LUISS

Dipartimento

di Impresa e Management

Cattedra di STATISTICA

Analisi delle Performance Sportive e *Cluster Analysis*: un'applicazione al mondo NBA

Prof.ssa Livia De Giovanni

RELATORE

Riccardo Del Prete Mat.239071

CANDIDATO

Anno Accademico 2021/2022

1 Introduzione

L'obiettivo principale di questo lavoro è quello di sottolineare il ruolo svolto dalla *Cluster Analysis* nell'ambito delle metodologie dirette all'analisi delle *sport performance*, soprattutto con riferimento agli sport professionistici ed in particolare alla NBA (National Basketball Association).

Nel Capitolo 2, prendendo le mosse da un excursus storico relativo allo sviluppo della data science nell'ambito del basket professionistico USA, vengono proposti i *key factors* (e le relative misure) comunemente utilizzati nella valutazione delle performance di atleti e team a fini predittivi o di supporto alle decisioni.

Una diffusa trattazione delle principali metodologie per la *Cluster Analysis* è oggetto del Capitolo 3, dove ci si sofferma sui punti di forza e di debolezza di ciascuno degli algoritmi più largamente utilizzati.

Infine, nel Capitolo 4, vengono illustrati gli esiti di una verifica empirica in cui le metodologie di clustering vengono utilizzate, attraverso l'ausilio del software R, per l'analisi delle prestazioni dei cestisti che hanno militato durante la stagione 2020-21 per le 30 squadre della lega NBA.

2 *Basketball analytics*: stato dell'arte

2.1 *Data science* e sport professionistici

La scienza dei dati, in inglese *data science*, è l'insieme di principi e tecniche multidisciplinari diretti a interpretare ed estrarre conoscenza dai dati. I metodi della scienza dei dati si basano su tecniche proveniente da varie discipline, principalmente da matematica, statistica, scienza dell'informazione, informatica, intelligenza artificiale.

Il crescente interesse verso la *data science* è peraltro in buona parte attribuibile al fenomeno *big data*, vale a dire alla disponibilità, praticamente in ogni campo dell'attività umana, di enormi moli di dati eterogenei strutturati e non, che richiedono, proprio in virtù della loro estensione in termini di volumi, velocità e varietà, l'adozione di tecnologie e metodi analitici specifici per l'estrazione di conoscenza.

Sotto questo profilo, il "matrimonio" tra *data science* e sport professionistici appare un esito del tutto naturale, e ciò soprattutto in contesti, come quello statunitense, dove la produzione di dati relativi ai team, agli atleti e alle loro prestazioni *game-by-game* ha sperimentato negli ultimi 20 anni una vera e propria esplosione anche grazie alle tecnologie di *tracking* utilizzate dai broadcaster e da società specializzate.

Non desta quindi meraviglia il fatto che in un recente report della società di ricerche *Grand View Research Inc.,* si stima che il mercato globale delle applicazioni e delle tecnologie di *sport analytics* raggiungerà nel 2028 il valore "monstre" di 3,44 Miliardi di $. Tale sviluppo sarà guidato essenzialmente dalle applicazioni di *on-field analytics*, vale a dire dai sistemi di supporto decisionale basati sull'analisi delle performance dei singoli atleti e dei team.

2.2 I sistemi di supporto dalle decisioni *on-field* nel basket professionistico USA

L'avvento della cultura e dei modelli di analisi dei dati nel basket professionistico USA (*NBA*), viene fatto tradizionalmente risalire al lavoro pioneristico di *Daryl Morey* e di *Dean Oliver*.

Daryl Morey, durante gli anni in cui ricoprì la carica di General Manager degli *Houston Rockets* (2007-2020), ha legato il proprio nome allo sviluppo della *basketball analytics*

promuovendo la disciplina anche attraverso collaborazioni e partnership accademiche (si veda la *Sports Analytics Conference* tenuta annualmente dalla *Sloan School of Management* del *MIT* che vede tra i suoi organizzatori lo stesso *Morey*).

A partire dai risultati di un'estesa analisi quantitativa, *Morey* sviluppò una celebre teoria, che è alla base di una ormai universalmente riconosciuta strategia di tiro, che si può riassumere in questi 4 concetti fondamentali:

- Alzare il ritmo del gioco: aumentando il numero di possessi per partita è alquanto probabile un incremento dei punti realizzati.
- Abolizione il tiro dal "*midrange*": il tiro da media distanza è il tiro più controverso nella *NBA*, vale "solamente" due punti, ed ha la più bassa percentuale di realizzazione della lega. Molto meglio fare 1 o 2 passi indietro e prendersi un tiro da tre smarcato.
- Il miglior tiro da tre è quello dagli angoli: il cosiddetto "*Corner Three*" ed è il tiro da tre punti più ravvicinato dato che viene scagliato da 6.72 metri dal canestro anziché dai canonici 7.25 metri.
- Il tiro migliore è il tiro libero: nella *restricted area* la percentuale di realizzazione è la più alta della lega e con maggiore frequenza si può guadagnare un fallo che ti porta a guadagnarti un tiro libero. Ogni viaggio in lunetta permette di mettere punti a referto a gioco fermo.

Tale strategia che è stata alla base di alcune fortunate stagioni degli *Houston Rockets*, è diventata oggi un vero e proprio benchmark per tutte le squadre *NBA* come è testimoniato dal fatto, ad esempio, che il tiro da 3 punti, introdotto nella lega nella stagione 1980, ha fatto registrare una crescita spettacolare nel suo utilizzo medio per partita (da 2,77 tentativi per partita nel 1980 a 34,6 nel 2021). La stessa costruzione di un team di successo come i *Golden State Warrior* (vincitori di 3 titoli *NBA* nel quadriennio 2014-2018) si può ritenere largamente ispirata alla strategia resa popolare da *Morey*, visto che *Stephen Curry*, la stella del team, è il massimo interprete del principio secondo cui "*three better than two*".

Lo status e lo sviluppo, anche professionale, raggiunto negli USA dalla *basketball analytics* come disciplina si deve però fondamentalmente al lavoro di *Dean Oliver*, uno statistico che ha ricoperto diversi incarichi nel basket professionistico americano, tra i quali, quello di *Director of Quantitative Analysis* per i *Denver Nuggets*.

Grazie a *Oliver* l'analisi quantitativa nel basket è ormai elemento fondamentale per il professionismo statunitense e oggigiorno tutte le franchigie *NBA* hanno un team di analisi statistica al proprio interno.

Attraverso la pubblicazione del suo fondamentale "*Basketball on paper: rules and tools for performance*" (Oliver, 2004), oggi considerata una sorta di bibbia dell'analisi statistica nella pallacanestro USA, e la direzione della rivista *Journal of Basketball Studies*, *Oliver* ha sviluppato un framework di riferimento per la valutazione di giocatori e team che resta probabilmente il più adottato dagli addetti ai lavori.

Nell'analisi di *Oliver* è centrale in concetto di *possesso*, dal momento che un aumentato numero di possessi fa crescere con ogni probabilità il numero dei tiri verso il canestro avversario e quindi, in definitiva, le performance realizzative della squadra.

Già *Dean Smith*, il leggendario coach della pallacanestro collegiale americana (*NCAA*) che viene ritenuto il padre della Statistica nel basket, aveva evidenziato (Smith, 1999) nel suo "*Basketball -- multiple offense and defense*" che la chiave di una buona analisi è la

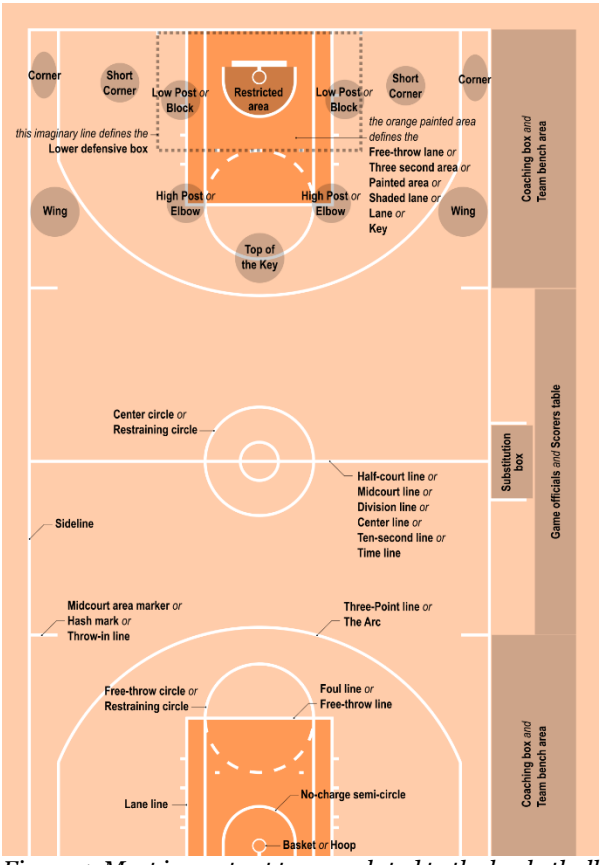

Figura 1. Most important terms related to the basketball court. Fonte:By Lencer - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=69364449

relativizzazione di ciascun aspetto del gioco alla quantità di possessi giocati dalle due contendenti.

Oliver fonda la sua analisi su una definizione operativa rigorosa di tale concetto secondo la quale il possesso inizia nel momento in cui la squadra guadagna il controllo della palla e termina nel momento in cui la squadra perde il controllo della stessa. In termini statistici il possesso può concludersi in soli 3 modi:

1) l'attacco realizza un canestro (un tiro dal campo o un tiro libero);
2) l'attacco perde il pallone (*turnover*);
3) la difesa cattura il rimbalzo difensivo dopo un canestro fallito (sia esso un tiro dal campo o un tiro libero).

Il dato relativo ai possessi così definiti non viene riportato nei tabellini (*box score*) relativi alle singole gare. Tuttavia *Oliver*, sulla base della definizione suddetta, ha formalizzato una relazione che permette di stimare il numero di possessi partendo da altre statistiche elementari usualmente prodotte (descritte nel seguito).

A Oliver si deve peraltro la definizione di alcuni indicatori relativi all'efficacia difensiva e offensiva dei team, nonché lo sviluppo dei cosiddetti 4 Fattori chiave di successo nel basket (Kubatko et al., 2007), cioè di alcune fondamentali metriche derivate dai box score per l'analisi del comportamento offensivo e difensivo di una squadra.

1) Percentuale dei tiri dal campo realizzati sui tiri tentati dal team e dalla squadra avversaria.
2) Percentuale di palle perse su possessi dal team e dalla squadra avversaria.
3) Percentuale dei rimbalzi offensivi conquistati dal team e dalla squadra avversaria.
4) Percentuale dei tiri liberi realizzati sul totale tentativi da tiro libero del team e della squadra avversaria.

Per Oliver questi 4 fattori chiave, che sarebbero alla base delle performance di ciascun team, non hanno la stessa importanza, tant'è che egli stima il peso approssimativo dei fattori nel modo che segue:

- Tiro (40%)
- Palle perse (25%)
- Rimbalzi (20%)
- Tiri Liberi (15%)

Le metriche sviluppate da Oliver si inseriscono in una ormai lunga tradizione che vede l'affermarsi anche presso il grande pubblico televisivo di indicatori analitici di performance legati al basket: dal *Total Basketball Proficiency Score* (Kay, 1966), *all'Individual Efficiency at Games* (Gómez Sánchez et al., 1980), al Magic Metric (Mays Consulting Group), all'*IBM Award* per il *Most Valuable Player* della stagione.

Sotto questo profilo, in letteratura sono stati individuate (Schumacher et al., 2010) 5 aree principali di applicazione della cosiddetta *basketball analytics*: l'analisi delle zone di tiro, la valutazione dell'efficacia dei singoli giocatori, analisi *plus/minus*, misurazione del contributo dei singoli giocatori nella vittoria del team, valutazione delle prestazioni "sotto pressione" (*clutch performance*).

L'enfasi, in questo caso è non tanto sull'efficienza di un team nel suo complesso, ma sull'efficienza personale dei singoli atleti, che viene valutata attraverso opportuni indici che rapportino quanto prodotto dall'atleta durante la gara alle opportunità avute dallo stesso.

John Hollinger, membro di spicco dell'*Association for Professional Basketball Research* (APBR), già vicepresidente dei *Memphis Grizzles*, nonché animatore della già citata conference annuale della *Sloan School* sulla *basketball analytics*, ha osservato a tal proposito che "*…l'efficienza è un concetto semplice, che però sfugge a molti. L'atleta è banalmente giudicato in relazione ai minuti che sta in campo, invece che alle reali opportunità avute per produrre gioco…".*

L'APBR ha pertanto definito un set di indicatori da utilizzare per valutare correttamente l'efficienza di un singolo giocatore all'interno di una gara o di una stagione. Di tali indicatori, quelli che seguono sono sicuramente i più utilizzati dagli addetti ai lavori:

- *EFFECTIVE FIELD GOAL PERCENTAGE* – percentuale di tiri realizzati "dal campo", corretta per il maggior valore di un tiro da tre punti rispetto ad uno da due.
- *TRUE SHOOTING PERCENTAGE* – percentuale di "punti fatti" per tiri tentati, tenuto conto del differente valore delle tre tipologie di tiro: libero, da due e da tre.
- *ASSIST RATIO* – numero di ASSIST per 100 possessi individuali giocati.
- *TURNOVER RATIO* – numero di "palle perse" per 100 possessi individuali giocati.
- *REBOUND RATE* – percentuale di "rimbalzi" presi sul totale disponibile mentre il giocatore era in campo. Può essere suddiviso in "rimbalzi difensivi" e "rimbalzi offensivi".

- *USAGE RATE* – percentuale stimata di "azioni concluse" da un giocatore sul totale di azioni giocate dalla squadra mentre l'atleta era in campo. Non necessariamente azioni concluse "positivamente".

2.3 Gli ambiti applicativi della *basketball analytics*

L'applicazione di metodologie quantitative e di diverse tecniche statistiche nell'analisi di team e giocatori, soprattutto con riferimento al mondo professionistico USA, ha prodotto ormai una sconfinata letteratura. Dall'esame di quest'ultima, i diversi obiettivi di tali applicazioni, sintetizzando, possono essere ricondotti a uno o più dei seguenti ambiti:

- Previsione dei risultati di una singola gara o di un torneo (West, 2008; Loeffelholz et al., 2009; Brown e Sokol, 2010; Gupta, 2015; Lopez e Matthews, 2015; Ruiz e Perez-Cruz, 2015; Yuan et al., 2015; Manner, 2016; Vračar et al., 2016).

2. Individuazione dei fattori discriminanti relativi ai team di successo (Trninić et al., 2002; Sampaio and Janeira, 2003; Ibáñez et al., 2003; De Rose, 2004; Csataljay et al., 2009; Ibáñezet al., 2009; Koh et al., 2011, 2012; García et al., 2013).

3. Analisi delle proprietà statistiche e dei patterns nei punteggi delle gare (Gabel e Redner, 2012; Schwarz, 2012; Cervone et al.,2016).

4. Analisi delle performance dei singoli giocatori e del loro impatto sulle probabilità di successo dei team (Page et al., 2007; Cooper et al., 2009; Sampaio et al., 2010a; Piette et al., 2010; Fearnhead and Taylor, 2011; Ozmen, 2012; Page et al., 2013; Erčulj e Štrumbelj, 2015; Deshpande e Jensen, 2016; Passos et al., 2016; Franks et al., 2016; Engelmann, 2017) anche con riferimento all'effetto "hot hand" (Gilovich et al., 1985; Vergin, 2000; Koehler and Conley, 2003; Tversky and Gilovich, 2005; Arkes, 2010; Avugos et al., 2013; Bar-Eli et al., 2006) e all'impatto nelle situazioni di gara caratterizzate da "alta pressione" (Madden et al., 1990, 1995; Goldman and Rao, 2012; Zuccolotto et al., 2018).

5. Analisi dei patterns di gioco con riferimento ai ruoli e alle posizioni in campo dei giocatori (Sampaio et al.,2006; Alagappan,2012; Bianchi et al., 2017).

6. Analisi cinetica dei movimenti del corpo dei giocatori in relazione alle performance realizzative degli stessi (Miller e Bartlett, 1996; Okubo e Hubbard, 2006; de Oliveira et al.,2006; Aglioti et al., 2008),

7. Analisi dei movimenti dei giocatori sul campo e delle traiettorie di gioco (Fujimura e Sugihara, 2005; Perše et al., 2009; Skinner, 2010; Therón e Casares, 2010).

8. Analisi delle tattiche de item e individuazione di strategie di gara ottimali (Annis, 2006; Zhang et al., 2013; Skinner and Goldman, 2017).

9. Analisi diretta all'individuazione di possibili errori sistematici nell'arbitraggio delle gare (Noecker e Roback, 2012).

10. Individuazione e misura di variabili psicologiche latenti e del loro impatto con le performance (Meyers e Schleser, 1980; Weiss e Friedrichs, 1986; Bourbousson et al., 2010a).

3 Alla ricerca della struttura nascosta dei dati: la Cluster Analysis

3.1 Concetti di base

La ricerca di una struttura "nascosta" in base alla quale i dati analizzati possano essere classificati e organizzati è al centro, da sempre, degli sforzi di ricerca della comunità scientifica in una pluralità di ambiti conoscitivi e applicativi, tant'è che tale ricerca, filosofica e scientifica al tempo stesso, si è andata configurando, nel corso della storia umana, come un'autonoma disciplina, la tassonomia. Il termine tassonomia, rimanda per l'appunto sia alla classificazione gerarchica di oggetti (animati e non) e di concetti, sia al principio stesso della classificazione.

In tale prospettiva, le tecniche e gli algoritmi di *data mining* e di analisi dei dati possono essere considerati come la risposta scientifica al problema della ricerca di uno schema tassonomico della realtà osservata, risposta basata sulla definizione di modalità e regole finalizzate all'assegnazione delle unità di osservazione a classi non definite a priori che dovrebbero in qualche modo riflettere la struttura delle entità che i dati rappresentano (Kaufman e Rousseeuw, 1990; Guyon et al., 2009).

La classificazione e il raggruppamento automatico (*unsupervised*, "senza supervisione") di casi individuali in gruppi il cui profilo emerge spontaneamente dai dati osservati è al centro della *Cluster Analysis*, una disciplina che annovera tecniche che differiscono tra loro in modo significativo sia per il concetto e la nozione di cluster che adottano sia per le modalità di ricerca dei gruppi.

La *Cluster Analysis* è quindi un insieme di tecniche di analisi multivariata il cui scopo principale è raggruppare oggetti in base alle caratteristiche che possiedono.

L'analisi multivariata si riferisce a tutte le tecniche statistiche che analizzano simultaneamente "misurazioni" multiple sulla realtà oggetto di osservazione. Pertanto, qualsiasi analisi simultanea di più due variabili potrebbe essere considerata a rigore un'analisi multivariata.

Molte tecniche multivariate sono estensioni dell'analisi univariata (analisi di distribuzioni a variabile singola) e dell'analisi bivariata (classificazione incrociata, correlazione, analisi della varianza e regressione semplice). Tuttavia, vi sono alcune tecniche di analisi il cui sviluppo è stato diretto in modo specifico ad affrontare problemi multivariati, come l'analisi fattoriale, che identifica la struttura alla base di un insieme di variabili, o l'analisi

discriminante, che differenzia i gruppi sulla base di un insieme di variabili, o per l'appunto la *Cluster Analysis*.

Le diverse tecniche raggruppate sotto la generale denominazione di *Cluster Analysis*, hanno avuto un'adozione estremamente diversificata trovando applicazione in campi quali psicologia, biologia, sociologia, economia, ingegneria, gestione aziendale, analisi delle performance sportive, ecc. . Aldilà delle differenze anche sostanziali tra le diverse tecniche, la caratteristica comune è la classificazione dei dati tratti dall'osservazione in base ai raggruppamenti "naturali" degli stessi.

La *Cluster Analysis* è paragonabile all'Analisi Fattoriale nel suo obiettivo di valutare la struttura dei dati e tuttavia differisce dalla stessa in quanto la *Cluster Analysis* raggruppa gli oggetti, mentre l'Analisi Fattoriale è principalmente interessata alle variabili di raggruppamento. Inoltre, mentre l'Analisi Fattoriale basa i raggruppamenti su metriche di variazione e covariazione (correlazione), la *Cluster Analysis* crea i raggruppamenti in base al concetto di distanza (prossimità).

In molti casi, tuttavia, il raggruppamento dei dati non costituisce il fine ultimo dell'analisi, ma un mezzo per facilitare lo sviluppo di modelli, soprattutto attraverso la possibilità di "riduzione della dimensionalità" (*data reduction*).

Un ricercatore può trovarsi infatti di fronte a un gran numero di osservazioni prive di significato se non classificate in gruppi gestibili. La *Cluster Analysis* può garantire una procedura oggettiva di *data reduction* riducendo le informazioni da un'intera popolazione o campione a informazioni su gruppi specifici. In questo modo, il ricercatore è in grado di fornire una descrizione comprensibile delle osservazioni con una minima perdita di informazioni.

Tuttavia, per quanto la *Cluster Analysis* possa essere (e venga) utilizzata come strumento di supporto nello sviluppo di concetti, modelli ed ipotesi sulla realtà, è assolutamente indispensabile per l'analista che l'utilizzo della stessa abbia come prerequisito un solido modello concettuale in relazione al fenomeno indagato.

L'approccio della *Cluster Analysis*, infatti, può essere criticato, paradossalmente perché funziona troppo bene, nel senso che in ogni caso produce risultati anche quando una base logica per i cluster non sia evidente. Pertanto, il ricercatore dovrebbe disporre di una solida base concettuale per rispondere a domande relative al motivo per cui i gruppi esistono e a

quali variabili spiegano logicamente perché le unità di osservazione ve vengono attribuite ai rispettivi raggruppamenti. Dal momento che la *Cluster Analysis* creerà sempre dei cluster, indipendentemente dall'esistenza effettiva di qualsiasi struttura nei dati, il ricercatore dovrebbe sempre ricordare che solo il fatto che i cluster possano essere trovati non ne convalida l'esistenza.

Per illustrare la natura dell'approccio che è alla base della *Cluster Analysis* e i suoi passaggi chiave, anche in termini di decisioni da assumere da parte del ricercatore, si farà ricorso a un semplice esempio tratto da *Joseph F. Hair et al.* (2020). L'esempio è relativo all'identificazione di diversi segmenti di clientela nel settore *retail*, in accordo a schemi di fedeltà (*loyalty*) osservati verso marchi (*brand loyalty*) e negozi (*store loyalty*). A tal fine, viene condotta l'analisi su un piccolo campione di sette intervistati e per ciascun intervistato vengono misurate i due tipi di fedeltà, V_1 (fedeltà al negozio) e V_2 (fedeltà al marchio), utilizzando in entrambe i casi una scala di valori compresi tra 0 e 10. V_1 e V_2 vengono assunte come variabili di clustering. I valori rilevati per ciascuno dei sette intervistati sono mostrati nella Figura 2 unitamente al diagramma a dispersione raffigurante la posizione di ogni intervistato rispetto alle due variabili.

Variabile di Clustering	*Intervistati*						
	A	**B**	**C**	**D**	**E**	**F**	**G**
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

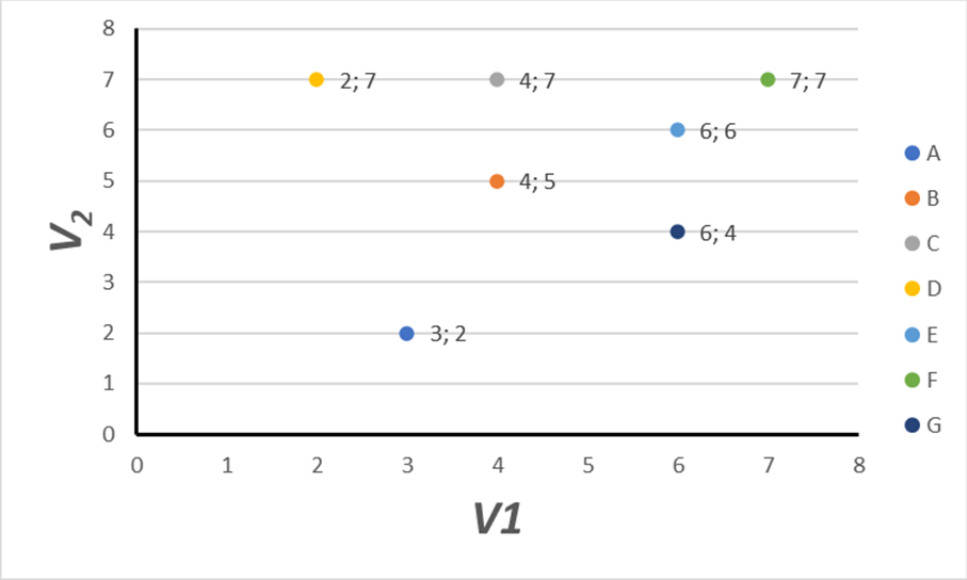

Figura 2. Distribuzione degli intervistati in base alle due variabili di clustering. Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

Con riferimento all'esempio in questione, l'obiettivo principale della *Cluster Analysis* è quello di definire la struttura dei dati inserendo le osservazioni più simili nello stesso gruppo. A tal fine, è necessario rispondere a tre fondamentali domande:

1) **Come misuriamo la somiglianza?** Abbiamo bisogno di un metodo per confrontare simultaneamente le osservazioni rispetto alle due variabili di clustering (V_1 e V_2). Sono possibili diversi metodi, inclusa la correlazione tra oggetti oppure una misura della loro vicinanza nello spazio bidimensionale assumendo che la distanza tra le osservazioni indichi somiglianza.
2) **Come si formano i cluster?** Indipendentemente da come viene misurata la somiglianza, la procedura deve raggruppare le osservazioni che sono più simili in uno stesso cluster dato l'insieme dei cluster che si sono formati. Il possibile numero dei cluster può assumere qualsiasi valore compreso tra due casi estremi (un unico cluster che raggruppi tutte le osservazioni, oppure un cluster per ciascuna osservazione).
3) **Quanti gruppi formiamo?** Occorre determinare la numerosità dei gruppi tenendo conto del trade-off tra economicità dell'analisi e grado di eterogeneità delle osservazioni presenti in ciascun gruppo. La scelta è tra un minor numero di cluster che comporta una minore omogeneità all'interno dei cluster, rispetto a un numero maggiore di cluster e una maggiore omogeneità all'interno dei singoli gruppi.

Per quanto riguarda la prima domanda (la misura della somiglianza), va innanzitutto precisato che quest'ultima rappresenta il grado di corrispondenza tra le osservazioni con riferimento all'insieme delle variabili di clustering. Spesso le misure di somiglianza sono in realtà misure di dissomiglianza in quanto sono inversamente legate al grado di somiglianza: un valore più grande della misura è indice di dissomiglianza.

Questo è il caso della maggior parte delle misure basate sulla "distanza", dove valori più grandi rappresentano maggiori distanze tra le osservazioni e quindi più dissomiglianza (cioè, meno somiglianza). Nel nostro esempio, la somiglianza deve essere determinata tra ciascuna delle sette osservazioni (intervistati A–G) e a tal fine si farà ricorso alla "*distanza euclidea*", la più comune misure di somiglianza tra due osservazioni che corrisponde alla

lunghezza del segmento tracciato tra i punti del piano corrispondenti alle rappresentazioni grafiche di ciascuna coppia di osservazioni.

Formalmente, la distanza euclidea è definita come:

$$d(\boldsymbol{x_i}, \boldsymbol{x_i'}) = \sqrt{\sum_{j=1}^{p}(x_{ij} - x'_{ij})^2} \qquad \textbf{(3.1)}$$

Dove $\boldsymbol{x_i}, \boldsymbol{x_i'}$ indicano i vettori relativi alle variabili osservate sulle unità ***i*** e ***i'***

L'insieme delle distanze così misurate è riportato nella Matrice di Prossimità di Fig.3.

		Osservazione						
Osservazione		*A*	*B*	*C*	*D*	*E*	*F*	*G*
	A							
	B	*3,162*						
	C	*5,099*	*2,000*					
	D	*5,099*	*2,828*	*2,000*				
	E	*5,000*	*2,236*	*2,236*	*4,123*			
	F	*6,403*	*3,606*	*3,000*	*5,000*	*1,414*		
	G	*3,606*	*2,236*	*3,606*	*5,000*	*2,000*	*3,162*	

Figura 3. Matrice di prossimità: distanza euclidea tra le osservazioni. Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

Per quanto attiene alla seconda domanda (procedura di formazione dei gruppi), va detto che nell'ambito della *Cluster Analysis*, vi sono varie tecniche alternative che possono essere utilizzate per la formazione dei cluster e di cui si darà conto nel seguito.

Al momento, per le necessità dell'esempio che si sta illustrando, basterà adottare un algoritmo (detto "*single linkage*" o metodo del vicino più prossimo) basato su una semplice regola in base alla quale vanno identificate le due osservazioni più simili (vicine) appartenenti a cluster diversi per combinarle poi in uno stesso cluster.

L'applicazione ripetuta di tale regola genera ad ogni iterazione una *cluster solution*, ovvero uno specifico numero di gruppi.

In avvio l'algoritmo prevede che tutti gli elementi siano considerati cluster a sé, il numero di cluster è pari al numero delle osservazioni (ogni osservazione è un cluster). Successivamente procede nelle iterazioni successive a combinare in base alla regola suddetta due cluster alla volta fino a quando la totalità delle osservazioni non risultino collocate all'interno di un unico cluster. L'algoritmo in questione pertanto fa uso di un metodo al contempo gerarchico

ed aggregativo, dal momento che procede in modo graduale alla formazione di un'intera gerarchia di *cluster solution* caratterizzata da un numero decrescente di gruppi ottenuti combinando i cluster esistenti.

Con riferimento al nostro esempio, il processo di clustering (Fig. 4) genera sei *cluster solution*, che vanno da sei cluster a un singolo cluster, attraverso i seguenti passi:

- Identificare nella matrice di prossimità le due osservazioni più vicine (E e F con distanza di 1,414) e combinarle in un gruppo passando così da sette a sei cluster.
- Trovare le successive coppie di osservazioni più vicine. In questo caso, tre coppie hanno la stessa distanza di 2,000 (E–G, C–D, e B–C). Supponiamo di scegliere la coppia E-G. Mentre G è un cluster a membro singolo, E è stato combinato nella fase precedente con F. Quindi, il cluster formato in questa fase ora ha 3 membri, (G, E, F) e la cluster solution generata dall'iterazione ha quindi 5 gruppi.
- Combinare i cluster a membro unico C e D in un unico cluster in modo da avere 4 gruppi.
- Combinare B con il gruppo formato al punto precedente. A questo punto avremo 3 cluster, cluster 1 (A), cluster 2 (B, C e D) e cluster 3 (E, F e G).
- La distanza minima successiva è 2,236 per tre coppie di osservazioni (E–B, B–G e C–E). Usando solo una di queste distanze, si perviene alla combinazione del cluster 2 e del cluster 3 del punto precedente. A questo punto avremo un cluster a sei membri e un cluster a membro singolo (A)
- Combinare l'osservazione A con il cluster a 6 membri (sei osservazioni) in un unico cluster.

	PROCESSO AGGLOMERATIVO			CLUSTER SOLUTION	
Step	Distanza Minima*	Coppia selezionata	Composizione dei cluster	Numero di cluster	Livello di Somiglianza**
	Soluzione Iniziale		(A) (B) (C) (D) (E) (F) (G)	7	0
1	1,4140	E-F	(A) (B) (C) (D) (E-F) (G)	6	1,414
2	2,0000	E-G	(A) (B) (C) (D) (E-F-G)	5	2,192
3	2,0000	C-D	(A) (B) (C-D) (E-F-G)	4	2,144
4	2,0000	B-C	(A) (B-C-D) (E-F-G)	3	2,234
5	2,2360	B-E	(A) (B-C-D-E-F-G)	2	2,896
6	3,1620	A-B	(A-B-C-D-E-F-G)	1	3,42

** Distanza euclidea tra le osservazioni*

*** Distanza media interna ai cluster*

Figura 4. Processo di clustering gerarchico e agglomerativo.

Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

Il processo di raggruppamento appena descritto può essere efficacemente illustrato anche graficamente facendo ricorso a due possibili modalità alternative di rappresentazione. Innanzitutto, dal momento che il processo è per sua essenza gerarchico, una prima rappresentazione può sintetizzare lo stesso attraverso una serie di raggruppamenti nidificati (Fig. 5).

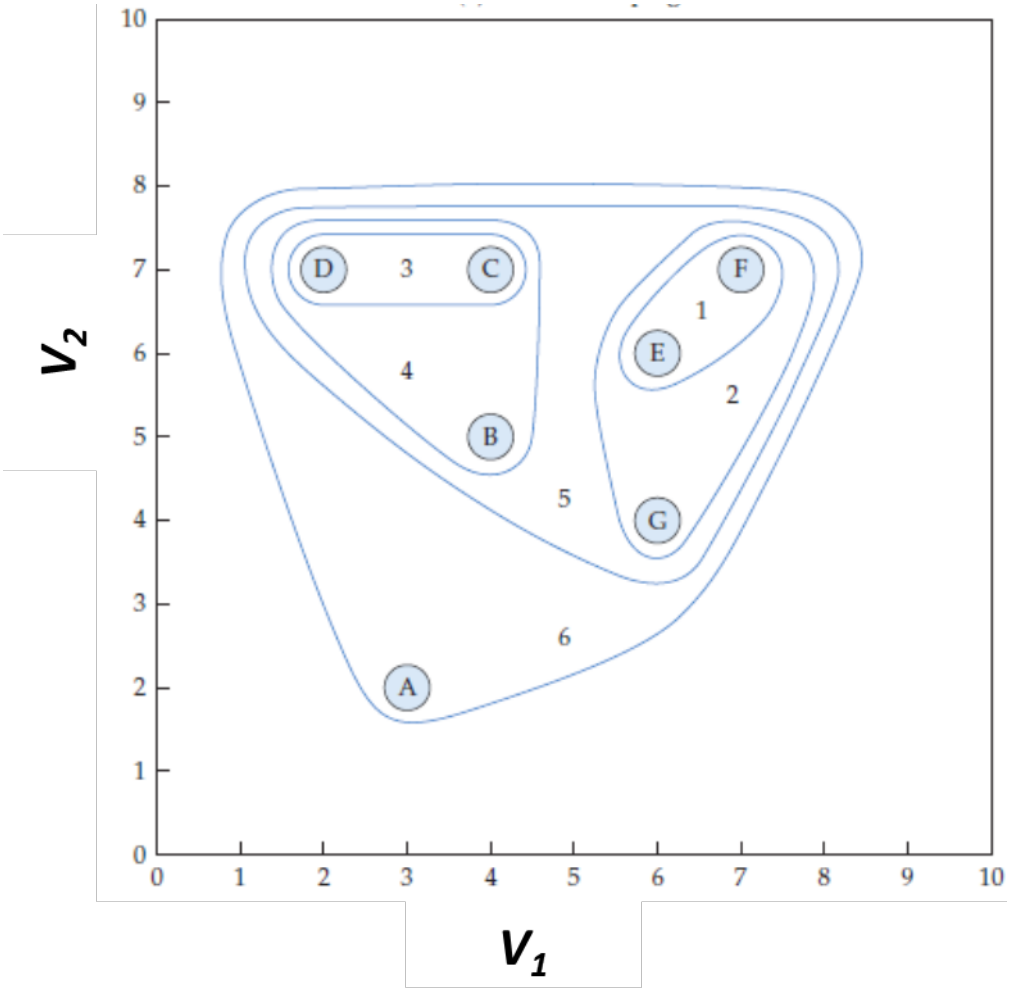

Figura 5. Diagramma a cluster nidificati. Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

Un approccio di più largo e comune utilizzo è quello della rappresentazione attraverso un *dendrogramma* (Fig. 6), un grafo ad albero posto in un sistema di assi cartesiani in cui sull'asse orizzontale vengono riportati i valori del coefficiente di agglomerazione (la distanza utilizzata per individuare i cluster da agglomerare). Questo approccio è peraltro particolarmente utile per identificare i valori anomali (*outliers*), come l'osservazione A del nostro esempio.

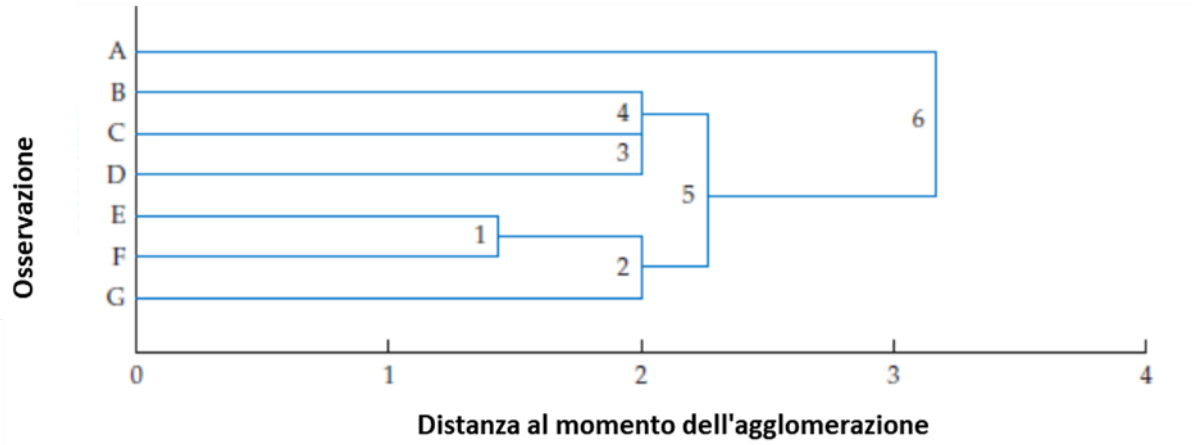

Figura 6. Dendogramma. Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

A questo punto occorre affrontare l'ultima delle tre questioni poste all'inizio del paragrafo, ovvero quella relativa alla scelta della *cluster solution* ritenuta più adeguata.

Si è evidenziato come nel passaggio da una *cluster solution* a membro unico (7 cluster) a una soluzione con un unico cluster, vi sia un inevitabile incremento del grado di eterogeneità tra le osservazioni progressivamente agglomerate.

Paradossalmente, dunque, la configurazione che garantisce la più elevata omogeneità è quella inziale dei 7 cluster. In tal modo, tuttavia, il nostro ricercatore di marketing non perverrebbe all'individuazione di segmenti significativi di clientela, che è invece l'obiettivo fondamentale della sua analisi. La soluzione dovrebbe essere ricercata facendo attenzione a bilanciare l'obiettivo della segmentazione con la necessità di controllare l'aumento del grado di eterogeneità che il processo di clusterizzazione comporta.

Tale necessità di controllo richiede la definizione di una opportuna metrica per la misurazione del grado di eterogeneità che rifletta il complessivo grado di diversità tra le osservazioni di tutti i cluster.

Con riferimento all'esempio, è possibile adottare come misura di eterogeneità la media di tutte le distanze tra le osservazioni contenute nei cluster. Nella tabella di Fig. 7 tale misura viene riportata per ciascuno step del processo di clusterizzazione. Come era lecito attendersi, dalla tabella si evince che il valore esibito dalla misura risulta crescente ad ogni successivo step di agglomerazione.

Step	Misura della somiglianza complessiva*	Variazione della somiglianza**	Crescita % dell'eterogeneità***
1	1,414	0,778	55,0%
2	2,192	-0,048	-2,2%
3	2,144	0,090	4,2%
4	2,234	0,662	29,6%
5	2,896	0,524	18,1%
6	3,420		

**Media delle distanze all'interno dei gruppi*

*** Variazione della distanza media tra uno step e il successivo*

**** Crescita dell'eterogeneità tra uno step e il successivo*

Figura 7. Variazione dell'eterogeneità. Tratto da: Joseph F. Hair et al., Multivariate Data Analysis, Cengage, 2019, p. 194-197

Infatti, mentre con l'iniziale soluzione di 7 cluster il valore della misura di somiglianza complessiva risulta nullo, dopo il primo step (6 cluster) la misura assume un valore pari a 1,414 (la distanza tra le due osservazioni raggruppate durante lo step 1). Lo Step 2 porta alla formazione di un cluster con tre membri (E, F, G) e quindi la misura di somiglianza è pari a 2,192, che corrisponde alla media delle distanze tra E e F (1,414), E e G (2,000), e F e G (3,162).

Nello step successivo, la formazione di un nuovo cluster con due osservazioni (C e D) che hanno una distanza euclidea pari a 2,000, produce ad una piccola diminuzione della misura complessiva di somiglianza (2,144). Negli ultimi 3 step si passa progressivamente da una *cluster solution* con tre gruppi (step 4), a una con due gruppi (Step 5) e infine a una con un solo gruppo (Step 6) in corrispondenza della quale la media di tutte le distanze è pari a 3,420.

La misura di eterogeneità adottata consente di comparare le diverse *cluster solution* generate dal processo. Infatti, una marcata crescita della misura nel passaggio da uno step al successivo, segnalerebbe un basso grado di somiglianza tra i due cluster aggregati durante lo step. Pertanto, per selezionare la *cluster solution* più adeguata, si procede all'esame delle variazioni relative della misura di eterogeneità tra uno step ed il successivo.

Dalla Fig. 6 si evince agevolmente che mentre nello Step 2 si registra una consistente crescita dell'eterogeneità rispetto allo step precedente, negli Step 3 e 4 tale misura non subisce sostanziali incrementi, il che indica che i nuovi cluster generati dal processo hanno un grado di eterogeneità in linea con i cluster già esistenti.

Il passaggio dallo Step 4 allo Step 5 fa invece registrare un incremento significativo (+29,6%) della misura di eterogeneità, incremento che deve essere attribuito alla combinazione dei

due cluster a tre membri dello Step 4 in un unico cluster che risulta marcatamente meno omogeneo.

Conseguentemente, l'analista considererà preferibile la *cluster solution* generata dallo Step 4 (tre cluster) alla *cluster solution* dello Step 5 (due cluster). Inoltre, la tabella evidenzia una crescita significativa della misura di eterogeneità nel passaggio da Step 5 a Step 6 dovuta all'aggregazione dell'osservazione A, il cui profilo statistico, con riferimento alle variabili di clustering adottate, risulta sicuramente anomalo rispetto alle restanti osservazioni. Per questa ragione sembra opportuno che tale osservazione, in quanto *outlier* e quindi indipendente dalle altre osservazioni, trovi collocazione all'interno di uno speciale gruppo a membro unico (*entropy group*).

3.2 Misure di somiglianza

Nella *Cluster Analysis*, come è evidente anche alla luce del semplice esempio proposto, il concetto di somiglianza riveste un ruolo centrale. La somiglianza tra le osservazioni è una misura empirica della corrispondenza, o somiglianza, tra le osservazioni oggetto di classificazione/raggruppamento.

La somiglianza può essere misurata in vari modi, ma vi sono tre metodi di misura che possono essere ritenuti dominanti nell'ambito delle applicazioni della *Cluster Analysis*: le misure basate sulla correlazione, le misure di distanza e le misure di associazione. Mentre le misure basate sulla correlazione e quelle basaste sulla distanza richiedono l'adozione di variabili di clustering quantitative, le misure di associazione sono adottate laddove le variabili siano di tipo qualitativo.

Ciò premesso, va comunque riconosciuto che con riferimento ai metodi di misura succitati, le misure di correlazione sono più raramente utilizzate nelle applicazioni della *Cluster Analysis* (a differenza di quanto accade nell'ambito di altre tecniche statistiche multivariate) delle misure basate sulla distanza.

Le misure di distanza, come si è già osservato, forniscono in realtà valori che riflettono il grado di dissomiglianza e che quindi vanno convertiti in misure di somiglianza attraverso una relazione inversa.

Una semplice illustrazione dell'utilizzo delle misure di distanza è stata mostrata nel nostro ipotetico esempio (Fig. 3), in cui i cluster sono stati definiti in base alla distanza euclidea tra ciascuna coppia di osservazioni.

La distanza euclidea, tuttavia, è solo una delle possibili misure di somiglianza basate sulla distanza che possono essere adottate nelle applicazioni di *Cluster Analysis*. Qui di seguito vengono illustrate le caratteristiche fondamentali delle misure di distanza più utilizzate:

- ***La distanza euclidea.*** È la misura di distanza più comunemente utilizzata. Equivale alla lunghezza del segmento che unisce due punti in un piano cartesiano (Fig. 8).

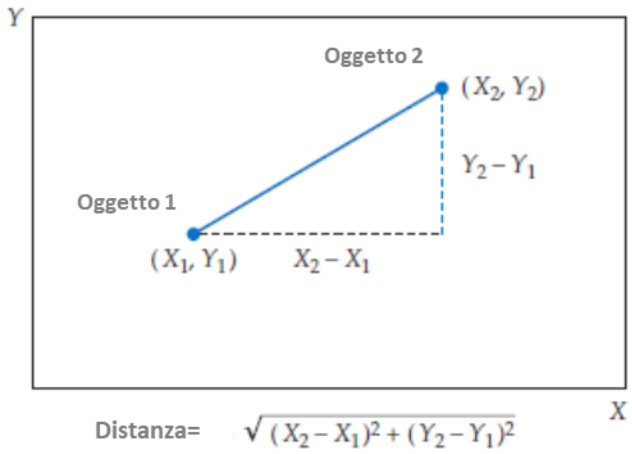

Figura 8. Distanza euclidea

La distanza euclidea tra i punti che rappresentano oggetti/osservazioni è la lunghezza dell'ipotenusa di un triangolo rettangolo, come calcolata dalla formula riportata in figura. Questo concetto è facilmente generalizzabile nel caso di più di due variabili.

- ***La distanza euclidea al quadrato*** (o ***assoluta***). È il radicando della distanza euclidea, una misura utilizzata in alcuni metodi di clustering come ad esempio quello di Ward.
- ***La distanza city-block.*** Non si basa sulla distanza euclidea, ma viene calcolata sommando le differenze tra le singole variabili (cioè sommando le misure dei due cateti del triangolo rettangolo di Fig. 8). È stato dimostrato che questa misura può portare alla formazione di cluster non validi nel caso in cui vi sia un'elevata correlazione tra le variabili di clustering, ma che può essere invece molto efficace al crescere della numerosità delle stesse variabili.

- ***Distanza di Mahalanobis (D^2)***. È una misura generalizzata di distanza che tiene conto della correlazione tra le diverse variabili attribuendo a ciascuna di esse lo stesso peso.

3.3 Procedure e algoritmi della *Cluster Analysis*

Un elemento chiave che differenzia le diverse possibili tecniche di *Cluster Analysis* risiede nelle diverse procedure di partizione che possono essere adottate, ossia nelle diverse metodologie che possono essere utilizzate per la formazione dei cluster.

In generale tutte le procedure di partizione condividono un comune principio di secondo cui gli oggetti/osservazioni andrebbero raggruppati in modo da massimizzare l'eterogeneità tra i gruppi (*between*) e al contempo minimizzare l'eterogeneità all'interno dei singoli gruppi (*within*), cioè le differenze tra i membri di ciascun gruppo (Fig. 9).

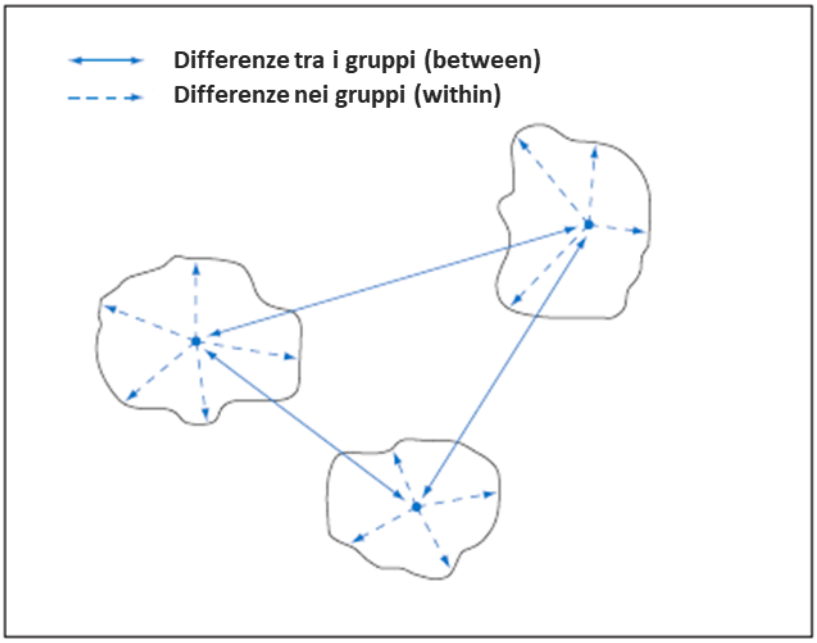

Figura 9. Variabilità "between" e "within"

Le procedure di partizione più utilizzate dai ricercatori possono essere ricondotte a due famiglie alternative: quella delle procedure gerarchiche e quella delle procedure non gerarchiche. Una procedura gerarchica è quella che è stata adottata nell'esempio illustrato nel paragrafo 3.1, dove partendo dal raggruppamento iniziale dei sette oggetti (gli intervistati) in sette cluster separati si è giunti progressivamente, agglomerando due cluster in ogni step del processo, all'unico cluster finale.

Pertanto, la caratteristica principale delle procedure di clustering gerarchico è che le stesse non portano a una singola partizione con un numero di cluster determinato in anticipo, ma invece producono una serie di partizioni ottenute in diversi passaggi. In tale processo le procedure gerarchiche richiedono necessariamente il supporto di una matrice delle distanze.

In una procedura non gerarchica, viceversa, il numero di cluster è specificato in anticipo dall'analista. L'approccio in questione è sicuramente più complesso e controintuitivo rispetto a quello gerarchico dal momento che richiede una scelta iniziale non semplice soprattutto quando non sono disponibili conoscenze preliminari sui dati. Tuttavia, il fatto che tali procedure, nonostante le difficoltà appena indicate, siano comunque largamente utilizzate, può essere dovuto a due importanti fattori:

1) Le procedure di tipo non gerarchico non richiedono il calcolo di una matrice di distanza e ciò è particolarmente rilevante nel caso di dataset di dimensioni elevate i quali il calcolo di una matrice di distanza di ordine (n × n) potrebbe essere estremamente dispendioso.
2) Nel clustering gerarchico, a differenza di quanto accade in quello non gerarchico, due unità unite in un cluster, non possono più essere divise durante i passaggi successivi del processo. In altre parole, le procedure di clustering gerarchico portano a una gerarchia di partizioni in cui la partizione in un dato passaggio dipende dalle partizioni ottenute nei passaggi precedenti

3.3.1 Procedure gerarchiche

Le procedure gerarchiche implicano una serie di ***n -1*** decisioni di clustering (dove **n** è uguale al numero di osservazioni) che combinano le osservazioni in una gerarchia che può essere rappresentata attraverso una struttura ad albero. Le procedure di raggruppamento gerarchico possono configurarsi come aggregative o divisive. Nelle procedure di tipo aggregative, ciascun oggetto/osservazione è inizialmente un cluster di per sé; ad ogni passo della procedura vengono quindi agglomerati i due cluster più vicini finché al termine della procedura si rimane con un singolo cluster. Viceversa, nelle procedure divisive tutte le osservazioni inizialmente appartengono ad un unico cluster e vengono successivamente divise (prima in due gruppi, poi tre e così via) fino ad ottenere cluster a membro singolo.

Il funzionamento di una procedura gerarchica può essere illustrato attraverso i passi di un processo di tipo aggregativo (più largamente adottato rispetto a quello divisivo):

- Ciascuna osservazione è un cluster a sé stante (cioè, ogni osservazione forma un cluster a membro singolo) in modo che il numero di cluster sia uguale al numero di osservazioni.
- Viene utilizzata una misura di somiglianza per combinare attraverso un algoritmo di clustering i due cluster più simili in un nuovo cluster.
- Tra i cluster esistenti, viene ripetuto il processo di cui al punto precedente.
- Il processo viene iterato complessivamente n-1 volte (n è il numero di osservazioni) fino a quando tutte le osservazioni risultano raggruppate in un unico cluster.

Un'importante caratteristica delle procedure di tipo gerarchico è che il risultato ottenuto dal processo in un certo stadio è sempre contenuto all'interno dei risultati ottenuti negli stadi successivi. In altre parole, dal momento che i cluster si formano esclusivamente dall'unione di cluster esistenti, per ogni membro di un determinato cluster può essere ricostruito il percorso ininterrotto che a ritroso conduce alla singola osservazione di partenza. Tale logica è alla base della rappresentazione fornita dai dendogrammi.

3.3.2 Algoritmi per il clustering gerarchico

Gli algoritmi di clustering all'interno delle procedure di tipo gerarchico definiscono le modalità attraverso le quali viene determinato il grado di somiglianza tra cluster a membri multipli.

Infatti, laddove nel primo step del processo occorre valutare la somiglianza tra singole osservazioni, successivamente il processo stesso genererà cluster che raggruppano un numero di osservazioni superiore a uno.

Gli algoritmi, per l'appunto, intervengono, nella misurazione della somiglianza tra questi cluster a membri multipli e ciascuno di essi è caratterizzato da un diverso approccio.

Tra i numerosi approcci utilizzati, sicuramente i più popolari tra i ricercatori fanno riferimento ai seguenti algoritmi:

- *Single-Linkage.*
- *Complete-Linkage.*
- *Average Linkage.*
- *Metodo del Centroide.*

- *Metodo di Ward.*

L'algoritmo single-linkage (anche detto metodo del "vicino più prossimo") definisce la somiglianza tra cluster (Fig. 10) in base alla più breve tra le distanze tra due osservazioni/oggetti appartenente a cluster diversi. Tale definizione, che è stata utilizzata

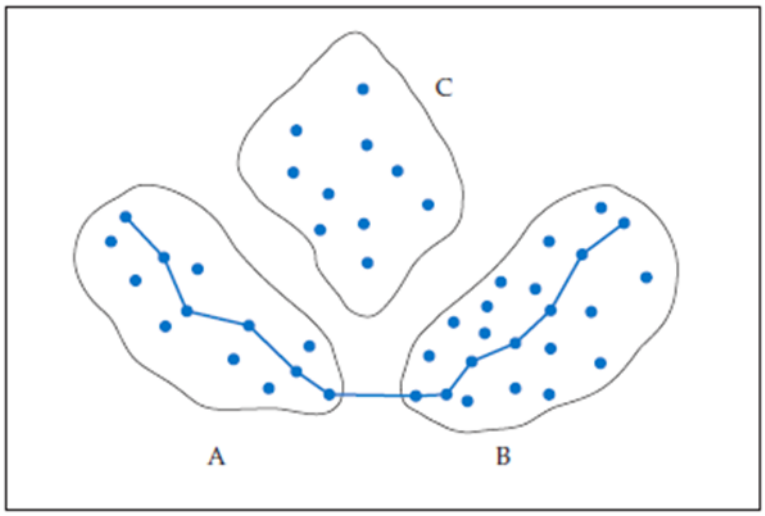

Figura 10. Esempio di applicazione del metodo single-linkage

nell'esempio guida analizzato all'inizio di questo capitolo, consente di impiegare l'originale matrice delle distanze tra le osservazioni senza la necessità di procedere al ricalcolo delle distanze. Sulla base della matrice, è sufficiente trovare le distanze tra tutte le osservazioni di due distinti cluster e selezionare come misura della somiglianza tra i cluster la più piccola di tali distanze. L'algoritmo single-linkage per facilità di utilizzo e flessibilità è probabilmente uno dei più popolari tra i ricercatori. Tuttavia presenta un fondamentale svantaggio che può essere illustrato riferendosi alla figura che segue (Fig. 11).

Nella figura vengono rappresentati 3 cluster (A, B, e C) tra i quali individuare la coppia di raggruppamenti da agglomerare. Dal momento che l'algoritmo single-linkage è esclusivamente basato, nella ricerca dei cluster più simili, sull'utilizzo della più breve delle distanze tra due osservazioni appartenenti a cluster diversi, il risultato del processo di clustering, in base a questa regola, sarà l'unione dei cluster A e B in un nuovo cluster che circonderà C.

Tuttavia, dall'esame della figura può evincersi facilmente che l'obiettivo di una maggiore omogeneità interna dei cluster (*within*) potrebbe essere perseguito più efficacemente unendo i cluster C e A oppure C e B.

In base ***all'algoritmo complete-linkage*** (anche noto come metodo del "vicino più lontano" o del "diametro") il grado di somiglianza tra cluster è basato sulla massima distanza tra le osservazioni di cluster diversi. In questo caso due cluster sono tanto più somiglianti quanto più piccolo è il diametro della sfera che ingloba la totalità delle osservazioni di entrambe. Il metodo viene denominato *complete-linkage* proprio perché tutti gli oggetti di un cluster hanno un comune legame con la misura della massima distanza.

L'algoritmo *complete-linkage* consente di evitare il principale svantaggio del metodo *single-linkage* evidenziato in precedenza.

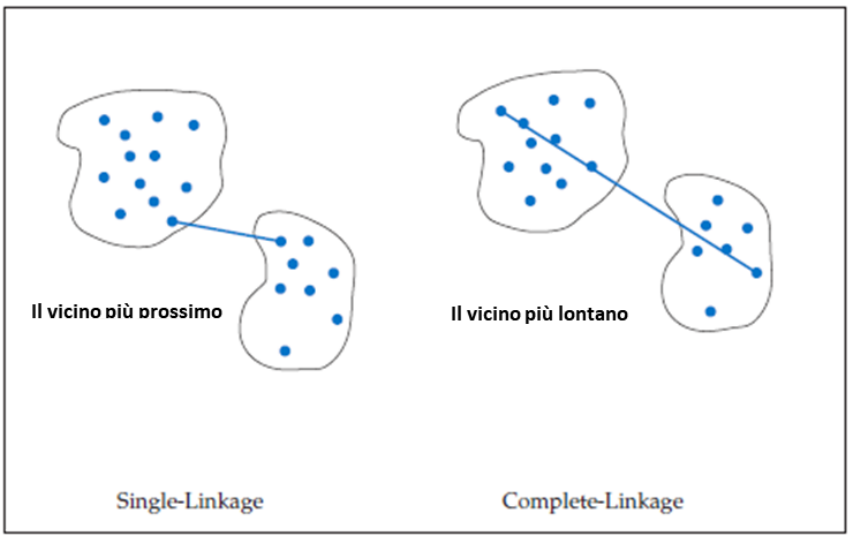

Figura 11. Single-Linkage vs. Complete-Linkage

L'algoritmo ***Average Linkage***, a differenza dei metodi precedenti, non affida la misura della somiglianza tra due cluster ai valori estremi (la coppia di osservazioni più vicine o più lontane in termini di distanza) bensì alla distanza media tra l'insieme delle osservazioni di un cluster e l'insieme delle osservazioni di un altro cluster. È un approccio che può essere considerato un compromesso tra i metodi *single* e *complete linkage* e che tende a produrre cluster che hanno una ridotta variabilità *within* approssimativamente uguale tra i diversi gruppi.

Per ***l'algoritmo del Centroide,*** il grado di somiglianza tra due cluster è determinato in base alla distanza tra i rispettivi centroidi, ossia dei punti che hanno per coordinate i valori medi, in ciascuno dei cluster, di ciascuna variabile di clustering. Questo metodo, molto popolare nelle scienze naturali (es. biologia), richiede ovviamente il calcolo di un nuovo

centroide ogni qual volta due cluster vengono aggregati. Il principale vantaggio del metodo consiste nel fatto che, come accade anche con l'*average-linkage*, il processo risulta meno influenzato dalla presenza di *outlier* rispetto agli altri algoritmi per il clustering gerarchico.

Infine, l'***algoritmo di Ward*** differisce da tutti i metodi precedenti dal momento che è basato, nella scelta dei gruppi da aggregare, sulla minimizzazione della varianza delle variabili di clustering. Il metodo di *Ward*, infatti, prevede la scomposizione della devianza totale in devianza entro i gruppi e devianza tra i gruppi. A ogni iterazione vengono considerati gli insiemi unione di tutte le possibili coppie di cluster e viene agglomerata la coppia che comporta il minore incremento della devianza nei gruppi e il maggiore incremento della devianza tra i gruppi. L'inconveniente principale del metodo è legato al fatto che l'applicazione dello stesso è limitato all'analisi delle sole variabili quantitative. Peraltro il metodo risulta molto oneroso dal punto di vista computazionale e tende a produrre gruppi di dimensioni pressoché analoghe.

3.3.3 Determinazione della partizione ottimale: il criterio dell'R^2

Si è evidenziato, anche con l'ausilio dell'esempio illustrato al punto 3.1, come gli algoritmi gerarchici producano un set di possibili partizioni (*cluster solution)* delle n osservazioni caratterizzato da un numero di cluster via via decrescenti da n a 1. Nell'esempio, la scelta della partizione ottimale è stata operata sulla base di una mera *rule of thumb*: per selezionare la *cluster solution* più adeguata, si procede all'esame delle variazioni relative della misura di eterogeneità tra uno step del processo di clusterizzazione ed il successivo e ci si arresta in corrispondenza della *cluster solution* dove viene evidenziata la crescita più significativa della misura di eterogeneità nel passaggio allo step successivo.

Tuttavia, i criteri (spesso indicati con la locuzione "*stopping rule*") ai quali si fa ricorso per determinare la partizione ottimale nelle applicazioni della *Cluster Analysis* gerarchica, sono in genere più rigorosi e affidabili della regola appena citata e sono accomunati dal comune obiettivo di individuare una partizione che concili l'esigenza di coesione interna dei gruppi (bassa variabilità *within*) con l'esigenza di garantire la massima distinzione tra i diversi gruppi (alta variabilità *between*).

I criteri adottati possono essere ricondotti alle seguenti famiglie:

a) Criteri basati su test statistici di separazione tra i cluster.
b) Criteri basati su indici sintetici.

I criteri basati sui test statistici sono diretti a valutare, attraverso l'utilizzo di distribuzioni statistiche note, la significatività statistica della misura di eterogeneità di ogni nuova *cluster solution*, ossia il grado di separazione tra i due cluster aggregati più recentemente dal processo.) Ad esempio, uno dei più utilizzati test è quello basato su una pseudo statistica T^2 (distribuzione di *Hotelling*) attraverso il quale si compara la *goodness-of-fit* di una partizione con k cluster con quella di una partizione con k-1 cluster. Qualora si osservasse un valore molto altro della statistica del test in corrispondenza della partizione con k-1 cluster, ciò segnalerebbe che la possibile soluzione sia quella con k cluster, e che quindi quella con k-1 cluster deve essere stata ottenuta aggregando due cluster molto diversi tra di loro.

I criteri basati su indici, come ad esempio l'indice R^2, l'indice *RMSSTD* (*Root Mean Square Standard Deviation*), l'indice *C* di *Calinski-Harabatz*, fanno generalmente ricorso a una misura della varianza/devianza (tra i gruppi e nei gruppi) di ciascuna possibile partizione. Sotto questo profilo, una partizione ottimale dovrebbe essere caratterizzata da una ridotta quota di devianza nei cluster e da un elevato valore della devianza tra i cluster.

In particolare, l'indice R^2 appare intuitivamente come un criterio adeguato per poter operare la scelta della partizione ottima. L'R^2 è definito come:

$$R^2 = \frac{B}{T} = 1 - \frac{W}{B} \qquad R \in [0,1] \qquad 3.2$$

dove B indica la devianza tra i gruppi (*between*), W la devianza interna ai gruppi (*within*) e T la devianza totale.

L'indice assumerà il valore massimo (1) quando i gruppi sono omogenei all'interno e ben separati all'esterno. In tal caso, B è uguale a T e il 100% della devianza totale risulta spiegato dalla devianza tra i gruppi. In questo caso, tuttavia, la partizione sarebbe del tutto priva di significato e utilità dal momento che sarebbe costituita da n gruppi, ciascuno formato da una sola osservazione.

Pertanto, andrebbe scelta come partizione ottima quella in corrispondenza della quale l'R^2 è prossimo a 1 poiché tale valore segnalerebbe l'omogeneità della classificazione, ossia che

le unità appartenenti ad un medesimo cluster sono simili tra loro (W ≅ 0) e i cluster sono ben separati (B ≅T).

3.3.4 Procedure non gerarchiche

A differenza di quanto accade con le procedure gerarchiche, il processo di clustering delle procedure non gerarchiche non è del tipo ad albero che caratterizza invece le prime. In questo caso, infatti, i diversi cluster ai quali vengono assegnate le osservazioni hanno una numerosità predeterminata dal ricercatore. Quindi, ad esempio, una *cluster solution* che preveda sei cluster, non viene originata dalla combinazione di due cluster presenti in una *cluster solution* di sette cluster, ma è prodotta unicamente dalla ricerca della migliore soluzione a sei cluster.

Le procedure e i programmi software di tipo non gerarchico generalmente prevedono due step:

- Specifica dei "*cluster seed*": per ciascun cluster va identificata un'osservazione di partenza. I *seed* possono essere predeterminati dal ricercatore oppure, come accade più frequentemente, possono essere selezionati attraverso un processo random.
- Assegnazione: dopo aver definito i *cluster seed*, si passa all'assegnazione a ciascuno di essi delle osservazioni più somiglianti tra quelle residue. Tale assegnazione può essere effettuata attraverso diversi approcci in alcuni dei quali, quando per effetto di un'assegnazione la composizione di un cluster varia, le osservazioni possono essere riassegnate a cluster diversi che risultino più somiglianti dei loro cluster originari.

Sebbene gli algoritmi di clustering non gerarchici che verranno discussi nel paragrafo 2.5 differiscono per le modalità adottate nell'assegnazione delle osservazioni ai cluster, essi risultano accomunati dal problema della selezione dei *seed*. La soluzione di questo problema, come si è anticipato, può seguire due differenti approcci.

Un primo approccio prevede che sia il ricercatore a prefissare i *seed point* sulla base di conoscenze preliminari sul fenomeno indagato derivanti da ricerche precedenti, che hanno già individuato i profili dei diversi cluster, oppure da altre tecniche di analisi multivariata (es. Analisi Fattoriale, PCA). Spesso si procede utilizzando un algoritmo di clustering gerarchico per stabilire il numero di cluster e generare *seed point* da questi risultati. In

questi casi il ricercatore non solo è a conoscenza del numero di cluster da formare, ma ha anche informazioni sul profilo di questi cluster.

Il secondo approccio consiste nel generare i *seed* attraverso l'estrazione di un campione sistematico o casuale dalle osservazioni. Un approccio basato sul campionamento sistematico prevede che la prima osservazione del data set (ammesso che non sia un *missing value*) corrisponde al primo *seed*; il secondo *seed* è la successiva informazione completa (*not missing value*) separata dal primo *seed* da una prefissata distanza minima, e così di seguito. Dopo aver selezionato i *seed*, la procedura assegna ogni osservazione ai cluster in base alla somiglianza con i *seed*. La procedura peraltro consente al ricercatore di scegliere se effettuare un aggiornamento dei *cluster seed* ogni qual volta una nuova osservazione viene aggiunta ad un cluster.

Qualunque sia il metodo adottato, il ricercatore dovrà avere piena consapevolezza di una limitazione potenzialmente grave legata al campionamento (sistematico o casuale) delle osservazioni per la selezione dei *seed*. In questo caso, infatti, la replica dell'analisi ai fini di convalida, potrebbe produrre risultati molto diversi dal momento che dipenderà da una nuova selezione dei *seed*.

3.3.5 Algoritmi per il clustering non gerarchico

In generale, la caratteristica distintiva degli algoritmi per il clustering non gerarchico risiede nelle modalità impiegate per l'assegnazione e la potenziale riassegnazione ai cluster delle osservazioni. Questo processo può seguire tre fondamentali approcci alternativi:

- *Metodo sequenziale della soglia*: prevede che sia selezionato un primo *seed* e che siano poi assegnate al relativo cluster le osservazioni che si trovano entro una certa distanza prefissata. Questo passaggio viene iterato a ogni selezione di un *seed*. Il principale svantaggio del metodo consiste nel fatto che quando un'osservazione viene assegnata ad un cluster, essa non potrà essere successivamente riassegnata ad un altro cluster anche se quest'ultimo avesse un *seed* più somigliante.
- *Metodo parallelo della soglia*: in questo caso i *seed* vengono selezionati simultaneamente e le osservazioni vengono assegnate al *seed* più prossimo in base a una determinata soglia di distanza.

- *Metodo di ottimizzazione*: questo metodo, a differenza dei due precedenti, prevede una procedura per la riassegnazione delle osservazioni tra i diversi cluster. Pertanto, se nel corso del processo di assegnazione un'osservazione diventa più vicina ad un cluster differente da quello al quale è assegnata, allora la procedura provvede alla riassegnazione.

L'algoritmo *k-means* è di gran lunga il più popolare tra gli algoritmi per il clustering non gerarchico e prevede che le osservazioni siano partizionate in un numero *k* predefinito di cluster e che siano poi iterativamente riassegnate fino a quando non viene soddisfatto un criterio numerico in base al quale risulti minima la distanza tra le osservazioni presenti all'interno di ciascun cluster e massima la distanza tra i cluster, ossia la distanza tra i rispettivi centroidi (punti medi).

L'algoritmo richiede quindi la determinazione dei centroidi sia in avvio, sia quando nuove osservazioni vengono assegnate ai cluster e perciò il suo utilizzo è limitato ai casi in cui le variabili in gioco siano di tipo quantitativo. Esistono tuttavia alcune varianti del *k-mean* che possono essere utilizzate nell'analisi di dati categoriali (ad esempio, l'algoritmo *k-medoids*).

L'utilizzo di *k-means* è così diffuso che tale denominazione è di fatto diventata sinonimo di clustering non gerarchico *tout court*.

Tale popolarità deriva soprattutto dal fatto che l'algoritmo processa le osservazioni in modo sequenziale e quindi ha la capacità di analizzare anche dataset di grandi dimensioni, laddove i metodi gerarchici richiedono in ogni step dei rispettivi processi la determinazione di una matrice delle distanze tra tutte le osservazioni.

Qui di seguito viene illustrata la sequenza dei passi del processo di partizionamento che è alla base dell'algoritmo.

- Selezione dei *k* centri iniziali dei cluster (*initial seeds*).
- Assegnazione di ciascuna osservazione al cluster più vicino, previa valutazione del distanza/dissomiglianza di quella osservazione dal centro di ciascun cluster.
- Calcolo dei centroidi di ciascuno dei cluster che sono stati creati. Come si è detto, i centroidi rappresentano i centri geometrici dei raggruppamenti e sono calcolati, per ogni cluster, mediando le *p* variabili di clustering sulle osservazioni del cluster stesso; in altre parole, le coordinate del centroide di un cluster vengono calcolate come media delle coordinate delle osservazioni presenti del gruppo.

- Ricalcolo delle distanze tra ciascuna osservazione e i centroidi e riassegnazione di ciascuna al suo cluster più prossimo.
- Proseguimento fino alla convergenza, cioè fino a quando i centroidi (e, di conseguenza, la composizione dei cluster) non risultino sostanzialmente stabili, in base a un determinato criterio prefissato (regola di arresto).

Molto spesso, il clustering *k-means* viene utilizzato adottando la distanza euclidea come misura di somiglianza dal momento che tale misura garantisce la convergenza dell'algoritmo. Una conseguenza di grande interesse derivante dall'adozione di tale misura è che, in questa situazione, l'algoritmo identifica la partizione dei dati in k gruppi che riducono al minimo la devianza all'interno dei cluster (WD). Il problema della ricerca della partizione che minimizza la WD viene risolto iterativamente dall'algoritmo dal momento che non può essere risolto "per enumerazione", ossia individuando prima tutte le possibili partizioni e procedendo poi alla selezione di quella che esibisce il valore minimo della WD. Lo sforzo computazionale sarebbe rilevante anche per i computer più performanti, tenendo conto ad esempio che 100 osservazioni possono essere raggruppate in k=5 gruppi in $6{,}57\times10^{67}$ modi diversi.

Una questione molto importante da sottolineare è che il clustering *k-means* può essere molto sensibile alla scelta dei *seed* iniziali. Per tale motivo, spesso viene condotta un'analisi preliminare di tipo gerarchico per determinare il numero k di cluster per il clustering *k-means* e per calcolare i centri dei cluster che possono poi essere utilizzati come *seed* iniziali nella procedura non gerarchica.

4 Cluster Analysis come strumento di supporto per la Sport Performance Analysis: un'applicazione al basket professionistico USA

4.1 Introduzione

Nel primo capitolo è stata fornita una breve ricostruzione, anche storica, dello status e della rilevanza che ha assunto la *data science* nell'ambito dell'analisi delle performance sportive soprattutto con riferimento al basket professionistico USA (NBA). Si è passati poi, nel secondo capitolo, a illustrare i fondamenti della *Cluster Analysis*, una tra le metodologie di analisi multivariata più promettente per i diversi ambiti applicativi della *sport analytics*. In questo capitolo dedicato alla verifica sperimentale, verranno illustrati i risultati dell'applicazione di tale metodologia alle prestazioni dei cestisti che hanno militato durante la stagione 2020-21 per le 30 squadre della lega, prestazioni che sono misurate attraverso i "*Four Factors of Basketball Success*" ai quali si è fatto diffusamente riferimento nel primo capitolo. L'analisi sarà diretta all'individuazione di raggruppamenti omogenei di giocatori rispetto all'insieme delle variabili di clusterizzazione (*Four Factor, Fig12*) attraverso l'utilizzo di un algoritmo di partizione gerarchico (il metodo di partizione che è stato adottato è quello di Ward, che è probabilmente il più utilizzato in questo tipo di applicazioni) per determinare il numero dei cluster e successivamente attraverso l'applicazione dal metodo *k-means*. L'analisi è stata svolta utilizzando il software R per l'analisi statistica dei dati.

Attribute type	Variables	mean [min; max]	Use
Numeric	**Four Factors** (per game)		community detection
	effective Field Goal Percentage (eFG)	52.3% [0.0; 76.3%]	
	Free Throw Attempt Rate (FTA)	18.2% [0.0; 57.9%]	
	Turnover Ratio (To Ratio)	10.1% [0.0; 71.4%]	
	Offensive Rebound Percentage (OREB)	4.2% [0.0; 16.8%]	
Numeric	**Other variables**		profiling
	Age	26.2 [19; 38]	
	Game played GP	44.2 [1; 69]	
	Wins W	22.1 [0; 50]	
	Losses L	22.1 [0; 51]	
	Minutes played MIN	21.4 [6.3; 37.2]	
	Offensive rating OFFRtg	108.3 [57.1; 125.0]	
	Defensive rating DEFRtg	109.8 [77.8; 161.5]	
	Net Rating NETRtg	-1.5 [-104.4; 29.8]	
	Defensive Rebound percentage (DREB%)	13.7 [0.0; 38.5]	
	Rebound percentage (REB%)	8.9 [2.3; 23.4]	
	Assist percentage AST%	13.8% [0.0; 47.6]	
	Assist to Turnover Ratio (AST/TO)	1.9% [0.0; 13.0]	
	True Shooting percentage TS%	55.4% [0.0; 76.3]	
	Usage percentage USG%	18.1% [4.8; 37.5]	
	PACE	101.4 [96.4; 112.7]	
	Player Impact Estimate PIE	9.0 [-26.0; 20.8]	
	Center C	77 players (16.4%)	
	Forward F	19 players (4.0%)	
	Guard G	21 players (4.5%)	
	Power Forward PF	78 players (16.6%)	
	Point Guard PG	86 players (18.3%)	
	Small Forward SF	86 players (18.3%)	
	Shooting Guard SG	103 players (21.9%)	

Figura 12. NBA: Four Factors e altre variabili utilizzate nelle analisi delle performance

4.2 Le variabili di clusterizzazione: i *Four Factors*.

I "*Four Factors of Basketball Success*" resi popolari e analizzati compiutamente da Dean Oliver, come si è visto mel capitolo 1, sono i seguenti:

1) *effective field goal percentage* (eFG%), misura che identifica la percentuale di realizzazione relativa ai tentativi effettuati dal campo da un giocatore (o da una squadra). La misura, che contiene una "correzione" che permette di considerare il maggior peso dei tentativi da 3 punti, si ottiene dall'applicazione della formula:

$$\boldsymbol{eFG\%} = \mathbf{100x}\frac{\boldsymbol{FGM} + \mathbf{0,5}\boldsymbol{x3PM}}{\boldsymbol{FGA}} \qquad \mathbf{4.1}$$

dove FGM è il numero di tiri dal campo realizzati, 3PM il numero dei tiri da 3 punti realizzati e FGA il numero dei tentativi.

2) *turnovers per possession* (TOt/POSSt), misura utilizzata per calcolare la percentuale di possessi che si conclude con una palla persa. L'obiettivo 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 di non perdere la palla. La misura viene calcolata attraverso la formula:

$$\boldsymbol{To\ Ratio\%} = \mathbf{100x}\frac{\boldsymbol{TO}}{\boldsymbol{FGA} + (\boldsymbol{FTAx}\mathbf{0,44}) + \boldsymbol{Assists} + \boldsymbol{TO}} \qquad \mathbf{4.2}$$

dove TO rappresenta il numero di palle perse da un giocatore, FGA il numero dei tentativi dal campo, FTA il numero di tiri dalla lunetta e Assists il numero di passaggi.

3) *Offensive Rebounding percentage* (*OREB%*), misura la percentuale di rimbalzi in attacco che un giocatore (o una squadra) ha conquistato. Viene calcolata come il numero di rimbalzi disponibili dopo un tentativo dal campo fallito. Se una squadra non è in grado di finalizzare con una realizzazione ogni possesso, il compito ottimale è quello di recuperare ogni tiro sbagliato e dare alla squadra una seconda opportunità. Un rimbalzo offensivo estende perciò il possesso e può consentire un nuovo tentativo di realizzazione. La misura viene calcolata attraverso la formula:

$$\boldsymbol{OREB\%} = \mathbf{100x}\frac{\boldsymbol{OREB}}{\boldsymbol{OREB} + \boldsymbol{DREB}} \qquad \mathbf{4.3}$$

dove OREB rappresenta il numero di rimbalzi offensivi conquistati da un giocatore mentre DREB il numero di rimbalzi difensivi.

4) *Free throw rate* (*FTMt/FGAt*), misura che rappresenta come la capacità di un giocatore di ottenere un fallo quando fa un tiro. Il *Throw Rate*, apparentemente, sembrerebbe una variabile non molto significativa; in realtà tale variabile incorpora un fattore molto importante dal momento che è in grado di esprimere la capacità/attitudine di un giocatore di conquistare un fallo ogni volta che effettua un tentativo di realizzazione. La formula utilizzata è:

$$\boldsymbol{FTA\% = 100x\frac{FTM}{FGA}} \qquad \mathbf{4.4}$$

A ciascuno di questi fattori, le cui rispettive distribuzioni di frequenza per l'insieme dei giocatori NBA considerati nell'analisi sono riportate nella figura che segue, Oliver ha attribuito uno specifico peso che riflette l'importanza relativa di ognuno di essi nel determinare il risultato di un incontro. I pesi sono i seguenti:

1) eFG=40%
2) To Ratio=25%
3) OREB=20%
4) FTA=15%

È ora possible presentare la relazione prodotta da Oliver che esprime il numero di possessi in relazione ai Four Factors.

$$POSS = (FGM + \lambda FTM) + \alpha[(FGA - FGM) + \lambda(FTA - FTM) - OREB] + (1 - \alpha)DREB - TO$$

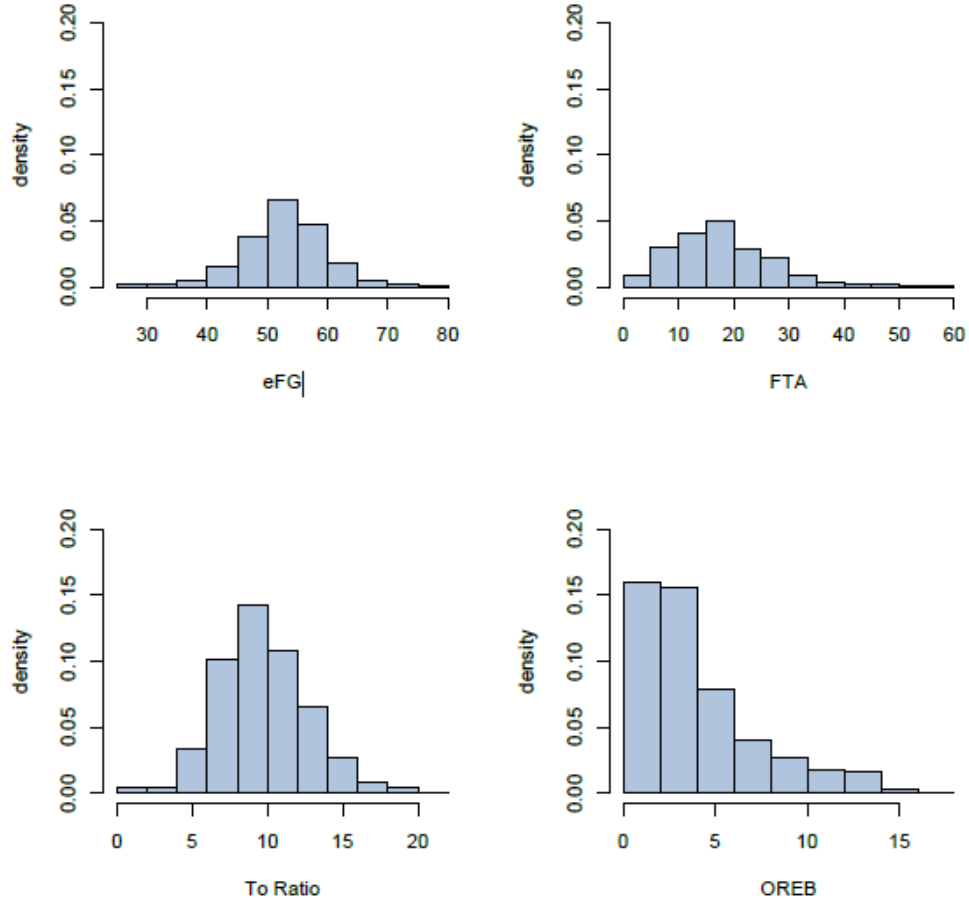

Figura 13. Distribuzioni di frequenza dei Four Factors

4.3 Risultati ottenuti dall'applicazione della metodologia di clustering

Come si è già anticipato, la metodologia adottata prevede un'analisi preliminare di tipo gerarchico (condotta attraverso l'algoritmo di *Ward*) per determinare il numero *k* di cluster per il *clustering k-means*. Dopo aver individuato il numero di cluster, l'algoritmo *k-means* partizionerà le osservazioni nel numero *k* predefinito di cluster, osservazioni che vengono iterativamente riassegnate fino a quando non viene soddisfatto un criterio numerico in base al quale risulti minima la distanza tra le osservazioni presenti all'interno di ciascun cluster e massima la distanza tra i cluster, ossia la distanza tra i rispettivi centroidi (punti medi).

I risultati ottenuti dall'applicazione della metodologia vengono di seguito rappresentati graficamente attraverso l'output prodotto dal software *R* che ha generato una possibile partizione di 480 giocatori appartenenti alle 30 squadre della lega, partendo dalla matrice delle distanze euclidee, calcolata sulla base della matrice dei dati originale. I 480 giocatori sono stati selezionati dai complessivi 536 giocatori della NBA, imponendo un filtro sulla media dei minuti giocati a partita.

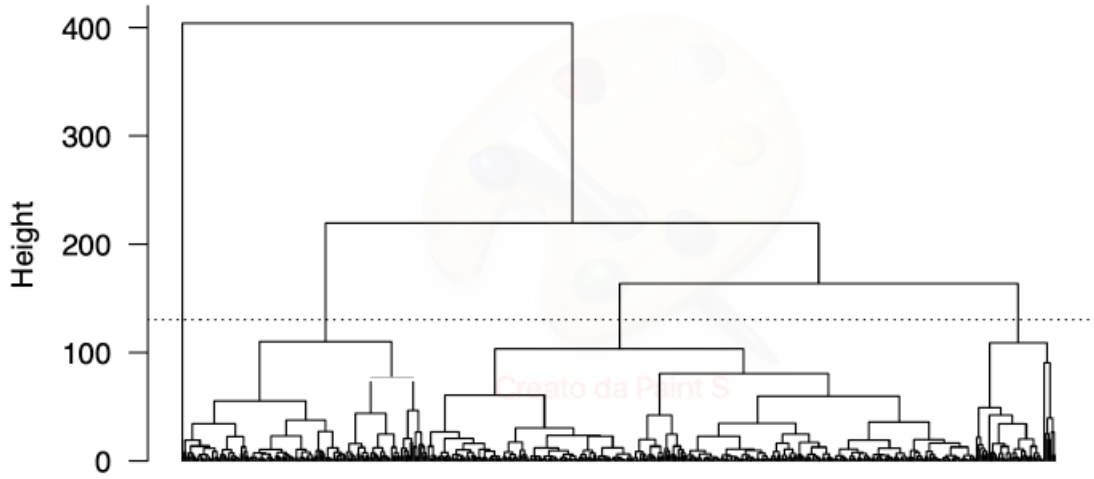

Figura 14. Dendogramma ottenuto con il medodo di Ward

Nella Fig. 14vengono riportate le *cluster solution* ottenute applicando il metodo di *Ward*. La linea di "taglio" indica che il numero di cluster ottimale in base al metodo è pari a 3.

L'applicazione dell'algoritmo *k-means* ha poi prodotto l'output che segue (Fig. 15) da cui si può evincere la numerosità di ciascun cluster, le coordinate dei centroidi degli stessi cluster espresse con riferimento alle variabili di clusterizzazione (*Four Factors*), i valori medi di ciascun fattore con riferimento alla popolazione totale delle osservazioni, la numerosità dei cluster, il valore del coefficiente R^2 che misura la devianza tra i cluster ottenuti.

	Four Factors					
	Numerosità	***OREB%***	***ToRatio%***	***EFG%***	***FTA Rate%***	***R^2***
Cluster1	304	4,35	9,49	55,83	17,45	
Cluster2	139	3,23	10,95	43,89	14,08	0,6246053
Cluster3	37	6,19	10,87	56,47	47,66	
Media		4,17	10,02	52,42	18,8	

Figura 15. Risultati ottenuti dall'applicazione di k-means

Il grado di presenza e la distribuzione di ciascuno dei *Four Factors* per ogni cluster risultante dall'analisi *k-means*, vengono rappresentati attraverso i box plot riportati nelle figure che seguono.

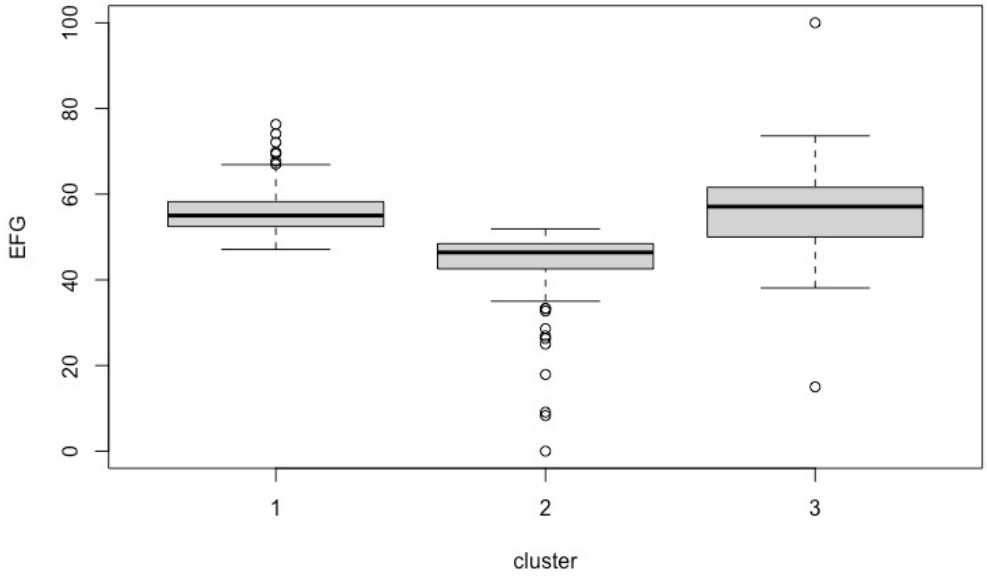

Figura 16. Distribuzione del fattore EFG%

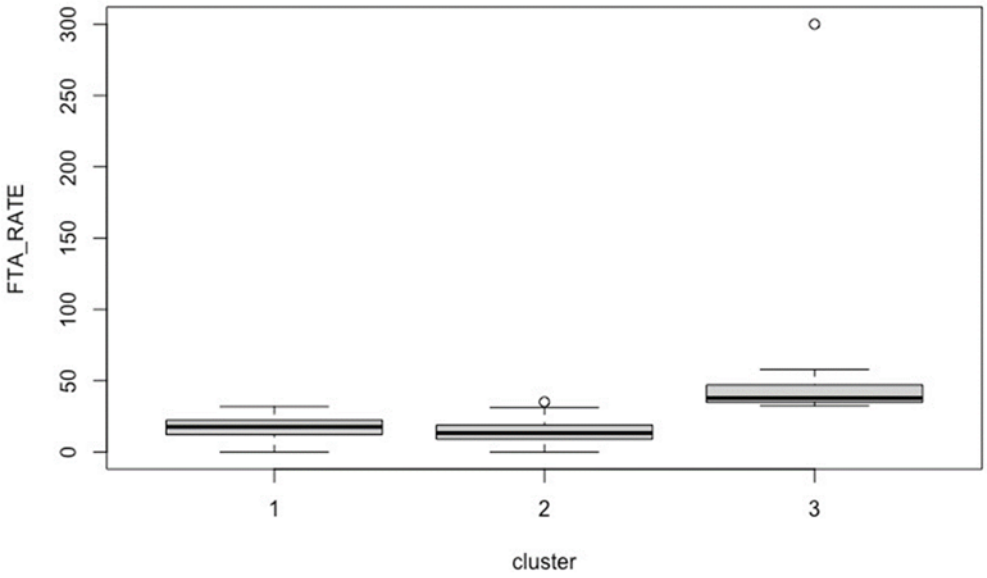

Figura 17.Distribuzione del fattore FTA%

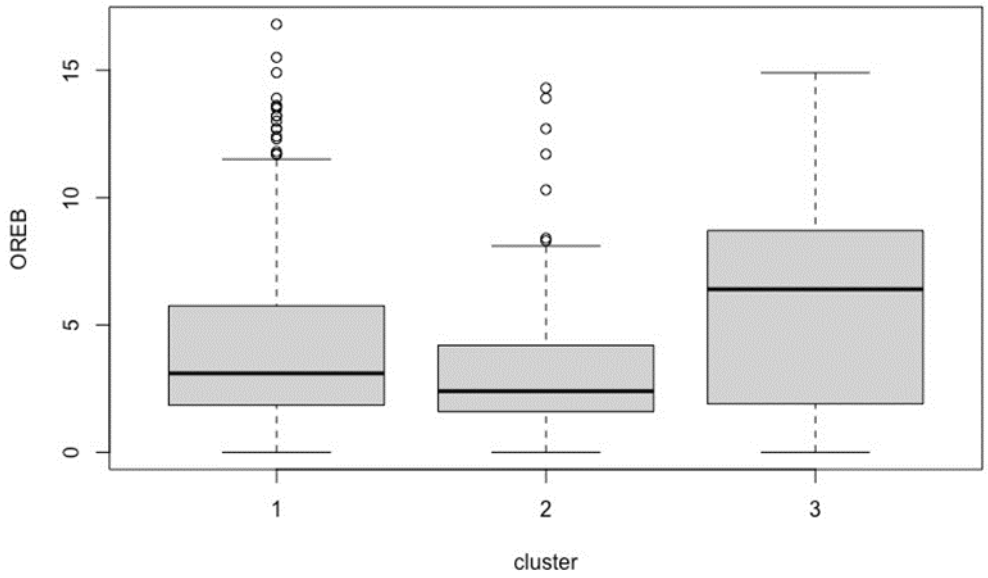

Figura 18. Distribuzione del OREB%

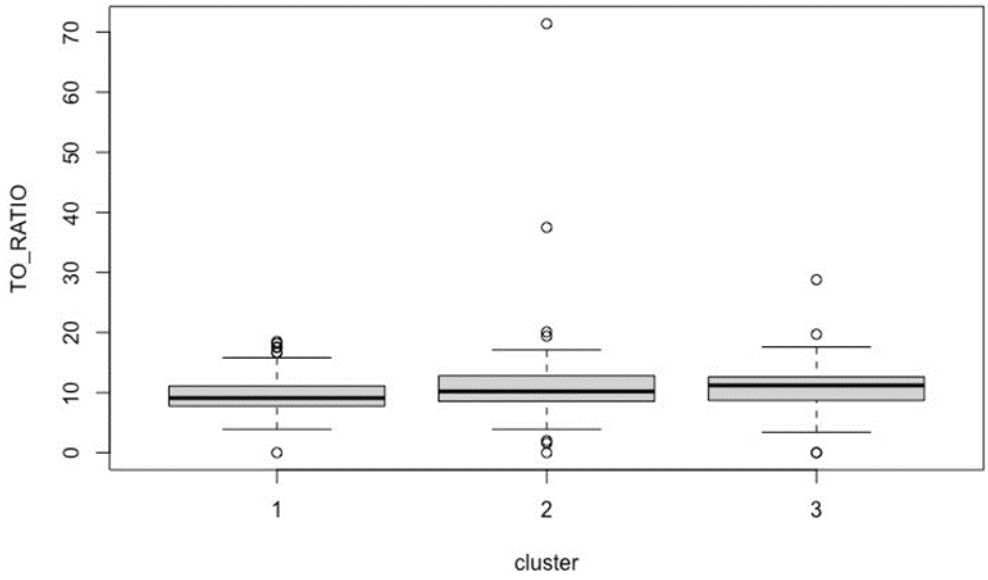

Figura 19. Distribuzione del TO-Ratio

L'analisi ha consentito di individuare tre distinti cluster di atleti ciascuno dei quali corrisponde a uno specifico *performance profile* caratterizzato dai valori medi assunti, all'interno del cluster, da ciascuno dei *Four Factors* (Figura 15).

In particolare:

1. Il Cluster 1, composto da 304 atleti, è quello dei *Big Shooters*, dal momento che risulta caratterizzato da valori medi del fattore FGA e del fattore OREB più elevati di quelli della lega. Si tratta tipicamente di giocatori "alti" che mostrano eccellenti performance nel tiro dal campo e nei rimbalzi offensivi, e performance più modeste nella riconquista della palla e nel tiro libero.
2. Il Cluster 2, composto da 139 atleti, è il cluster delle *Difensive Star*, dal momento che il valore medio del *Turn Over Ratio* nel gruppo (10,95%) è superiore sia a quello medio della lega (10,02%) che a quello dei cluster 1 e 3 (rispettivamente 9,49% e 10,87%). È un cluster fortemente specializzato sulla difesa, dal momento che i valori medi di tutte le altre variabili di clusterizzazione sono inferiori ai valori medi della lega.
3. Il Cluster 3, composto da 37 atleti, potrebbe essere etichettato come *Super Stars*, dal momento che risulta caratterizzato da un valore medio di 3 variabili (OREB, FGA e FTA) di clusterizzazione su 4 più elevato di quello rilevato negli altri cluster. Per quanto riguarda il TO Ratio, il valore medio è di poco inferiore a quello del Cluster 2 (leader per questo fattore), ma comunque nella media della lega.

Ovviamente, l'analisi svolta avrebbe consentito di ottenere risultati più solidi e affidabili sui *performance profile*, se avesse avuto come input le osservazioni relative a un arco temporale più significativo di una singola stagione. Pur tuttavia, i risultati ottenuti mostrano, pur scontando il limite suddetto, il valore, l'utilità e il ruolo che la *Cluster Analysis* può avere come supporto decisionale nella gestione dei team.

4.4 Conclusioni

L'analisi sviluppata ha mostrato l'efficace utilizzo dell'analisi dei dati nella valutazione delle performance dei giocatori di basket la possibilità di classificare i giocatori in gruppi in base al loro rendimento rappresenta il primo passo per il management di ogni squadra nel valutare come plasmare il *Roster* (cioè la lista di giocatori che fanno parte della squadra) sia a livello tecnico che finanziario. Tale possibilità appare particolarmente rilevante in sede di

selezione e scambio di giocatori (*Trades*), specialmente in occasione del cosiddetto *NBA Draft*, l'evento annuale in cui le trenta squadre della lega possono scegliere nuovi giocatori, di solito le migliori giovani promesse provenienti generalmente dal campionato dei *college* (*National Collegiate Athletic Association*, *NCAA*).

BIBLIOGRAFIA ESSENZIALE

P. Giordani et al., *An introduction to clustering with R*, Springer, 2020.

Joseph F. Hair et al., *Multivariate Data Analysis*, Cengage, 2019.

C. Keith Harrison, *Sport business analytics: using data to increase revenue and improve operational efficiency,* CRC Press, 2017.

L. Kaufman, e P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, New York, 1990.

J. Kubatko et al., *A starting point for analyzing basketball statistics*, Journal of Quantitative Analysis in Sports, 3(3):1‑22, 2007.

D. Oliver, *Basketball on Paper: Rules and Tools for Performance Analysis*, Dulles: Potomac Books Inc., 2014.

R.P. Schumaker et al., *Sports Data Mining*, Springer, 2010.

Paola Zuccolotto e Marica Manisera , *Basketball Data Science: With Applications in R*, CRC Press, 2020.