facendo crescere un albero di regressione per ciascun campione fino a quando questo non raggiunge la dimensione minima dei nodi. La sola modifica fatta rispetto al classico algoritmo di crescita sta nel fatto che le variabili candidate per la scelta del punto di suddivisione sono solo m delle p variabili disponibili, selezionate in modo casuale. Questa scelta serve a ridurre la correlazione presente tra gli alberi. Il problema principale del bagging sta appunto nel fatto che, soprattutto quando solo alcune delle variabili prese in considerazione sono significative per la previsione, gli alberi che vengono costruiti risultano molto correlati, abbassando l'impatto che hanno sulla riduzione della varianza rispetto a un insieme di alberi non correlati. Infatti, se avessimo una media di B variabili casuali indipendenti e identicamente distribuite, ciascuna con varianza σ^2, la varianza di tale media sarebbe uguale a $\frac{\sigma^2}{B}$. Con B elevato dunque, tale varianza si ridurrebbe sensibilmente. Tuttavia, nel nostro caso ci troviamo davanti a una media di B variabili casuali identicamente distribuite, con varianza σ^2 e con correlazione positiva ρ, che risulta avere varianza pari a:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{2.19}$$

Scegliere casualmente m variabili permette di abbassare il termine ρ e quindi di provocare la decorrelazione degli alberi, garantendo quindi stime migliori. Una volta ottenuti B alberi di regressione, la previsione di un nuovo punto viene effettuata calcolando la media dei valori previsti per quel punto da ciascuno degli alberi. L'algoritmo viene descritto in 1.

Algorithm 1 Random forest per la regressione

1. Per $b = 1$ fino a B:
   (a) Crea una replicazione Bootstrap di dimensione N dai dati di addestramento.
   (b) Crea un albero del random forest T_b a partire dai dati bootstrap, ripetendo ricorsivamente i seguenti passaggi per ogni nodo terminale dell'albero, fino a quando non è raggiunta la dimensione minima n_{min} di un nodo.
      i. Seleziona m variabili casuali dalle p disponibili.
      ii. Seleziona il miglior punto di di taglio considerando la miglior variabile tra le m disponibili.
      iii. Dividi il nodo in due nodi figli
2. Per effettuare una previsione di un nuovo punto x a partire dall'insieme di alberi $\{T_b\}_1^B$ appena ottenuto, calcola la media delle previsioni: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$.

Capitolo 3

Clustering

In questo capitolo verrà applicato il modello mistura sui dati a nostra disposizione. Per giocatori e squadre si procederà a selezionare le variabili rilevanti, il modello e il numero di gruppi secondo i criteri descritti nel capitolo precedente. Seguirà quindi una descrizione delle caratteristiche di ciascun cluster. Infine verranno descritti la costruzione delle soft-lineups e il calcolo del net rating bayesiano.

3.1 Clustering dei giocatori

Anzitutto, decidiamo di standardizzare i dati relativi ai giocatori, in modo da dare un peso equo a ciascuna variabile selezionata. Successivamente, tramite la funzione clustvarsel() del pacchetto clustvarsel effettuiamo la selezione delle variabili e del modello utilizzando i criteri descritti nel Paragrafo 2.1.4. In particolare con tale funzione, per ciascuno dei $14 * 9 = 126$ modelli possibili (14 sono le possibili strutture della matrice di varianza-covarianza descritte nel Paragrafo 2.1.1, 9 sono i possibili valori di G, inteso come numero di gruppi, che vanno da 1 a 9) effettuiamo la scelta delle variabili, calcoliamo il valore di 2.12 e selezioniamo il migliore tra essi. Le variabili rilevanti indicate dall'utilizzo di tale funzione sono THREE FG AST PCT, THREE FGA PCT, CORNER3, THREE FG PCT, e come modello più plausibile abbiamo 'VEV,9', ovvero 9 gruppi con matrice di varianza-covarianza che hanno volume e orientamento variabili e forma elipsoidale ed uguale per ciascuna componente. Interessante notare come tutte le variabili rilevanti siano indicatori

della capacità e della frequenza con cui i giocatori tirano da 3 punti, fattore che abbiamo analizzato nel Capitolo 1. A seguito di tale scelta, tramite la funzione Mclust() del pacchetto mclust è stato eseguito il clustering dei giocatori. Tale funzione, inseriti in input il numero di gruppi, il tipo di modello e le variabili da utilizzare per effettuare la clusterizzazione, stima i parametri del modello tramite i criteri descritti nel Paragrafo 2.1.2 utilizzando l'algoritmo EM. Qui però ci si è accorti di una problematica: a ogni lancio della funzione, la classificazione risultava significativamente diversa. Ciò è dovuto al fatto che l'inizializzazione dell'algoritmo EM, che si basa sulla classificazione gerarchica dei modelli, nella funzione Mclust() avviene utilizzando un campione di 2000 osservazioni sulle 3100 a disposizione, provocando probabilmente problemi di massimo locale per alcuni dei campioni utilizzati. Come soluzione sono state quindi prodotte 100 classificazioni diverse, memorizzando il valore del BIC di ciascuna di esse e calcolando per ogni coppia di classificazioni la loro similarità tramite la funzione adjustedRandIndex(). Tale funzione permette di confrontare due classificazioni secondo il metodo proposto da Hubert e Arabie, 1985, che fornisce un valore da 0 e 1 per valutare il grado di equivalenza di due configurazioni, dove 1 rappresenta l'equivalenza e 0 invece la totale dissonanza. Poichè la media di quest'utlimo valore risultava molto bassa (0.45), sono state prese le 10 classificazioni con BIC più alto. Considerando solo queste 10 configurazioni il valore medio della similarità è cresciuto (0.67) e si è scelta quindi la configurazione con BIC maggiore per effettuare il clustering.

A ciascuna osservazione viene quindi assegnato un vettore che indica con quale probabilità un giocatore appartiene a un certo gruppo. Tale giocatore viene quindi inserito nel cluster per cui tale probabilità ha valore maggiore. Nella Figura 3.1 è riportato il grafico che rappresenta la distribuzione delle probabilità massime presente nei vettori . Il grafico risulta soddisfacente: l'84 percento delle osservazioni ha valori superiori allo 0.9.

Nella Figura 3.2 è riportato il numero di giocatori presente in ciascun cluster.

Per valutare le caratteristiche di ciascun cluster, guardiamo in maniera preliminare i boxplot delle variabili rilevanti di ciascun gruppo (Figura 3.3). Il cluster 6 è particolarmente rappresentativo, e comprende giocatori che tirano poco o niente da 3 punti. Il grafico comunque non è sufficiente a comprendere la composizione dei 9 gruppi, perciò andremo a valutare sia i boxplot delle altre variabili non comprese nel processo di clustering, sia i giocatori inclusi nei cluster. Nei paragrafi successivi ogni cluster viene descritto a uno a uno.

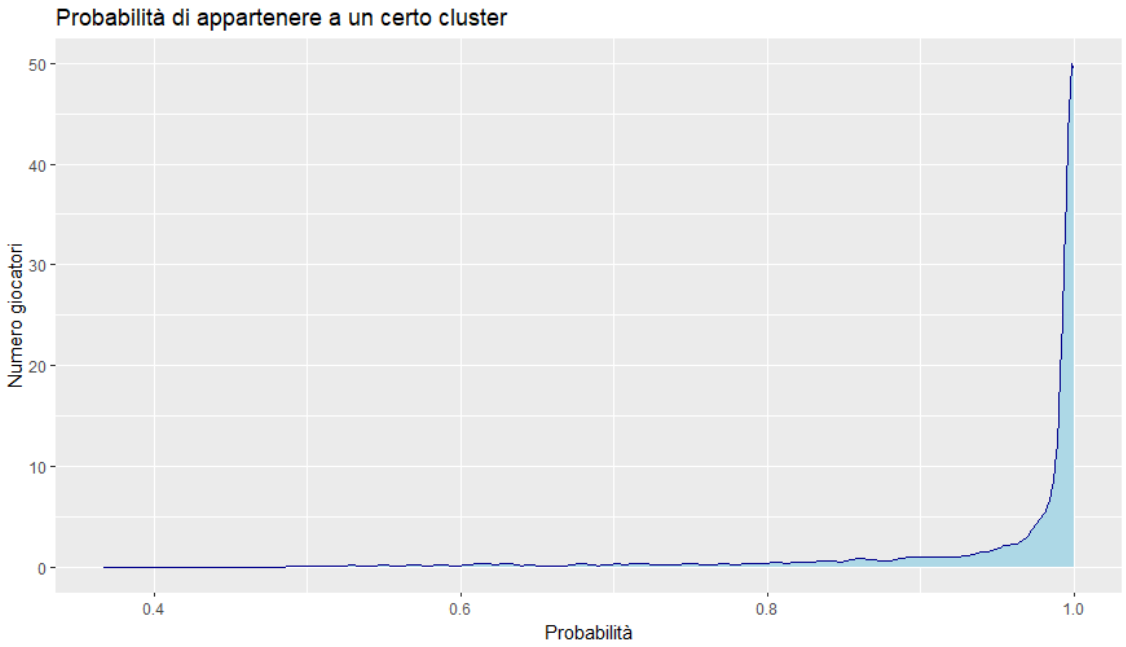

Figura 3.1: Distribuzione della probabilità massima di appartenenza ai cluster dei giocatori.

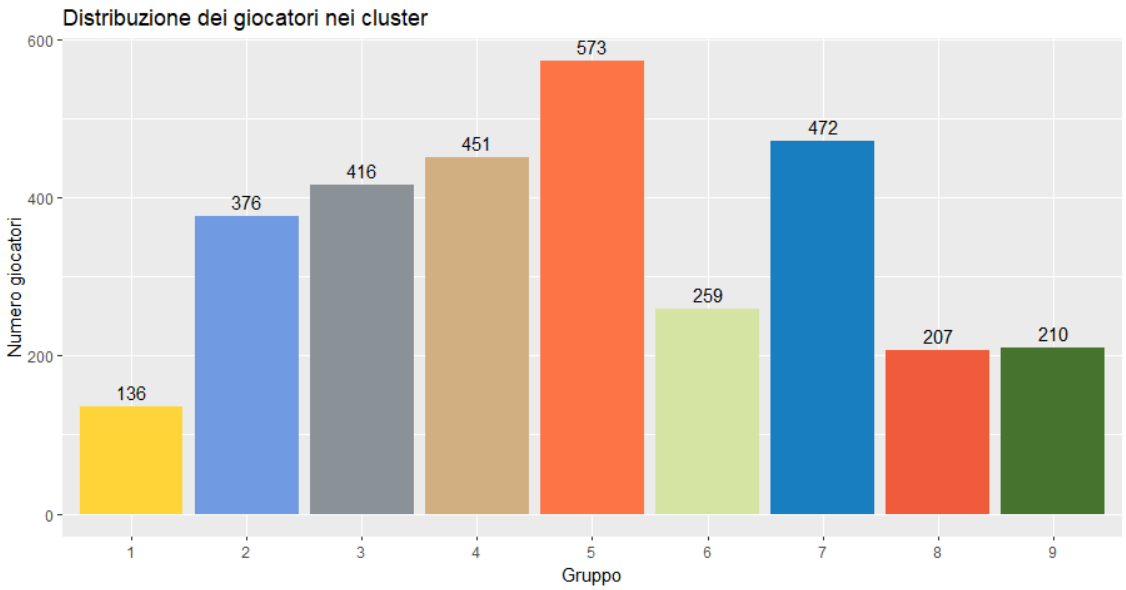

Figura 3.2: Numero di giocatori in ciascun cluster.

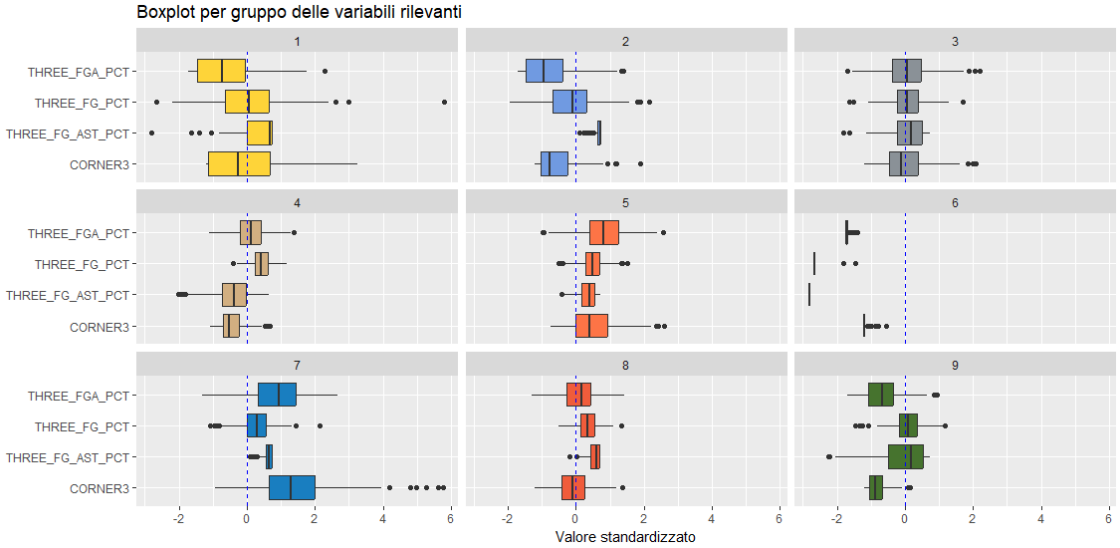

Figura 3.3: Boxplot delle variabili rilevanti per ciascun cluster individuato.

3.1.1 Cluster 1: Role players

Il cluster 1 è il meno numeroso tra i 9 trovati (136 osservazioni). Definiamo i giocatori che lo compongono come role players in quanto, guardando ai boxplot delle variabili, nessuna di queste risulta particolarmente sopra o sotto la media. I giocatori dunque non eccellono in nessuna statistica. Inoltre, sono presenti sia giocatori più perimetrali che lunghi che giocano vicino al ferro.

Esempio giocatori: Andrea Bargnani ’16, Austin Reaves ’23.

3.1.2 Cluster 2: Lunghi versatili

Il secondo cluster comprende i lunghi versatili, ovvero giocatori in genere molto alti, con alti valori per quanto riguarda statistiche difensive come stoppate e rimbalzi e bassi valori per quanto concerne il tiro da 3 punti, sebbene non in maniera estrema come nel caso del cluster 6. Sono versatili poichè comprendono giocatori in grado di spaziare il campo e di attaccare il ferro anche partendo lontano da canestro. Da notare anche la presenza di Ben Simmons in questo gruppo, giocatore che di fatto è un playmaker sovradimensionato ma che paga la sua scarsa abilità da 3 punti. Nella Figura 3.4 è possibile visualizzare le principali caratteristiche di tale cluster.

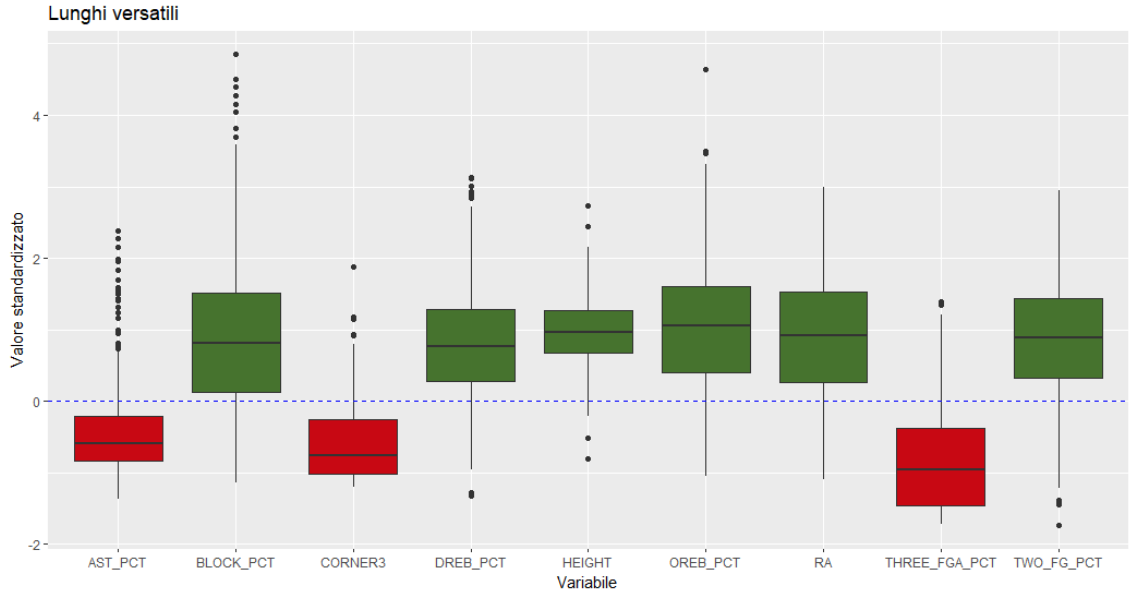

Figura 3.4: Boxplot delle variabili notevoli per il secondo cluster.

Esempio giocatori: Draymond Green '16, Domantas Sabonis '23.

3.1.3 Cluster 3: Generali in campo

I generali in campo sono giocatori il cui obiettivo primario è quello di mettere in ritmo i compagni, servendo soprattutto assist più che tentando soluzioni personali. Sono generalmente piccoli e dunque con bassi numeri per quanto riguarda i rimbalzi e le stoppate, come mostrato in Figura 3.5

Esempio giocatori: Ricky Rubio '17, Rajon Rondo '18.

3.1.4 Cluster 4: Guardie offensive

Il quarto cluster risulta composto da alcuni dei miglior giocatori NBA. In particolare sono esterni (quindi giocatori non troppo alti) che hanno molto spesso in mano il pallone (come si evince dall'alto valore dell'usage percentage - Figura 3.6), e che quindi hanno libertà di attaccare sia per crearsi un tiro sia per fornire un assist ai compagni.

Esempio giocatori: Kyrie Irving '17, Stephen Curry '18.

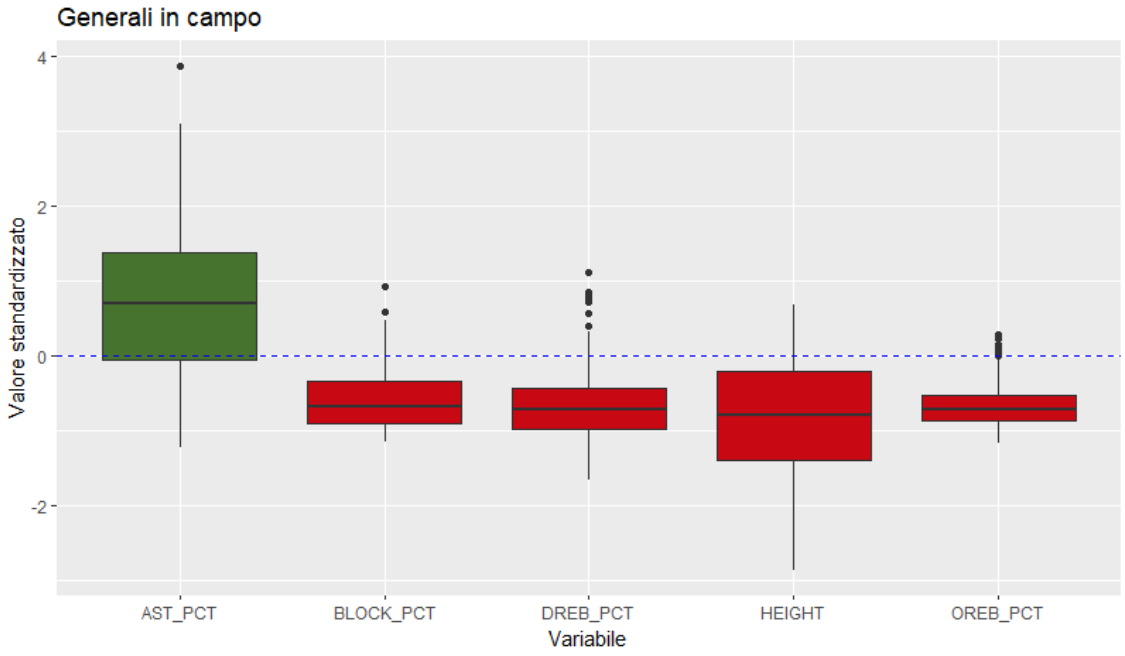

Figura 3.5: Boxplot delle variabili notevoli per il terzo cluster.

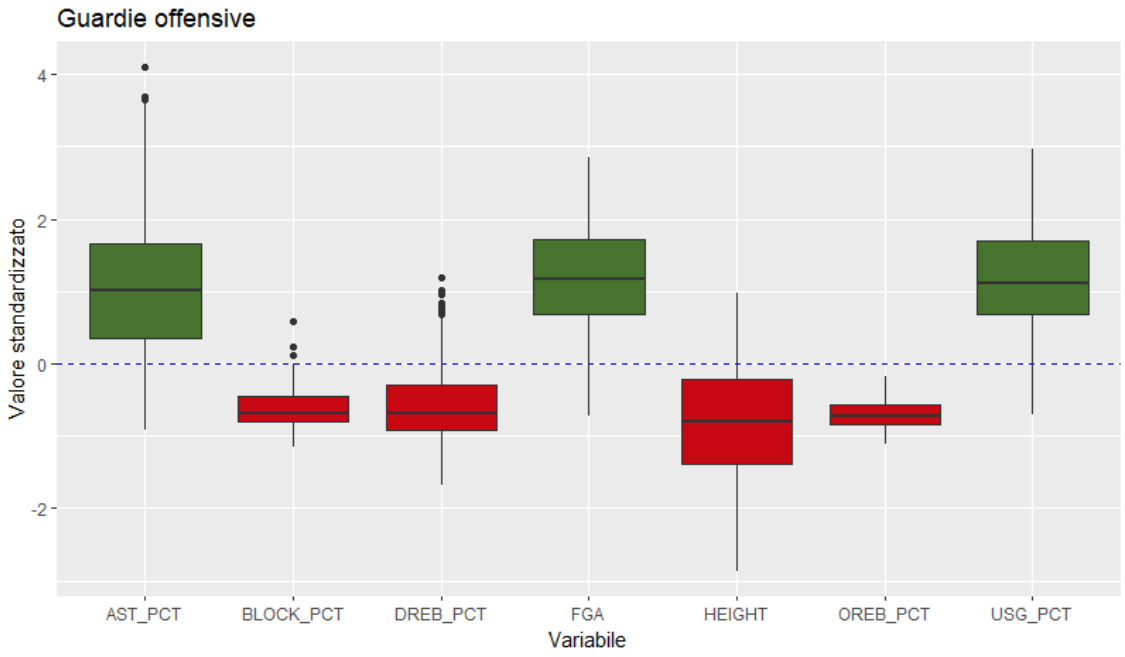

Figura 3.6: Boxplot delle variabili notevoli per il quarto cluster.

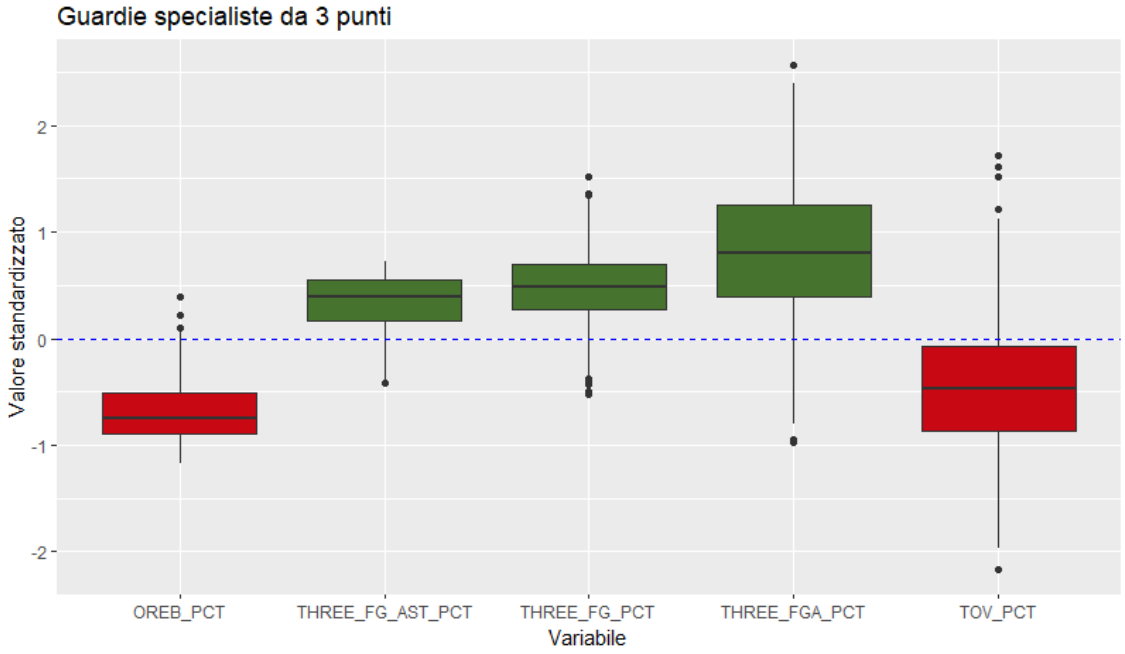

Figura 3.7: Boxplot delle variabili notevoli per il quinto cluster.

3.1.5 Cluster 5: Guardie specialiste da 3 punti

Le guardie specialiste da 3 punti sono esterni il cui compito è, per l'appunto, stare sul perimetro per punire la difesa con un tiro da tre punti piazzato. Hanno generalmente un basso di numero di palle perse (Figura 3.7) proprio perchè raramente devono gestire il pallone.

Esempio giocatori: Marco Belinelli '16 , Klay Thompson '19.

3.1.6 Cluster 6: Lunghi classici

Come era plausibile aspettarsi dal grafico in Figura 3.3, il sesto cluster è formato da tutti quei lunghi "vecchia scuola" il cui compito è quello di proteggere il ferro, prendere rimbalzi e, soprattutto, segnare da vicino canestro (esplicativa la Figura 3.8). Particolare il fatto che nessuna delle osservazioni presente in questo gruppo abbia un minimo grado di incertezza: ogni giocatore è inserito in questo cluster con probabilità 1. Quello del lungo classico è un ruolo che sta via via scomparendo: come mostra la Figura 3.9, negli ultimi 8 anni il numero

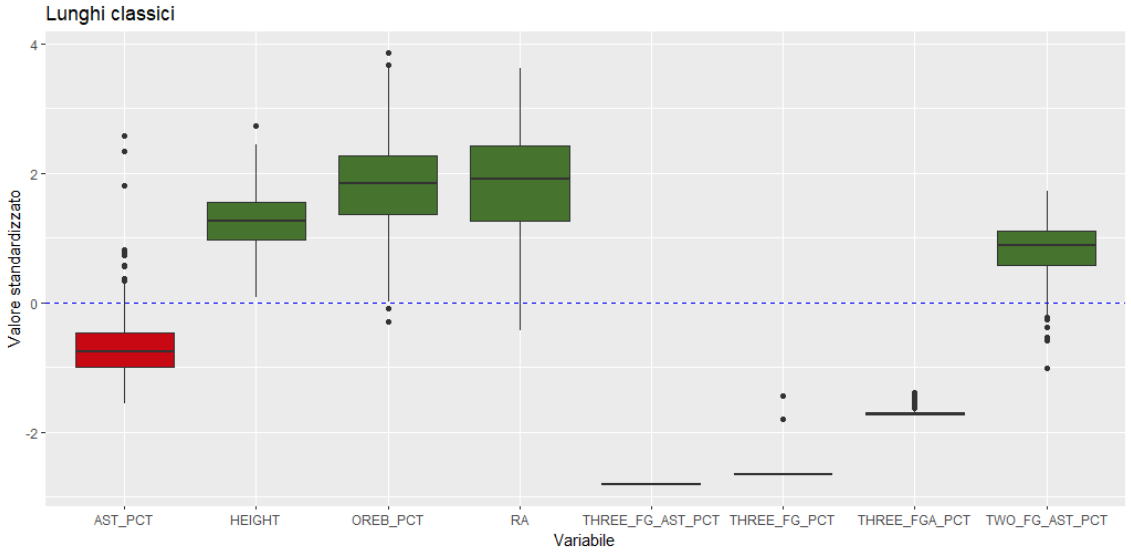

Figura 3.8: Boxplot delle variabili notevoli per il sesto cluster.

di interpreti di tale ruolo è più che dimezzato, altro segno di una pallacanestro che richiede giocatori sempre più versatili e capaci di aprire il campo.

Esempio giocatori: Clint Capela ’20, Rudy Gobert ’21.

3.1.7 Cluster 7: Stretch forwards

Le stretch forwards sono giocatori simili alle guardie specialiste da 3 punti per quanto riguarda il tiro da oltre l’arco, ma sono generalmente più alte e hanno un usage percentage molto basso. Puniscono la difesa soprattutto dagli angoli (Figura 3.10).

Esempio giocatori: Reggie Bullock ’22, Robert Covington ’21.

3.1.8 Cluster 8: Skilled forwards

Il cluster 8 è composto da ali capaci sia di attaccare il ferro che di tirare da tre punti se messi in ritmo (Figura 3.11), sebbene in maniera meno efficace delle stretch forwards. Sono in grado di sfruttare la loro stazza in situazioni di rimbalzo e sono spesso una delle principali opzioni offensive della squadra.

Esempio giocatori: Chris Bosh ’16, Danilo Gallinari ’16.

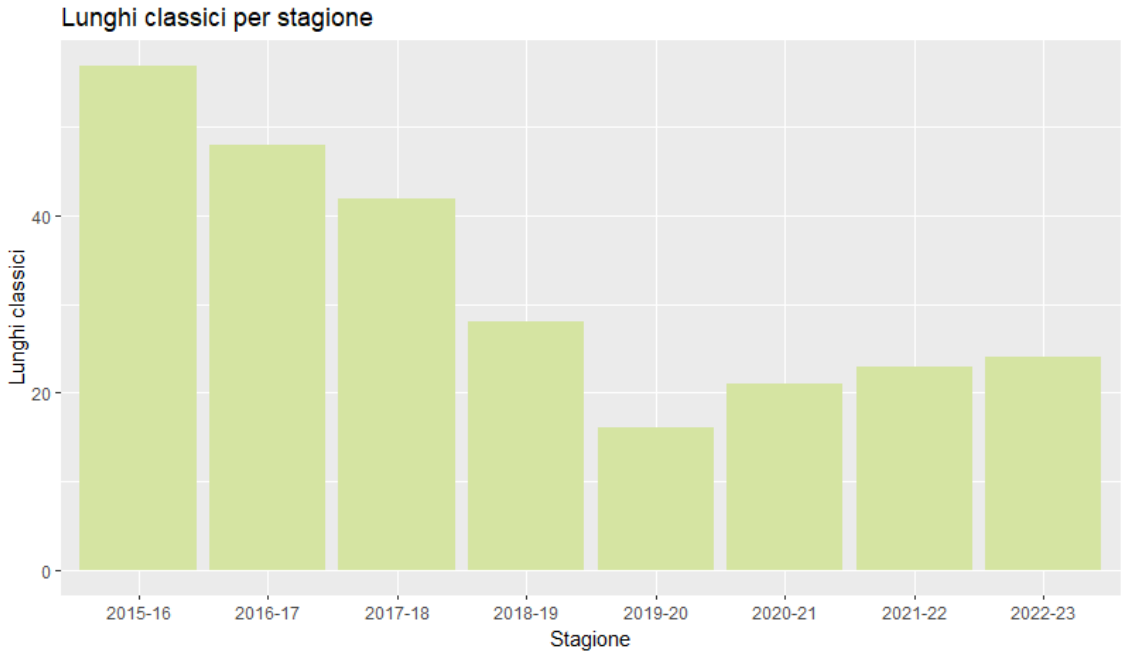

Figura 3.9: Lunghi classici presenti in ciascuna stagione.

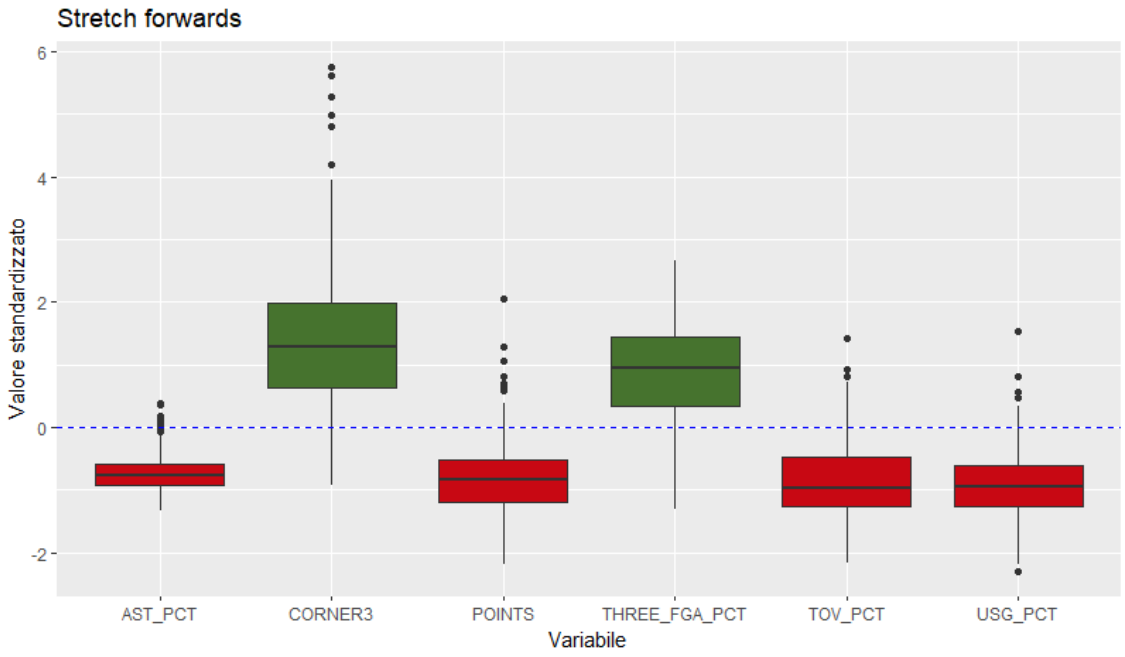

Figura 3.10: Boxplot delle variabili notevoli per il settimo cluster.

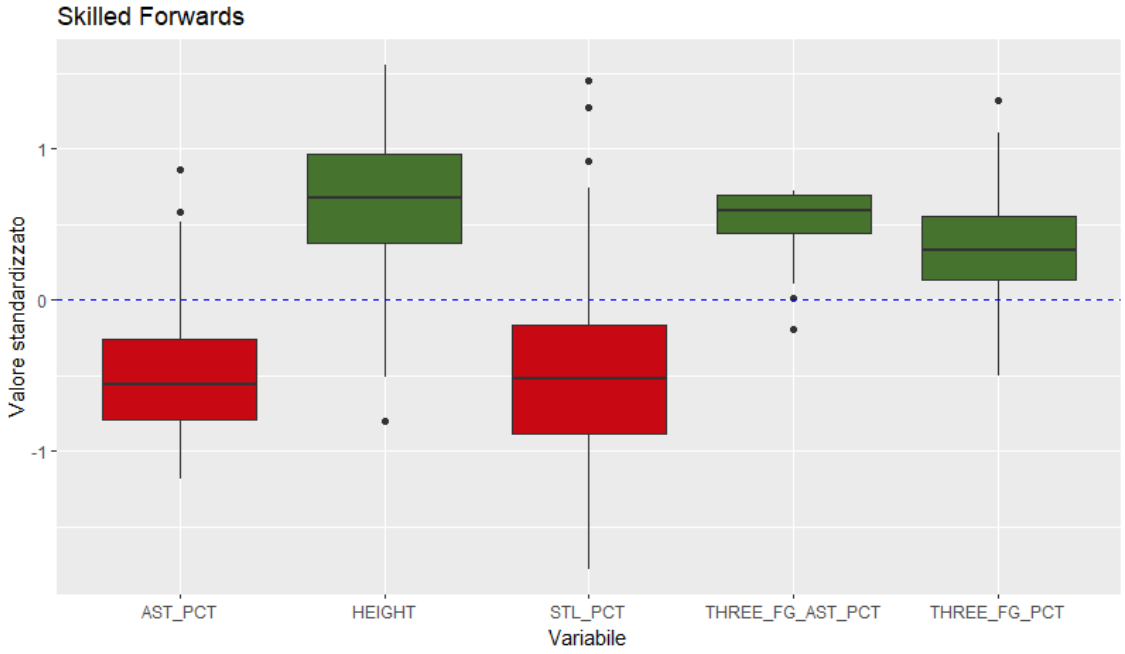

Figura 3.11: Boxplot delle variabili notevoli per l'ottavo cluster.

3.1.9 Cluster 9: Marcatori

I marcatori sono i protagonisti indiscussi nell'attacco loro squadra. Sono particolarmente efficaci e sono leader per punti segnati, PIE e tentativi dal campo (Figura 3.12) . Rispetto alle guardie offensive, hanno valori più bassi per quanto riguarda l'assist percentage. Questo cluster, insieme alle guardie offensive, rappresenta l'elite dei giocatori NBA.

Esempio giocatori: LeBron James '20, Giannis Antetokounmpo '21.

3.2 Clustering delle squadre

Analogamente a quanto fatto per i dati sui giocatori, standardizziamo i dati sulle squadre e utilizziamo la funzione Clustvarsel() per poter decidere la forma della matrice di varianza-covarianza, il numero di gruppi e le variabili rilevanti. Il modello finale ha identificativo "EVE", presenta 3 cluster e ha come variabili rilevanti Opp Mid, Isolation, Opp LC3, DREB, Opp ITP. Da notare che solo Isolation rappresenta una variabile che descrive l'attacco delle squadre, mentre le altre variabili rappresentano le zone del campo da cui

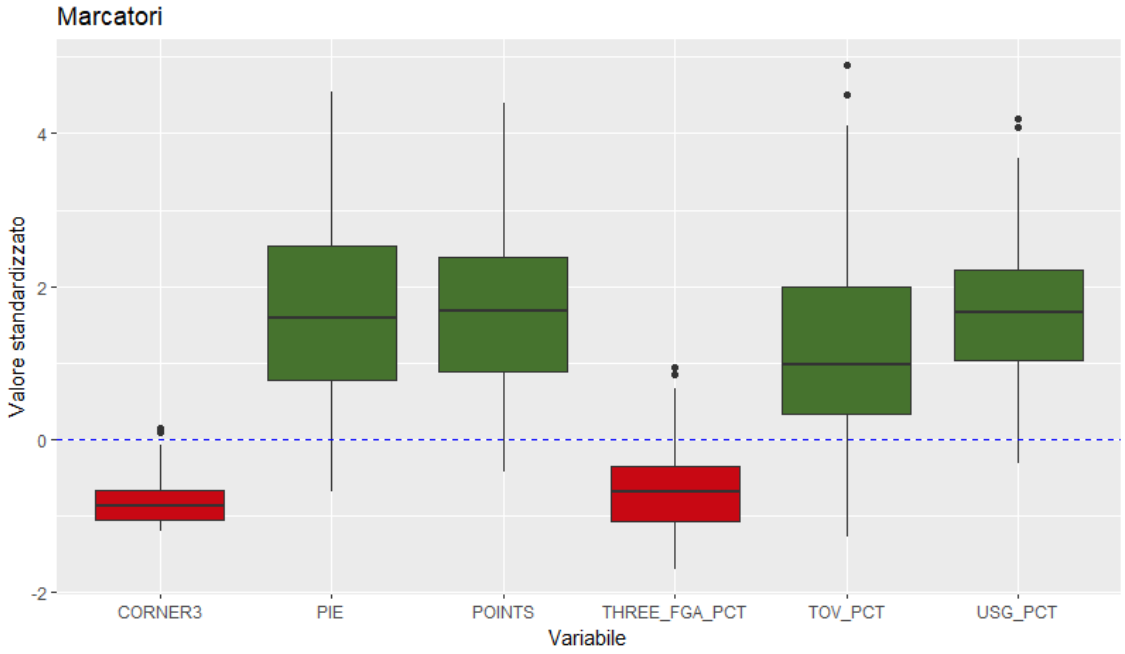

Figura 3.12: Boxplot delle variabili notevoli per il nono cluster.

le squadre concedono più o meno tiri, insieme al numero di rimbalzi difensivi presi (che è indicatrice di quanti tiri sbagliano i propri avversari). Come per i giocatori, una volta ottenuti il miglior modello e le variabili rilevanti, è stata utilizzata la funzione mclust() per effettuare il clustering. In questo caso non ci sono stati problemi per quanto riguarda l'inizializzazione dell'algoritmo EM, dato che le osservazioni risultano essere solamente 240.

La Figura 3.13 mostra la distribuzione della probabilità massima per le squadre di appartenere a un certo cluster. Anche in questo caso, la gran parte delle squadre ha un grado di incertezza sostanzialmente nullo.

La Figura 3.14 rappresenta la distribuzione delle squadre per stile di gioco e stagione. E' interessante notare come ci sia una netta divisione temporale tra il triennio 2016-2018 (dove a farla da padrone è il cluster 1) e il periodo del 2018-2023 (con una supremazia sostanzialmente totale del cluster 3). Potrebbe effettivamente essere questo un segno del cambiamento nello stile di gioco di cui si parlava nell'introduzione.

La Figura 3.15 mostra i boxplot delle variabili rilevanti per ciascun gruppo. In linea con le analisi esplorative del primo capitolo, in cui si evidenziava l'andamento in picchiata del mid-range, è lampante che la principale differenza tra il gruppo 1 e il gruppo 3 stia

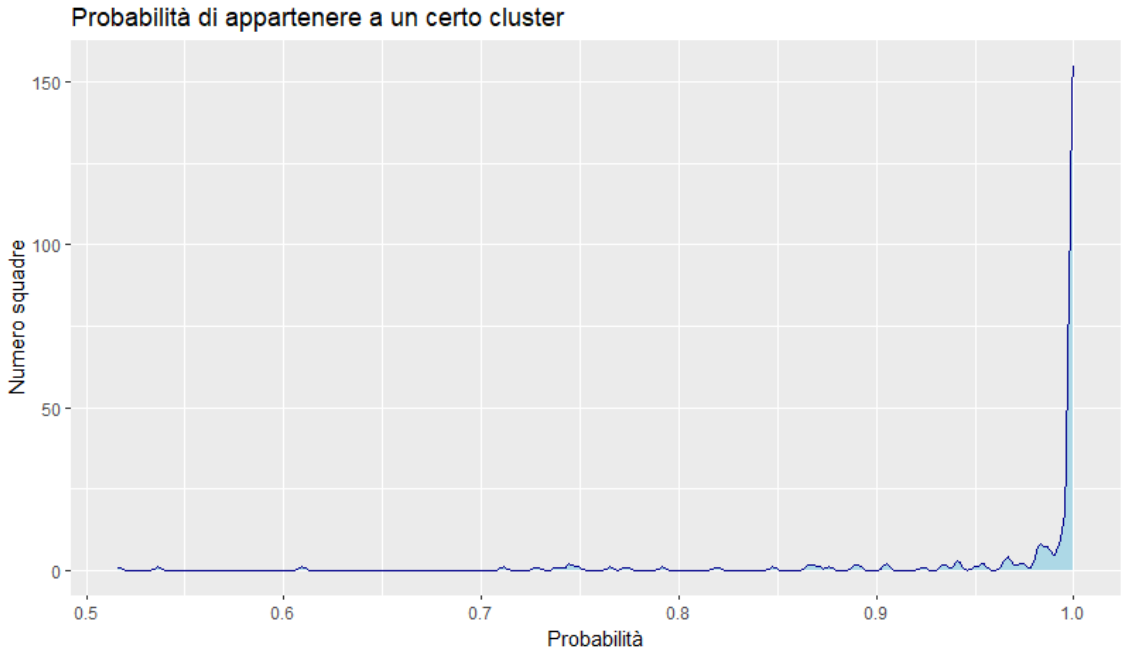

Figura 3.13: Probabilità massima delle squadre di appartenere a un certo cluster.

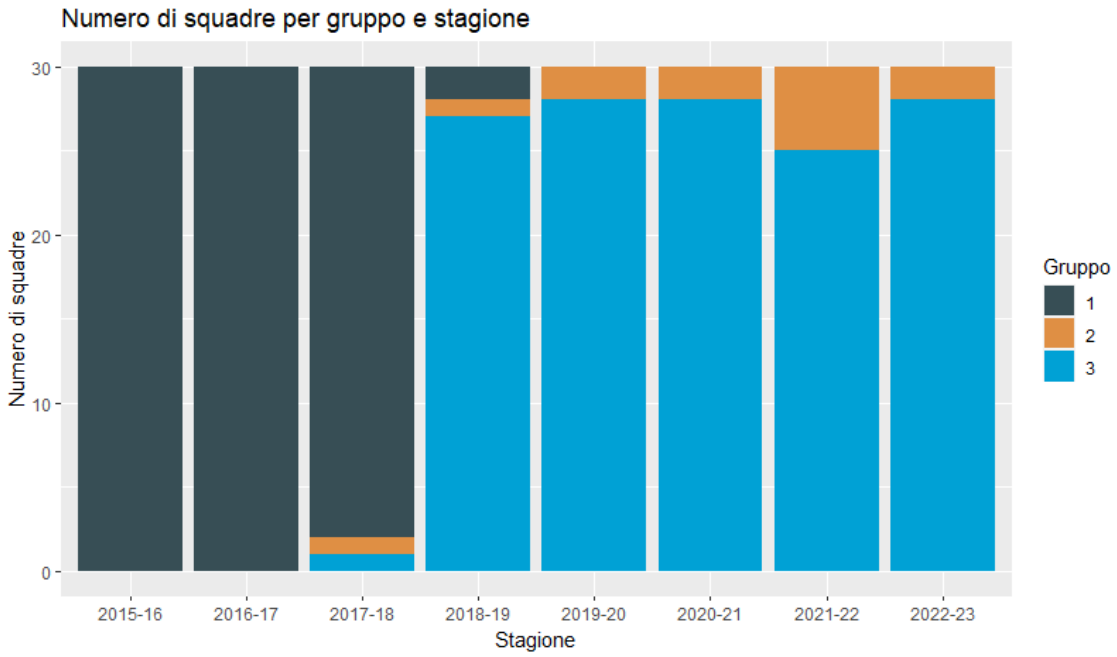

Figura 3.14: Istogramma delle squadre per stagione e gruppo di appartenenza.

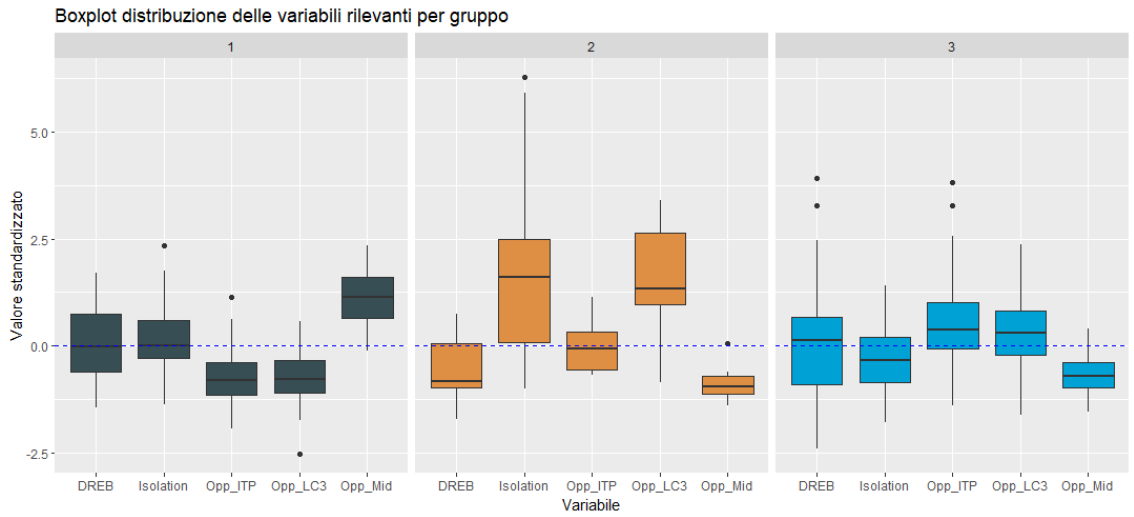

Figura 3.15: Boxplot delle variabili rilevanti per gruppo.

nel numero di tiri concessi dalla media. In crescita nel gruppo 3 sono i tiri concessi da tre punti e quelli nel pitturato, ovvero i tiri più efficaci. Il gruppo 2, che è il meno numeroso dei tre (appena 13 osservazioni), si distingue per un alto utilizzo dell'isolamento e per un alto numero di tiri concessi da 3 punti.

La Figura 3.16 mostra le differenze tra i vari cluster basandosi sui grafici di dispersione di ciascuna coppia di variabili rilevanti.

Nella Tabella 3.1 sono riportate le squadre che hanno vinto il titolo nelle rispettive stagioni, il loro cluster di appartenenza e il vettore delle probabilità associato.

Una volta eseguito il clustering per giocatori e team, si procede con la costruzione delle soft lineups e con la previsione del net rating.

3.3 Soft-lineups

Coerentemente con quanto detto nel Capitolo 1, rimuoviamo dai 16000 quintetti iniziali le osservazioni in cui compaiono i giocatori con meno di 30 partite giocate in una determinata stagione: rimangono dunque 13881 quintetti. Aggiungiamo al dataset così ottenuto anche un'altra variabile: il TEAM NET RATING, che indica il differenziale di punti per 100 possessi della squadra di cui il quintetto fa parte. Altra modifica necessaria si ha nel calcolo del net rating per i quintetti che non hanno giocato un alto numero di possessi:

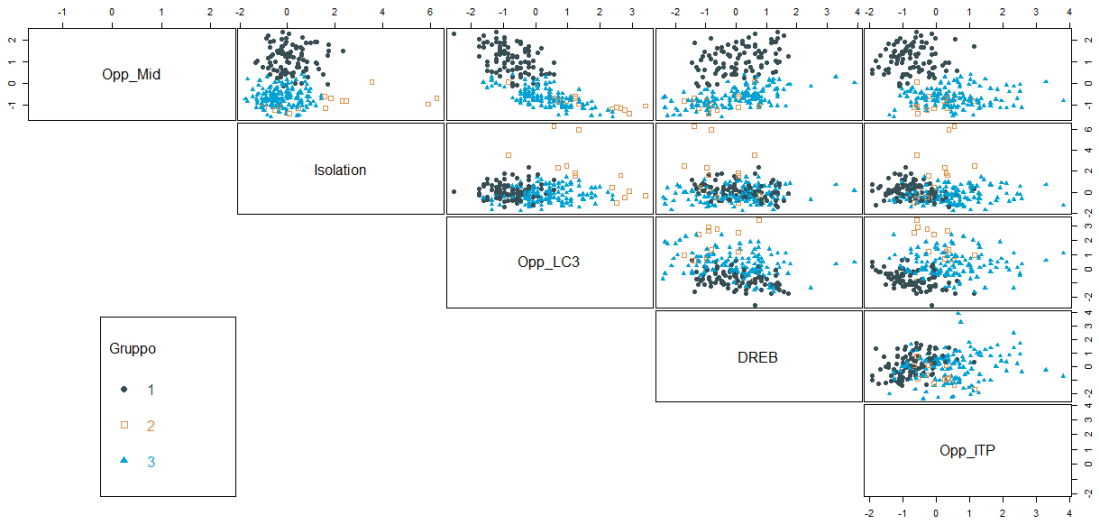

Figura 3.16: Scatter plot per ciascuna coppia di variabili rilevanti.

Stagione	Squadra	Cluster	Vettore delle probabilità
15-16	CLE	1	1, 0, 0
16-17	GSW	1	1, 0, 0
17-18	GSW	1	0.905, 0, 0.095
18-19	TOR	3	0.025, 0, 0.975
19-20	LAL	3	0.003, 0, 0.997
20-21	MIL	3	0, 0.009, 0.991
21-22	GSW	3	0, 0, 1
22-23	DEN	3	0, 001, 0.999

Tabella 3.1: Cluster e probabilità di appartenenza per cluster delle squadre che hanno vinto il titolo.

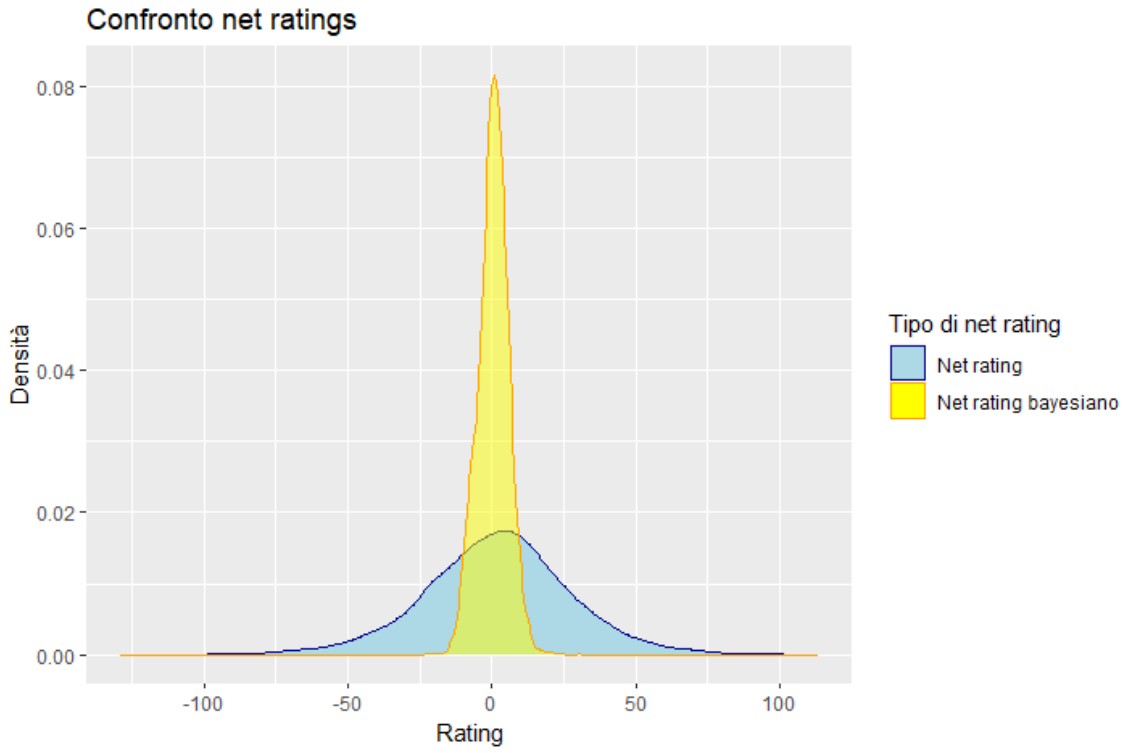

Figura 3.17: Confronto tra le distribuzioni del net rating originale e del net raing bayesiano.

come proposto da Kalman e Bosch si decide di utilizare ciò che loro chiamano "Net rating bayesiano'". In particolare si sceglie di mantenere il net rating originale per le lineup con più di 600 possessi (poco più dell' 1,3 percento delle osservazioni) e di modificarlo per tutti gli altri quintetti secondo questa formula:

$$Net\ rating\ bayesiano = (\frac{POSS}{600}) * Net\ rating + (1 - \frac{POSS}{600}) * Team\ Net\ rating. \quad (3.1)$$

Nella Figura 3.17 è possibile vedere la differenza nella distribuzione dei due tipi di net rating. Valori come +50 di net rating risultano molto irrealistici da mantenere per un quintetto durante tutta la stagione, e il net rating bayesiano permette di pesare in maniera diversa e più plausibile le lineup con pochi possessi giocati, alla luce del rendimento generale della squadra.

Una volta aggiustata quella che nel nostro modello sarà la variabile risposta, si procede dunque con la costruzione delle soft-lineups. Una soft-lineups è composta dalla somma dei

Stagione	Squadra	Team Soft	Player Soft
15-16	OKC	1,0,0	0.01, 0.99, 0, 0, 0, 1, 1, 0, 2

Tabella 3.2: Esempio di composizione di una soft lineup.

vettori delle probabilità di appartenere a un certo cluster dei 5 giocatori che compongono il quintetto e dal vettore di probabilità della squadra di cui fanno parte. Prendiamo ad esempio il quintetto degli Oklahoma City Thunder della stagione 2015-2016, formato da Kevin Durant, Russell Westbrook, Serge Ibaka, Andre Roberson e Steven Adams. In particolare, questa lineup è composta nel modo seguente:

- Kevin Durant: marcatore.
- Russell Westbrook: marcatore.
- Serge Ibaka: role player (1 percento), lungo versatile (99 percento).
- Andre Roberson: stretch forward.
- Steven Adams: lungo classico.
- Oklahoma City Thunder: gruppo 1.

La soft lineup risultante sarà dunque quella rappresentata nella Tabella 3.2.

Ottenuto il dataframe con tutte le 13881 soft-lineup, utilizzeremo tali dati per procedere con la previsione del net rating. Vedremo tale applicazione e i risultati nel prossimo capitolo.

Capitolo 4

Risultati

In questo capitolo viene applicato l'algoritmo random forest ai dati sulle soft-lineups. In particolare, per verificare l'efficacia di tale algoritmo, si è scelto di selezionare un campione casuale di 1800 quintetti (circa il 15 percento delle osservazioni) come test set e i restanti 12081 come training set.

4.1 Scelta del numero di alberi e delle variabili di split

Come illustrato nella Sezione 2.2, è necessario scegliere alcuni parametri di regolazione per la costruzione del random forest: il numero B di alberi, il numero m di variabili tra cui scegliere per effettuare le suddivisioni e la dimensione minima di cascun nodo n_{min}. Per iniziare, utilizziamo l'algoritmo random forest (1) impostando come variabile risposta il net rating bayesiano, come variabili esplicative i $p = 12$ valori dati da ciascuna soft lineup (3 per il team di cui il quintetto fa parte, 9 per la composizione del quintetto stesso) e come valori di default $B = 500$, $m = p/3 = 4$ e $n_{min} = 5$. Dal random forest così prodotto, verifichiamo l'andamento dell'errore quadratico medio. Si noti che tale errore è ottenuto calcolando la media del quadrato dell' errore oob (out-of-bag). Per la previsione di ciascun osservazione appartenente al training set infatti, vengono utilizzati solo gli alberi che, nella replicazione bootstrap necessaria alla costruzione dell'albero, non presentano tale osservazione. La stima dell'errore oob così ottenuta risulta quasi identica a quella ottenuta

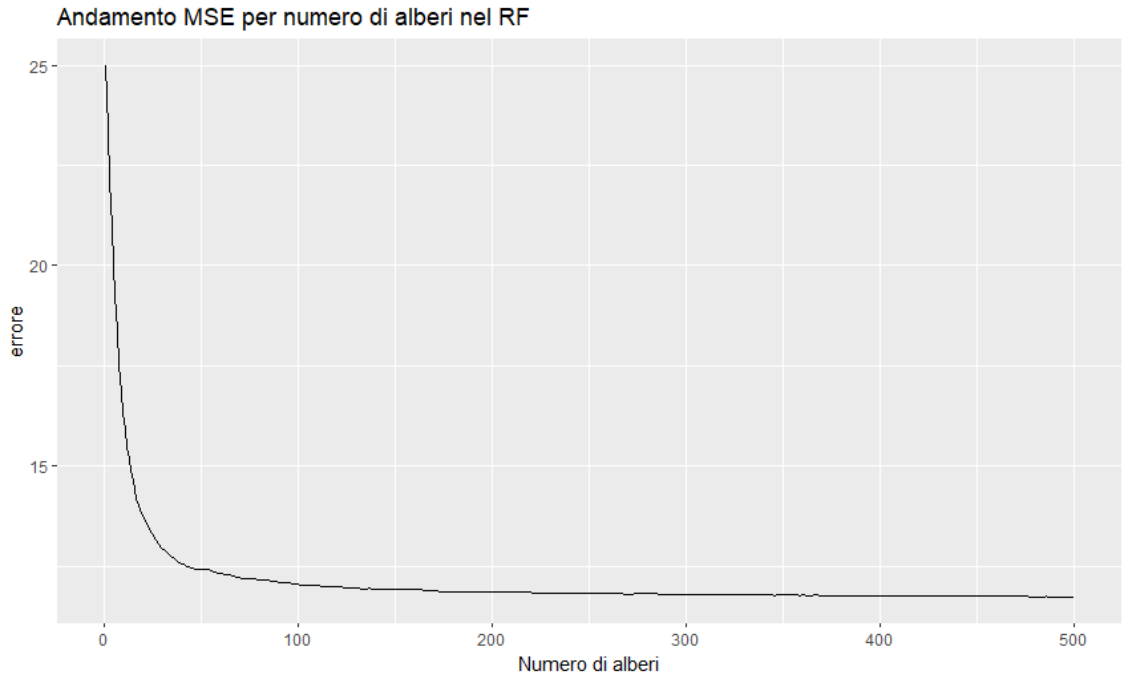

Figura 4.1: Valore del MSE al variare del numero di alberi.

utilizzando la N fold cross-validation (Hastie, Tibshiriani e Friedman, 2009). Come si evince dalla Figura 4.1, l'errore si stabilizza intorno ai 200 alberi.

La differenza tra l'errore tra 200 e 500 alberti risulta essere di appena 0.11. Per questo motivo decidiamo di impostare $B = 300$, in modo da avere un grado di complessità del modello non troppo elevato, mantenendo però una buona precisione. Fissato B, è necessario scegliere quale valore di m sia maggiormente efficace per la previsione. Per ogni $m = 1, \ldots, 12$ implementiamo l'algoritmo random forest e calcoliamo l'errore oob e l'errore sul test set. Alla luce di quanto riportato dal grafico in Figura 4.2, si sceglie 4 come valore di m.

Per quanto riguarda n_{min}, si è scelto di mantenere il valore di default 5, poichè una varianzione di tale valore non migliorava sensibilmente l'errore di previsione. Una volta quindi trovati i parametri di regolazione dell'algoritmo random forest, procediamo con la costruzione del modello finale e analizziamo i risultati.

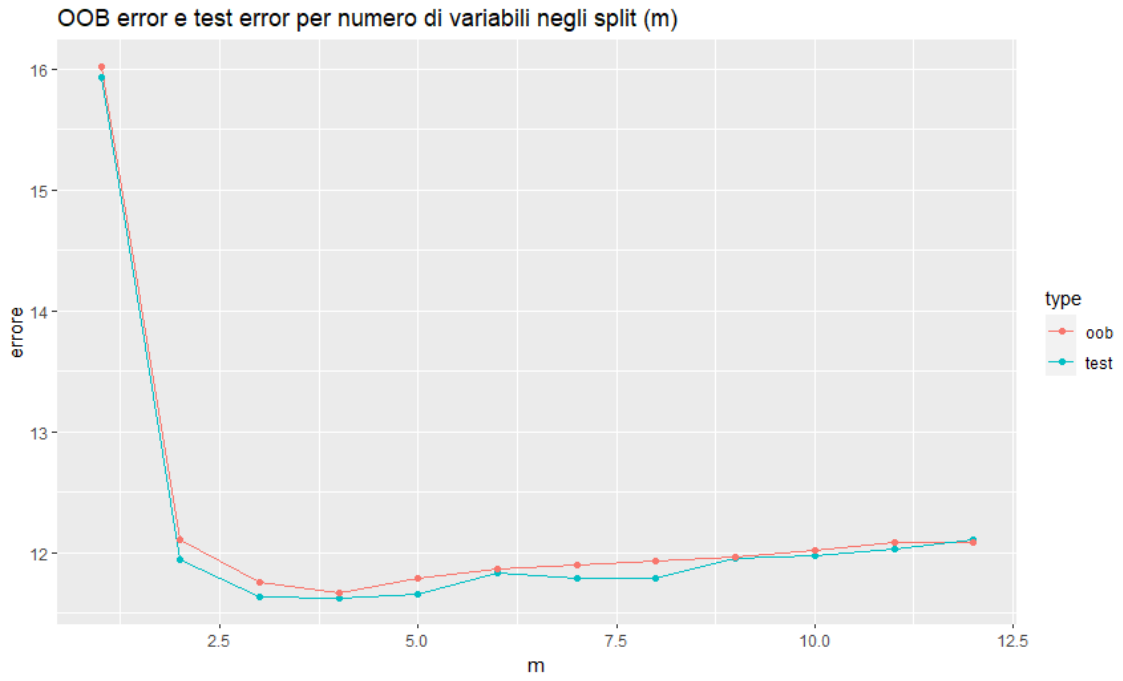

Figura 4.2: Confronto tra errore OOB ed errore sul test-set al variare del numero di variabili per lo split.

4.2 Modello finale

Coerentemente con quanto detto sopra, costruiamo un random forest con $B = 300$, $m = 4$, $n_{min} = 5$ e confrontiamo i valori predetti con quelli osservati. La Figura 4.3 mostra tale confronto. Poichè i valori vicini alla bisettrice rappresentano previsioni più accurate, il modello sembra nel complesso avere una buona capacità predittiva.

Figura 4.3: Confronto tra valori previsti e osservati per train set e test set.

Nella Figura 4.4 è possibile vedere la distribuzione del root mean square error misurato sul test set.

Il modello finale sembra tutto sommato convincente: più del 70 percento delle previsioni ha rmse minore di 2, segno di una buona affidabilità. Nelle Tabelle 4.1 e 4.2 sono riportate rispettivamente le cinque migliori e le cinque peggiori previsioni basate sull'errore assoluto tra valori previsti e osservati (arrotondato alla 3 cifra decimale). Interessante notare come, tra le peggiori stime, ci sia quella relativa al quintetto formato da Durant, Green, Thompson, Curry e Iguodala, da molti definito come uno dei migliori quintetti della storia NBA. Sebbene la previsione sia comunque molto alta (11.3 il net rating bayesiano previsto), l'al-

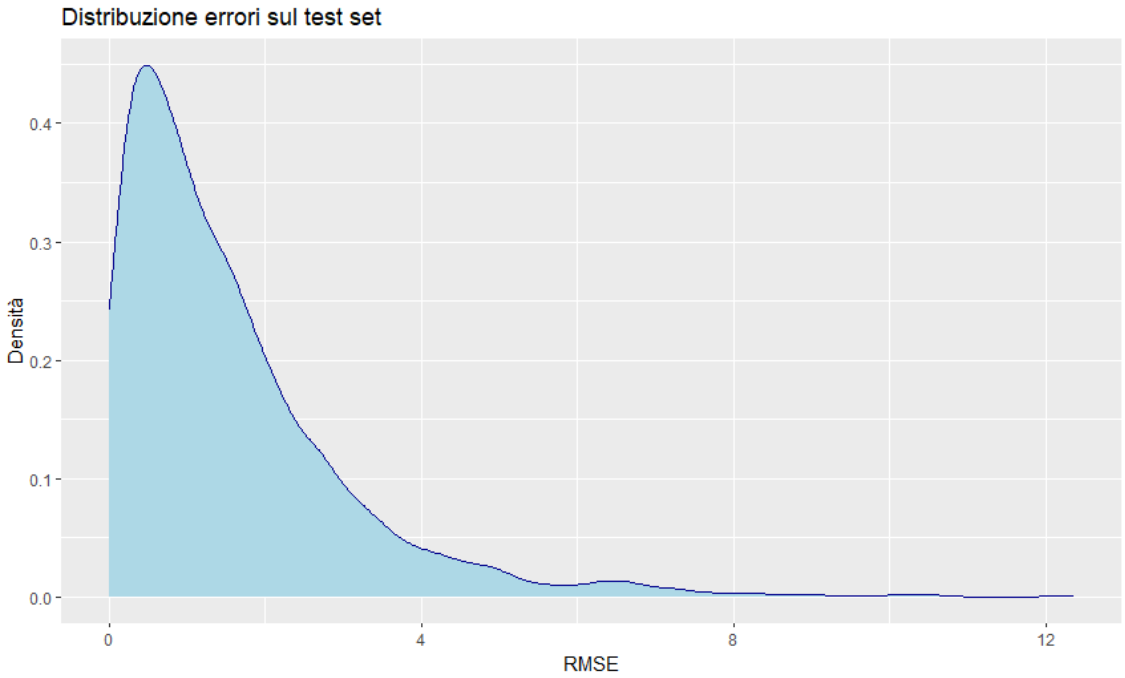

Figura 4.4: Distribuzione del root mean square error sul test-set per il modello finale.

goritmo non è riuscito a cogliere tutta l'efficacia che questa formazione ha dimostrato in campo, con un impressionante 21.3 per quanto riguarda il net rating bayesiano osservato.

4.3 Conclusioni

Alla luce di quanto visto, il clustering basato sul modello mistura risulta appropriato per i dati che sono stati utilizzati. I gruppi individuati per i giocatori sono infatti, per quanto riguarda la loro composizione e le loro caratteristiche, coerenti anche dal punto di vista cestistico e colgono molto bene giocatori "ibridi" grazie al soft-assignment garantito dal modello mistura. Per quanto riguarda invece la composizione dei gruppi relativi alle squadre, l'obiettivo primario era quello di individuare dei cluster che si caratterizzassero per diversi stili di gioco. Tale obiettivo non è stato raggiunto pienamente, in quanto principalmente i 3 gruppi ottenuti mostravano, più che modi diversi di giocare, l'evoluzione generale e la direzione che sta prendendo la NBA, ovvero l'utilizzo sempre più massiccio del tiro da tre punti. Potrebbe essere interessante valutare altre variabili, diverse da quel-

Stagione	Squadra	Quintetto	Errore
16-17	UTA	Favors, Hayward, Gobert, Hood, Exum	0
16-17	MIN	Rush, Rubio, Bjelica, Wiggins, Towns	0
17-18	DET	Tolliver, Moreland, Galloway, Johnson, Kennard	0.003
21-22	LAC	Ibaka, George, Morris, Jackson, Zubac	0.005
15-16	DET	Morris, Harris, Jackson, Drummond, Johnson	0.006

Tabella 4.1: I 5 quintetti del test set con minor errore di previsione in valore assoluto.

Stagione	Squadra	Quintetto	Errore
15-16	LAL	World Peace, Bass, Russell, Nance Jr, Huertas	12.357
16-17	GSW	West, Livingston, Durant, Thompson, Clark	10.44
15-16	CHA	Williams, Hawes, Lin, Lamb, Kaminsky	10.39
18-19	GSW	Iguodala, Durant, Curry, Thompson, Green	10.05
22-23	CLE	LeVert, Mitchell, Allen, Garland, Mobley	9.46

Tabella 4.2: I 5 quintetti del test set con maggior errore di previsione in valore assoluto.

le proposte, per capire se sia possibile catturare maggiormente le differenze nel modo di giocare delle squadre.

L'utilizzo del Random Forest per la previsione del net rating bayesiano è risultato abbastanza efficace, e potrebbe essere uno strumento utile per GM e allenatori NBA nei processi di costruzione della squadra e di analisi delle prestazioni, provando a valutare che impatto potrebbe avere un determinato quintetto in un certo tipo di squadra. Ciò non toglie che, mantenendo il concetto di soft-lineup, si potrebbero utilizzare approcci differenti da quello visto qui per verificare la presenza di modelli più precisi.

Bibliografia

Azzalini, Adelchi e Bruno Scarpa (2012). *Data Analysis and Data Mining: An Introduction*. Oxford University Press.

Banfield, Jeffrey D. e Adrian E. Raftery (1993). «Model-Based Gaussian and Non-Gaussian Clustering». In: *Biometrics* 49, pp. 803–821. URL: https://doi.org/10.2307/2532201.

Biernacki, Christophe, Gillex Celeus e Gerard Govaert (2003). «Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models». In: *Computational Statistics Data Analysis* 41, pp. 561–575.

Bouveryon, Charles et al. (2019). *Model-Based Clustering and Classification for Data Science. With applications in R*. Cambridge University Press.

Breiman, Leo (2001). «Random Forests». In: *Machine Learning* 45, pp. 5–32. URL: https://doi.org/10.1023.

Dempster, Arthur, Nan Laird e Donald Rubin (1977). «Maximum Likelihood from Incomplete Data via the EM Algorithm». In: *Journal of the Royal Statistical Society* 39.1, pp. 1–38.

Goldsberry, Kirk (2019). *Sprawlball: A Visual Tour of the New Era of the NBA*. Mariner Books.

Hastie, Trevor, Robert Tibshiriani e Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Haughton, Dominique M. A. (1988). «On the Choice of a Model to Fit Data from an Exponential Family». In: *Annals of Statistics* 16, pp. 342–355. URL: https://www.jstor.org/stable/2241441.

Hubert, Lawrence e Phipps Arabie (1985). «Comparing Partition». In: *Journal of Classification* 2, pp. 193–218. URL: https://doi.org/10.1007/BF01908075.

Kalman, Samuel e Jonathan Bosch (2020). *NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball and modelling lineup efficiency of soft lineup aggregate*.

Kass, Robert E. e Adrian E. Raftery (1995). «Bayes Factors». In: *Journal of the American Statistical Association* 90, pp. 773–795. URL: https://sites.stat.washington.edu/raftery/Research/PDF/kass1995.pdf.

Keribin, Christine (1998). «Consistent estimate of the order of mixture models». In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 326, pp. 243–248. URL: https://doi.org/10.1016/S0764-4442(97)89479-7.

Leroux, Brian G. (1992). «Consistent Estimation of a Mixing Distribution». In: *Annals of Statistics* 20, pp. 1350–1360. URL: https://doi.org/10.1214/aos/1176348772.

Mariott, F. H. C. (1975). «Separating Mixtures of Normal Distributions». In: *Biometrics* 31, pp. 767–769. URL: https://doi.org/10.2307/2529563.

Maugis, Cathy, Gilles Celeux e Marie-Laure Martin-Magniette (2009). «Variable Selection for Clustering with Gaussian Mixture Models». In: *Biometrics* 65, pp. 701–709. URL: https://www.jstor.org/stable/20640567.

McLachlan, Geoffrey J. e Thriyambakam Krishnan (1997). *The EM algorithm and extensions*. Wiley.

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

Ripley, Brian D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.