

I ruoli dei giocatori NBA stanno scomparendo?

Ball Brothers

Daniel Dummer

Eloisa Pozzi

Federico Tresoldi

Analisi di clustering e classificazione comparativa tra il 2025 e il 2015

Negli ultimi anni il basket professionistico, in particolare nella **NBA**, ha subito una profonda trasformazione strutturale. La tradizionale distinzione dei ruoli in *Guard*, *Forward* e *Center*, storicamente fondata su compiti ben definiti, appare oggi sempre meno rigida. In questo contesto, le etichette di ruolo potrebbero non riflettere più strutture nette e ben separate, ma rappresentare una semplificazione di una realtà continua e multidimensionale.

L'obiettivo di questo lavoro è valutare in che misura i ruoli NBA siano identificabili a partire dalle statistiche di gioco individuali e se tale identificabilità sia cambiata nel tempo. A tal fine, vengono analizzati due dataset, relativi a due stagioni diverse, tratti da *Basketball-Reference*, costruiti in modo omogeneo e contenenti statistiche individuali standardizzate per 36 minuti di gioco, così da ridurre l'eterogeneità legata alla diversa mole di utilizzo dei giocatori, consentendo un confronto diretto tra profili di performance indipendentemente dai minuti disputati. Le due stagioni considerate sono distanti dieci anni: la stagione più recente e la stagione 2014/15, quest'ultima coincidente con l'affermazione di Stephen Curry e dei Golden State Warriors e con l'inizio della più recente rivoluzione tattica legata all'uso sistematico del tiro da tre punti. Questo confronto temporale consente di investigare se e in che misura l'evoluzione dello stile di gioco abbia modificato la struttura dei ruoli.

Alla luce di tali cambiamenti, la domanda di ricerca che guida lo studio è la seguente: *i ruoli tradizionali nel basket NBA esistono ancora come cluster ben separati nei dati di performance? La loro struttura è divenuta progressivamente più sfumata nel tempo?*

1 I dati e preprocessing

I dataset analizzati includono le statistiche individuali di tutti i giocatori scesi in campo nelle stagioni considerate. Per garantire l'affidabilità dell'analisi, sono stati esclusi i giocatori con meno di 20 partite disputate, meno di 500 minuti complessivi e quelli caratterizzati da valori mancanti, così da concentrarsi su profili di gioco stabili e rappresentativi.

Il criterio di selezione delle variabili è stato guidato da considerazioni di dominio e di interpretabilità statistica. Le variabili scelte coprono dimensioni diverse del gioco: efficienza realizzativa, presenza difensiva e gioco vicino al ferro, capacità di creazione e gestione del pallone e orientamento perimetrale. Non sono state incluse variabili di volume, quali punti segnati o tentativi complessivi, poiché fortemente dipendenti dal tempo di utilizzo. Allo stesso modo, si è evitato l'impiego di metriche composite o avanzate, che incorporano ipotesi di modello e pesature non direttamente osservabili, selezionando invece statistiche semplici, interpretabili e direttamente riconducibili alle componenti fondamentali del gioco. Sulla base dei suddetti criteri, sono state selezionate le seguenti variabili:

| Variabile | Descrizione | Media (Var) - 25 / 15 |
|-----------|---------------------------------------------------------------------------------------------|-----------------------------|
| FG% | Percentuale di tiri realizzati sul totale dei tentativi. | 0.46 (0.003) / 0.44 (0.002) |
| BLK | Numero di stoppate, indicatore della presenza difensiva. | 0.71 (0.34) / 0.61 (0.32) |
| TRB | Rimbalzi totali, misura dell'impatto nelle fasi di gioco prossime al ferro. | 6.5 (7.38) / 5.9 (3.8) |
| AST/TOV | Variabile costruita ad hoc che sintetizza l'efficienza nel playmaking. | 2.02 (0.5) / 1.6 (1.4) |
| AST/TRB | Indicatore della propensione alla creazione di gioco rispetto al contributo sotto canestro. | 0.73 (0.29) / 0.73 (0.37) |
| X3PA/FGA | Rapporto tra tentativi da tre punti e tentativi totali (orientamento perimetrale). | 0.45 (0.03) / 0.29 (0.04) |

Tabella 1: Descrizione delle variabili selezionate e statistiche descrittive per le stagioni 2015 e 2025.

Prima del clustering è stata condotta un'analisi esplorativa della struttura di correlazione tra le variabili, al fine di valutare la presenza di dimensioni latenti e il grado di separazione dei profili di gioco.

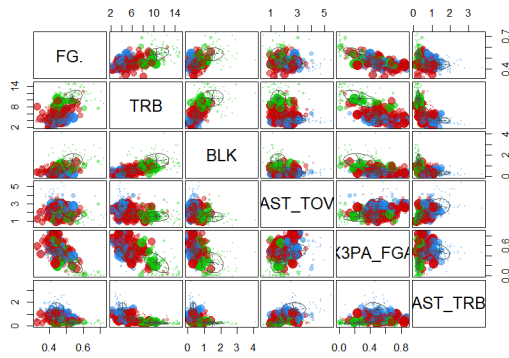


Figura 1: Confronto delle matrici di correlazione

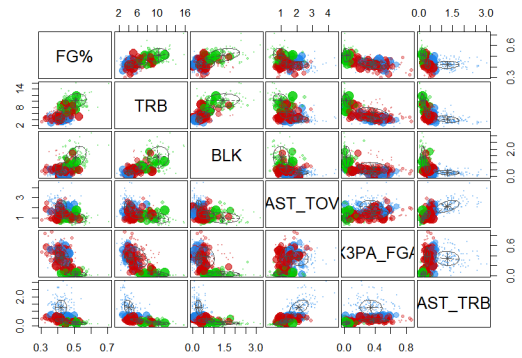
Le matrici di correlazione evidenziano nel 2015 una struttura più modulare, con associazioni più forti all'interno dei blocchi di variabili riconducibili al gioco interno (BLK, TRB, FG%) e perimetrale (AST/TRB, AST/TOV, X3PA/FGA). Nel 2025 tale struttura risulta leggermente attenuata, suggerendo una maggiore sovrapposizione tra le dimensioni latenti e una riduzione della separazione tra profili di gioco.

2 CLUSTERING

Abbiamo utilizzato la tecnica del *Model-Based clustering* sulle variabili selezionate. Per entrambi i modelli è stato fissato $G = 3$, in coerenza con l'ipotesi dei ruoli tradizionali e supportato dal criterio ICL. Per il 2025 viene selezionato un modello EVE, mentre per il 2015 un modello EVV.



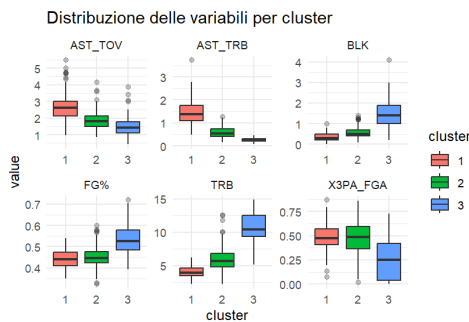
(a) Stagione 2025



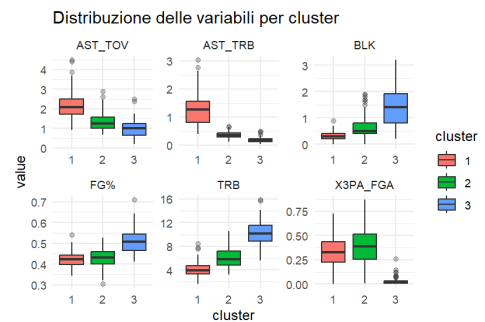
(b) Stagione 2015

Figura 2: Confronto dei grafici di incertezza

Il confronto tra i grafici di incertezza mostra minore entropia media delle posterior probability nel 2015, con aree di sovrapposizione tra i cluster più contenute. Nel 2025 l'incertezza risulta invece più diffusa, soprattutto nelle regioni intermedie dello spazio delle variabili, indicando una maggiore continuità tra i profili di gioco. Possiamo visualizzare il profilo dei cluster per ogni variabile.



(a) Profili medi dei cluster - 2025



(b) Profili medi dei cluster - 2015

I profili medi dei cluster risultano coerenti con un'interpretazione funzionale dei ruoli. Il cluster identificato come G presenta valori più elevati nelle variabili di playmaking (AST, AST/TOV), a conferma di un ruolo prevalentemente perimetrale e orientato alla creazione di gioco. Il cluster C si distingue per valori più alti di BLK, TRB e FG%, delineando un profilo maggiormente associato al gioco vicino al canestro. Il cluster F assume una posizione intermedia nella maggior parte delle variabili, fungendo da ponte tra gli altri due gruppi. La differenza più marcata si osserva nella variabile X3PA/FGA: nel 2015 essa assume valori prossimi allo zero per il cluster dei centri.

2.1 Bontà del clustering

Per verificare la bontà dei clustering possiamo utilizzare vari indici:

Tabella 2: Indici di bontà del clustering

| Stagione | CER | ARI | Commento |
|----------|------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2025 | 0.32 | 0.28 | Entrambi gli indici mostrano una qualità discreta del clustering, coerente con l'ipotesi di una struttura dei ruoli più continua e meno nettamente separata. |
| 2015 | 0.23 | 0.48 | La qualità del clustering risulta sensibilmente migliore, a confermare una separazione tra le componenti più marcata . |

Tabella 3: Confronto tra le matrici di confusione

| Stagione 2025 | | | | Stagione 2015 | | | |
|------------------------|----|-----|----|------------------------|----|----|-----|
| Prediction \ Reference | C | F | G | Prediction \ Reference | C | F | G |
| C | 56 | 23 | 0 | C | 37 | 31 | 0 |
| F | 5 | 108 | 88 | F | 6 | 97 | 30 |
| G | 0 | 7 | 86 | G | 0 | 17 | 141 |

Le matrici di confusione evidenziano una diversa separabilità dei cluster nelle due stagioni. Nel 2025 la maggior parte degli errori di classificazione si concentra tra Guardie e Ali, indicando una forte sovrapposizione tra questi due profili; al contrario, i Centri risultano complessivamente meglio identificati. Nel 2015 la struttura appare più definita: pur permanendo una certa ambiguità nella classe delle Ali, questo confronto conferma che dieci anni fa i profili di gioco risultavano più distinti.

Il confronto tra le matrici di distanza KL (Tabella 4) evidenzia nel 2015 una separazione più netta tra le componenti associate ai ruoli di guardie, ali e centri. Nel 2025 le distanze che coinvolgono le ali si riducono, indicando un avvicinamento sia alle guardie sia ai centri.

Tabella 4: Confronto tra le matrici di distanza KL

| Stagione 2025 | | | | Stagione 2015 | | | |
|---------------|--------|-------|--------|---------------|-------|-------|-------|
| | G | F | G | | G | F | C |
| G | 0 | 14.46 | 232.05 | G | 0 | 57.07 | 192.4 |
| F | 14.46 | 0 | 28.03 | F | 57.07 | 0 | 42.14 |
| C | 232.05 | 28.03 | 0 | C | 192.4 | 42.14 | 0 |

3 CLASSIFICATION

L’approccio EDDA, tramite `mixmodLearn`, ha identificato il modello $\mathcal{N}(p_k, L, C)$ come ottimale per entrambi i dataset. Il confronto dei parametri di validazione evidenzia una maggiore separabilità dei ruoli nel 2015, con un errore di CV di 0.19 e un’accuracy di 0.79, rispetto ai valori di 0.23 e 0.76 riscontrati nel 2025. Tale riduzione delle prestazioni predittive nel periodo recente conferma una crescente sovrapposizione tra le classi, indicativa di una maggiore fluidità nei ruoli dei giocatori.

| Confusion matrix 2025 | | | | | Confusion matrix 2015 | | | | |
|------------------------|----|----|----|--|------------------------|---|----|----|--|
| Prediction \ Reference | C | F | G | | Prediction \ Reference | C | F | G | |
| C | 10 | 3 | 0 | | C | 7 | 4 | 0 | |
| F | 2 | 15 | 2 | | F | 2 | 20 | 8 | |
| G | 0 | 6 | 37 | | G | 0 | 2 | 29 | |

Tabella 5: Confronto delle matrici di confusione per i modelli di classificazione.

L’errore di classificazione principale riguarda le classi *Guard* e *Forward*, che rappresentano i ruoli più prossimi in termini di profilo statistico. La loro distinzione si basa spesso su differenze sottili nelle variabili considerate, come il grado di coinvolgimento nella gestione del pallone e la propensione al tiro da tre punti.

3.1 Rappresentazione della classificazione

I grafici (Figura 4) riportano la proiezione delle osservazioni nello spazio bidimensionale definito dalle variabili `TRB` e `X3PA_FGA`. Le aree colorate rappresentano le regioni di classificazione indotte dal modello stimato, ossia l’assegnazione a ciascun cluster sulla base della massima probabilità a posteriori, mentre i punti corrispondono alle osservazioni reali, colorate in base al cluster di appartenenza.

In entrambi i grafici emerge una struttura di classificazione ben definita lungo la dimensione `TRB`, che risulta essere la variabile maggiormente discriminante tra i cluster. Il confronto tra i modelli mostra come l’introduzione di una maggiore flessibilità nella struttura di covarianza conduca a frontiere di classificazione più curve e adattive, soprattutto nelle regioni di sovrapposizione tra i cluster. Emerge come, nel 2015, le classi

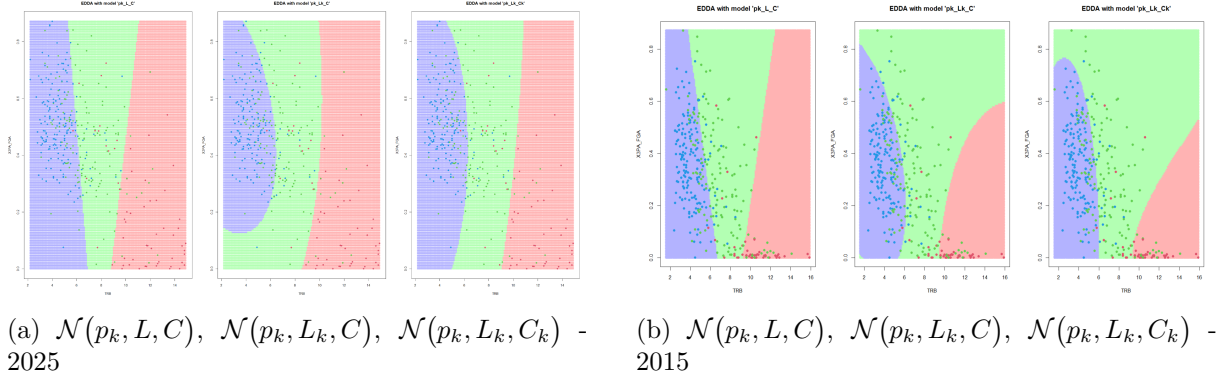


Figura 4: Confronto delle regioni di classificazione

risultino complessivamente più definite e meglio separate rispetto all'analisi condotta sul dataset più recente. In particolare, le regioni di classificazione appaiono meno sovrapposte e i gruppi presentano una maggiore compattezza nello spazio delle variabili considerate, suggerendo una più elevata omogeneità interna ai cluster.

3.2 Classificazione con *Mixture Discriminant Analysis models*

La determinazione del numero ottimale di componenti G per la classificazione MDA è stata effettuata tramite una *20-fold cross-validation* testando l'intervallo $G \in \{1, \dots, 10\}$ con struttura di covarianza VVV. Il criterio di minimizzazione del *misclassification error* medio ha identificato $G = 2$ come valore ottimale per entrambe le annualità, registrando un errore di circa 0.27 nel 2025 e 0.20 nel 2015. Il minor errore riscontrato nel 2015 conferma statisticamente una struttura dei ruoli maggiormente polarizzata e meglio separata nello spazio delle covariate, riflettendo una stratificazione tecnica tra profili specialistici come *playmaker* puri, difensori *3&D* e *rim protectors* che risulta sensibilmente più sfumata nel contesto cestistico attuale.

4 Conclusioni

L'uso congiunto di metodi di clustering basati su modelli di mistura e di classificazione supervisionata ha consentito di analizzare la struttura dei ruoli NBA sia in ottica esplorativa sia predittiva. Il confronto tra le stagioni 2014/15 e 2024/25 evidenzia una riduzione della separabilità statistica tra i profili di performance individuale, come indicato dal peggioramento degli indici di bontà del clustering e dall'aumento dell'errore di classificazione nel periodo più recente. In particolare, la crescente sovrapposizione tra le classi di guardia e ala suggerisce una convergenza dei rispettivi vettori di caratteristiche osservate, mentre il ruolo di centro mantiene una maggiore stabilità relativa. Nel complesso, i risultati supportano l'ipotesi di una transizione da una struttura dei ruoli discretamente separata a una rappresentazione più continua, coerente con il fenomeno del *positional-less basketball*.