LUISS Guido Carli
LIBERA UNIVERSITÀ INTERNAZIONALE DEGLI STUDI SOCIALI

Dipartimento di Impresa e Management

Cattedra di Statistica

LA STATISTICA NEL BASKET: LA NUOVA SCIENZA DELLA VITTORIA ATTRAVERSO LA CLUSTER ANALISYS

RELATORE
Livia De Giovanni

CANDIDATO
Alfonso Gebbia 218711

ANNO ACCADEMICO 2019/2020

INDICE

INTRODUZIONE

Il mondo in cui viviamo oggi, è un ambiente influenzato principalmente da due fattori: incertezza e dinamicità; proprio per questo aspetto incerto, molte industrie in diversi settori hanno sentito l'esigenza di adottare approcci più analitici al processo decisionale, in modo da poter ridurre i rischi connessi ad esso. Tra i vari settori, nessuno ha in corso gli stessi tipi di iniziative analitiche del settore degli sport professionistici. L'industria dello sport è infatti un settore in continua crescita, ed oggi più che mai, si trova a implementare processi e strumenti innovativi, adattando i dati e le pratiche di analisi, per aggiungere valore alle diverse aree delle loro attività.

"Management of data and application of predictive models and the use of information systems to gain a competitive advantage (adapted from Alamar, 2013)" [1].

Lo scopo di questo elaborato è, pertanto, dimostrare come l'implementazione dell'analisi sportiva possa migliorare i risultati sul campo, analizzando le performance delle squadre di basket della NBA (National Basketball Association) tramite la *Cluster Analysis*, un insieme di tecniche di analisi multivariata dei dati, in grado di raggruppare elementi omogenei in un insieme (*cluster*) di dati. Nello specifico, le unità statistiche sono le 30 squadre NBA, mentre le variabili scelte sono i "*Four Factors of Basketball Success*".

Questo elaborato è diviso in tre capitoli, di cui il primo definisce il fenomeno dell'analisi sportiva identificandone i campi di applicazione e descrivendo i metodi utilizzati. Il fenomeno viene descritto partendo dalle prime teorie nel mondo del baseball di Bill James fino ad arrivare al mondo del basket con Dean Oliver.
Il secondo capitolo, invece, introduce la metodologia della *Cluster Analysis*, focalizzando l'attenzione sulla scelta delle variabili da adottare ed i criteri di misurazione connessi ad esse, applicati attraverso i metodi di raggruppamento. Infine, utilizzando opportuni indici, vi è la valutazione dei risultati derivanti dai campionamenti.
Per ultimo, il terzo capitolo, applica il metodo non gerarchico del k–means sui 30 team NBA, cercando di analizzare il loro andamento durante la stagione 2019/20 per effettuare delle previsioni di vittorie e sconfitte, fornendo opportuni risultati.
Al termine, vi saranno le conclusioni dell'elaborato.

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

CAPITOLO 1 – L'Analisi Sportiva

1.1 Definizione

L'analisi sportiva, si riferisce all'uso di dati e statistiche avanzate, rappresentando al tempo stesso uno strumento in grado di ottenere un vantaggio sportivo competitivo, il cui obbiettivo non è quello di sostituire i giocatori o gli allenatori, ma aumentarne la loro prestazione.
Il settore degli sport professionistici ha in comune determinati attributi che permettono di effettuare un confronto:

1. I clienti sono analitici sul prodotto offertogli dal settore
2. L'esistenza di svariati fattori analitici da prendere in considerazione come la prestazione del gioco/giocatori, le relazioni con i clienti e la gestione aziendale.
3. L'industria dispone di canali di output multipli per l'analisi
4. Il lavoro del settore dell'analisi è stato celebrato in articoli/film (vedi "*Moneyball*" attraverso teorie di Bill James)
5. La quantità di dati è in continuo aumento
6. La rotazione manageriale tra le squadre ha portato ad una facile diffusione del know-how.

A differenza di come si ci potrebbe aspettare, gli investimenti in team di analisi sportiva sono però notevolmente inferiori. Ciò è dovuto, in primis perché molte volte i direttori o gli allenatori sono troppo legati all'esperienza durante il proprio processo decisionale; ed in secundis perché le squadre sportive possono essere paragonate a delle piccole-medie imprese, le quali preferiscono impiegare il loro capitale nell'acquisto di macchinari o tecnologie, piuttosto che nell'assunzione di un team di esperti, per problemi legati ai costi.

1.2 Le categorie di analisi

L'analisi sportiva si divide in 3 grandi aree:

1. Analisi delle prestazioni del giocatore e del gioco/dei risultati
2. Analisi della salute e degli infortuni dell'atleta
3. Analisi dei business sportivi.

1.2.1 Analisi dei business sportivi

Di queste tre tipologie di analisi, sicuramente l'ultima è quella meno analitica e focalizzata sul campo di gioco, poiché si propone di sottoporre ad analisi diversi ambiti che fanno da contorno ad un evento sportivo, ma che sono importanti per la sua buona riuscita e per il business della squadra. Questi ambiti includono:

a) Andamento dei prezzi dei biglietti/tickets
b) Valore dei fan *(Fan Equity)*
c) Social media *(SM).*

Sono svariati, i metodi che le squadre possono utilizzare nella fissazione del prezzo dei biglietti per ottenere entrate aggiuntive. Nella MLB (Major League Baseball) per esempio statisticamente 26 squadre su 30 utilizzano il metodo *flexible pricing,* consistente nell'offerta di prezzi variabili, in modo che alcuni biglietti costino più degli altri, a seconda che, la partita sia di maggiore interesse o meno. Il *flexible pricing,* tuttavia, è stato abbandonato in vista di un nuovo metodo più sofisticato, il *dynamic pricing,* consistente nell'offrire prezzi dinamici che cambiano durante la stagione in base a fattori quali il rendimento della squadra, il successo dell'avversario della squadra e anche fattori come il tempo meteorologico. Indipendentemente dal metodo utilizzato, lo scopo principale è sempre la soddisfazione del cliente, ponendolo al centro del processo decisionale. L'obbiettivo finale, allora, non è trovare il giusto fan per il tipo di biglietto ma quest'ultimo per il giusto fan.

Gli Orlando Magic (Franchigia del basket NBA) per esempio:

"decision tree models that bucket subscribers into three categories: most likely to renew, least likely, and fence-sitters. The fence-sitters then get the customer service department's attention come renewal time" [2].

Il secondo aspetto legato alla business analysis è il valore dei fan, ovvero la loro fidelizzazione verso la squadra ed i valori che essa porta con sé.
La *Fan equity* è composta da tre elementi:

a) *Value*, che si riferisce alla qualità del servizio ed al valore percepito da parte del fan/cliente
b) Brand
c) *Relationship,* che include valori etici come la fiducia, lealtà, supporto e passione per la squadra.

[2] https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf

Il terzo ed ultimo aspetto fa invece riferimento ai Social Media (*SM*), i quali costituiscono un efficace metodo per la misurazione del sentimento dei fan ed il loro attaccamento alla squadra. I Social Media, permettono di essere presenti in un luogo, come una partita, pur non essendo presenti fisicamente, e quindi offrono la possibilità di interagire via remoto attraverso strumenti come i "*like*", "*commenti*" e "*hashtag*" (es. #*Streightinumbers* dei Golden State Warriors) che permettono di essere parte di un movimento e collegare i fan gli uni con gli altri.
La raccolta di queste informazioni permette alle squadre sportive di stilare dati per fini monetari e non monetari. Quelli non monetari, vengono rappresentati maggiormente dalla "*Sentiment Analysis*" (una forma più veloce di feedback), e le metriche usate per misurare il successo sono rappresentate da strumenti come le "*impressions*", la "*copertura*" ed i "*followers*", le quali individuano quanto il cliente ha interagito con il post, esprimendo come i contenuti del marchio risuonano con il loro pubblico. Quelli monetari invece, possono identificarsi principalmente nella "*conversion*", ovvero un procedimento per fare mercato attraverso i Social Media. Questi ultimi, diventano così uno strumento per fare generare profitti, e se i Social Media generano profitti allora il *ROI* (*Return of Investment*), utilizzato per misurare il valore dell'investimento, può essere misurato in diversi modi nei Social Media. Per fare ciò, bisogna rispettare 6 passaggi:

1) Impostare gli obbiettivi
2) Selezionare il pubblico
3) Definire l'investimento (come i salari ed i costi)
4) Benchmark
5) Selezionare gli strumenti giusti per misurare gli obbiettivi
6) Trasformare i dati raccolti in decisioni pratiche

L'aspetto più importante, è sicuramente la scarsa capacità dei dirigenti di essere flessibili ai possibili mutamenti dei Social Media, ed inoltre l'essere consapevoli dell'enorme rilevanza che questi hanno nel mondo odierno. Molti senior dirigenti, sono cresciuti nella generazione dove i Social Media non esistevano, e per tale motivo, si sono dovuti adattare agli ingenti investimenti e alla creazione di team appositi per gestire tali aree all'interno delle società sportive.

Un'altra pratica di frontiera è l'ottimizzazione della spesa e dei programmi di marketing per le promozioni orientate ai fan. Uno degli aspetti più difficili dell'ottimizzazione del marketing è appunto la raccolta di tutti i dati rilevanti. Gli Orlando Magic, per esempio, hanno esplorato quest'applicazione, attraverso il corretto uso dei dati wi-fi per comprendere al meglio il coinvolgimento del fan, installando una rete wi-fi interna all'arena sportiva.

L'Analisi sui business sportivi, ci pone allora, di fronte a due obbiettivi:

1) La *fidelizzazione del cliente*, attraverso la *brand awareness*
2) Il *risultato delle entrate*, attraverso le vendite dei biglietti/tickets e le vendite e-commerce del merchandising.

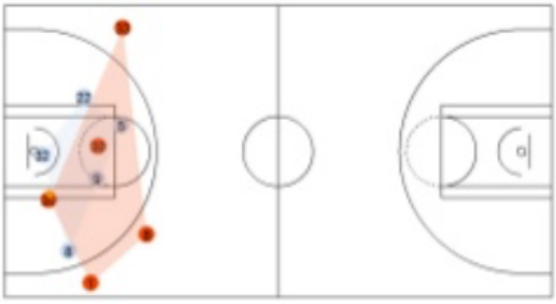
Entrambi, come abbiamo visto, presentano evidenti limiti e problematiche, le quali attraverso l'implementazione dell'analisi possono essere studiate ed interpretate per migliorare le prestazioni.

1.2.2 Analisi della salute e degli infortuni dell'atleta

Il secondo tipo di analisi dello sport, che affronta questo elaborato, è l'analisi della salute e delle lesioni. Per raggiungere i propri traguardi, una squadra non può fare a meno delle sue componenti, ovvero i giocatori, ed è importante che essi siano in salute per un massimo rendimento sul campo. Questo è il motivo principale per cui la prevenzione di infortuni e malattie, è un focus naturale per l'uso di dati e analisi. Sebbene quest'area sia fonte di studio da decenni, ci sono pensieri contrastanti fra i fisiologi su quale sia la metodologia più adatta per la previsione di lesioni. Da un lato c'è chi demanda gli infortuni alla non corretta alimentazione e alle ore di riposo fisico dell'atleta; dall'altro c'è invece chi si concentra sull'allenamento con i pesi o su altre attività fisiche.
Con il progredire della tecnologia e degli investimenti in campo tecnologico, si è capito che una grande quantità di dati che supportano le prestazioni del giocatore e della squadra, possono anche essere utilizzati per l'analisi della salute dell'atleta. I dati, infatti, possono essere utilizzati per valutare e monitorare il livello di attività fisica e lo stress legato ad esse, il quale potrebbe causare lesioni in futuro. Pratica ormai diffusa è l'utilizzo da parte dei team sportivi di veri e propri sistemi *GPS (Global Positioning Systems*), tecnologie in grado di monitorare i movimenti, velocità, schemi di sonno ed altri indicatori biometrici dei singoli giocatori, ottenendo una vasta gamma di metriche. Come ha affermato un allenatore dei Jacksonville Jaguars, queste metriche possono portare a spunti interessanti:

"We look at total odometer, high intensity yards, accelerations, decelerations, contact load, and PlayerLoad..[it shows] the lack of any true speed in our game. It really just shows our game is truly acceleration and deceleration..." [3]

[3] https://www.catapultsports.com/blog/jacksonville-jaguars



Nell'analisi della salute, le statistiche descrittive dell'attività dei singoli giocatori basate su video, *GPS* o dispositivi biometrici sono le analisi di posta in gioco per gli sport professionistici. Data la quantità di dati disponibili ora, le evidenti analisi di frontiera sono quelle che coinvolgono allerta e previsione degli infortuni.

Caso MilanLab

Un caso esemplare di ciò è rappresentato dalla squadra di calcio italiana AC Milan, la quale attraverso la creazione del centro MilanLab, ha iniziato ad utilizzare le tecniche predittive attraverso l'analisi della salute ed il condizionamento dei giocatori. Il MilanLab ha adottato dal 2002 ad oggi una grande varietà di tecnologie sportive, incluso il video sportivo (*Programma SportVU Israeliano*). Il programma tiene traccia di circa 60.000 punti dati per ogni giocatore ed i dati mentali, biochimici e musco-strutturali, vengono raccolti ogni due settimane su otto apparecchiature scientifiche, le quali creano degli avvisi se i dati di un giocatore non rientrano nell'intervallo previsto. I dati e le analisi misurano l'idoneità di ciascun giocatore e vengono utilizzati per prevedere la probabilità di lesioni gravi.

Dai dati raccolti e dall'analisi di previsione, si è stimato che l'AC Milan abbia subito una riduzione del 90% degli infortuni nel 2003, e da allora in poi i casi sono rimasti bassi. Il caso MilanLab non è il solo nel panorama sportivo, ma anche diverse squadre di altri sport stanno iniziando ad esplorare la biomeccanica. La squadra di baseball di San Francisco Giants, ad esempio, utilizza l'analisi video della biomeccanica per aiutare la riabilitazione del giocatore infortunato, mettendo a confronto le analisi rilevate con quelle precedenti o con altri giocatori simili.

Nonostante il lavoro di spicco di squadre come l'AC Milan e i San Francisco Giants, l'analisi della salute e degli infortuni dei giocatori è ancora agli inizi, ma le fonti di dati si stanno espandendo rapidamente e presto consentiranno analisi predittive ed una descrizione ancor più accurata.

1.2.3 Analisi delle prestazioni del giocatore e del gioco/dei risultati

La terza e ultima categoria di analisi, fa riferimento alle prestazioni del giocatore e del risultato sportivo della squadra. In effetti, quando la maggior parte dei fan pensa all'analisi negli sport, è proprio questo

l'ambito di riferimento. L'analisi delle prestazioni sportive è composta da numerosi aspetti ognuno molto legato dall'altro ed i cui principali ambiti di applicazione sono da ricercare:

a) Nella selezione dei migliori giocatori possibili
b) Nello schierare le migliori squadre possibili
c) Nel prendere le migliori decisioni possibili sul campo.

Questi tre ambiti, oltre a costituire il raggio d'azione dell'analisi, identificano allo stesso tempo gli obbiettivi principali che si propongono di raggiungere i team sportivi che adottano questi schemi.

1.3 Baseball

1.3.1 Bill James

Molti studiosi e statistici hanno dedicato gran parte del loro lavoro accademico, nell'approfondimento delle teorie statistiche che stanno alla base dei risultati sportivi e tra questi, merita di essere citato in questa schiera George William (Bill) James, precursore dell'analisi sportiva moderna e padre di alcune delle teorie statistiche che hanno rivoluzionato il modo di pensare e prendere decisioni nell'intero panorama del mondo del baseball. Il suo approccio, definito "*Sabermetrica* (analisi empirica del baseball)", analizza e studia scientificamente il baseball attraverso l'uso di dati statistici, nel tentativo di determinare perché le squadre *vincono o perdono*. Il libro annuale, intitolato "*The Baseball Abstract*", largamente criticato e combattuto nei primi anni della pubblicazione, rappresenta il vero e proprio punto di partenza dell'analisi sportiva moderna. In questi scritti, lo statista, si proponeva l'obbiettivo di fornire un quadro chiaro e semplice agli appassionati del baseball, analizzando ben 18 categorie di informazioni statistiche sui punteggi delle partite delle stagioni precedenti.

"*Science is like a clean slate, and that's what makes it effective, "James said in an interview. "You can also be a physics graduate and think that Einstein was wrong, but if you bring a thesis supported by concrete facts, people will listen to you. And that's exactly what I tried to do with baseball: you can be an expert as much as you want, but the facts are clear."* [4] *(James)*

In verità, il Baseball è sempre stato uno sport di numeri, ma a differenza del passato, James e altri studiosi si sono chiesti quali fossero i numeri che importassero davvero, ed un particolare articolo è considerato il pilastro fondante delle sue teorie. In questo scritto, lo statista si interrogava su "quali lanciatori o catcher consentivano ai corridori di *rubare la maggior parte di basi"*, giungendo alla formulazione del suo algoritmo vincente, ovvero la prima citata "*Sabermetrica*" e definita dallo stesso James come la "*ricerca per*

[4] James, Bill. The 1985 Baseball Abstract. Ballantine Books, 1985.

la conoscenza oggettiva sul baseball". L'algoritmo sviluppato, dall'ex studente di Kansas, viene eseguito valutando i giocatori in ogni aspetto del gioco, in particolare in battuta (*batting*), lancio (*pitching*) e *fielding*, e tali misure di valutazione sono generalmente formulate in termini di vittorie di squadra in quanto le statistiche precedenti furono ritenute inefficaci.

Traditional measuraments

La misura tradizionale delle prestazioni in battuta era considerata formata dal rapporto tra colpi e numero totale di battute (*Batting Average*), ma James scoprì che essa era difettosa, in quanto ignorava qualsiasi altro modo in cui un battitore poteva raggiungere la base oltre ad un colpo.

Batting measurements

Da questo presupposto, James giunse alla creazione della "*On-base percentage*", ovvero la percentuale sulla base, la quale prende in considerazione le camminate (*walks*) e le piazzole (*hit-by-pitches*).

Formula On-base percentage

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

- *H = Hits*
- *BB = Based on Balls (walks)*
- *HBP = Times Hit By a Pitch*
- *AB = At bats*
- *SF = Sacrifice Flies*

Un altro problema con la misura tradizionale della battura media (Batting Average) era l'impossibilità di distinguere se i colpi (*hits*) erano singoli, doppi, tripli oppure home runs, attribuendo indistintamente lo stesso valore ad ognuno di essi. Pertanto, James pensò di creare una misura che fosse in grado di distinguere questi quattro risultati, ovvero la "*Slugging percentage*" (*percentuale di rallentamento*), il cui risultato è dato dal rapporto tra il numero totale di basi di tutti i colpi ed il numero totale di "*time at bat*" (*AB*).

Formula Slugging percentage

$$SLG\% = \frac{TB}{AB} = \frac{1B + (2 \times 2B) + (3 \times 3B) + (4 + HR)}{AB}$$

- *TB = Total Bases* [5]
- *AB = At Bat*

Queste due misure sabermetriche migliorate, costituiscono aspetti importanti da misurare in una battuta, e sono state combinate per creare la moderna statistica *OPS* (*On-base plus slugging*) rappresentata dalla somma della *On-base percentage* e della *Slugging percentage*, diventata utile per confrontare giocatori prevedendo i tiri segnati da un certo giocatore.

Formula OPS (On-base plus slugging)

$$OBP = \frac{[H + Base\ on\ balls\ (BB) + Hit\ by\ pitch\ (HBP)]}{[AB + BB + HBP + Sacrifice\ flies\ (SF)]}$$

$$SLG = \frac{[Total\ Bases\ (TB)]}{[AB]}$$ [6]

Pitching measurements

Per quanto riguarda il "lancio (*pitching*)", la misura tradizionale delle prestazioni, era considerata la *earned run average (La media della corsa guadagnata)*, la quale veniva calcolata dividendo il numero di corse guadagnate per il numero di innings lanciati e moltiplicando per 9 a causa dei 9 innings, fornendo il numero di tiri consentiti da un lanciatore per partita. James tuttavia, trovò qualcosa di difettoso anche in questa statistica, argomentando che il problema risiedeva nel non separare l'abilità del lanciatore dalle abilità dei lanciatori con cui si trova a giocare. La *earned run average*, non rappresenta però l'unica statistica legata al lancio che lo studioso mise in crisi, ma ve ne fu una seconda, la cosiddetta *"percentuale di vincita",* confutando il problema nel fatto che essa dipendeva dalle prestazioni dei compagni di squadra del lanciatore (*fielders*) sul campo.

I sabametrici hanno tentato di trovare diverse misure delle prestazioni di lancio che non includano le prestazioni dei lanciatori coinvolti. Una delle prime sviluppate, e delle più usate, è la *walks plus hits per inning pitched (WHIP)*, che sebbene non sia completamente indipendente dalla difesa, tende ad indicare quante volte è probabile che un lanciatore metta un giocatore sulla base, e quindi l'efficacia dei battitori contro un particolare lanciatore *nel raggiungere la base*.

[5] TB è ponderata per il tipo di hit (x2 for 2B, x3 for 3B, x4 for HR).
[6] TB è ponderata per il tipo di hit x2 for 2B, x3 for 3B, x4 for HR.

1.3.2 Moneyball (L'arte di vincere) – Il caso Oakland A.

Le teorie di Bill James, e della Sabermetrica, hanno rappresentato per molti appassionati del baseball e non solo, un nuovo modo di vedere e concepire lo sport, tanto che alcune delle statistiche e teorie implementate da James furono riprese da alcuni dirigenti della MLB (Major League Baseball) nel processo di selezione dei giocatori. Questo è il caso degli Oakland A, squadra della Major League, dipinta dal famoso film Moneyball (L'arte di vincere)[7], che nella stagione sportiva 2001/02, conquistarono il record di vittorie in campionato. A quel tempo, la squadra degli Oakland Athletics, non disponeva di un grande budget per finanziare la sua campagna acquisti e non era neppure ai primi posti nel ranking della lega per potersi accaparrare le migliori scelte al draft di inizio anno; vi erano quindi tutti i presupposti per una stagione deludente. La svolta avvenne, quando l'allora general manager della squadra, Billy Beane, invece di affidarsi all'esperienza ed all'intuizione degli osservatori, caratteristiche conservatrici che rendono ancora oggi difficile lo sviluppo dell'analisi, decise di adottare per la selezione dei giocatori quasi esclusivamente l'approccio sabermetrico, ed in particolare l'utilizzo della *OBP* (*On-base percentage).* Trovando giocatori con un'alta *OBP*, ma con caratteristiche che avevano portato gli osservatori (*scouters*) delle altre squadre a scartarli, Bean riuscì a mettere insieme una squadra con molto più potenziale rispetto a quella che il budget gli prospettava. Nonostante un inizio di campionato sottotono, i risultati non tardarono ad arrivare e gli Oakland Athletics in quella stagione ottennero un record di 19 vittorie consecutive, dimostrando a tutto il panorama sportivo che le teorie di James, tanto criticate, avessero dei solidi fondamenti e cosa più importante che portavano a risultati concreti. L'esempio degli Oakland A. rappresenta soltanto uno di svariati casi in cui vennero riprese le teorie di James, le quali non si fermarono all'applicazione solamente nel mondo del baseball ma coinvolsero altri ambiti sportivi.

Un analista di NFL scriveva:

"I am working against a culture of indifference toward analytics. Despite that, I am trying to find the one or two things the coaches will use. Every time I engage them—and that's a struggle in itself--I throw out several things. If they accept one, I consider myself successful. Football is a good old boy culture that sees security in the status quo, and it has been hard for analytics to make a dent in it". [8]

Sport come il calcio ed il basket, ad esempio, hanno fatto tesoro delle scoperte di James e le hanno utilizzate come punto di partenza nello sviluppo di nuovi algoritmi.

La prima sfida, che si dovette affrontare, per lo sviluppo di tali teorie, fu sicuramente quella dell'*accettazione*. Come avviene nel mondo degli affari, così nel mondo dello sport il ruolo della *leadership*

[7] Lewis, Michael. Moneyball: The Art of Winning an Unfair Game. New York: W.W. Norton, 2003. Print.
[8] Intervista tenuta da un analista della NFL

risulta essere il singolo fattore comune nel rendere un team di successo attraverso l'analisi. Il processo applicativo, tuttavia, si è differenziato nei due sport prima citati.

1.4 Calcio

Nello specifico caso del calcio, un altro aspetto che rese difficile lo sviluppo dell'analisi, fu la cultura dell'allenatore, il quale si dimostrava nella maggioranza dei casi *"conservatore"*. Questo sport inoltre, presenta un "*handicap"* nella valutazione e nell'analisi, ovvero la bassa frequenza del punteggio; quindi fattori tattici come il possesso ed i passaggi completati, diventano strumenti chiave e metriche per il processo d'analisi. Tali premesse, costituiscono per appunto le linee guida che spiegano l'obbiettivo primario di tale analisi, ossia l'attenzione principale sull'attività fisica e sull'idoneità dei giocatori, concentrandosi su ciò che è accaduto, piuttosto che sul perché è accaduto o su ciò che potrebbe accadere in seguito.

Nonostante tali difficoltà applicative, alcune squadre europee come ad esempio il West Ham United, appartenente alla Premier League (1° campionato inglese), sono relativamente avanzate in termini di analisi delle prestazioni ed hanno persino reso disponibili i dati sulle prestazioni ai fan per analisi open source. Queste squadre, sono infatti, state in grado di superare la sfida dell'accettazione, ed abbracciando tutti i giocatori e gli allenatori l'analisi, hanno ottenuto i vantaggi ad essa connessa.

1.4.1 Il caso West Ham United

Questo è stato l'approccio adottato dalla squadra di calcio inglese *West Ham United*, il cui manager *Sam Allardyce* [9], è stato uno dei primi ad adottare l'analisi in questo sport. Allardyce, durante la stagione sportiva del 2011/12, guidò il club dal secondo livello dei campionati di calcio inglese alla Premier League e lo ha mantenuto lì da allora. Lo studio del manager, è partito da una panoramica della squadra, per poi andare ad esaminare e discutere i singoli profili dei giocatori e le loro metriche specifiche, mettendo insieme strategie per posizioni e gruppi specifici (*cluster analysis*) in un piano di gioco e un playbook unificati; giunse così alla conclusione che la preparazione e la pianificazione di una varietà di opzioni tattiche è la *chiave del successo*. Il suo team è stato per lo più, uno dei primi ad adottare le allora nuove tecnologie, tra cui i dispositivi *GPS*, video e monitoraggio del sonno. Ci sono tre parole che Allardyce cita come chiave per il successo della promozione del cambiamento analitico, per influenzare positivamente le prestazioni dei giocatori:

1) Fede
2) Prove

[9] https://it.uefa.com/memberassociations/news/newsid=538357.html

3) Determinazione.

Fede significa la fiducia ed il sostegno dell'organizzazione sportiva. Prova significa l'uso di dati e analisi per fornire prove basate sui fatti per approcci e decisioni manageriali. La Determinazione infine, è qualcosa che è necessario durante i periodi in cui i risultati sono influenzati da circostanza esterne/attenuanti (situazione analoga al caso Oakland A.).

"...fortune is temporary, but knowledge is permanent." [10]

Con il West Ham, Allardyce ed il suo staff sono così stati in grado di implementare dati ed analisi in diverse aree:

a) Supporto per la strategia di gioco individuale
b) Approfondimenti sulle prestazioni dei giocatori per fornire suggerimenti
c) Modellistica dei modelli di infortuni con l'obbiettivo finale di previsione e prevenzione

Infine, il manager è anche famoso per una precisa analisi del campionato, suddividendo la stagione in fasi chiaramente definite, impostando gli obbiettivi del team e dei giocatori per ogni fase in base alle metriche analiticamente definite. Sostiene: *"It's what we know, not what we think, that matters!".*

1.5 Basket

Rispetto al calcio, in cui a differenza di poche eccezioni, il processo analitico non è ancora stato accettato del tutto, nel mondo del basket, l'analisi è un qualcosa di consolidato, e da cui molte squadre sportive dipendono per il raggiungimento dei risultati. L'analisi cestistica, tiene traccia di analisi descrittive come tocchi della palla, rimbalzi (contestati e non contestati) ed altre variabili che ampliano notevolmente il campo di ricerca. Ciò nonostante, le tecnologie utilizzate per la rilevazione dei dati sono pressoché simili a quelle degli altri sport, come dispositivi di localizzazione e biometrici, i quali si riservano utili nel valutare la qualità di attività fisica di un particolare giocatore. A differenza degli altri sport, nel basket, l'analisi ha raggiunto un ulteriore gradino poiché, anche se sconsigliato da alcuni allenatori, si è deciso di coinvolgere i giocatori stessi nell'uso dell'analisi. Come ha detto un giocatore NBA (National Basketball Association) in una notizia della *CBS* sull'utilizzo di dispositivi *GPS* nelle arene di pallacanestro:

[10] http://www.dailymail.co.uk/sport/football/article-2430369/Sam-Allardyce-100th-West-Hammatch-Sportsmail-reveals-life-Hammers-numbers-game-Big-Sam.html

"If you're thinking about the camera up there while you're playing, then Chris Paul might go by you for a layup or, you know, Blake Griffin might," disse Garrett Temple [giocatore dei Washington Wizards] [11].

1.5.1 Daryl Morey

A tal proposito, non mancano le eccezioni. Daryl Morey, general manager degli Houston Rockets (squadra NBA), ha commentato che uno dei suoi giocatori (Shane Battier), è stato l'unico giocatore che ha incontrato in grado di assorbire e agire sull'analisi relativa alla sua performance.
Il General Manager dei Rockets, è comunemente descritto come il Billy Bean (GM dei Oakland A.) della NBA[12]. Come Bean, è un forte sostenitore del processo decisionale analitico, ma a differenza del suo predecessore, è diventato un operatore di analisi dello sport in generale. Morey è noto per quattro risultati in termini di prospettiva analitica:

1) Identificare i giocatori cosiddetti *free agent,* ovvero coloro che sono in scadenza di contratto ed al termine della stagione possono quindi decidere liberamente con quale squadra della lega firmare a parametro zero. Tra questi possiamo ricordare superstar del calibro di James Harden e Dwight Howard. Gli stili di gioco di entrambi i giocatori sposano infatti a perfezione la strategia di gioco dei Rockets improntata da Morey.
2) Costruire un forte gruppo di analisti e stagisti.
3) Sviluppare una strategia disciplinata per la selezione dei tiri di squadra. I Rockets non tirano quasi mai dalla media distanza (*jumper*), ma le scelte di tiro coincidono con la strategia, ovvero prevalentemente tiri da vicino (*lay-up*, vedi Howard) oppure tiri da 3pt (vedi Harden), i quali hanno un valore atteso più alto. Dalle statistiche della lega, emerge infatti, che i Rockets hanno quasi sempre la più bassa percentuale di tiri nella cosiddetta "terra di nessuno" compresa fra i 16 ed i 23 piedi dal canestro.
4) Ricerca intensiva di dati distintivi o proprietari. Morey, ha usato il suo team per studiare il comportamento degli atleti in tutta la lega NBA, diventando il primo GM della NBA a siglare un accordo con *SportVU*, per mettere le videocamere sul campo in modo da catturare i movimenti di tutti i giocatori.

Altro concetto fondamentale del pensiero di Morey, è l'acquisizione di una o due superstar, indispensabile secondo il general manager per qualsiasi squadra.
Nonostante c'è chi lo definisce come "*the smartest GM in the league",* non manca chi critica il suo operato o semplicemente chi si è concentrato sull'analisi di altri aspetti del gioco del basket.

[11] Testimonianza del giocatore NBA Garrett Temple sull'uso dei dispositivi *GPS* durante le partite di basket.
[12] Daryl Morey, "Success Comes from Better Data, Not Better Analysis," Harvard Business Review blog post.

1.5.2 Dean Oliver

Dean Oliver, come Daryl Morey, è stato ed è tutt'oggi anch'esso un general manager di una squadra NBA (National Basketball Association), i Denver Nuggets. Nella sua analisi, si focalizza per l'appunto su altri aspetti che influiscono sull'evento sportivo, ed in particolare sul concetto di "possesso". Tale concetto, è fondamentale per l'analisi del basket, poiché all'aumentare del numero di possessi, aumentano le possibilità di effettuare tiri e quindi implicitamente le probabilità di segnare un tiro. Dean Oliver, ha cercato così di studiare questa grande statistica/metrica attraverso alcuni concetti e metodi, tra cui valutazioni offensive e difensive, statistiche al minuto, percentuali del tiro reale, tassi di rimbalzo, utilizzo del possesso individuale, efficienza individuale, metodo pitagorico e metodo della curva di Bell.
L'analisi dello statista parte con la definizione di cosa sia un "*possesso*"[13], attestando che un possesso inizia nel momento in cui una squadra ottiene il controllo del pallone e termina quando quella squadra ne perde il controllo. Ci sono diversi modi attraverso i quali una squadra può perdere il possesso, tra cui:

a) canestri o tiri liberi realizzati che portano l'altra squadra a rimettere la palla del fondo del campo, iniziando quindi un nuovo possesso
b) rimbalzi difensivi
c) palle perse.

Bisogna notare bene, che un rimbalzo offensivo non costituisce un nuovo possesso ma una nuova "*giocata*", in quanto il possesso del pallone non è cambiato da una squadra all'altra ma semplicemente dopo un tiro sbagliato, la squadra in attacco ha recuperato il pallone e può quindi attaccare nuovamente.
La scelta del variabile "*possesso*", è stata ritenuta valida poiché essi sono approssimativamente uguali per le due squadre in una partita, e quindi forniscono una base utile per valutare l'efficienza di squadra ed individuale. Come avviene in tutti gli sport, anche nel basket la squadra che fa più punti vince, ed è questo il motivo per cui si cerca di segnare più punti per possesso rispetto agli avversari.
L'analisi di Oliver, non si allontana più di tanto da quella effettuata dal citato Bill James nel baseball. I possessi sono infatti analoghi agli "outs" [14] del baseball, dove le squadre hanno di solito 27 out per superare i propri avversari.

Stima dei possessi
I possessi per squadra (*POSSt*) possono essere stimati utilizzando i dati ricavabili dai registri *play-by-play* [15], oppure utilizzando una formula generale:

[13] Oliver D. *Basketball on paper. Rules and tools for performance analysis.* Washinton, D.C.: Brassey's, Inc.; 2004
[14] James, Bill. The *Baseball Abstract*.
[15] Nessun possesso viene conteggiato alla fine di un periodo in cui sono rimasti meno di quattro secondi.

$$POSS_t = (FGM_t + \lambda FTM_t) + \alpha[(FGA_t - FGM_t) + \lambda(FTA_t - FTM_t) - OREB_t] + (1 - \alpha)DREB_o + TO_t$$

dove:

- *FGA (t) is field goal attempts for team (t)*
- *FTM (t) is free throw made for team (t)*
- *FGM (t) is field goals made for team (t)*
- *FTA (t) is free throw attempts for team (t)*
- *OREB (t) is offensive rebounds for team (t)*
- *DREB (t) is defensive rebounds for opponent (o)*
- *TO (t) is turnovers for team (t)* [16]
- *λ is the fraction of free throws that and the possession*
- *α is a parameter between zero and one.*

Secondo l'equazione sopra citata, un *FGM (t),* ovvero un tentativo di canestro dal campo realizzato, così come *FTM (t),* ovvero un tiro libero segnato ed un *TO (t)* ovvero una palla persa, costituisce l'inizio di un nuovo "*possesso*", cioè ha un valore di possesso ("*possession value*") pari ad 1. I tentativi di tiro dal campo sbagliati ed i tiri liberi sbagliati contribuiscono al possesso con i rimbalzi offensivi (*OREB*). I tentativi dal campo persi ed i tiri liberi che terminano con il possesso, ottengono una quota α del possesso, mentre il rimbalzo difensivo (*DREB*) ottiene una quota (1 – α). I rimbalzi offensivi infatti, annullano i tentativi dal campo ed i tentativi di tiro libero a fine possesso sbagliati, poiché come detto in precedenza generano una nuova "*giocata*" (*play*), quindi il loro valore di possesso è – α.

La formula precedente può essere riscritta nel seguente modo:

$$POSS_t = FGA_t + 0.44 \times FTA_t - OREB_t + TO_t$$ [17]

In cui si assume α = 1 e λ = 0,44

Questa formula viene definita come "*possession lost",* in cui si presuppone che i *DREB (t)* non hanno *"possesion value".*

[16] *TO (t)* include le palle perse di squadra, come violazioni di cinque secondi per la rimessa o 24 secondi per l'azione di gioco, i quali non vengono attribuiti a nessun giocatore ma alla squadra.
[17] L'assunzione di α = 0 risulta da $POSS_t = FGA_t + 0.44 \times FTA_t - OREB_t + TO_t$

Un'altra formulazione, presuppone invece che i *OREB (t),* i *FGA (t*) ed i *FTA (t*) sbagliati, non possiedono "*possesion value",* quindi con $\alpha = 0$.

Invece di assumere particolari valori per α e λ, possiamo stimare il "*possesso*", come detto prima, attraverso l'utilizzo del registro *play-by-play*.

$$POSS_t = \beta_0 + \beta_1 FGA_t + \beta_2(FGA_t - FGM_t) + \beta_3 FTA_t + \beta_4(FTA_t - FTM_t) + \beta_5 OREB_t + \beta_6 DREB_o + \beta_7 TO_t + \varepsilon$$

Da questa equazione, non è possibile stimare perfettamente i possessi utilizzando i box score data, poiché alcuni quarti della partita terminano con i *OREB (t)* senza colpi di follow-up (ovvero *tap-in*) e non tutti i rimbalzi sono attribuiti ai giocatori nei punteggi delle caselle del box score.

Alcuni rimbalzi inoltre, nello specifico in seguito a tiri sbagliati o stoppati (*blocked shots),* vengono registrati come rimbalzi di squadra. Per questa ragione, Oliver ha sentito il bisogno di stimare con la precedente relazione, quelle variabili come i rimbalzi di squadra che non sarebbero stimate nel modello, utilizzando una media stimata di entrambe le squadre anziché il numero delle singole.

La seguente tabella, fornisce un chiaro esempio di analisi del possesso, prendendo come riferimento l'intervallo di tempo 2002/06 e le relative partite (5.178) disputate dalle squadre NBA (National Basketball Association).

Possession Formula	**Correlation**	**Mean**
(1) Actual possessions	1.0000	91.67
(2) Possessions from specification (1) of Table 1	0.9806	91.67
(3) Possessions from specification (2) of Table 1	0.9733	91.67
(4) $FGA_t + 0.44 \times FTA_t - OREB_t + TO_t$	0.9729	93.88
(5) $FGA_t + 0.44 \times FTA_t - OREB_t + TO_t$, own team	0.9488	93.88
(6) $FGA_t + 0.5 \times FTA_t - OREB_t + TO_t$	0.9727	95.40
(7) $FGA_t + 0.4 \times FTA_t - 1.07 \times OREBMiss_t + TO_t$	0.9766	91.28

Figura 1.1: Studio dei possessi nell'intervallo di tempo 2002/06

In queste stagioni, le squadre hanno registrato una media di circa 91,7 possessi per partita. Si può notare inoltre, che le formulazioni che utilizzano $\lambda = 0{,}44$ in media presentano 1,5 possessi per giocata troppo alti e quelli con $\lambda = 0{,}5$ sono invece in media 3,1 possessi per giocata troppo alti. Ciò allora suggerisce che, una buona formula per possessi di gioco nella NBA è la seguente, mediata su entrambe le squadre.

$$POSS_t = 0.976 \times (FGA_t + 0.44 \times FTA_t - OREB_t + TO_t)$$[18]

Offensive and Defensive rating

Dopo aver definito la nozione di possesso e le misure ad esso legato, l'analisi di Dean Oliver fa un ulteriore passo in avanti, utilizzando ciò che aveva descritto in precedenza per valutare l'*efficienza,* ed in particolare il "*rating*". Questo termine specifico è usato per indicare *l'efficienza per possesso*. In particolare, i punti segnati per 100 possessi sono stati chiamati rispettivamente *offensive rating* e *defensive rating* per squadra[19].

$$Offensive\ Rating\ (ORtg_t) = \frac{PTS_s}{POSS_t} \times 100$$

$$Defensive\ Rating\ (DRtg_t) = \frac{PTS_o}{POSS_o} \times 100$$

Queste due misure, riflettono sia l'efficienza di squadra sia il *ritmo* con cui gioca, ovvero la *velocità*. Dato che, in ogni partita, il numero dei possessi è dettato da entrambe le squadre ed è approssimativamente uguale per entrambe, l'efficienza del gioco quando si ha la palla è ciò che alla fine vince; quindi questa misura isola meglio la qualità dell'attacco e della difesa di una squadra. In quanto tale, la maggior parte degli studi delle prestazioni del team è soggetta a tali "*ratings*". Il seguente grafico illustra le classifiche offensive e difensive del 2005/06 per le squadre della NBA. Nel quadrante in alto a destra ci sono squadre che presentano buon rating di attacco e scarsa difesa. Nel quadrante in basso a sinistra ci sono squadre con scarsa offensiva buona difesa.

[18] Si noti che questa formula è calibrata per l'attuale NBA. Gli adeguamenti dovrebbero probabilmente essere effettuati per periodi precedenti nell'NBA e per il basket ad altri livelli.

[19] La differenza tra valutazioni offensive e valutazioni difensive viene spesso definita *net efficiency ratings*.

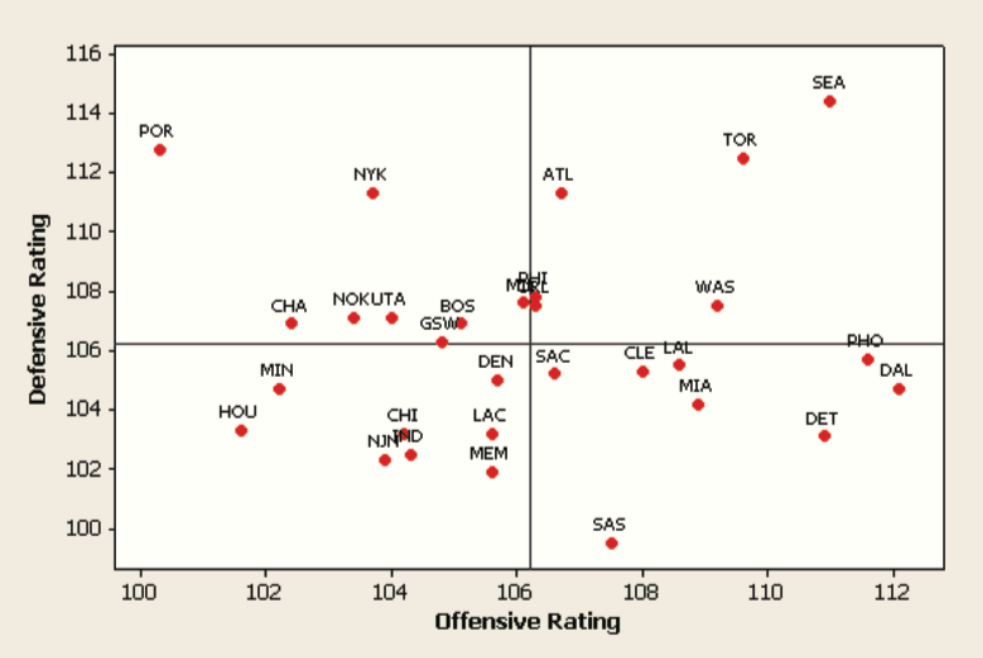


Figura 1.2: Valutazione Offensiva e Difensiva 2005/06.

Storicamente e nell'intervallo 2005/06, vi è pochissima correlazione tra valutazioni offensive e difensive; le buone squadre offensive non tendono ad essere migliori o peggiori in difesa.

Ciò che invece è cambiato storicamente sono i punteggi medi della lega. Negli anni '70 i punteggi erano piuttosto bassi, e si è assistito ad un aumento costante. Alla fine degli anni '80 e all'inizio degli anni '90, i punteggi in media erano nel loro punto più alto, ma ancora erano ben distanti dai valori registrati negli anni recenti, in cui sono registrati valori ben superiori ai 108 degli anni '90. Questo ci indica, che negli ultimi si da sempre più enfasi alla fase offensiva che alla fase difensiva, e soprattutto che il livello del singolo giocatore e la sua efficienza si è notevolmente alzata.

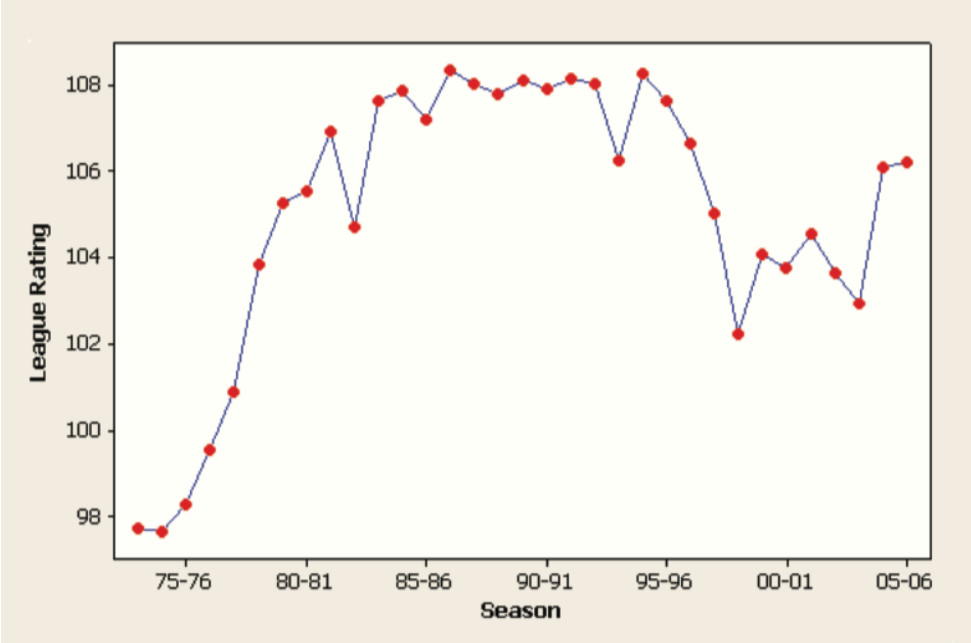


***Figura 1.3*:** Valutazioni medie della lega dalle stagioni 1973-74 al 2005-06.

Dal seguente grafico, può essere estrapolata anche un ulteriore metrica, *"floor percentage",* ovvero la percentuale di possessi in cui viene segnato almeno un punto. Questa percentuale minima può allora, secondo Oliver, servire come probabilità caratteristica in una distribuzione binomiale che descrive la sequenza dei punteggi su un campo.

Plays

Durante la sua capillare analisi, Oliver si preoccupò di chiarire una volta per tutte il concetto di azioni di gioco (*plays*), il quale era comunemente confuso con quello di possesso. Le giocate sono infatti simili ai possessi, tranne per una particolare differenza, ovvero per il fatto che i rimbalzi offensivi costituiscono una nuova azione di gioco, ma non un nuovo possesso. Una squadra può tirare, sbagliare, ottenere un rimbalzo del proprio tiro più volte, dando luogo a più giocate per una squadra, senza concedere una giocata agli avversari. Per tanto, le giocate non sono approssimativamente uguali per le due squadre, come avviene invece per i possessi, e non sono quindi utilizzabili come base per valutare l'efficienza della squadra. A causa di questo potenziale fraintendimento, Dean Oliver ha stimato così anche il numero di giocate (*plays*).

$$PLAYS_t = FGA_t + 0.44 \times FTA_t + TO_t$$

Analogamente alla *"floor percentage",* la percentuale di giocate su cui viene segnato almeno un punto è chiamata *"play percentage"*.

Per- minute statistics

Avanzando nell'analisi delle squadre NBA, Oliver è andato ancora più a fondo nel dissezionare le singole partite, scorporandole minuto per minuto. Si è giunti così alla scoperta delle statistiche calcolate su base al "*minuto (giocato)",* le quali tendono ad essere abbastanza coerenti anche quando i minuti giocati da un singolo giocatore sono variabili, aspetto fondamentale perché permette un confronto fra i giocatori che partono in quintetto (*starters*) e quelli che partono dalla panchina. A volte le statistiche al minuto sono indicate come "*rates*", la percentuale di punteggio per un giocatore (*p*), ad esempio, è punti segnati in 40 minuti (*PTS40p*) [20]:

$$PTS40_p = \frac{PTS_p}{MIN_p} \times 40$$

Si può notare così come l'analisi di Oliver, iniziata da un punto di vista più generale sulla squadra, si è spostata all'analisi capillare di ogni aspetto che possa spiegare in maniera dettagliata il singolo giocatore. La percentuale dei tiri dal campo (*%FG*) secondo Oliver, non tiene conto dei tiri da 3pt (*3PTM*) oppure dei tiri liberi (*FTM)* segnati. A riguardo, lo studioso ha introdotto altre due misure che li analizzano:

a) *effective field goal percentage* (*eFG%):* la quale tiene conto solo dei tiri da 3pt segnati

$$eFG\% = (FGM + 0.5 \times 3PM)/FGA$$

b) *true shooting percentage* (*TS%):* la quale tiene invece conto sia dei *3PTM* che dei *FTM*

$$TS\% = \frac{\left(\frac{PTS}{2}\right)}{FGA + 0.44 \times FTA}$$

La true shooting percentage, fornisce una misura dell'efficienza totale in relazione ai tentativi di tiro, mentre l'effective field goal percentage isola l'efficienza di tiro di un giocatore dal campo. Entrambe le metriche sono così appropriate per misurare in situazioni diverse altrettanti aspetti.

Nel periodo compreso fra il 1996 ed il 2006, queste erano le medie (e le deviazioni standard) per le due misure del singolo giocatore:

$$eFG\% = 47.9\% \ (4.4\%)$$

$$TS\% = 52.3\% \ (4.5\%)$$

[20] *Journal of Quantitative Analysis in Sports*, A Starting Point for Analyzing Basketball Statistics, Volume 3, Issue 3 2007 Article 1

Rebound rate

Come anticipato precedentemente, Oliver poneva molta enfasi sul concetto di ritmo e velocità di gioco. Questo aspetto non influenza solamente l'efficienza offensiva e difensiva ma anche la capacità di effettuare tiri e costringere gli avversari a sbagliare. Tutto ciò pone le basi per la definizione di un'ulteriore metrica, il rimbalzo. Il rimbalzo viene per appunto valutato meglio dalla percentuale di tiri sbagliati che un giocatore recupera quando è in campo. Questo è noto come *"rebound rate" per singolo giocatore p (percentuale di rimbalzo), REB%p.*

$$(REB\%)p = (REBp/(REBt + REBo))/(MINp/MINt)$$

Questa percentuale viene stimata prendendo in esame i rimbalzi al minuto della squadra aggiunti ai rimbalzi dell'avversario al minuto e moltiplicati per i minuti in cui il giocatore è stato sul terreno di gioco.

La *REB%p*, può essere scisso in due ulteriori misure a seconda che il rimbalzo sia offensivo *(OREB%)* oppure difensivo *(DREB%):*

$$OREB\%p = (OREBp/(OREBt + DREBo))/(MINp/MINt)$$

$$DREB\%p = (DREBp/(OREBo + DREBt))/(MINp/MINt)$$

Sempre sull'onda dell'analisi individuale, Oliver osservò che i possessi a livello di squadra sono importanti per comprendere le prestazioni del gruppo, ed ha senso considerarli a livello individuali. Insieme ad Holliger, un altro studioso del gioco, introduce così il concetto di "*individual possession rate*", che misura l'intensità con cui i giocatori usano i possessi attraverso i tentativi dei field goal, dei tiri liberi, degli assist, delle palle perse, e (nel caso di Dean) i rimbalzi offensivi. Oliver normalizza anche i tassi di possesso individuali in modo che i singoli giocatori, in media, utilizzino 1/5 dei possessi della squadra mentre sono in campo. Lo statista, introduce anche il concetto di "*individual offensive rating*" (*valutazione offensiva individuale*), cioè la versione a livello individuale della valutazione offensiva della squadra, la quale misura quanto sono efficienti i giocatori con i loro possessi. Pertanto, sia "*l'individual possession rate"* che "*l'indivdual offensive rating*", forniscono un buon metodo per scomporre la valutazione offensiva della squadra.

Four Factors

L'unione delle metriche precedentemente descritte, forniscono un riepilogo delle prestazioni complessiva di una squadra in base al possesso, ed al tempo stesso danno forma a quella che è considerata una delle

statistiche più importanti del pensiero di Oliver, i *"Four Factors of Basketball Success"* [21]. Questa metrica, deriva dai punteggi del box-score, ed identifica i punti di forza e di debolezza strategici di una squadra, sia in termini offensivi che difensivi, facendo affidabili previsioni di vittorie e sconfitte delle squadre NBA. Secondo Oliver, l'efficienza nei tiri, nelle palle perse, nei rimbalzi e nei tiri liberi può portare vittorie alle squadre. In particolare, ve ne sono 4 per i *DREB%* e 4 per i *OREB%.*

I "*Four Factors*" sono i seguenti:

1) *effective field goal percentage (eFG%)*
2) *turnovers per possession (TOt/POSSt)*
3) *Rebounding percentage (REB%)*
4) *Free throw rate (FTMt/FGAt),* misura che rappresenta come la squadra riesce ad ottenere un fallo quando fa un tiro *("ability to make foul shot").*

A questi fattori, Oliver ha assegnato un peso specifico [22], in relazione all'importanza che ognuno di essi ha nell'influenzare un esito come vincente o meno:

1) *eFG%* = 40%
2) $TO_t/POSS_t$ = 25%
3) *REB%* = 20%
4) $FTMt/FGAt = 15\%$

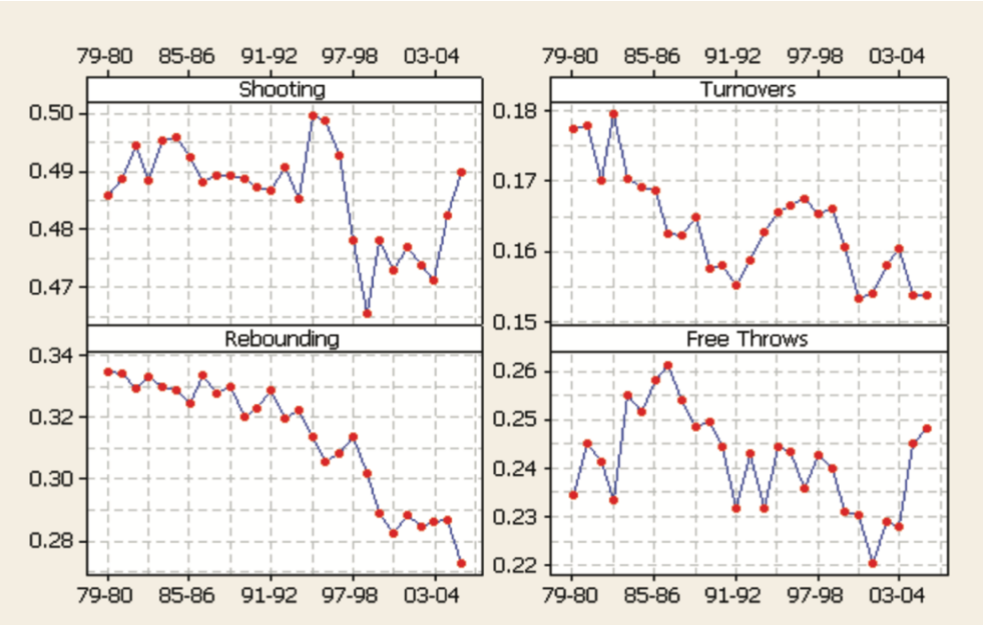


Figura 1.4: Valori medi della lega dei Four Factors dal 1979-80 al 2005-06.

[21] https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/
[22] https://www.basketball-reference.com/

Offensivamente una squadra vuole minimizzare le palle perse (TO_t) per possesso e massimizzare gli altri aspetti. Difensivamente invece accade l'opposto.

Pythagorean Winning Percentage

Lo studio di Dean Oliver, si muove in maniera simbiotica a quello di Bill James; la sua analisi parte dal concetto di possesso, ripreso dall'outs teorizzati da James, e chiude il suo cerchio con un'invenzione proprio dello stesso analista del baseball. L'invenzione in questione è il "*modello di Pitagora"*, il quale si basa sulla consapevolezza che le percentuali di vincita delle squadre sono strettamente correlate ai punti segnati ed ai punti consentiti. La percentuale vincente di Pitagora ($PYTH_t$) è formulata nel seguente modo:

$$PYTH_t = \frac{{PTS_t}^x}{({PTS^x}_t + {PTS^x}_o)}$$

Dove il pedice "t" indica ancora la squadra (*team*), il pedice "*o*" indica l'avversario (*opponent*) e l'apice "*x*" è un esponente determinato empiricamente. Nel baseball, James ha scoperto empiricamente che *x = 2*; nel basket NBA, il valore di "*x*" è stato stimato in un valore compreso fra 13 e 17. Il valore, tuttavia, varia in base all'epoca, questo è il motivo per cui *ESPN* (*fornitore di dati sportivi)* insieme ad Oliver nel 2007 gli ha attribuito un valore di 16,5.

Bell Curve Method

Oltre al modello di Pitagora, vi è un secondo e ultimo modello, relativamente più teorico, per mettere in relazione i punti segnati e concessi con la percentuale di vincita di una squadra, il "*Bell Curve method*". A differenza del primo, in questo metodo le distribuzioni per i punti della squadra segnate e concesse sono normalmente distribuite e possono essere sottratte le une con le altre per formare un'altra variabile casuale normalmente distribuita, i "*net points*". Questi punti netti vengono normalizzati dividendo per la deviazione standard del punto netto per la squadra specifica, formando una statistica Z. La "*probabilità di vincita"* stimata, è quindi data dalla probabilità che una normale casuale distribuita standard assume un valore inferiore a questa statistica Z. La formula per prevedere la percentuale di vincita per la squadra t ($WIN\%_t$) usando questo metodo è:

$$WIN\% = NORMSDIST[\frac{(PPG_t - PPG_o)}{StDev(PPG_t - PPG_o)}]$$ [23]

I vantaggi rispetto al metodo pitagorico è che non si deve fornire una valutazione empirica alla "*x*".

[23] Questa è la funzione NORMSDIST in MS Excel. Si noti che il denominatore, incorpora un peso extra come varianza aggiuntiva.

CAPITOLO 2 – La Cluster Analysis

2.1 Definizione

L'analisi di un nuovo fenomeno, presuppone la necessità da parte dell'uomo di una base di conoscenze, le quali possono derivare da due fonti:

1) Una nuova scoperta nell'ambito che ci si accinge a studiare
2) Il raggruppamento di elementi di conoscenze e know-how simili che provengono da ambiti di applicazione diversi, e che per il loro grado di similarità possono aiutare a comprendere il nuovo fenomeno.

In statistica, una tecnica che presenta il comune obbiettivo di effettuare raggruppamenti di unità statistiche omogenee, descritte da un insieme di variabili, è la *Cluster Analysis o "analisi dei grappoli"*, i quali dovrebbero essere caratterizzati da una omogeneità fra gli elementi al loro interno ed una disomogeneità con gli elementi degli altri grappoli, in modo che siano più simili fra loro che non agli elementi appartenenti ad altri gruppi. A differenza dei metodi di classificazione supervisionati, come l'analisi discriminante, in cui i gruppi sono noti a priori, l'analisi dei cluster è un metodo esplorativo che mira a riconoscere i gruppi naturali che compaiono nella struttura dei dati.

L'implementazione di tale metodologia si articola in alcune fasi fondamentali che prescindono dalla scelta dell'algoritmo per l'analisi:

1) La scelta delle variabili di classificazione
2) La scelta di una misura di similarità o di distanza esistente fra le unità statistiche
3) La scelta di un metodo di raggruppamento.
4) La procedura di valutazione dei risultati ("*validità dei cluster*").

2.2 Le variabili di classificazione

Il punto di partenza di un'analisi cluster, è sicuramente la selezione delle variabili da includere nell'analisi stessa. Questa scelta, dovrebbe in primis essere supportata da considerazioni concettuali, ed in secundis, dovrebbe rispecchiare lo scopo che ci si è prefissi di raggiungere tramite questa metodologia. Di solito, se disponibile, è inclusa una pluralità di variabili, in modo tale che l'eliminazione di una di esse o l'aggiunta di una nuova variabile non modifichi in modo sostanziale la struttura identificata dai gruppi. Infine, le variabili devono presentare un carattere discriminatorio, poiché, se la scelta prendesse ad esame variabili troppo

simili fra di loro, si correrebbe il pericolo di formare gruppi il cui il carattere della disomogeneità verrebbe meno; ed altrettanto succederebbe nel caso contrario venendo meno il carattere dell'omogeneità all'interno dei gruppi.

Successivamente alla scelta ed alla rilevazione delle variabili, si procede alla loro disposizione in una matrice di dati X di dimensione *nxp,* la quale può assumere la forma generica del tipo

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ik} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{bmatrix}$$

il cui termine generico, x_{ik} , fornisce la misurazione *k-esima* per l'unità *i*; oppure, può assumere la forma a blocchi, il cui blocco generico è rappresentato da un vettore *1 x p.*

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_n]' .$$

2.3 Le misure di similarità o di distanza

Una volta che le variabili sono disposte sulla matrice, nella maggior parte dei casi devono essere standardizzate, trasformandole in variabili senza unità (*con media zero e varianza dell'unità*), per evitare la dipendenza dei risultati dalla misurazione dell'unità e l'entità delle variabili da input originali.

In generale, l'ordinamento di coppie di unità è sensibile alla misura selezionata, quindi una fase importante nell'analisi dei cluster è la scelta di una misura di similarità o di distanza tra le unità statistiche (cioè tra le righe della matrice dei dati). A seconda del tipo di dati, si hanno infatti misure diverse. Per dati numerici (*quantitativi),* si hanno misure di distanza; per dati categoriali (*qualitativi)* è preferibile utilizzare misure "*matching-type*", cioè di associazione *(similarità o dissimilarità).*

2.3.1 Misure di distanza

Come detto in precedenza, bisogna ricordare che i suoi (*n*) vettori riga, rappresentano le *n* unità statistiche, ed implicitamente ciascuna di essa è quindi un vettore di *p-elementi,* i quali contengono i valori da essa assunti sulla prima, la seconda, la *j-esima* e la *p-esima* variabile. Supponendo che tali valori siano numeri, ovvero che le p variabili siano quantitative, possiamo definire la distanza tra due unità statistiche, i ed h, in diversi modi, ognuno dei quali deve godere delle seguenti proprietà:

- $d_{ij} > 0$ (*non negatività*)
- $d_{ii} = 0$
- $d_{ij} = d_{ji}$ (*simmetria*)
- $d_{ij} \leq d_{im} + d_{mj}$ (*disuguaglianza triangolare)*

Esistono diversi metodi per il calcolo della distanza, tra cui:

1) Metrica di Mahalanobis
2) Metrica euclidea
3) Metrica di Minkowski (analizzata in questo elaborato)

Quest'ultimo metodo, a differenza delle precedenti, non fornisce una singola misura, ma una famiglia di misure distanza, generalizzando sia la distanza euclidea sia quella di Manhattan, che si ottengono dalla seguente espressione:

$$d_{ij} = [\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda}]^{1/\lambda}$$

al variare di $\lambda(\lambda > 0)$. Le diverse misure che compongono tale famiglia, si formano a seconda del valore che assume il coefficiente λ. Avremo quindi:

a) Se $\lambda = 1 \Rightarrow$ distanza della città a blocchi (o distanza di Manhattan)
b) Se $\lambda = 2 \Rightarrow$ distanza euclidea
c) Se $\lambda = \infty \Rightarrow$ distanza della scacchiera (o distanza di Lagrange)

Una possibile applicazione pratica, volta a chiarire i concetti fin qui esposti, è quella dell'urbanistica. Immaginando di dover attuare un progetto di costruzione di una strada che passi esattamente a metà tra due punti A e B e tale che ogni punto di questa strada sia sempre equidistante dai due punti A e B, possiamo illustrare in base al valore assunto da λ la distanza euclidea e quella della città a blocchi (o di Manhattan). Se il coefficiente assumesse il valore della distanza euclidea, la strada che si dovrebbe costruire è quella lungo la linea continua. Tuttavia, se fossero presenti degli edifici all'interno dei quadrati della figura, l'unica possibilità di costruire la strada è rappresentata dalla linea tratteggiata, ed il coefficiente assumerà il valore della città a blocchi.

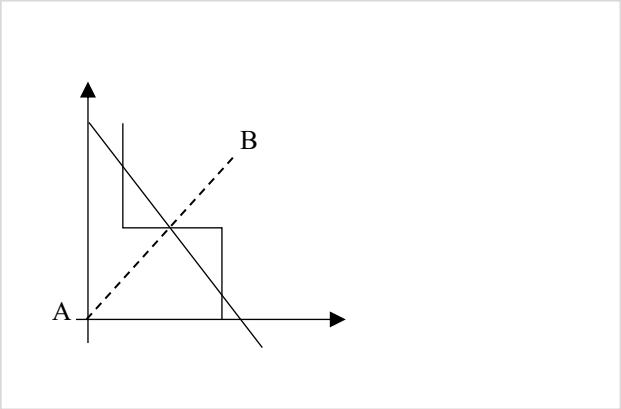


Figura 2.1: Illustrazione grafica del rapporto fra distanza euclidea e distanza di Manhattan.

2.3.2 Misure di similarità

A differenza dei dati numerici, per quelli categoriali, è preferibile adottare misure di associazione. Tali misure, sono volte a confrontare *p* attributi, ciascuno dei quali può essere presente o assente in una generica unità statistica. Supponendo di voler confrontare la similarità tra una statistica x_i ed un'altra x_h, la matrice dei dati assumerà la seguente forma:

ATTRIBUTI						
	1	**2**	**…**	***j***	**…**	***p***
i	0	1	1	0	1	1
h	1	0	1	0	0	1

in cui il valore 1 indica "*presenza*", mentre il valore 0 indica "*assenza*" dell'attributo j.

2.4 I metodi di raggruppamento

Dopo aver effettuato la scelta della misura da utilizzare, si pone il problema della scelta del metodo di clustering e dell'eventuale criterio di aggregazione/suddivisione. I metodi di raggruppamento si dividono in:

1) Metodi non gerarchici
2) Metodi gerarchici, i quali a loro volta, si suddividono in:
   a) Aggregativi

b) Divisivi.

2.4.1 Metodi gerarchici

I metodi gerarchici, realizzano attraverso il loro procedimento fusioni o divisioni successive di dati. Nel caso dei metodi aggregativi (o "*agglomerativi*"), i dati vengono fusi in gruppi man mano più grandi seguendo un procedimento di tipo *bottom–up*, ovvero dal basso verso l'alto; nel caso dei metodi divisivi (o "s*cissori*"), invece, vengono formate delle suddivisioni sempre più piccole dell'insieme iniziale, secondo un procedimento di tipo *top*–down, che porta alla formazione di clusters contenenti ciascuno un elemento. La principale differenza che vi è con quelli non gerarchici è l'irrevocabilità dell'assegnazione al cluster di riferimento. Ovvero, una volta che un elemento è entrato a far parte di un cluster, non può essere più rimosso.

Metodi gerarchici aggregativi

Nel raggruppamento aggregativo, il punto di partenza è rappresentato da tanti cluster quante unità, le quali sono dotate di una misura di dissimilarità, come ad esempio la distanza. Dopo essersi disposte in una prima matrice di distanza, le singole unità *(n)* si aggregano ognuna con la rispettiva vicina, dando vita ad un cluster. Una volta formatosi il cluster, durante il passaggio successivo, una terza unità entra all'interno del cluster, oppure, le due unità precedentemente unitesi, vengono fuse per formare un diverso cluster. Questo procedimento continua finché non viene formato un unico cluster contenente tutte le unità *(n),* le quali condividono sia la misura adottata, sia il criterio di assegnazione ai cluster. Tale criterio, non è esclusivo, ma ne esistono diversi possibili, così come diversi sono anche gli algoritmi aggregativi, e differiscono tra di loro per la modalità di calcolo della distanza fra i gruppi.

Metodo del legame singolo

Nel metodo del legame singolo, la distanza tra i gruppi viene misurata partendo da un criterio di distanza minima, dalla più piccola esistente tra le unità di un gruppo e quelle di un altro. A livello teorico, supponendo di disporre di 4 unità: *A, B, C, D,* le quali siano definite da una misura di distanza tra le unità *(n),* cosi che $(d_{AB}, d_{AC}, \dots, d_{CD})$; ponendo a condizione che le unità *(n) A* e *B* si fondano in un solo cluster, la distanza tra il cluster (*AB*) e l'unità *C*, viene definita nel seguente modo:

$$d_{(A,B)C} = \min(d_{AC}, d_{BC})$$

Dopo essersi formato il primo cluster di *n–1* gruppi, e posto che le unità *(n) C* e *D* si fondano, come nel precedente caso, nel cluster (*CD*), si calcola una nuova matrice di distanza fra gli n–1 gruppi e si aggregano i

due cluster aventi distanza minima, procedendo su tale strada fino ad avere un cluster unico contenente (*n*) unità.

$$d_{(AB)(CD)} = \min(d_{AC}, d_{AD}, d_{BC}, d_{BD})$$

A livello pratico, prendendo in esame 5 unità (*n*) *A, B, C, D, E*, la cui matrice di distanza è rappresentata nel modo seguente:

$$\begin{array}{c} \\ (A) \\ (B) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{ccccc} (A) & (B) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & & \\ \boxed{2} & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{pmatrix} \end{array}$$

la coppia di unità che presenta la distanza minima è *(AB),* le quali si fondono in un unico gruppo. Il passaggio successivo è determinare, sulla base della più piccola distanza con le unità che compongono il gruppo, la distanza tra il gruppo *(AB),* appena formatosi, e le unità rimanenti, ovvero tra *(AB)* e *(C)*.

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

A questo punto, si ottiene la nuova matrice di distanza, formatasi dall'unione delle unità *(D)* e *(E)* nel gruppo *(DE)*.

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 5 & 0 & \\ 8 & \boxed{4} & 0 \end{pmatrix} \end{array}$$

La nuova matrice presenta un'evidente distanza minima (4) tra i gruppi (C) e (DE), che si fondono ottenendo:

$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{5} & 0 \end{pmatrix} \end{array}$$

Il passaggio finale, vede la fusione dei due gruppi in uno solo, contenente tutte le unità:

Iterazione	Gruppi	Livello di distanza
0	$(A)(B)(C)(D)(E)$	-
1	$(AB)(C)(D)(E)$	2
2	$(AB)(C)(D)(E)$	3
3	$(AB)(CDE)$	4
4	$(ABCDE)$	5

Metodo del legame completo

A differenza del legame singolo, quello completo, non prende a riferimento la più piccola, ma la massima distanza, la quale coincide e rappresenta il diametro della sfera contenente tutti i punti appartenenti ai due gruppi. Condizione necessaria, anche in questo caso, è la definizione di una misura di distanza tra le unità *(n).* Supponendo, nuovamente, che le unità *A* e *B* vengano fuse in un unico cluster, la distanza fra il cluster *(AB)* e l'unità *(C),* viene espressa come:

$$d_{(A,B)C} = \max d_{AC}\,, d_{BC}$$

mentre la distanza fra il cluster (AB) e quello (CD) viene espressa come:

$$d_{(AB)(CD)} = \max(d_{AC}\,, d_{AD}, d_{BC}, d_{BD}).$$

Ritornando all'esempio numerico di prima, pur essendo il primo passaggio che porta alla formazione del gruppo *(AB)* identico, le differenze con il legame singolo si notano nel calcolo della distanza tra il gruppo *(AB)* e le altre unità *(n).* Ad esempio, quella tra (AB) e (C) verrà calcolata prendendo a riferimento i valori più grandi, ovvero $d_{AC} = 6\ e\ d_{BC} = 5$.

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 6 & 0 & & \\ 10 & 4 & 0 & \\ 9 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

Come avveniva nel legame singolo, le unità (D) e (E) vengono fuse nel gruppo (DE) formando la seguente matrice:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 6 & 0 & \\ 10 & \boxed{5} & 0 \end{pmatrix} \end{array}$$

fondendo nuovamente i gruppi con distanza minima (5), ovvero (C) e (DE).

$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{10} & 0 \end{pmatrix} \end{array}$$

L'ultimo passaggio, prevede l'unione dei diversi gruppi per formare un unico cluster, il quale da come si può notare differisce da quello del caso precedente; ciò che cambia rispetto al legame singolo sono i livelli di distanza.

Metodo del legame medio

Il metodo del legame medio, si trova a metà fra i due precedentemente descritti, in cui distanza tra gruppi viene definita come la media aritmetica delle distanze fra tutte le unità che compongono i due gruppi. Dati 2 cluster A e B, che all'interno hanno rispettivamente $n_A \; e \; n_B$ unità, la loro distanza è la seguente:

$$d_{A,B} = \frac{1}{n_A n_B (\sum_i \times \sum_h \times \mathrm{d}_{i,h})}$$

in cui l'indice *i* e l'indice *h*, rappresentano rispettivamente un elemento generico del cluster A e del cluster B.

Con riferimento all'esempio numerico, la distanza tra il gruppo (AB) e (C) è la media aritmetica tra i valori $d_{AC} = 6 \; e \; d_{BC} = 5$

$$\begin{array}{c} \\ (AB) \\ (C) \\ (D) \\ (E) \end{array} \begin{array}{c} \begin{array}{cccc} (AB) & (C) & (D) & (E) \end{array} \\ \begin{pmatrix} 0 & & & \\ 5.5 & 0 & & \\ 9.5 & 4 & 0 & \\ 8.5 & 5 & \boxed{3} & 0 \end{pmatrix} \end{array}$$

I passaggi successivi, forniscono le seguenti matrici:

$$\begin{array}{c} \\ (AB) \\ (C) \\ (DE) \end{array} \begin{array}{c} \begin{array}{ccc} (AB) & (C) & (DE) \end{array} \\ \begin{pmatrix} 0 & & \\ 5.5 & 0 & \\ 9 & \boxed{4.5} & 0 \end{pmatrix} \end{array}$$

$$\begin{array}{c} \\ (AB) \\ (CDE) \end{array} \begin{array}{c} \begin{array}{cc} (AB) & (CDE) \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{7.25} & 0 \end{pmatrix} \end{array}$$

le quali, a differenza dei due modelli precedenti, sono soggette ad una fusione intermedia.

Il processo di fusione ed i vari livelli di aggregazione vengono rappresentati attraverso il *"deondogramma"*.

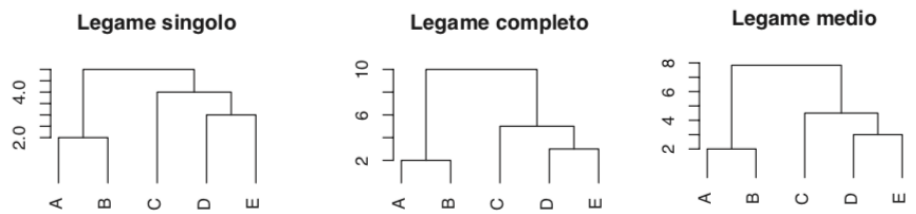


Figura 2.2: Esempi di dendrogramma per i vari metodi aggregativi.

In sintesi, permette di visualizzare l'intero processo di aggregazione, composto da una gerarchia di partizioni, le quali possono essere scomposte "tagliando" il deondogramma in ogni livello dell'indice di distanza della gerarchia.

Metodo del centroide

Il metodo del centroide, si applica solo a variabili quantitative e concentra la sua analisi specialmente sui singoli vettori, tanto che la distanza tra i gruppi è pari alla distanza dei centroidi, ovvero i valori medi calcolati sugli individui appartenenti ai gruppi. Il procedimento del calcolo della distanza, non prescinde come nei casi precedenti, dalla matrice dei dati originaria X, la quale subisce fenomeni "gravitazionali", in cui i gruppi piccoli tendono ad essere attratti da quelli più grandi.

Metodo di Ward

Come il metodo precedente, anche quello di Ward si applica solamente alle variabili quantitative, ed è principalmente diretto alla minimizzazione della varianza all'interno dei gruppi. Ad ogni passo questo algoritmo tende ad ottimizzare la partizione ottenuta dalle possibili coppie di cluster, fondendo la coppia che presenta all'interno dei cluster la varianza più bassa. Tramite questo procedimento di selezione e fusione, si formano cluster che presentano un numero di osservazioni simile.

Metodi gerarchici divisivi

Nei metodi gerarchici divisivi, il punto di partenza è rappresentato dal raggruppamento di tutte le unità statistiche in un unico cluster, il quale viene successivamente diviso in partizioni sempre più piccole. Tale procedimento inizia suddividendo l'insieme di unità *(n)* in due gruppi, ognuno dei quali adotta delle restrizioni per giungere ad una soluzione praticabile, scegliendo le due unità più distanti come nodi (metodo nodale). Questi nodi, rappresentano così i poli ai quali pervengono tutte le altre unità statistiche a seconda della vicinanza all'uno o all'altro. Dopo aver fatto la prima, si passa alla successiva divisione seguendo

sempre lo stesso criterio e si continua il processo fin quando ogni singola unità costituisce un gruppo a sé stante.

Facendo riferimento all'esempio numerico precedente, si può notare come i punti che distano in una maggior misura siano *A e D* ($d_{AD} = 10$), e la prima suddivisione sarà:

$$\begin{array}{c} \\ A \\ B \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 0 & \\ \boxed{2} & \end{pmatrix} \end{array}$$

Le unità *A e B*, della prima suddivisione, andranno, secondo il criterio nodale, a costituire altri due gruppi separati, scegliendo come punti nodali le unità statistiche *C ed E*, ottenendo la seconda suddivisione: *A, B, (CD), E*

$$\begin{array}{c} \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccc} C & D & E \end{array} \\ \begin{pmatrix} 0 & & \\ 4 & 0 & \\ \boxed{5} & 3 & 0 \end{pmatrix} \end{array}$$

Il processo si vedrà concluso con la terza ed ultima suddivisione, in cui vi sarà la costituzione di un gruppo per ogni unità statistica.

2.4.2 Metodi non gerarchici

I metodi gerarchici, sono generalmente onerosi in termini di calcoli che richiedono, così un modo per ridurre tale quantità, è di determinare a priori il numero di cluster. Nel caso dei metodi non gerarchici, l'algoritmo non cercherà quindi ad ogni passaggio di ottenere la migliore scissione o aggregazione tra i cluster, ma dividerà le unità statistiche in un numero di gruppi precedentemente fissato basandosi sull'ottimizzazione del criterio scelto, in cui l'assegnazione di un'unità ad un cluster non è irrevocabile.

Un algoritmo non gerarchico è cosi composto dalle seguenti fasi:

1) Inizializzazione: in cui si scelgono i *(g)* centri in cui avviene una prima suddivisone provvisoria. Il criterio usato per questa prima partizione tiene conto della minima distanza dell'unità da uno di questi centri *(g)*.
2) Primo step: i nuovi centri provvisori assumono la forma dei baricentri, e vengono nuovamente riallocate le unità secondo la distanza minima. Dopo ciò, si calcolano i nuovi baricentri delle classi (cd. *individui medi*).

3) Step successivi: ad ogni passaggio successivo, le partizioni ottenute in precedenza vengono cancellate e si ripete il processo di aggregazione, in cui vengono assunti come nuovi centri provvisori i baricentri dello step precedente.
4) Fine del processo: nel caso in cui non si siano generate nuove riallocazioni.

I principali algoritmi dei metodi non gerarchici sono due:

1) Metodo delle aggregazioni dinamiche
2) Metodo k–means (più popolare)

Nel primo metodo, la scelta dei centri provvisori *(g)* avviene attraverso un'estrazione casuale delle unità statistiche *(n);* mentre nel secondo metodo la scelta non è casuale ma ricade sulle prime *(k)* unità.

Metodo k–means

Dopo aver scelto i centri provvisori, bisogna assegnare ciascuna unità al cluster più vicino. Tale assegnazione, viene fatta utilizzando ancora una volta il criterio della distanza, che molto spesso coincide con quello della distanza euclidea, poiché ciò garantisce la convergenza dell'algoritmo. Il passaggio successivo, consiste nel calcolare i centroidi (ovvero i centri geometrici del cluster, calcolati facendo la media delle coordinate delle unità nel gruppo) e verificare che ciascuna unità sia assegnata al cluster che possiede il centroide più vicino, in modo da poter effettuare degli spostamenti se ciò non avviene naturalmente, e poter procedere così alla verifica della soluzione ottenuta, la quale può essere considerata valida quando rimane approssimativamente stabile al cambiare degli algoritmi.

Esempio numerico

Il punto di partenza è la seguente matrice di dati, contenente cinque unità statistiche descritte dalle variabili, *X e Y.*

UNITÀ	***VARIABILE X***	***VARIABILE Y***
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

Supponiamo che *A e C* siano casualmente scelti come centri iniziali, e poniamo *k=2*. Si calcola la distanza tra ciascuno dei centri del cluster e tutti gli altri punti.

UNITA	***DISTANZA DA A***	***DISTANZA DA C***
A	0	1,4
B	1	2,2
C	1,4	0
D	3,2	2,8
E	4,5	4,2

Il passaggio successivo, prevede l'assegnazione di ciascuna unità al cluster con distanza minima. Al polo *A*, viene quindi associata l'unità *B*; mentre al polo *C* vengono assegnate le unità *D* ed *E*, formando i seguenti gruppi:

gruppo 1 = {*A, B*}
gruppo 2 = {*C, D, E*}

Successivamente, si calcolano i centroidi ($C_1 e\ C_2$) dei gruppi, facendo la media aritmetica delle coordinate delle unità nel gruppo.

$C_1 = (1; 0{,}5)$
$C_2 = (1{,}7; 3{,}7)$

Dopo aver formato i centroidi, bisogna ricalcolare la distanza di ogni unità da quest'ultimi ($C_1 e\ C_{2)}$, e la matrice che ne verrà fuori avrà la seguente forma:

	CENTROIDE 1	***CENTROIDE 2***
A	0,5	2,7
B	0,5	3,7
C	1,8	2,4
D	3,6	0,5
E	4,9	1,9

Si ripete nuovamente l'assegnazione, che in questo caso avrà ad oggetto come poli i centroidi formati. In questo caso, le unità *A*, *B* e *C* faranno parte del centroide C_1, mentre le unità *D* ed *E* faranno parte del centroide C_2, formando i due nuovi gruppi:

gruppo 1 = {*A, B e C*}
gruppo 2 = {*D ed E*}

Osservando i due gruppi, è facile notare che la soluzione non è rimasta approssimativamente stabile al cambiare degli algoritmi, ma saranno necessarie successive riallocazioni per giungere ad una soluzione valida.

2.5 La validità dei cluster

All'inizio del capitolo, si sono delineati gli obbiettivi di questa metodologia di analisi. La *Cluster Analysis*, mira infatti ad effettuare raggruppamenti di unità statistiche, che presentino al loro interno caratteri di similarità ed al loro esterno invece caratteri di disomogeneità. La variabilità di tali caratteri deve così, essere misurati una volta ottenuti i risultati della partizione, ed il modo migliore per farlo è attraverso la varianza, della quale molto spesso risulta essere più comodo considerare solo il numeratore, ovvero la devianza. Tale indicatore, può essere a sua volta scorporato in: devianza interna *(within),* la quale è adatta misurazione della similarità, e devianza esterna (*between*)*,* la quale misura efficientemente la disomogeneità.

$$Dev(Y) = Dev(W) + Dev(B)$$

la cui forma estesa è:

$$Dev(Y) = \sum_i \times \sum_j (y_j - \mu_{Y|x_i})^2 n_{ij} + \sum_i (\mu_{Y|x_i} - \mu_Y) \sum_j n_{ij}$$

Il secondo membro dell'equazione, può essere inoltre scomposto nel seguente modo:

$$Dev(Y) = \sum_i \big(Dev(Y|X = x_i)\big) + \sum_i (\mu_{Y|x_i} - \mu_Y)^2 n_i$$

il quale descrive, con la devianza *within* la variabilità "i*nterna*" ai gruppi, ovvero la somma delle variabilità della Y in ciascun gruppo; mentre con la devianza *between* la variabilità "*tra*" i gruppi, ovvero la variabilità delle medie parziali di Y rispetto alla media generale.
La spiegazione di ciò è che, quanto più i gruppi sono discriminati tanto maggiore è la componente di variabilità esterna rispetto a quella interna, implicando che la variabile X dimostra il comportamento della Y.

Dopo aver misurato la variabilità, attraverso una proporzione fra quest'ultima e la correttezza del modello statistico utilizzato, viene calcolato il coefficiente di determinazione R^2. Tale coefficiente, indica così la correttezza della partizione, e viene espresso nel modo seguente:

$$R^2 = \frac{Dev(B)}{Dev(Y)} = 1 - \frac{Dev(W)}{Dev(Y)}$$

Secondo tale equazione, all'aumentare del valore del coefficiente aumenta il grado di coesione interna e di separazione esterna, e viceversa.

CAPITOLO 3 – La statistica nella NBA

Se nel primo capitolo, si è cercato di descrivere il fenomeno dell'importanza dell'analisi in ambito sportivo ed i risultati connessi ad essa, e nel secondo si ci è focalizzati sulla metodologia adatta a descrivere ed interpretare tali risultati; in questo terzo capitolo, si ci porrà l'obbiettivo di applicare tale metodologia al fenomeno descritto, ed in particolare in ambito cestistico. La metodologia usata, vedrà un'analisi gerarchica ed una non gerarchica. Nella prima verrà usato il metodo di Ward per ottenere la partizione; mentre quella non gerarchica, in cui si entrerà nel vivo dell'analisi, vedrà l'applicazione del *k–means clustering*, la quale prenderà ad oggetto le 30 squadre NBA (National Basketball Association) e come variabili, i *"Four Factors of Basketball Success",* studiandone l'andamento attraverso l'utilizzo del software R.

3.1 Le variabili scelte

Nel definire la metodologia utilizzata, avevamo indicato come punto di partenza della *Cluster Analysis*, la selezione delle variabili di classificazione. Tale scelta, come detto poc'anzi, deve rispettare non solo lo scopo che ci si è prefissi di raggiungere tramite l'analisi, ma anche il carattere discriminatorio, fondamentale per la composizione dei gruppi. Le variabili scelte in questo elaborato, non prescindono da questi due fattori. I *"Four Factors of Basketball Success"* [24]*,* descritti nel secondo capitolo, in verità, non sono un'invenzione di Oliver, poiché sono state introdotte nel mondo cestistico da alcuni allenatori risalenti ai primi anni '90. Tuttavia, la novità attribuibile ad Oliver, è sicuramente quella di comprendere questi fattori attraverso un approccio analitico, nel tentativo di far luce sull'enfasi di ciascun fattore. I "*Four Factors*", sono i seguenti:

a) *effective field goal percentage (eFG%),* misura corretta in scala che identifica la percentuale di obbiettivo sul campo per una squadra. La correzione in scala deve tenere conto dei tiri da 3pt realizzati dal campo. Tramite, questa variabile, si ottiene così la migliore misurazione relativa per i

[24] https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/

punti per tentativo di "*field goal*" (*FG*), ovvero un canestro segnato su qualsiasi tiro diverso da un tiro libero, che può valere due o tre a seconda della distanza del tentativo dal canestro; semplicemente moltiplicando per due.

b) *turnovers per possession (TOt/POSSt),* misura che serve semplicemente a calcolare la percentuale di possessi che si è conclusa con una palla persa. L'obbiettivo di ogni squadra, dopo aver preso un rimbalzo offensivo, è quello di riuscire nuovamente ad effettuare un tiro o di ottenere un tiro libero in seguito ad un fallo, cioè non perdere la palla.

c) *Offensive Rebounding percentage (OREB%),* la quale misura la percentuale di rimbalzi in attacco che si riescono a prendere. Questa quantità viene calcolata come il numero di rimbalzi disponibili dopo un tentativo di canestro in campo fallito. Da come si può evincere, se una squadra non è in grado di segnare su ogni possesso, il compito ottimale è quello di afferrare ogni tiro sbagliato e dare alla squadra una seconda opportunità. Un rimbalzo offensivo estende infatti il possesso e consente un secondo tentativo a canestro su azione. Questo è effettivamente un "*do-over*" per le squadre e, se strategicamente corretto, può essere un piano di attacco mortale per una squadra.

d) *Free throw rate (FTMt/FGAt),* misura che rappresenta come la squadra riesce ad ottenere un fallo quando fa un tiro *("ability to make foul shot").* Una seconda chance di fare canestro, oltre ad un rimbalzo offensivo, si può ottenere tramite i tiri liberi. Raggiungere la linea del tiro libero, ha per l'appunto, due scopi in una partita. Innanzitutto, garantisce i tentativi di segnare punti. In secondo luogo, avvicina sempre più gli avversari all'essere fuori dal gioco, superando il limite di falli permessi (in NBA sono per un massimo di 6 a giocatore). Arrivare alla linea del tiro libero, tuttavia, non è abbastanza, ma la parte importante è trasformare il fallo in punto, realizzando il tiro libero. Misurare la quantità di una squadra nell'arrivare alla linea di fallo, significa così identificare il numero di tiri liberi effettuati per ogni tentativo di tiro in campo, ovvero il cosiddetto (*FTRate*).

Pur sembrando scollegati fra loro, tali elementi, presentano un singolo fattore in comune che li rende strettamente correlati, ovvero la cessione di un possesso per squadra. Al contrario, si è deciso di non includere tra le variabili, misure come *"dead ball rebounds"*, *"end of period"*, e *"specialized fouling situations",* poiché raramente incidono sulla cessione di un possesso per la singola squadra. Quindi ci concentreremo sui quattro principali modi per terminare un possesso. Nel basket, il successo non costituito dal numero di punti segnati, ma piuttosto dal differenziale di punteggio alla fine di una partita; chiamato vittoria (*win*) se il segno di quest'ultimo è positivo. Quindi, il successo, è definito come vincere una partita. Come anticipato precedentemente nel primo capitolo, Oliver, ha posizionato un certo set di pesi su ciascuna

delle quattro variabili, in modo da poter accertare il valore di ciascun fattore in relazione ad una vittoria. I pesi in questione sono [25]:

a) $eFG\%$ = 40%
b) $TO_t/POSS_t$ = 25%
c) $REB\%$ = 20%
d) $FTMt/FGAt$ = 15%

In ambito applicativo, le variabili appena descritte, si identificano all'interno della matrice di dati X di dimensione nx*p,* nelle colonne (*p*), mentre le righe (*n*) costituiscono le unità statistiche oggetto di studio della nostra analisi, ovvero le 30 squadre NBA. Di seguito, la matrice dei dati X:

TEAM	EFG_pct	FTARATE_ptc	TOV_pct	OREB_pct
"Milwaukee Bucks"	55,30	26,30	14,10	24,10
"Los Angeles Lakers"	54,80	26,80	14,90	28,40
"Toronto Raptors"	53,60	25,60	14,20	25,90
"LA Clippers"	53,20	29,30	14,40	28,40
"Boston Celtics"	52,90	25,50	13,60	28,40
"Denver Nuggets"	53,20	23,00	13,90	29,40
"Utah Jazz"	55,20	26,90	15,00	25,80
"Miami Heat"	54,90	29,80	14,90	25,90
"Houston Rockets"	53,90	28,70	14,10	26,10
"Oklahoma City Thunder"	53,40	29,20	13,50	23,90
"Indiana Pacers"	53,30	21,70	13,20	24,60
"Philadelphia 76ers"	53,00	25,20	14,20	27,80
"Dallas Mavericks"	54,80	25,70	12,70	27,50
"Memphis Grizzlies"	53,00	23,40	14,80	27,40
"Brooklyn Nets"	51,50	26,80	15,10	28,40
"Orlando Magic"	50,30	24,80	12,70	26,80
"Portland Trail Blazers"	53,00	23,90	12,80	26,30
"New Orleans Pelicans"	53,80	25,20	15,40	28,80
"Sacramento Kings"	53,10	23,20	14,40	25,60
"San Antonio Spurs"	52,90	25,40	12,10	22,80

[25] https://statathlon.com/four-factors-basketball-success/

"Phoenix Suns"	52,80	27,30	14,80	26,10
"Washington Wizards"	52,80	27,20	13,50	26,00
"Charlotte Hornets"	50,40	25,20	15,00	28,00
"Chicago Bulls"	51,50	23,10	15,30	26,70
"New York Knicks"	50,10	26,30	14,30	30,00
"Detroit Pistons"	52,90	26,10	15,50	27,70
"Atlanta Hawks"	51,50	25,80	15,50	26,70
"Minnesota Timberwolves"	51,40	27,70	14,60	26,40
"Cleveland Cavaliers"	52,20	22,70	16,50	29,60
"Golden State Warriors"	49,70	26,40	14,60	25,50

Figura 3.1: Matrice dei dati non standardizzata. Indici di gioco per squadra NBA.

Come è possibile vedere, la matrice in questione non è standardizzata, infatti le variabili, sono tutte espresse in unità percentuale, ad eccezione della prima che fornisce i nomi dei team. Dopo aver formato la matrice dei dati, prima di passare all'analisi gerarchica ed entrare nel vivo della Cluster Analysis, si può fornire un'ulteriore analisi descrittiva delle variabili, attraverso il *box-plot*. Quest'ultimo, è un diagramma a scatola e baffi, che fornisce una rappresentazione grafica della distribuzione di un campione di dati, tramite semplici indici di dispersione e di posizione. La rappresentazione può avvenire in due modalità:

1) La prima, utilizza come strumento di rappresentazione un rettangolo diviso in due parti, il quale è delimitato dal primo e dal terzo quartile ($q_{1/4}$ *e* $q_{3/4}$), e diviso al suo interno dalla mediana ($q_{1/2}$), che divide la popolazione in due parti di uguale numerosità. Il primo quartile, è quel valore che lascia a sinistra il 25% delle osservazioni, ovvero l'estremo inferiore del rettangolo; mentre il terzo quartile, è quel valore che lascia a sinistra il 75% delle osservazioni, ovvero l'estremo superiore del rettangolo. Da quest'ultimo, fuoriescono due segmenti (*"baffi"*) che sono delimitati dal minimo (*"baffo inferiore"*) e dal massimo dei valori (*"baffo superiore"*). Attraverso questa modalità, vengono così rappresentati i quattro intervalli delimitati dai quartili.

2) La seconda modalità, racchiude un insieme di alternative di rappresentazione, ognuna delle quali è concordante sui tre quartili per rappresentare il rettangolo, ma differiscono le une della altre per la lunghezza dei baffi, solitamente scelti di lunghezza minore per evitare valori troppo estremi.

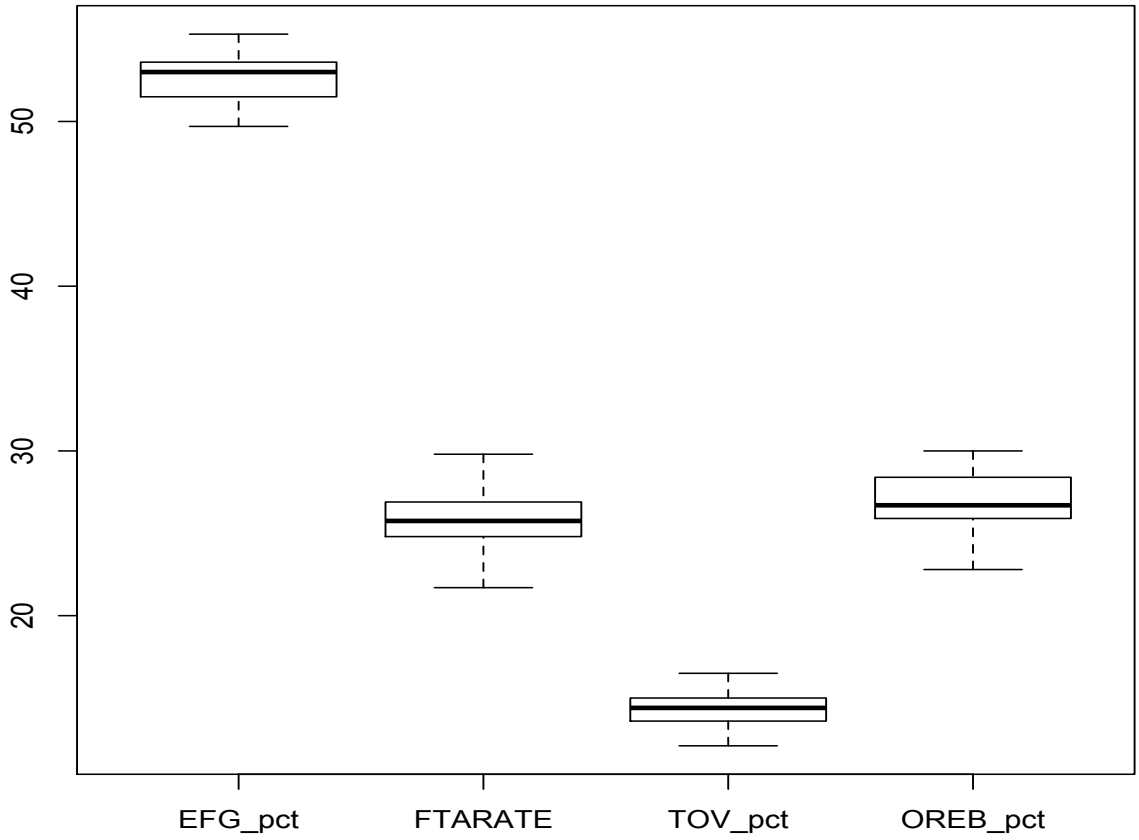


Figura 3.2: Box-plot dei valori riportati da ciascun team per ciascuno dei quattro indici considerati.

Confrontando tra loro le lunghezze dei due baffi (che rappresentano le distanze tra q_1 e il minimo e tra q_3 e il massimo) e le altezze dei due rettangoli che costituiscono la scatola (che rappresentano le distanze tra q_1 e mediana, e tra quest'ultima e q_3) si riesce a verificare se la distribuzione è asimmetrica o simmetrica. Partendo da quest'ultima, una distribuzione si dice simmetrica, se i due baffi che protendono dalla scatola hanno la stessa lunghezza, e la mediana centri perfettamente la scatola. Al contrario, si dice asimmetrica, quando i due baffi non presentano la medesima lunghezza. A riguardo, si può avere, una distribuzione asimmetrica positiva o negativa. In quella positiva, il secondo baffo (ovvero quello che protende da q_3) è più lungo del primo baffo (ovvero quello che protende da q_1), ed abbiamo un eccesso nella parte destra della distribuzione (coda a destra). Se invece, il primo baffo è più lungo del secondo, allora la distribuzione è asimmetrica negativa, ed abbiamo un eccesso nella parte sinistra della distribuzione (coda a sinistra). Dal grafico si può notare come *Offensive Rebounding percentage (OREB%)*, sia la dimensione caratterizzata da più ampia variabilità. Ciò è visibile dall'altezza della scatola, data dalla distanza tra q_3 *e* q_1, detto anche *scarto interquartile*. Inoltre, questa variabile, è l'unica che presenta un eccesso nella parte destra della distribuzione, e quindi un'asimmetria positiva. Al contrario, *effective field goal percentage (eFG%) e turnovers per possession (TOt/POSSt),* presentano un eccesso nella parte sinistra della scatola, e quindi

hanno un'asimmetria negativa. La seconda delle due, inoltre, è la distribuzione che presenta minore variabilità, da come è desumibile dalla lunghezza della scatola. Infine, *Free throw rate (FTMt/FGAt)*, si differenzia dalle altre per la propria simmetricità, avendo i due baffi della stessa lunghezza, e la mediana al centro della scatola.

Il box-plot, si rivela un ottimo strumento per rappresentare una distribuzione in modo sintetico: con poche informazioni, si riesce a comprendere la forma della distribuzione, asimmetrica o simmetrica che sia. Infine, caratteristica ancor più importante, è che dà una rappresentazione univoca della distribuzione.

3.2 Applicazione della metodologia: analisi gerarchica

Dopo aver descritto le variabili scelte, in questo elaborato, l'analisi entra nel vivo attraverso l'applicazione dei metodi gerarchici. I metodi adottati sono:

a) Metodo del legame singolo;
b) Metodo del legame composto;
c) Metodo del legame medio;
d) Metodo del centroide.
e) Metodo di Ward

I risultati ottenuti, sono stati rappresentati graficamente attraverso l'utilizzo del *software* R, il quale ha generato una partizione delle 30 squadre NBA, partendo dalla matrice delle distanze euclidee, calcolata sulla base della matrice dei dati originale.

Di seguito, vengono riportati i dendrogrammi rappresentativi delle partizioni di ogni metodo utilizzato:

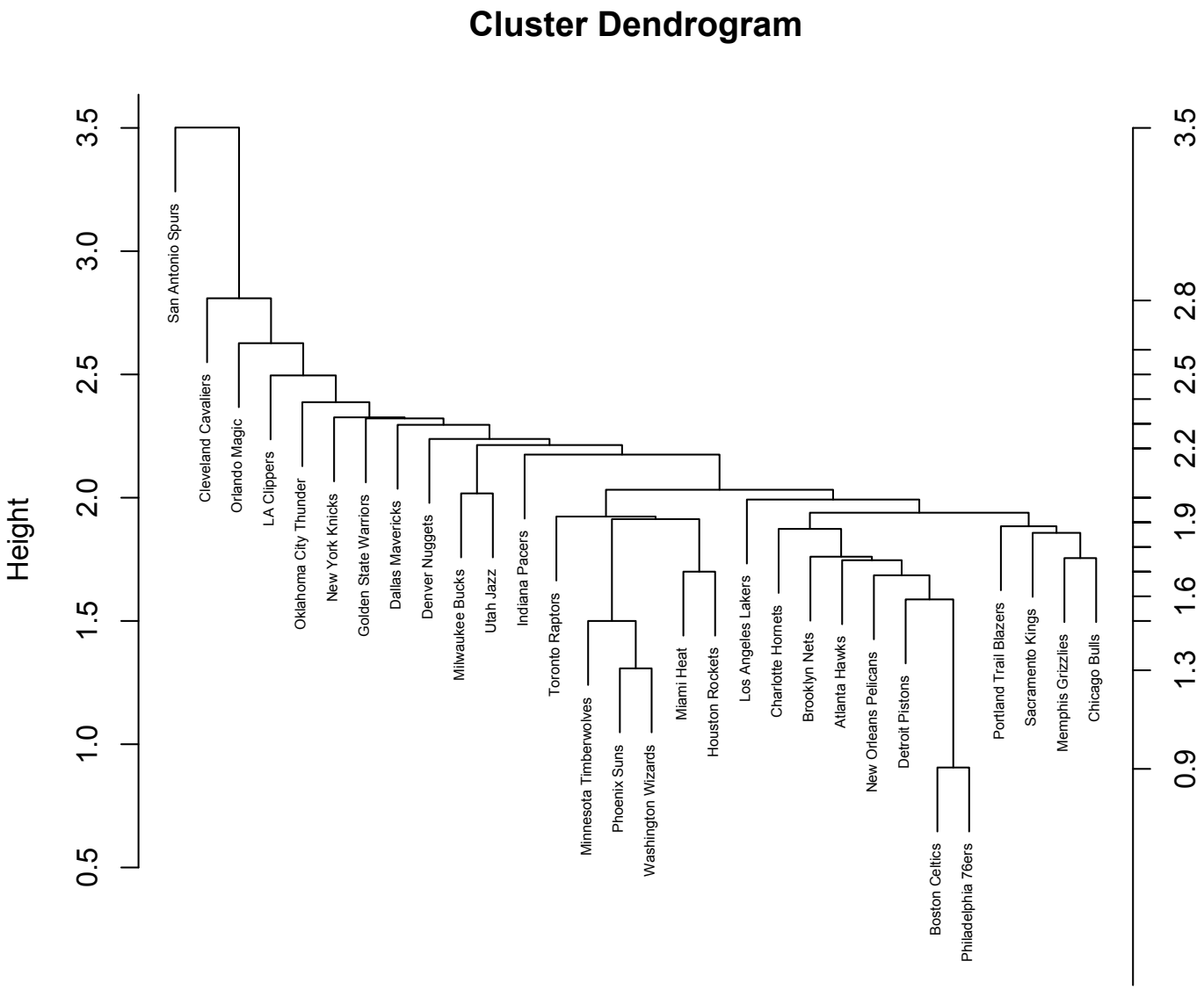


Figura 3.3: Illustrazione grafica del metodo legame singolo.

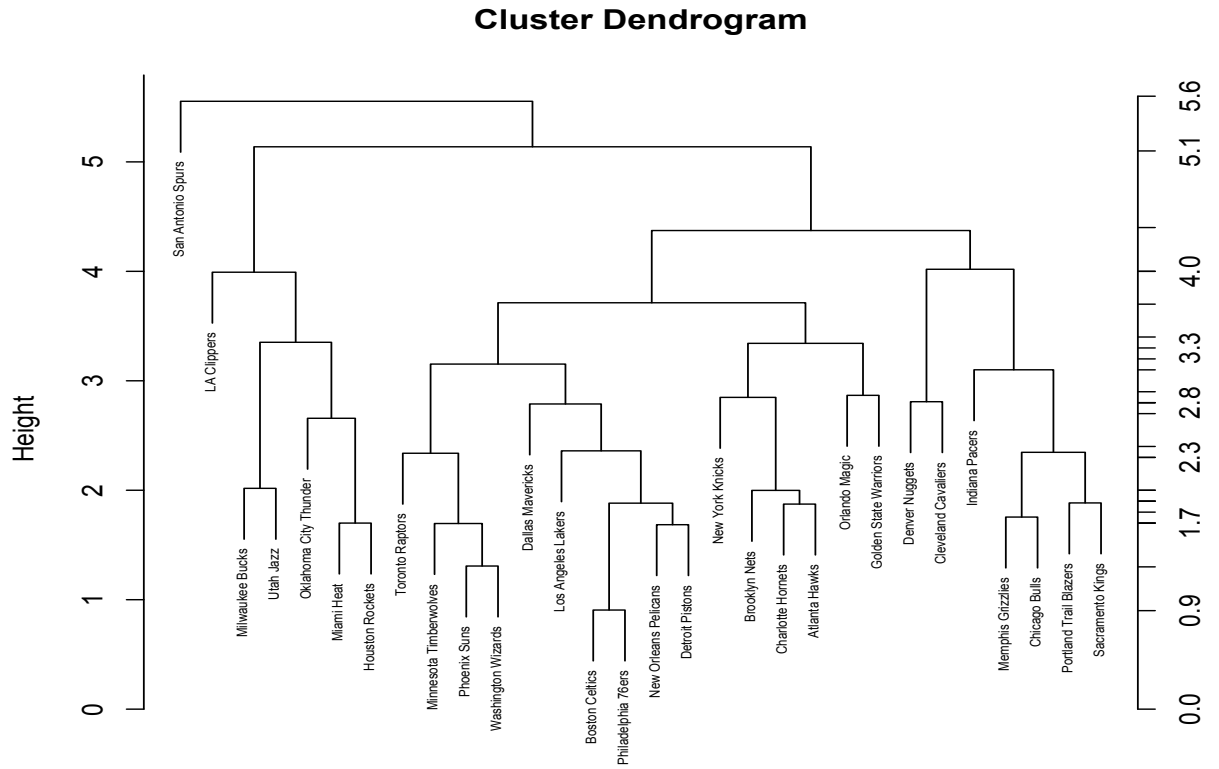


Figura 3.4: Illustrazione grafica del metodo legame composto.

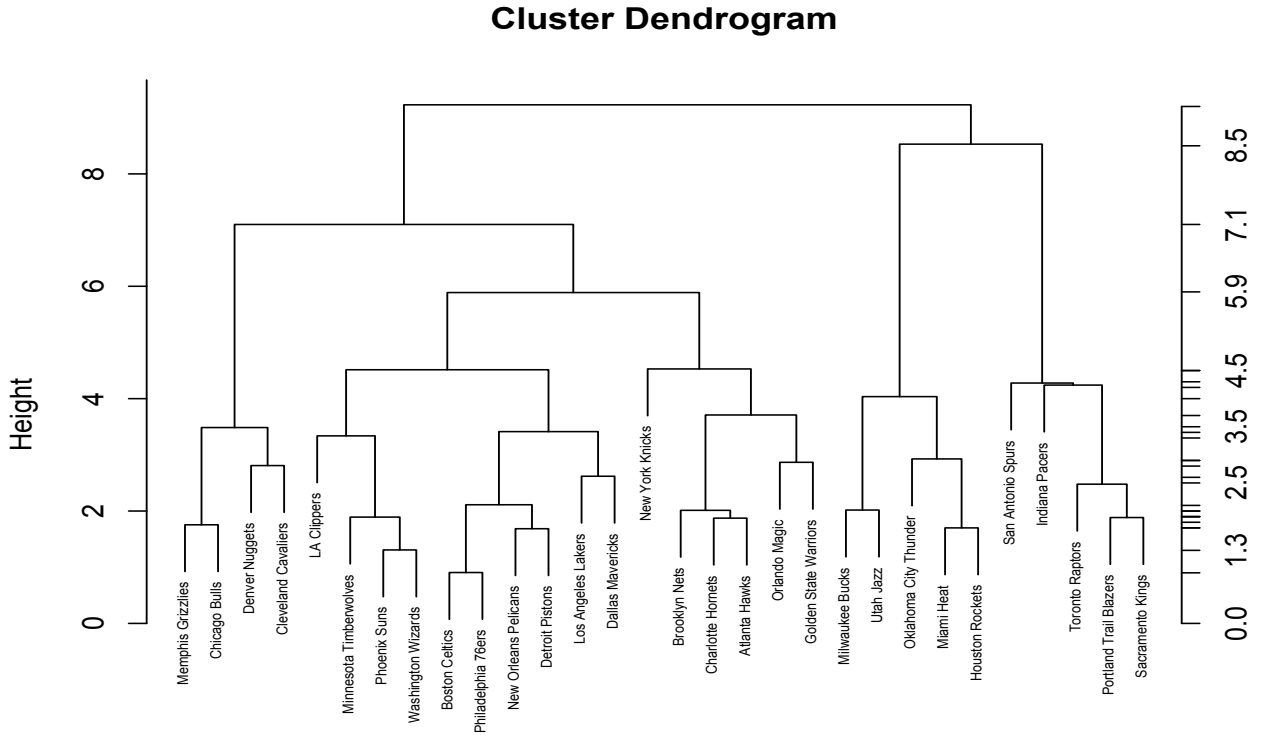


Figura 3.5: Illustrazione grafica del metodo legame medio.

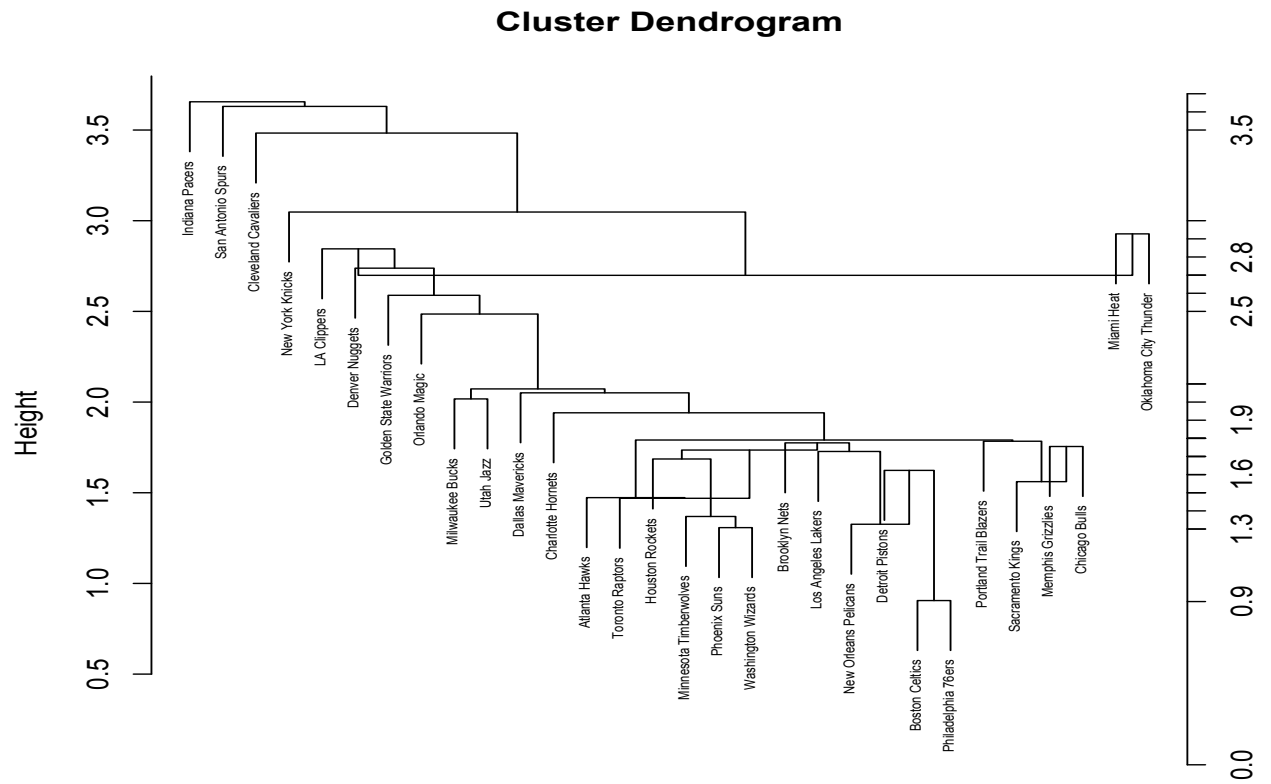


Figura 3.6: Illustrazione grafica del metodo del centroide.

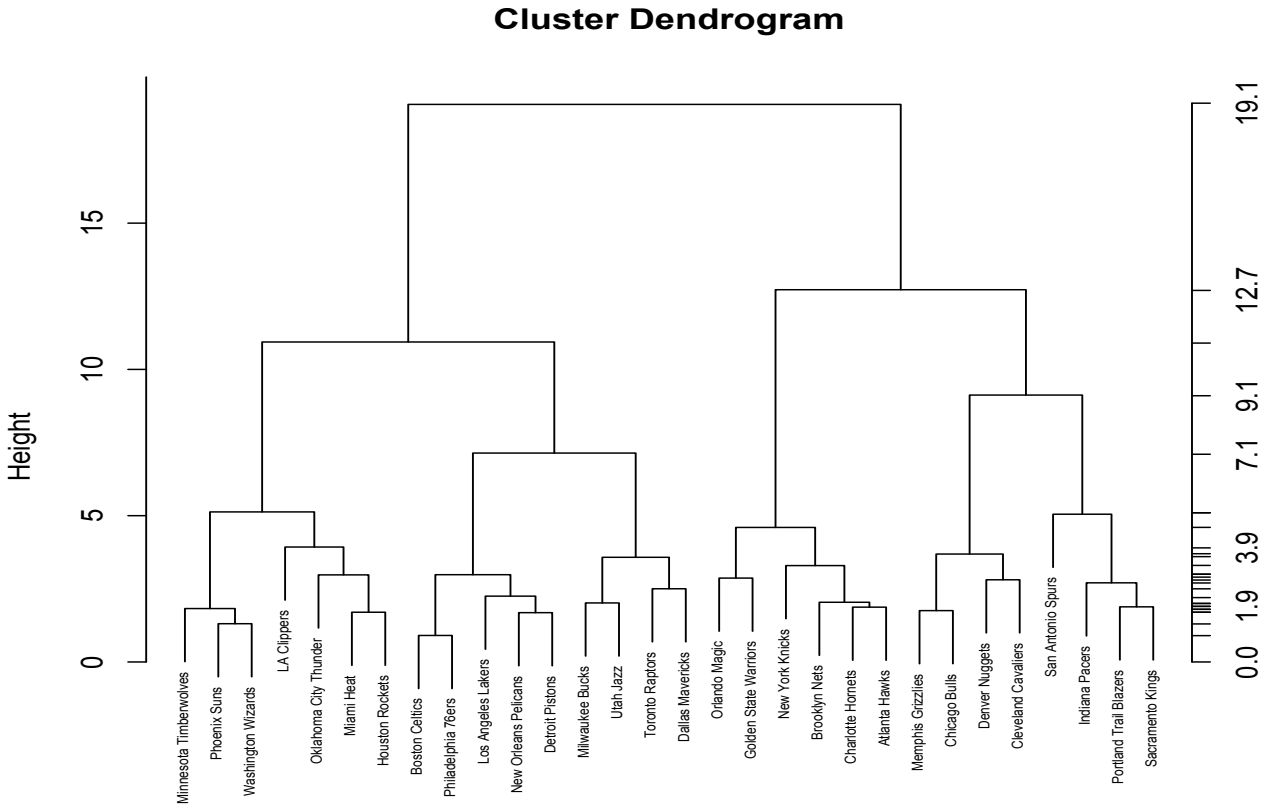


Figura 3.7: Illustrazione grafica del metodo di Ward.

Osservando i grafici, è possibile notare come pur essendoci differenze nelle partizioni, vi siano degli elementi di somiglianza comuni a tutti i metodi adottati. In particolare, vi sono alcune coppie di unità statistiche che tendono a rimanere inalterate pur cambiando metodo di aggregazione. Le coppie in questione sono le seguenti:

a) Memphis Grizzlies e Chicago Bulls;
b) Boston Celtics e Philadelphia 76ers;
c) Phoenix Suns e Washington Wizards, le quali tendono a legarsi con un'altra coppia, ossia quella formata dai Detroit Pistons e i New Orleans Pelicans;
d) Milwaukee Bucks e Utah Jazz.

Il metodo completo e di Ward, inoltre, dividono il dendrogramma e le sue 30 unità statistiche in 2 gruppi: Nella fattispecie del metodo completo, i gruppi formatisi sono i seguenti:

1) Memphis Grizzlies, Chicago Bulls, Denver Nuggets, Cleveland Cavaliers, LA Clippers, Minnesota Timberwolves, Phoenix Suns, Washington Wizard, Boston Celtics, Philadelphia 76ers, New Orleans Pelicans, Detroit Pistons, Los Angeles Lakers, Dallas Mavericks, New York Knicks, Brooklyn Nets, Charlotte Hornets, Atlanta Hawks, Orlando Magic, Golden State Warriors.

2) Milwaukee Bucks, Utah Jazz, Oklahoma City Thunder, Miami Heat, Houston Rockets, San Antonio Spurs, Indiana Pacers, Toronto Raptors, Portland Trail Blazers, Sacramento Kings.

Da una prima visione, è chiaro che il primo gruppo è nettamente superiore in termini di numerosità del secondo. Al loro interno, questi due gruppi, presentano degli elementi di diversificazione. La lega dell'NBA, infatti, divide le squadre al suo interno in due *conference*, ognuna delle quali ha tre *division*, ed ogni *division* ha cinque squadre. Le due conference sono per l'appunto, la *Eastern Conference* e la *Western Conference*. La prima fa riferimento al raggruppamento delle squadre appartenenti agli stati orientali; mentre la seconda raccoglie invece le squadre appartenenti agli stati occidentali. Il primo gruppo, presenta una maggioranza di team provenienti dalla Eastern Conference, ben 11, contro i 9 provenienti dalla Western Conference; inversamente, il secondo gruppo, è composto da 6 squadre della Western Conference e solamente 4 della Eastern Conference.

Il metodo di Ward ha composto altrettanti gruppi, che però differiscono nella loro composizione:

1) Minnesota Timberwolves, Phoenix Suns, Washington Wizards, LA Clippers, Oklahoma City Thunder, Miami Heat, Houston Rockets, Boston Celtics, Philadelphia 76ers, Los Angeles Lakers,

New Orleans Pelicans, Detroit Pistons, Milwaukee Bucks, Utah Jazz, Toronto Raptors, Dalla Mavericks

2) Orlando Magic, Golden State Warriors, New York Knicks, Brooklyn Nets, Charlotte Hornets, Atlanta Hawks, Memphis Grizzlies, Chicago Bulls, Denver Nuggets, Cleveland Cavaliers, San Antonio Spurs, Indiana Pacers, Portland Trail Blazers, Sacramento Kings

Guardando al dendrogramma, è facile da vedere, come il primo gruppo, elencato in precedenza, presenta una quantità maggiore di squadre provenienti dagli stati occidentali. I team della Western Conference sono 9, a discapito di quelli provenienti dalla Eastern Conference che sono in percentuale minore, ed uguali a 7. Al contrario, guardando al secondo gruppo, tale trend cambia notevolmente, avendo una predominanza dei team provenienti dagli stati orientali, uguali a 8, contro i 6 dei team degli stati occidentali.
Come precedentemente affermato, le due *conference*, si dividono a loro volta in tre *division*, per formare un totale di 6 gruppi. Atlantic, Central e Southeast Division compongono la Eastern Conference, mentre Northwest, Pacific e Southwest Division sono nella Western Conference. La divisione è sommariamente geografica, pur ammettendo delle eccezioni dovuta alla storia dei team, che nel corso degli anni sono nati, scomparsi o cambiato città. Emblematica, è per esempio, la situazione della Nortwest Division, con i Minnesota Timberwolves molto più vicini geograficamente a qualsiasi squadra della Central Division piuttosto che alle altre della propria. Rimanendo all'interno della Northwest, si può fare lo stesso discorso in riferimento alla distanza fra gli Oklahoma City Thunder e la Southwest Division. Di seguito vengono elencate le 6 divisioni:

1) Atlantic (Brooklyn Nets, Boston Celtics, New York Knicks, Philadelphia 76ers, Toronto Raptors)
2) Central (Cleveland Cavaliers, Chicago Bulls, Detroit Pistons, Indiana Pacers, Milwaukee Bucks)
3) Southeast (Charlotte Hornets, Atlanta Hawks, Miami Heat, Orlando Magic, Washington Wizards)
4) Pacific (LA Clippers, Golden State Warriors, Los Angeles Lakers, Phoenix Suns, Sacramento Kings)
5) Southwest (Houston Rockets, Dallas Mavericks, Memphis Grizzlies, New Orleans Pelicans, San Antonio Spurs)
6) Northwest (Minnesota Timberwolves, Denver Nuggets, Oklahoma City Thunder, Portland Trail Blazers, Utah Jazz)

I due metodi analizzati, uniti a quello del legame medio, presentano, a differenza degli altri due (singolo e del centroide) delle partizioni più stabili. Il metodo del legame singolo, sebbene permette di individuare gruppi di qualsiasi forma, purché ben separati, d'altro canto è soggetto all'*effetto di concatenamento*, cioè ad ogni fusione le unità statistiche non ancora classificate, tendono ad unirsi a gruppi già esistenti piuttosto che formare nuovi gruppi. Dal seguente grafico, è possibile intuire quanto detto:

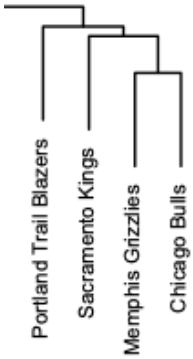


Figura 3.8: Effetto di concatenamento osservato nel metodo del legame singolo.

I Sacramento Kings si fondono alla coppia già esistente, formata dai Memphis Grizzlies e Chicago Bulls. Analogamente, anche i Portland Trail Blazers, seguono lo stesso procedimento.

Il metodo del centroide, è invece caratterizzato da un fenomeno gravitazionale, ossia quello dell'*inversione.* In questo metodo aggregativo, la distanza di fusione, può infatti aumentare e diminuire, rendendo oscura l'interpretazione della partizione, come si evince dal grafico seguente:

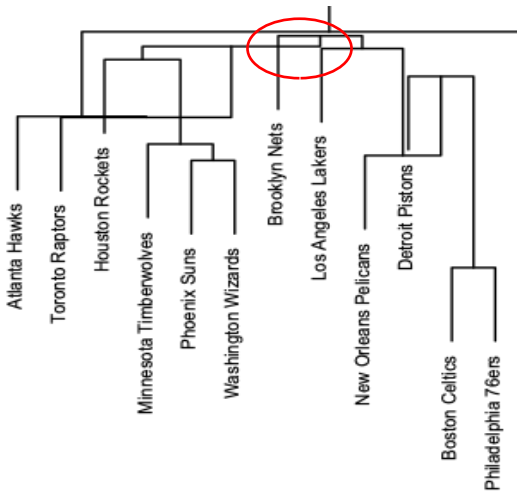


Figura 3.9: Fenomeno dell'inversione osservato nel metodo del centroide.

3.3 Scelta dei gruppi

La distanza di fusione, citata poc'anzi, si rivela utile per effettuare la scelta del numero dei cluster. Questo passaggio, costituisce uno step fondamentale della nostra analisi, in quanto, il suo risultato erigerà le basi per l'applicazione del metodo *k-means*. Per determinare tale numerosità, spesso si ricorre all'utilizzo dei cosiddetti "*scree plot"*, grafici cartesiani in cui vengono allocati nell'asse delle ordinate il numero di gruppi e nell'asse delle ascisse la distanza di fusione. Tali grafici, assumono la forma di una spezzata con pendenza sempre negativa (vedi sotto):

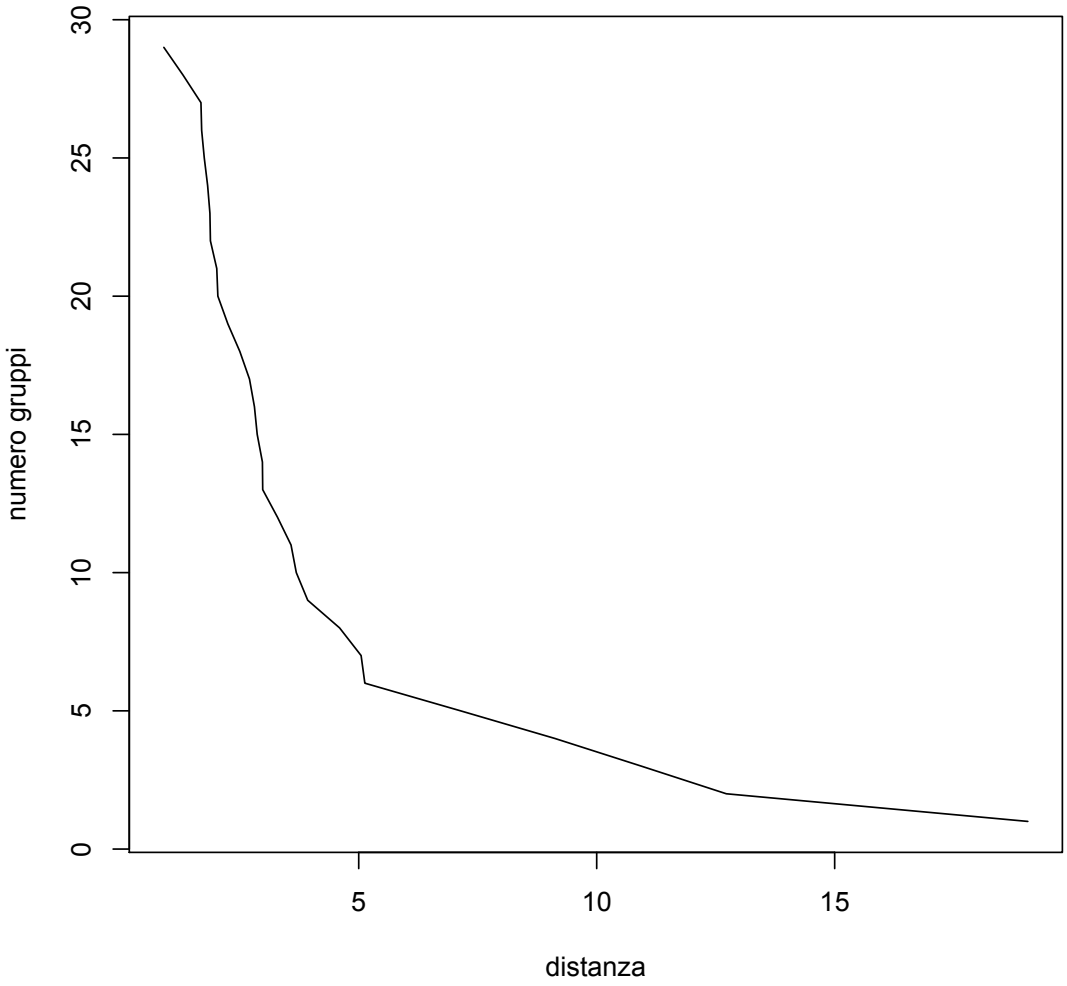


Figura 3.10: Illustrazione grafica della variazione della distanza di fusione all'aumentare del numero di gruppi con metodo di Ward.

Guardando al grafico, si può notare come ad una iniziale pendenza molto alta, corrispondente ad un numero elevato di gruppi, segua una brusca variazione di pendenza nella partizione da 4-5 gruppi. Andando avanti nell'asse delle ascisse, si nota come la distanza di fusione aumenta e con essa anche i gruppi formati da unità lontane tra di loro. Questa osservazione, ci fornisce un'importante indicazione sul taglio del dendrogramma, rispecchiando allo stesso tempo, l'obbiettivo della *Cluster Analysis*, ovvero effettuare raggruppamenti di unità statistiche, che presentino al loro interno caratteri di similarità ed al loro esterno invece caratteri di disomogeneità. La partizione scelta, attraverso la selezione grafica, è composta da 5 gruppi, riflettendo una partizione né con un numero troppo limitato di gruppi, né eccessivamente ampia.

Dopo aver individuato la migliore partizione, la nostra analisi può procedere con il taglio del dendrogramma, in modo da poter discernere la struttura dei gruppi. Se la validità della partizione attraverso la selezione grafica, ha evidenziato la scelta di 5 cluster, questa numerosità si rivela una delle basi per il taglio. Attraverso una semplice istruzione, il *software R*, ripartisce così i gruppi:

- Gruppo 1: Milwaukee Bucks, Lon Angeles Lakers, Toronto Raptors, Boston Celtics, Utah Jazz, Philadelphia 76ers, Dallas Mavericks, New Orleans Pelicans, Detroit Pistons.
- Gruppo 2: LA Clippers, Miami Heat, Houston Rockets, Oklahoma City Thunder, Phoenix Suns, Washington Wizard, Minnesota Timberwolves.
- Gruppo 3: Denver Nuggets, Memphis Grizzlies, Chicago Bulls, Cleveland Cavaliers.
- Gruppo 4: Indiana Pacers, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs.
- Gruppo 5: Brooklyn Nets, Orlando Magic, Charlotte Hornets, New York Knicks, Atlanta Hawks, Golden State Warriors.

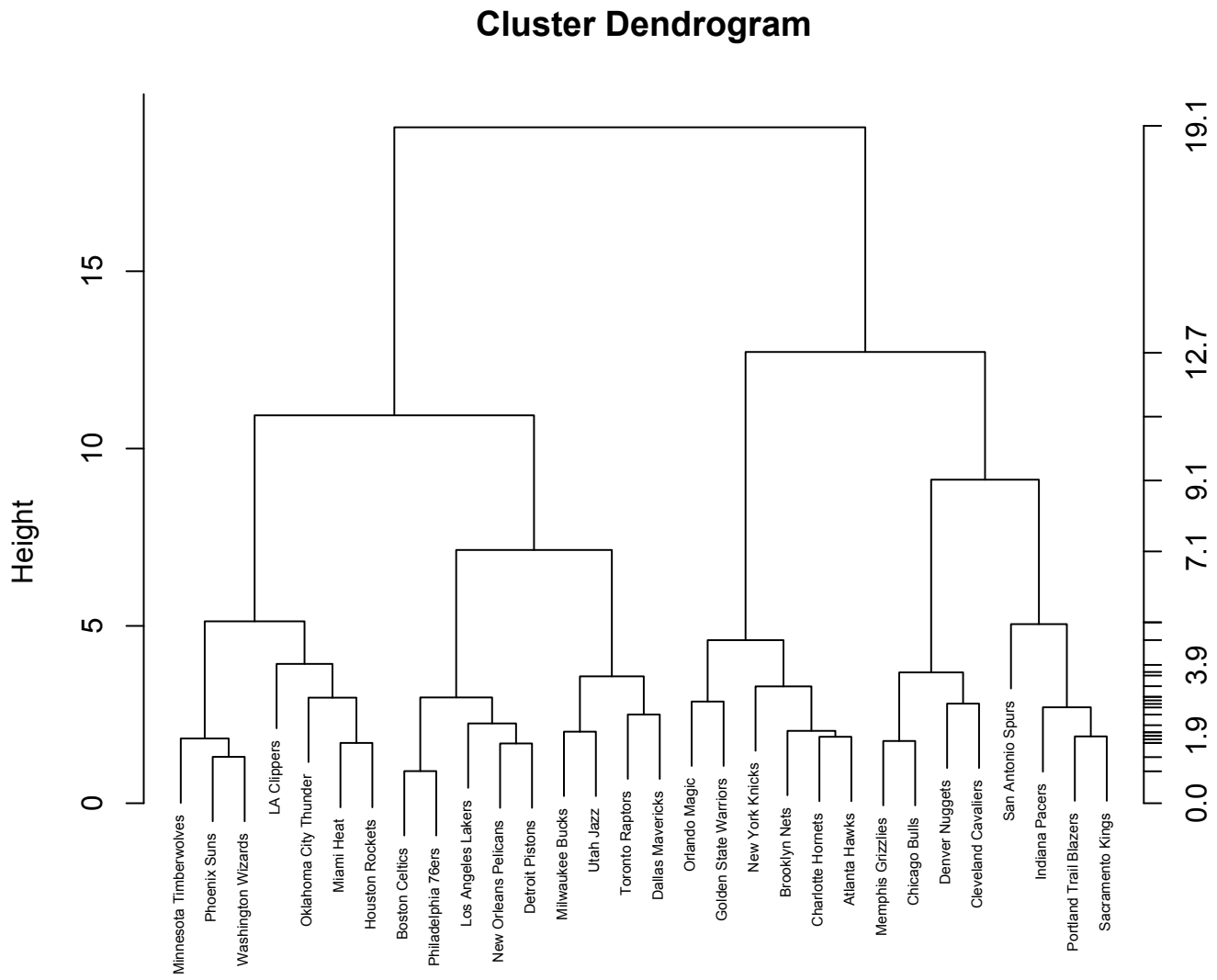


Figura 3.11: Taglio del dendrogramma ricavato dal metodo di Ward.

Di seguito viene presentata la matrice dei dati, la quale racchiude la media di ogni gruppo per le quattro variabili:

Gruppo 1	EFG_ptc	FTARATE_ptc	TOV_pct	OREB_pct
1	54,03	25,92	14,40	27,16
2	53,20	28,46	14,26	26,11
3	52,48	23,05	15,13	28,28
4	53,08	23,55	13,13	24,83
5	50,58	25,88	14,53	27,57

Figura 3.12: Matrice delle medie delle variabili per gruppo.

3.4 Analisi non gerarchica: metodo *k-means*

Dopo aver applicato la metodologia dell'analisi gerarchica ed in particolare il metodo di Ward, attraverso il quale si è sviluppato il dendrogramma ed il taglio ad esso connesso, l'analisi di questo elaborato, procede con l'applicazione del metodo non gerarchico del k-*means*. L'algoritmo, dividerà le unità statistiche nel un numero di gruppi precedentemente ottenuto e fissato dal metodo gerarchico di Ward, basandosi sull'ottimizzazione del criterio scelto, che nella maggior parte dei casi, come specificato nel secondo capitolo ed anche all'interno di questa analisi, coincide con quello della distanza euclidea, distanza che, garantisce convergenza all'algoritmo. I gruppi formatisi sono per l'appunto 5, e seguono tale partizione:

- Gruppo 1: Indiana Pacers, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs
- Gruppo 2: Los Angeles Lakers, LA Clippers, Miami Heat, Houston Rockets
- Gruppo 3: Milwaukee Bucks, Toronto Raptors, Utah Jazz, Oklahoma City Thunder, Dallas Mavericks, Phoenix Suns, Washington Wizards
- Gruppo 4: Brooklyn Nets, Orlando Magic, Charlotte Hornets, New York Knicks, Atlanta Hawks, Minnesota Timberwolves, Golden State Warriors
- Gruppo 5: Boston Celtics, Denver Nuggets, Philadelphia 76ers, Memphis Grizzlies, New Orleans Pelicans, Chicago Bulls, Detroit Pistons, Cleveland Cavaliers.

Di seguito viene presentata la matrice dei dati contenente i centroidi dei gruppi:

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc
1	53,08	23,55	13,13	24,83
2	54,20	28,65	14,58	27,20
3	53,99	26,89	13,97	25,61
4	50,70	26,14	14,54	27,40
5	52,81	24,28	14,90	28,23

Figura 3.13: Matrice dei dati rappresentativa dei centroidi dei gruppi formatisi con il metodo *k-means.*

Nel caso specifico di questo metodo applicativo, il calcolo dell'indice R^2, a differenza di quello gerarchico, è automatico, e di seguito si riportano i valori della devianza e dell'indice R^2:

- Devianza Within: 116,94
- Devianza Between: 183,95
- Devianza Totale: 300,88
- Indice R^2: 61,13%

3.5 Analisi dei gruppi

Prima di entrare nel vivo con l'analisi dei gruppi, è utile riportare la matrice dei dati che mostra l'andamento medio dei gruppi alle quattro variabili statistiche, e la rappresentazione grafica ad essa connessa.
Quest'ultima, si potrà rivelare utile nell'analisi specifica dei gruppi:

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	53,08	23,55	13,13	24,83	4
2	54,20	28,65	14,58	27,20	4
3	53,99	26,89	13,97	25,61	7
4	50,70	26,14	14,54	27,40	7
5	52,81	24,28	14,90	28,23	8
MEDIE GENERALI	52,81	25,81	14,32	26,83	30

Figura 3.14: Matrice rappresentativa dell'andamento medio dei gruppi per le quattro variabili.

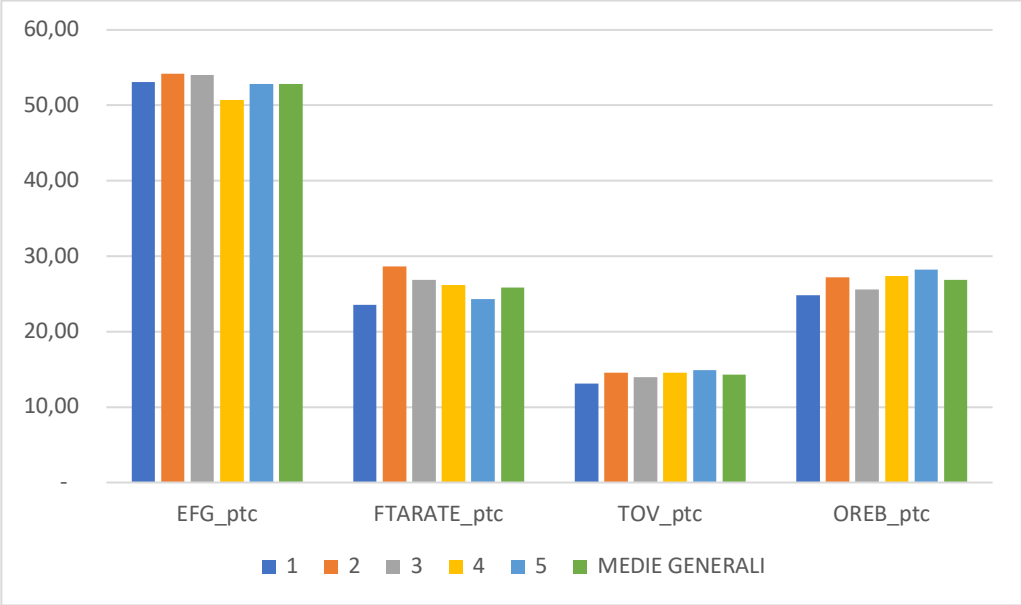


Figura 3.15: Rappresentazione grafica dei valori medi dei gruppi per le quattro variabili.

3.5.1 Gruppo 1: Indiana Pacers, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs

Il primo gruppo, si compone di quattro elementi, ovvero 4 squadre NBA (Indiana Pacers, Portland Trail Blazers, Sacramento Kings, San Antonio Spurs), ricoprendo il 13,3% dell'intero campionario. La composizione all'interno del gruppo, è abbastanza omogenea, sia in termini statistici che geografici. Se volessimo fare una prima distinzione fra *Western* ed *Eastern Conference*, è visibile ad occhio nudo, la netta maggioranza dei team occidentali su quelli orientali, con l'unica eccezione degli Indiana Pacers, in rapporto 3:1. Per quanto riguarda l'analisi statistica, questo gruppo, presenta solamente una variabile delle quattro sopra i valori delle medie generali. La variabile in questione è l'*effective field goal percentage (eFG%)* con il 53,08%, che come anticipato precedentemente rappresenta quella con più peso specifico delle quattro (40%). Nonostante ciò, questo non basta ad annoverare questo gruppo tra la schiera dei migliori, anzi, per ben due percentuali statistiche *(Free throw rate* e *Offensive Rebounding percentage (OREB%))*, presenta i risultati peggiori, ovvero 23,55% e 24,83%, in rapporto alle medie generali. Quanto detto, è particolarmente denotabile dal seguente grafico:

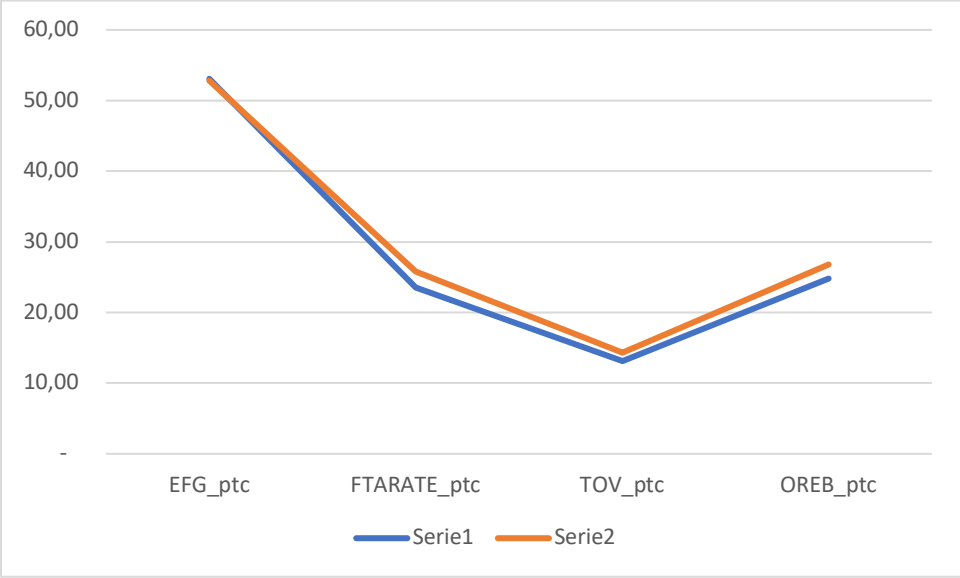


Figura 3.16: Confronto fra le medie generali ed i *Four Factors* del Gruppo 1.

In particolare, i San Antonio Spurs, squadra della Western Conference, non solo in termini relativi ricoprono il posto più basso in classifica (20°), e quindi rappresentano il peggiore dei quattro team; ma anche, in termini assoluti, sono il team peggiore della NBA per *Free throw rate,* presentando il minimo valore di 12,10%. L'analisi individuale del gruppo, si rispecchia anche a livello generale. Confrontando i centroidi del gruppo con le medie generali, si è in grado di stabilire chi presenta valori accettabili e chi no. Nella tabella seguente, vengono evidenziati in verde i valori del gruppo sopra la media ed in rosso quelli sotto la media:

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	0,26	- 2,26	- 1,20	- 2,01	4
2	1,39	2,84	0,25	0,37	4
3	1,17	1,08	- 0,35	- 1,22	7
4	- 2,11	0,34	0,22	0,57	7
5	-	- 1,53	0,58	1,39	8

Figura 3.17: Tabella illustrativa del rapporto fra centroidi del Gruppo 1 e le medie generali.

Guardando alla tabella, è facile da notare, come il Gruppo 1, è l'unico a presentare solamente un centroide al di sopra delle medie generali, e questo potrebbe assegnarli il titolo di gruppo peggiore.
La nostra analisi, tuttavia non si ferma, e per avere un'ulteriore prova, in questo elaborato, si procede con lo studio delle Last 10, ovvero le ultime dieci partite giocate dalle squadre, in modo da poter avere uno sguardo più recente alla composizione.

	LAST 10	
TEAM	W	L
"Indiana Pacers"	7,00	3,00
"Portland Trail Blazers"	4,00	6,00
"Sacramento Kings"	7,00	3,00
"San Antonio Spurs"	5,00	5,00
MEDIE LAST 10 GENARALI	5,00	5,00
MEDIE LAST 10 GRUPPO 1	5,75	4,25

Figura 3.18: Andamento dei team appartenenti al Gruppo 1 nelle ultime 10 partite.

Riguardo alle vittorie e sconfitte delle ultime 10 partite, il gruppo presenta degli accennati miglioramenti, rispetto a quello emerso dall'analisi dei centroidi e le medie generali. Per l'appunto, due team (Indiana Pacers e Sacramento Kings) detengono un record positivo, con 7 vittorie e 3 sconfitte all'attivo; di contro, i Portland Trail Blazers, sono l'unico team dei 4 che presenta un record negativo con 4 vittore e 6 sconfitte. Caso speciale, è ancora quello dei San Antonio Spurs, che non solo ha un eguale numero di vittorie e sconfitte, ma inoltre tale rapporto, coincide perfettamente con le medie generali delle ultime 10 partite di tutti i team. A livello generale, il Gruppo 1 ha una media di 5,75 vittorie e 4,25 sconfitte nelle Last 10, numeri che sono visibilmente migliori rapportati a quelli generali, che vedono una parità fra vittorie e sconfitte di 5 e 5. Dall'analisi di questo gruppo, si può concludere che, sebbene questo, presenti dei valori che potrebbero classificarlo tra i peggiori (se non il peggiore in assoluto), si intravede un possibile miglioramento guardando alle ultime 10 partite, che potrebbe presagire una scalata nella classifica generale, se dovesse continuare ad ottenere questi risultati.

3.5.2 Gruppo 2: Los Angeles Lakers, LA Clippers, Miami Heat, Houston Rockets

Il secondo Gruppo è, rispetto al primo, uguale per numerosità ma non per composizione. I team che lo compongono (Los Angeles Lakers, LA Clippers, Miami Heat, Houston Rockets) ricoprono il 13,3% della partizione. In termini geografici, come verificatosi nel gruppo precedente, vi è una predominanza dei team di Ovest (75%) contro l'unico di Est (25%), rappresentato dai Miami Heat. Dal punto di vista statistico invece, il Gruppo 2, è l'unico dei 5, che presenta tutte le variabili sopra la media, e quindi si proietta ad essere il migliore tra quelli analizzati. Per le prime due variabili, *effective field goal percentage (eFG%)* e *Free throw rate,* detiene le migliori percentuali, con il 54,20 ed il 28,65, a discapito di quelle generali che raggiungono i valori di 52,81 e 25,81. Questo rapporto è rappresentato dal grafico sottostante:

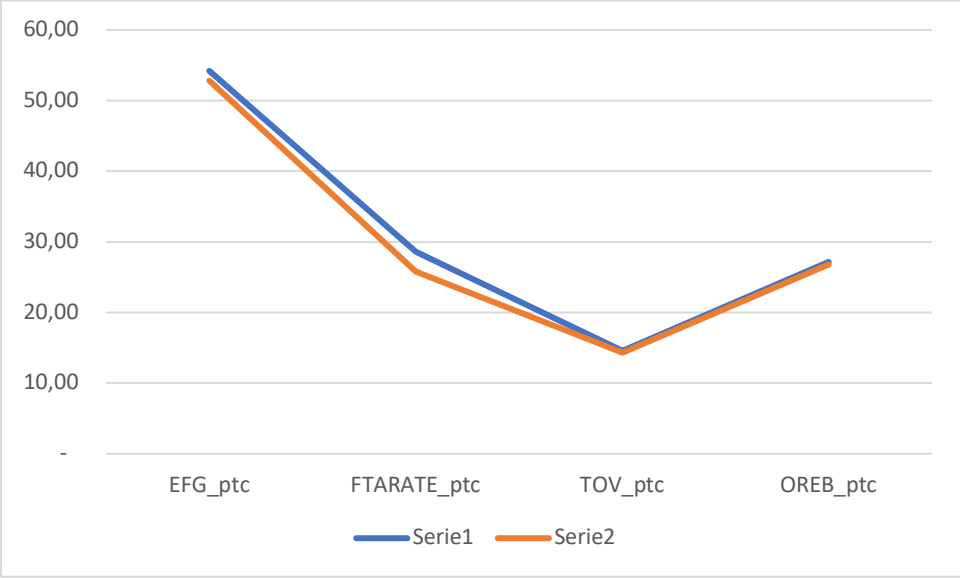


Figura 3.19: Confronto fra le medie generali ed i *Four Factors* del Gruppo 2.

Nello specifico, i Miami Heat, unica squadra della *Eastern Conference,* detengono i migliori risultati in termini di *Free throw rate* (29,80%) dell'intera classifica. Storicamente, le squadre degli stati occidentali, sono state sempre più forti ed ottenuto migliori risultati, ma questo insieme ad altri dati, potrebbe indicare un cambio di rotta nella storia della franchigia. Come dimostrato successivamente:

ANNO	TEAM	CONFERENCE
2010	Los Angeles Lakers	Western
2011	Dallas Mavericks	Western
2012	Miami Heat	Eastern
2013	Miami Heat	Eastern
2014	San Antonio Spurs	Western
2015	Golden State Warriors	Western
2016	Cleveland Cavaliers	Eastern
2017	Golden State Warriors	Western
2018	Golden State Warriors	Western
2019	Toronto Raptors	Eastern

Figura 3.20: Storico delle vittorie del campionato nell'ultima decade.

Nelle ultime 10 stagioni, il campionato è stato sei volte vinto da un team della *Western Conference*, contro i 4 vinti dai team della *Eastern Conference,* ma ciò potrebbe cambiare in futuro.

Come fatto per il primo gruppo, all'analisi individuale, segue quella generale, comparando i centroidi con le medie generali.

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	0,26	- 2,26	- 1,20	- 2,01	4
2	1,39	2,84	0,25	0,37	4
3	1,17	1,08	- 0,35	- 1,22	7
4	- 2,11	0,34	0,22	0,57	7
5	-	- 1,53	0,58	1,39	8

Figura 3.21: Tabella illustrativa del rapporto fra centroidi del Gruppo 2 e le medie generali.

Da tale rapporto, si potrebbe ipotizzare, che il Gruppo 2, sia il migliore dei 5, anche per il fatto che presenta come detto, l'*effective field goal percentage (eFG%)* migliore, ovvero la variabile con più peso specifico delle quattro.

Per completare l'analisi di questo gruppo, viene presentato l'andamento delle Last 10 sotto forma di tabella, nella quale vengono evidenziate i successi e gli insuccessi delle singole squadre.

	LAST 10	
TEAM	W	L
"Los Angeles Lakers"	8,00	2,00
"LA Clippers"	7,00	3,00
"Miami Heat"	6,00	4,00
"Houston Rockets"	6,00	4,00
MEDIE LAST 10 GENERALI	5,00	5,00
MEDIE LAST 10 GRUPPO 2	6,75	3,25

Figura 3.22: Andamento dei team appartenenti al Gruppo 2 nelle ultime 10 partite.

Da un primo sguardo alla tabella, si può notare che tutti gli elementi che compongono il gruppo, hanno ottenuto nel corso delle ultime 10 partite più successi che insuccessi. In questa particolare classifica, guidano i Los Angeles Lakers, con 8 vittorie su 10 partite, per concludere con i Miami Heat e gli Houston Rockets a pari merito con 6 e 4. Analizzando più a fondo l'andamento degli altri gruppi, il Gruppo 2 risulta l'unico ad avere al suo interno tutti gli elementi con record positivi (Last 10), ed implicitamente quello con la media di

vittorie più alta, con 6,75 su 10 partite. Dall'analisi individuale, si era già denotato qualche segnale riguardo l'eccellenza di questo gruppo, e se questo non fosse bastato, l'analisi a livello generale e l'andamento del gruppo nelle Last 10, hanno confermato quanto si ci aspettava, ovvero il primato del Gruppo 2.

3.5.3 Gruppo 3: Milwaukee Bucks, Toronto Raptors, Utah Jazz, Oklahoma City Thunder, Dallas Mavericks, Phoenix Suns, Washington Wizards

Il Gruppo 3, a differenza dei primi due, ha una numerosità più alta, comprendente ben 7 squadre (Milwaukee Bucks, Toronto Raptors, Utah Jazz, Oklahoma City Thunder, Dallas Mavericks, Phoenix Suns, Washington Wizards), le quali ricoprono il 23,3% delle unità. Da un punto di vista geografico, ancora una volta, pur essendosi affievolita rispetto ai casi precedenti, si vede una supremazia dei team provenienti da Ovest (4) contro quelli dell'Est (3). Da un punto di vista analitico, invece, il Gruppo 3, guardando a prima vista l'*effective field goal percentage (eFG%),* sembra il gruppo che potrebbe contendere con il secondo. Esso infatti presenta un *eFG%* pari a 53,99%, ampiamente sopra la media (52,81%) e di poco lontano dal valore del secondo gruppo (54,20%). Anche il *Free throw rate,* contiene valori incoraggianti con (26,86%), essendo solamente secondo al valore del Gruppo 2. Andando avanti nell'analisi tuttavia, ai buoni risultati precedentemente raccolti, se ne alternano alcuni che compromettono la corsa al titolo di questo gruppo. Nello specifico, *Free throw rate* e *Offensive Rebounding percentage (OREB%)* e *turnovers per possession,* rappresentano i centroidi al di sotto delle medie generali, con valori rispettivamente di 13,97% e 25,61%, come dimostrato graficamente:

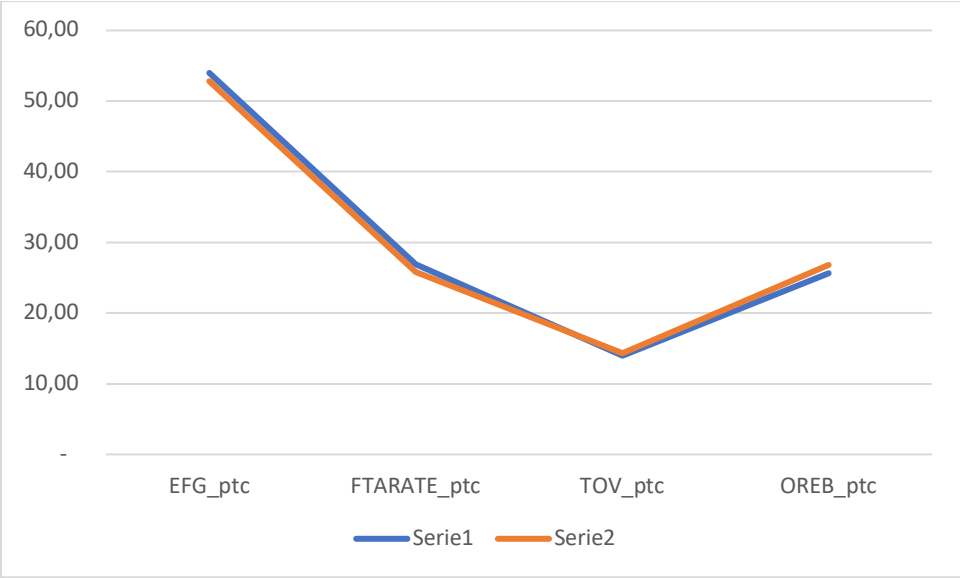


Figura 3.23: Confronto fra le medie generali ed i *Four Factors* del Gruppo 3.

Nonostante nel complesso i risultati non siano i migliori, singolarmente il Gruppo 3, presenta alcune squadre, ed in particolare i Milwaukee Bucks, atipiche per il gruppo d'appartenenza. Questa squadra, non

solo presenta il più alto valore registrato di *eFG%* (55,30%), ma inoltre guida la classifica nel rapporto tra vittorie e sconfitte. Questo riprendere in parte quanto anticipato prima; se i Miami Heat rappresentavano un caso singolare perché proveniente da Est, la nostra tesi si avvalora maggiormente con quanto analizzato di un'altra squadra proveniente dagli stati orientali, come Milwaukee.

A seguire, una tabella illustrativa, che mette in rapporto i team sulla base dei centroidi e delle medie generali.

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	0,26	- 2,26	- 1,20	- 2,01	4
2	1,39	2,84	0,25	0,37	4
3	1,17	1,08	- 0,35	- 1,22	7
4	- 2,11	0,34	0,22	0,57	7
5	-	- 1,53	0,58	1,39	8

Figura 3.24: Tabella illustrativa del rapporto fra centroidi del Gruppo 3 e le medie generali.

Infine, l'analisi di questo gruppo, passa al focus sulle ultime 10 partite ed i risultati ad esse connessi.

	LAST 10	
TEAM	W	L
"Milwaukee Bucks"	6,00	4,00
"Toronto Raptors"	6,00	4,00
"Utah Jazz"	5,00	5,00
"Oklahoma City Thunder"	8,00	2,00
"Dallas Mavericks"	6,00	4,00
"Phoenix Suns"	4,00	6,00
"Washington Wizards"	4,00	6,00
MEDIE LAST 10 GENERALI	5,00	5,00
MEDIE LAST 10 GRUPPO 3	5,57	4,43

Figura 3.25: Andamento dei team appartenenti al Gruppo 3 nelle ultime 10 partite.

Anche in questo caso, le vittorie del gruppo sono più delle sconfitte, con valori di 5,57 e 4,43 (Last 10). Questo risultato, è in gran parte supportato, dalla squadra degli Oklahoma City Thunder, i quali a differenza degli altri team, che vedono un'alternanza dei valori 6 e 4 tra vittorie e sconfitte (con la solo eccezione di Utah), hanno vinto 8 delle ultime 10 partite, balzando al 10° posto in classifica. Guardando all'analisi individuale che a quella generale, possiamo affermare che, il Gruppo 3, occupa un ruolo intermedio, se non addirittura da contender, in questa particolare classifica tra i cluster.

3.5.4 Gruppo 4: Brooklyn Nets, Orlando Magic, Charlotte Hornets, New York Knicks, Atlanta Hawks, Minnesota Timberwolves, Golden State Warriors

Il Gruppo 4, si presenta agli occhi dell'analisi, con la stessa numerosità, ma con sostanziali differenze del gruppo precedente. Tale gruppo, è composto da 7 team (Brooklyn Nets, Orlando Magic, Charlotte Hornets, New York Knicks, Atlanta Hawks, Minnesota Timberwolves, Golden State Warriors), e ricopre il 23,3% dei campioni studiati in questo elaborato. Guardando alla disposizione geografica, notiamo come la predominanza dei team di Ovest, in questo gruppo cambia radicalmente. Per l'esattezza, questa volta 5 team su 7 provengono dalla Eastern Conference, ed a rappresentare la Western vi sono solamente i Minnesota Timberwolves ed i Golden State Warriors. Oltre all'aspetto geografico, anche le caratteristiche analitiche di questo gruppo, sono singolari rispetto ai gruppi analizzati fin ora. Questo cluster, è in assoluto quello con l'*effective field goal percentage* minore (50,70%), un valore che non fa ben sperare, a maggio ragione considerando il peso specifico della variabile in questione (40%). Il grafico illustra quanto detto:

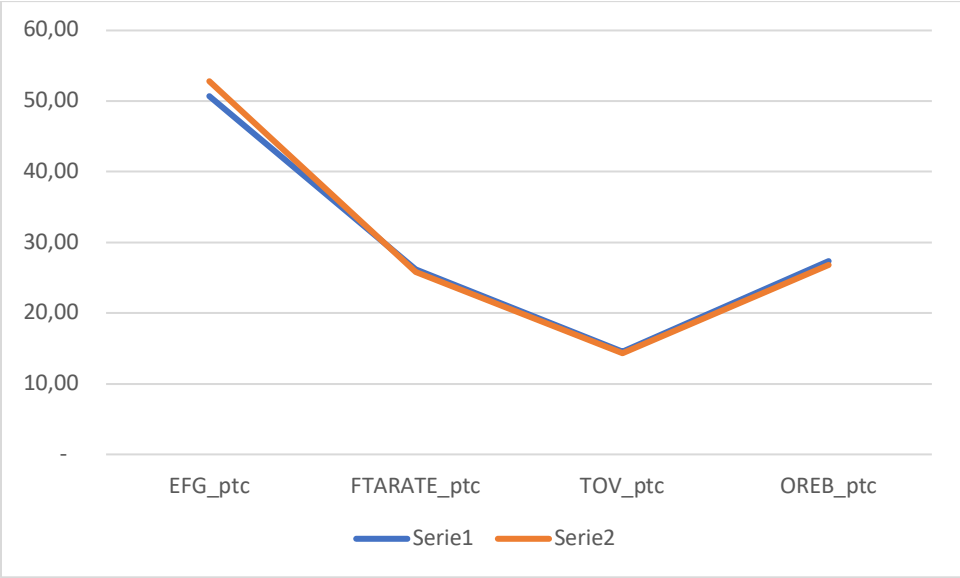


Figura 3.26: Confronto fra le medie generali ed i *Four Factors* del Gruppo 4.

Da un punto di vista individuale, si può notare che, nessuna delle squadre al suo interno, occupa una posizione nella prime 10, con i Brooklyn Nets, che risiedono esattamente a metà classifica (15°). Seguono

poi gli Orlando Magic (16°), e per concludere con i Golden State Warriors, i quali occupano l'ultimo posto in classifica. Queste indicazioni, ci forniscono delle previsioni utili per quando verrà analizzato l'andamento delle Last 10, e la percentuale di vittorie ad esse connesse. Di seguito la tabella che mette in relazione i centroidi del gruppo con le medie generali:

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	0,26	- 2,26	- 1,20	- 2,01	4
2	1,39	2,84	0,25	0,37	4
3	1,17	1,08	- 0,35	- 1,22	7
4	- 2,11	0,34	0,22	0,57	7
5	-	- 1,53	0,58	1,39	8

Figura 3.27: Tabella illustrativa del rapporto fra centroidi del Gruppo 4 e le medie generali.

Potremmo definire questo gruppo, come l'antagonista del Gruppo 1. Il Gruppo 4, è per l'appunto, l'unico in cui *(eFG%),* ha un valore negativo rispetto alle medie generali (- 2,11%), e presenta invece valori positivi per le altre tre variabili, nella tabella evidenziati in verde. Quanto visto sino ad ora, è in completamente agli antipodi con quanto avviene nel Gruppo 1, in cui l'unica variabile ad essere sopra la media è proprio *(eFG%),* mentre le tre rimanenti sono sotto la media, ossia presentano valori negativi. Come anticipato poc'anzi, le indicazioni dedotte dall'analisi del gruppo, possono dare un breve anticipo di ciò che afferma l'analisi delle Last 10:

	LAST 10	
TEAM	W	L
"Brooklyn Nets"	5,00	5,00
"Orlando Magic"	6,00	4,00
"Charlotte Hornets"	4,00	6,00
"New York Knicks"	4,00	6,00
"Atlanta Hawks"	4,00	6,00
"Minnesota Timberwolves"	3,00	7,00
"Golden State Warriors"	3,00	7,00
MEDIE LAST 10 GENERALI	5,00	5,00
MEDIE LAST 10 GRUPPO 4	4,14	5,86

Figura 3.28: Andamento dei team appartenenti al Gruppo 4 nelle ultime 10 partite.

Guardando alla tabella, possiamo confermare le nostre intuizioni. Ad esclusione degli Orlando Magic, i quali hanno un record positivo di 6/4, ed i Brooklyn Nets, il cui risultato si colloca ancora una volta nelle medie della classifica; tutti gli altri team hanno un record negativo, che contribuisce notevolmente ai valori osservati precedentemente. I risultati peggiori sono quelli ottenuti dai Minnesota Timberwolves e dai Golden State Warriors, squadra in cui gli infortuni in questa stagione, hanno decimato le speranze di poter continuare la dinastia vincente degli ultimi 5 anni, in cui hanno vinto 3 titoli NBA.

3.5.5 Gruppo 5: Boston Celtics, Denver Nuggets, Philadelphia 76ers, Memphis Grizzlies, New Orleans Pelicans, Chicago Bulls, Detroit Pistons, Cleveland Cavaliers

Il Gruppo 5, si presenta come l'ultimo cluster che si propone di analizzare questo scritto, ed inoltre il gruppo più numeroso, con 8 unità, rappresentanti il restante 26,6% della partizione. Da un punto di vista geografico, come avvenuto nel gruppo precedente, vi è una maggioranza dei team provenienti dagli stati orientali (5), contro i 4 provenienti da occidente. Dal punto di vista statistico invece, guardando alla classifica, si nota che, il gruppo è abbastanza eterogeneo al suo interno, con 4 squadre nella prima metà ed altrettante nella seconda metà. Difficilmente da come si ci potrebbe aspettare, sono proprio le squadre della seconda metà di classifica, a migliorare il gruppo con i propri risultati.

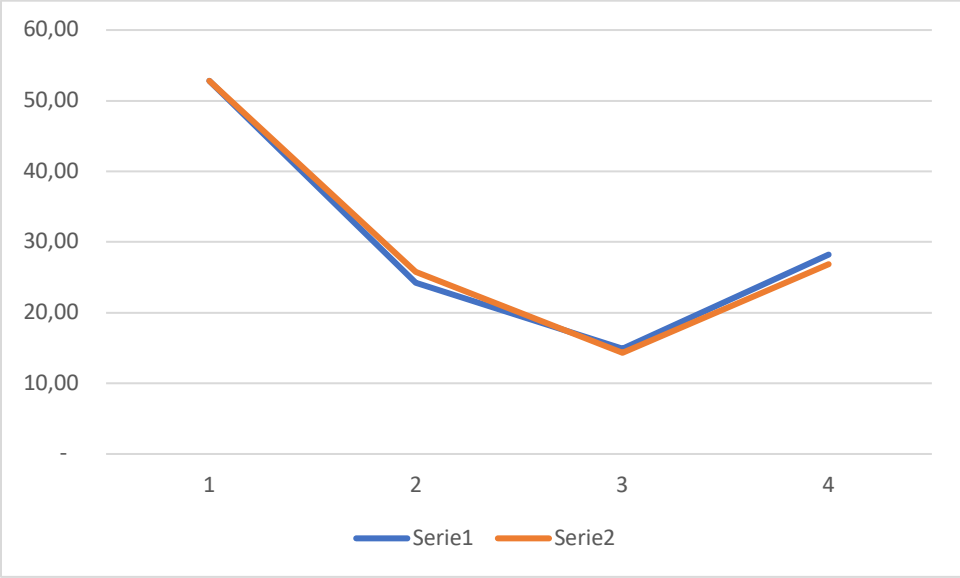


Figura 3.29: Confronto fra le medie generali ed i *Four Factors* del Gruppo 5.

Figura di spicco è occupata dai Cleveland Cavaliers, i quali non soltanto, possiedono il valore più alto all'interno del gruppo per *Offensive Rebounding percentage* (29,60%), ma soprattutto, guidano la classifica assoluta per *turnovers per possession (TOt/POSSt),* con il valore di (16,50%).

Nonostante questi risultati, a livello generale, il Gruppo 5 non si prospetta come uno dei migliori. Per l'appunto tale gruppo, presenta una *eFG%* coincidente alla media generale, e come dimostra la tabella sottostante, un *Free throw rate,* ben al di sotto della media (-1,53%):

GRUPPO	EFG_ptc	FTARATE_ptc	TOV_ptc	OREB_ptc	NUM.
1	0,26	- 2,26	- 1,20	- 2,01	4
2	1,39	2,84	0,25	0,37	4
3	1,17	1,08	- 0,35	- 1,22	7
4	- 2,11	0,34	0,22	0,57	7
5	-	- 1,53	0,58	1,39	8

Figura 3.30: Tabella illustrativa del rapporto fra centroidi del Gruppo 5 e le medie generali.

La nostra previsione, è confermata dallo studio delle ultime 10 partite giocate, in cui emergono i seguenti risultati:

	LAST 10	
TEAM	W	L
"Boston Celtics"	5,00	5,00
"Denver Nuggets"	5,00	5,00
"Philadelphia 76ers"	5,00	5,00
"Memphis Grizzlies"	4,00	6,00
"New Orleans Pelicans"	5,00	5,00
"Chicago Bulls"	3,00	7,00
"Detroit Pistons"	1,00	9,00
"Cleveland Cavaliers"	4,00	6,00
MEDIE LAST 10 GENERALI	5,00	5,00
MEDIE LAST 10 GRUPPO 5	4,00	6,00

Figura 3.31: Andamento dei team appartenenti al Gruppo 5 nelle ultime 10 partite.

La tabella, mostra un caso emblematico mai verificatosi prima, ovvero un gruppo in cui nessuna delle squadre detiene un record positivo di vittorie nelle ultime 10 partite. Singolare, è inoltre, il caso dei Detroit Pistons, i quali sono l'unica squadra dell'intera classifica ad aver vinto solamente 1 partita negli ultimi 10 eventi, andamento che li ha visti scendere notevolmente di posizioni in classifica; e cosa più importante ai

fini della nostra analisi, hanno contribuito in modo sostanziale nel determinare la peggiore percentuale di vittorie (Last 10), detenuta appunto dal Gruppo 5.

Dopo aver analizzato i singoli gruppi, ed avere fornito una prospettiva sull'andamento dei diversi team che li compongono, l'analisi di questo elaborato, si conclude con la stima della percentuale di vittoria del rappresentante delle squadre di ogni singolo gruppo (Win%). Questa stima, a differenza di come fatto fino ad ora, tiene conto dei singoli pesi specifici delle variabili, dei quali se ne fornisce un rapido elenco:

a) $eFG\%$ = 40%
b) $TO_t/POSS_t$ = 25%
c) $REB\%$ = 20%
d) $FTMt/FGAt = 15\%$

Attraverso questi pesi, si procede a calcolare i valori dei centroidi dei singoli team e la percentuale di vittoria a loro connessa, in modo da poter decretare quali siano i gruppi migliori e quali i peggiori:

GRUPPO	**EFG_ptc (1)**	**FTARATE_ptc (1)**	**TOV_ptc (1)**	**OREB_ptc (1)**	**NUM.**	**WIN%**
1	21,23	5,89	2,63	3,72	4	0,33
2	21,68	7,16	2,92	4,08	4	0,36
3	21,59	6,72	2,79	3,84	7	0,35
4	20,28	6,54	2,91	4,11	7	0,34
5	21,13	6,07	2,98	4,23	8	0,34
MEDIE GENERALI	52,81	25,81	14,32	26,83	30,00	

Figura 3.32: Tabella illustrativa della percentuale di vittoria (*Win%)* calcolata sulla base dei valori dei centroidi dei singoli team.

Dalla tabella, si evince chiaramente che il miglior gruppo sia il 2 con una percentuale di vittoria del 36%, seguito del Gruppo 3 con una percentuale di poco inferiore (35%) e dai Gruppi 3 e 4 che condividono la stessa percentuale di vittoria (34%), infine abbiamo il Gruppo 1, il quale presenta la percentuale più bassa con il 33%.

CONCLUSIONI

Questo elaborato, si è preposto l'obbiettivo di spiegare i benefici della statistica all'interno dello sport, dapprima fornendo opportune testimonianze e descrivendo il fenomeno assai complesso, e successivamente è entrato nel vivo attraverso la *Cluster Analysis*. Questa metodologia analitica, ha portato alla luce ed ha permesso di studiare i punti di forza e debolezza delle unità statistiche scelte, ovvero le 30 squadre della NBA. Utilizzando il metodo di Ward, attraverso la formazione di due gruppi, si è evidenziata graficamente una frattura fra le due *division: Eastern e Western Conference.* Il primo gruppo, presentava una maggioranza di team provenienti dalla Eastern Conference; inversamente, il secondo gruppo, era maggiormente composto da squadre della Western Conference, rispetto a quelle della Eastern Conference. Dopo aver osservato il dendrogramma ed aver calcolato l'indice di concentrazione R^2, si è scelta una partizione a cinque gruppi. Tale composizione, è stata generata dal metodo *k-means,* metodo non gerarchico, che ci ha permesso di ottimizzare il criterio inizialmente scelto, ovvero quello della distanza euclidea. L'analisi dei gruppi, ha portato alla luce gli elementi di forza e debolezza insiti all'interno delle variabili scelte. Senza dubbio, tra le cinque analizzate, le performance migliori sono riconducibili al secondo, al terzo ed al quarto gruppo; mentre quelle peggiori sono attribuibili al primo ed al quinto gruppo, cluster questi, che presentano statistiche sotto la media sia a livello generale che individuale. I risultati ottenuti dall'analisi, esprimono l'andamento dei *Four Factors of Basketball Success,* i quali uniti fra di loro, ci hanno permesso di calcolare, tenendo conto dei loro specifici pesi, la percentuale di vittoria *(Win%)* del centroide di ogni gruppo, in modo da poter stilare una classifica dei gruppi studiati. Nonostante ciò, se le vittorie si ottengono attraverso l'implementazione di una strategia di gioco, il nostro fine ultimo non potrà essere una mera analisi descrittiva, ma in quest'ultima parte del nostro elaborato, si ci proporrà di tradurre sul campo, i dati osservati fino ad ora, facendo un'analisi predittiva delle vittorie di ogni singola squadra, nella stagione in corso. Attraverso vari algoritmi che sono in grado di prevedere il record atteso, è possibile costruire modelli che forniscono stime. I *Four Factors,* sono la base di molti di essi. Di seguito, verrà eseguito un algoritmo, la cui equazione è la seguente:

$$Projected\ wins = 40 * Team\ eFG\% - 25 * Team\ TOV\% + 20 * OREB\% + 15 * TeamFTR - 40 *$$
$$Opp\ eFG\% + 25 * OppTOV\% + 20 * Opp\ OREB\% - 10 * Opp\ FTR$$

Partendo dalla seguente matrice di dati:

TEAM	EFG%	FTA RATE	TOV%	OREB%	OPP EFG%	OPP FTA RATE	OPP TOV%	OPP OREB%
Milwaukee Bucks	55,30	26,30	14,10	24,10	48,60	22,00	13,50	22,40
Los Angeles Lakers	54,80	26,80	14,90	28,40	50,90	26,00	15,40	26,60
Toronto Raptors	53,60	25,60	14,20	25,90	50,20	26,00	16,60	28,50
LA Clippers	53,20	29,30	14,40	28,40	50,30	27,50	13,80	26,60
Boston Celtics	52,90	25,50	13,60	28,40	51,00	27,80	15,20	26,80
Denver Nuggets	53,20	23,00	13,90	29,40	52,60	25,60	14,50	27,30
Utah Jazz	55,20	26,90	15,00	25,80	51,40	23,40	12,10	25,30
Miami Heat	54,90	29,80	14,90	25,90	52,20	26,70	13,70	24,40
Houston Rockets	53,90	28,70	14,10	26,10	52,80	25,50	15,20	28,60
Oklahoma City Thunder	53,40	29,20	13,50	23,90	52,00	20,60	14,00	27,20
Indiana Pacers	53,30	21,70	13,20	24,60	51,30	24,50	14,40	27,90
Philadelphia 76ers	53,00	25,20	14,20	27,80	52,20	28,40	14,10	24,60
Dallas Mavericks	54,80	25,70	12,70	27,50	51,90	22,70	12,20	26,20
Memphis Grizzlies	53,00	23,40	14,80	27,40	52,00	27,00	14,10	27,00
Brooklyn Nets	51,50	26,80	15,10	28,40	50,70	23,70	12,30	26,30
Orlando Magic	50,30	24,80	12,70	26,80	53,50	22,10	15,00	25,20
Portland Trail Blazers	53,00	23,90	12,80	26,30	52,30	26,40	12,50	29,00
New Orleans Pelicans	53,80	25,20	15,40	28,80	53,10	26,50	13,80	26,90
Sacramento Kings	53,10	23,20	14,40	25,60	54,00	27,90	15,10	26,50
San Antonio Spurs	52,90	25,40	12,10	22,80	54,20	25,20	12,70	24,90
Phoenix Suns	52,80	27,30	14,80	26,10	54,30	28,30	15,60	27,20
Washington Wizards	52,80	27,20	13,50	26,00	56,00	29,70	15,70	29,20
Charlotte Hornets	50,40	25,20	15,00	28,00	54,60	21,00	14,80	29,40
Chicago Bulls	51,50	23,10	15,30	26,70	54,60	31,50	18,10	28,10
New York Knicks	50,10	26,30	14,30	30,00	54,10	29,90	13,80	25,80
Detroit Pistons	52,90	26,10	15,50	27,70	54,10	23,90	14,20	28,30
Atlanta Hawks	51,50	25,80	15,50	26,70	54,30	30,30	14,30	29,20
Minnesota Timberwolves	51,40	27,70	14,60	26,40	54,10	27,90	14,70	26,90
Cleveland Cavaliers	52,20	22,70	16,50	29,60	56,00	21,40	13,00	26,20

Golden State Warriors	49,70	26,40	14,60	25,50	55,30	24,70	15,40	27,80

Figura 3.33: Matrice dei dati originale.

Utilizzando i diversi pesi assegnati, l'equazione restituisce i seguenti risultati:

TEAM	GP	W	L	WIN %	PARTITE RIMANENTI	PROJECTED WINS
Milwaukee Bucks	65	53	12	0,815	17	13,86
Los Angeles Lakers	63	49	14	0,778	19	14,78
Toronto Raptors	64	46	18	0,719	18	12,94
LA Clippers	64	44	20	0,688	18	12,38
Boston Celtics	64	43	21	0,672	18	12,10
Denver Nuggets	65	43	22	0,662	17	11,25
Utah Jazz	64	41	23	0,641	18	11,54
Miami Heat	65	41	24	0,631	17	10,73
Houston Rockets	64	40	24	0,625	18	11,25
Oklahoma City Thunder	64	40	24	0,625	18	11,25
Indiana Pacers	65	39	26	0,600	17	10,20
Philadelphia 76ers	65	39	26	0,600	17	10,20
Dallas Mavericks	67	40	27	0,597	15	8,96
Memphis Grizzlies	65	32	33	0,492	17	8,36
Brooklyn Nets	64	30	34	0,469	18	8,44
Orlando Magic	65	30	35	0,462	17	7,85
Portland Trail Blazers	66	29	37	0,439	16	7,02
New Orleans Pelicans	64	28	36	0,438	18	7,88
Sacramento Kings	64	28	36	0,438	18	7,88
San Antonio Spurs	63	27	36	0,429	19	8,15
Phoenix Suns	65	26	39	0,400	17	6,80
Washington Wizards	64	24	40	0,375	18	6,75
Charlotte Hornets	65	23	42	0,354	17	6,02
Chicago Bulls	65	22	43	0,338	17	5,75
New York Knicks	66	21	45	0,318	16	5,09
Detroit Pistons	66	20	46	0,303	16	4,85

Atlanta Hawks	67	20	47	0,299	15	4,49
Minnesota Timberwolves	64	19	45	0,297	18	5,35
Cleveland Cavaliers	65	19	46	0,292	17	4,96
Golden State Warriors	65	15	50	0,231	17	3,93

Figura 3.34: Tabella rappresentante le Projected wins per le 30 unità statistiche.

Osservando la tabella, è possibile proiettare le statistiche di ogni singolo team fino alla fine del campionato, in modo da poter stilare una classifica ipotetica. Questo è ciò che rende affidabili i *Four Factors:* ovvero la possibilità di una squadra, di poter fare affidamento su di loro, per realizzare i suoi punti di forza e di debolezza.

Di seguito verrà stilata la classifica aggiornata basandosi sulle previsioni effettuate:

TEAM	GP	W	L	WIN %	PARTITE RIMANENTI	PROJECTED WINS	WIN TOT.
Milwaukee Bucks	65	53	12	0,815	17	13,86	67
Los Angeles Lakers	63	49	14	0,778	19	14,78	64
Toronto Raptors	64	46	18	0,719	18	12,94	59
LA Clippers	64	44	20	0,688	18	12,38	56
Boston Celtics	64	43	21	0,672	18	12,10	55
Denver Nuggets	65	43	22	0,662	17	11,25	54
Utah Jazz	64	41	23	0,641	18	11,54	53
Miami Heat	65	41	24	0,631	17	10,73	52
Houston Rockets	64	40	24	0,625	18	11,25	51
Oklahoma City Thunder	64	40	24	0,625	18	11,25	51
Indiana Pacers	65	39	26	0,600	17	10,20	49
Philadelphia 76ers	65	39	26	0,600	17	10,20	49
Dallas Mavericks	67	40	27	0,597	15	8,96	49
Memphis Grizzlies	65	32	33	0,492	17	8,36	40
Brooklyn Nets	64	30	34	0,469	18	8,44	38
Orlando Magic	65	30	35	0,462	17	7,85	38
Portland Trail Blazers	66	29	37	0,439	16	7,02	36
New Orleans Pelicans	64	28	36	0,438	18	7,88	36

Sacramento Kings	64	28	36	0,438	18	7,88	36
San Antonio Spurs	63	27	36	0,429	19	8,15	35
Phoenix Suns	65	26	39	0,400	17	6,80	33
Washington Wizards	64	24	40	0,375	18	6,75	31
Charlotte Hornets	65	23	42	0,354	17	6,02	29
Chicago Bulls	65	22	43	0,338	17	5,75	28
New York Knicks	66	21	45	0,318	16	5,09	26
Detroit Pistons	66	20	46	0,303	16	4,85	25
Atlanta Hawks	67	20	47	0,299	15	4,49	24
Minnesota Timberwolves	64	19	45	0,297	18	5,35	24
Cleveland Cavaliers	65	19	46	0,292	17	4,96	24
Golden State Warriors	65	15	50	0,231	17	3,93	19

Figura 3.35: Proiezione della classifica finale della stagione in corso 2019/20.

Guardando alla tabella, è facile da notare, che pur variando il totale delle vittorie dei singoli team, la classifica rimane invariata. Questa classifica, tuttavia, non tiene conto né dei possibili scontri diretti tra le squadre che hanno un egual punteggio nè dalla differenza canestri, aspetti questi, che potranno cambiare la classifica e le sorti dell'intero campionato. La statistica ed i dati connessi ad essa, potranno quindi aiutare i singoli team nello sviluppo del loro sistema di gioco, curandone ogni singolo dettaglio. Come disse Kobe Bryant: *"Determination wins games, but detail wins championship".*

BIBLIOGRAFIA

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

Albert, J., Glickman, M. E., Swartz, T. B., and Koning, R. H. (2017). *Handbook of Statistical Methods and Analyses in Sports*. CRC Press.

Araújo, D. and Esteves, P. (2010). The irreducible variability of decision making in basketball. *Aportaciones teóricas y prácticas para el baloncesto del futuro.*

Arendt, E. and Dick, R. (1995). Knee injury patterns among men and women in collegiate basketball and soccer: NCAA data and review of literature. *The American Journal of Sports Medicine*, 23(6):694–701.

Basketball Data Science with applications in R, Paola Zuccolotto, Marica Manisera.

Bianchi, F. (2016). *Towards a new meaning of modern basketball players positions*. Master Degree Thesis in Computer Science and Multimedia, University of Pavia, Italy.

Heeren, Dave. *The Basketball Abstract*. 1988. Prentice Hall, Englewood Cliffs, NJ, Prentice Hall.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Handbook of Cluster Analysis*. CRC Press.

International Journal of Performance Analysis in Sport

James, Bill. The 1985 *Baseball Abstract*. Ballantine Books, 1985.

Journal of Human Kinetics volume 36/2013, 163-170 Section III – Sports Training

Journal of Quantitative Analysis in Sports, A Starting Point for Analyzing Basketball Statistics, Volume 3, Issue 3 2007 Article 1

Journal of Sports Sciences (2002)

Kay, H. K. (1966). *A statistical analysis of the profile technique for the evaluation of competitive basketball performance*. PhD thesis, University of Alberta
Key Performance Indicators: Developing, Implementing, and Using Winning KPIs, 4th edition (David Parmenter)
Lewis, Michael. Moneyball: The Art of Winning an Unfair Game. New York: W.W. Norton, 2003. Print.
Loeffelholz, B., Bednar, E., and Bauer, K. W. (2009). Predicting NBA games using neural networks. *Journal of Quantitative Analysis in Sports*, 5(1):1–15.

McGuire, Frank. *Defensive Basketball*. 1959. Prentice Hall, Englewood Cliffs, NJ.

Oliver D. *Basketball on paper. Rules and tools for performance analysis.* Washinton, D.C.: Brassey's, Inc.; 2004

Severini, T. A. (2014). *Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports*. Chapman & Hall/CRC.

Sport Business Analytics using data to increase revenue and improve operational efficiency a cura di C. Keith Harrison, Scott Bukstein

Strauss, Factor, Laing & Lyons. (2005). *What Wins Basketball Games, a Review of "Basketball on Paper: Rules and Tools for Performance Analysis" By Dean Oliver*.

Zuccolotto, P., Manisera, M., and Sandri, M. (2018). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching*, 13(4):569–589.

SITOGRAFIA

http://www.82games.com/comm30.htm

http://www.rawbw.com/~deano/methdesc.html#pyth

https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/

https://statathlon.com/four-factors-basketball-success/

https://stats.nba.com/teams/four-factors/?sort=W_PCT&dir=-1

https://www.basketball-reference.com/

https://www.espn.com/nba/story/_/id/20225286/projected-records-win-totals-standings-every-nba-team-2017-18-season

https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf