Marco Marinoni n°matricola 812405

CLUSTER ANALYSIS

INDICE **pag.**

IN COSA CONSISTE LA CLUSTER ANALYSIS?

La cluster analysis (in italiano “analisi dei gruppi”) è un metodo statistico per la classificazione delle unità in gruppi omogenei. Si tratta di una procedura tipicamente esplorativa: essa consiste nella ricerca nelle *n* osservazioni attribuite a diverse variabili, di gruppi di unità tra loro simili, ipotizzando che tali gruppi omogenei esistano effettivamente nel data set. Ripartizioni di questo tipo assumono una certa rilevanza solo nei casi in cui nei dati siano evidentemente presenti delle strutture di gruppo, le quali vengono immediatamente individuate dalla metodologia statistica. Obiettivo della cluster analysis è quindi quello di suddividere l’insieme di dati in gruppi “naturali”, che siano coesi internamente (le unità assegnate ad un medesimo gruppo devono essere tra loro simili) e separati esternamente (i gruppi devono essere il più possibile distinti), consentendo sia una più moderata descrizione che una più semplice interpretazione dei risultati ottenuti.

Prima di procedere alla classificazione delle osservazioni di una data matrice di n elementi e p variabili, è necessario e fondamentale definire una misura della diversità delle osservazioni stesse. A seconda del tipo di dati esistono misure diverse: per dati quantitativi si utilizzano misure di distanza, mentre per dati qualitativi si applicano misure di associazione.

Successivamente si procede all’identificazione del metodo o dell’algoritmo più idoneo per la ripartizione; i metodi di classificazione più comuni sono:

- Metodi gerarchici
- Metodi non gerarchici

I metodi gerarchici si distinguono in procedure aggregative e divisorie. Per quanto riguarda il primo tipo, gli elementi di una matrice vengono fusi in gruppi via via più ampi, fino ad arrivare a raggruppare tutti gli elementi in unico gruppo. Mentre per il secondo, vengono definite partizioni a mano a mano sempre più fini dell’insieme iniziale, fino a giungere a classi composte da un solo elemento. Caratteristica che contraddistingue i metodi gerarchici è l’irrevocabilità dell’assegnazione di un oggetto ad un cluster, cioè una volta che un oggetto è stato attribuito ad uno specifico gruppo, non può essere rimosso.

I metodi non gerarchici sono solo di tipo aggregativo, e generano un’unica partizione. Realizzano attribuzioni successive delle unità tra i gruppi definiti a priori, fino alla partizione ritenuta “ottima” sulla base di un criterio predefinito.

DESCRIZIONE DEL DATA SET E DELLE TECNICHE DI ANALISI

Il data set che ho scelto di analizzare è stato estratto dal sito web Sports Reference, che è un sito internet che offre statistiche sportive relative a diversi sport, quali: basket, football americano, hockey, baseball e calcio. Siccome sono un appassionato di pallacanestro, più precisamente del campionato statunitense di basket, ho deciso di focalizzare la mia attenzione sulle medie di tutti i giocatori della stagione attuale, relativamente ad alcuni aspetti del gioco. L'obiettivo è quello di classificare i giocatori in gruppi in base al loro rendimento, il quale 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. Le variabili scelte per confrontare i giocatori sono:

-Nome giocatore: variabile testuale

- Punti per partita: variabile numerica, nel dataset espressa con il codice "PTS".

- Assist per partita: variabile numerica, nel dataset espressa con il codice "AST".

- Rimbalzi totali per partita: variabile numerica, tiene conto sia dei rimbalzi offensivi, sia quelli difensivi. Nel dataset espressa con il codice "TRB".

- Palle rubate per partita: variabile numerica, nel dataset espressa con il codice "STL".

- Stoppate per partita: variabile numerica, nel dataset espressa con il codice "BLK".

- Minuti giocati partita: variabile numerica, nel dataset espressa con il codice "MP".

I valori registrati da ciascun giocatore, per ogni variabile, sono espressi sottoforma di media aritmetica, determinati rapportando la somma delle cifre, per ciascun campo di osservazione, che i giocatori hanno registrato fino al giorno 28 febbraio 2020, diviso per il numero di partite che ogni giocatore ha disputato fino a quella data.

Dato che il dataset presenta dimensioni abbastanza grandi, ho deciso di riassumere in queste due tabelle alcune delle informazioni essenziali della popolazione nella sua interezza.

	Statistiche Descrittive							
	N	Minimo	Massimo	Media	Errore Std.	Deviazione Std.	Varianza	Asimmetria
PTS	511	.0	35.2	8.611	.2904	6.564	43.091	1.106
TRB	511	.0	15.3	3.609	.1133	2.562	6.563	1.377
AST	511	.0	10.6	1.887	.0798	1.804	3.254	1.751
STL	511	.0	2.1	.607	.0186	.4206	.177	.668
BLK	511	.0	3.1	.399	.0187	.4224	.178	2.291
MP	511	.5	36.9	19.431	.4261	9.633	92.785	-0,067

Il giocatore medio in NBA ha una media punti a partita pari a 8,611, raccoglie 3,609 rimbalzi totali a partita, effettua 1,887 assist a partita, recupera 0,607 palloni a partita, stoppa 0,399 tiri a partita e gioca 19,431 minuti a partita. Le medie evidenziano valori abbastanza bassi, per il semplice motivo che il numero di giocatori osservati è abbastanza elevato. Inoltre, anche i valori delle varianze sono abbastanza alti, soprattutto per le variabili punti e minuti, e il motivo è sempre legato all'ampio dataset.

	PERCENTILI						
	5	10	25	50	75	90	95
MEDIA	1.000	1.500	3.800	7.300	11.800	18.700	21.380
	.500	.900	1.800	3.200	4.800	6.600	8.780
	.100	.300	.700	1.400	2.300	4.380	6.200
	.000	.100	.300	.600	.900	1.200	1.340
	.000	.000	.100	.300	.500	.900	1.240
	3.700	5.320	11.500	19.100	28.100	32.500	34.400

In questa seconda tabella vediamo che, solo il 5% del totale dei giocatori in NBA non registra né una palla rubata né una stoppata e il 10% della popolazione non effettua neanche una stoppata. Il 50 % dei giocatori presenta valori, in termini di rimbalzi totali, palle rubate e stoppate, molto simili a quelli espressi dalla media sull'intera popolazione. Infine, possiamo notare che il 50 % dei giocatori non registra la doppia cifra di punti, cioè non ha una media punti per partita pari o superiore a 10.

Prima di procedere alla cluster analysis vera e propria, ho voluto verificare che le variabili scelte per il dataset non fossero eccessivamente correlate fra di loro. Infatti, l'alta correlazione è un problema, siccome determina sovrabbondanze nei dati che vengono contate nel processo di classificazione, provocando alterazioni nei risultati.

Matrice di correlazione							
		PTS	TRB	AST	STL	BLK	MP
Correlazione	PTS	1,000	0,649	0,742	0,635	0,351	0,876
	TRB	0,649	1,000	0,382	0,478	0,691	0,694
	AST	0,742	0,382	1,000	0,646	0,081	0,696
	STL	0,635	0,478	0,646	1,000	0,275	0,740
	BLK	0,351	0,691	0,081	0,275	1,000	0,393
	MP	0,876	0,694	0,696	0,740	0,393	1,000

Dalla tabella qui sopra, possiamo vedere che non è presente una significativa correlazione fra le variabili, perché nessun valore si avvicina troppo a 1 o -1. Questo ci consente di procedere con l'analisi senza dover ridurre il numero di variabili da considerare, evitando predisporre un'analisi delle componenti principali.

CLUSTER ANALYSIS MEDIANTE SPSS

L'analisi è stata affrontata attraverso l'impiego del software SPSS. Tra le diverse modalità di classificazione che il software offre, ho deciso di puntare per il mio studio, su un metodo gerarchico, più precisamente il Metodo di Ward, il quale è un criterio molto utilizzato perché permette la minimizzazione della varianza all'interno dei gruppi e poi, per la sua applicazione è possibile utilizzare qualsiasi distanza.

Quindi, nella scelta della distanza da dover utilizzare, ho proceduto con la distanza euclidea quadratica, che è quella consigliata dal software.

La tabella sottostante riporta l'agglomerazione dei dati.

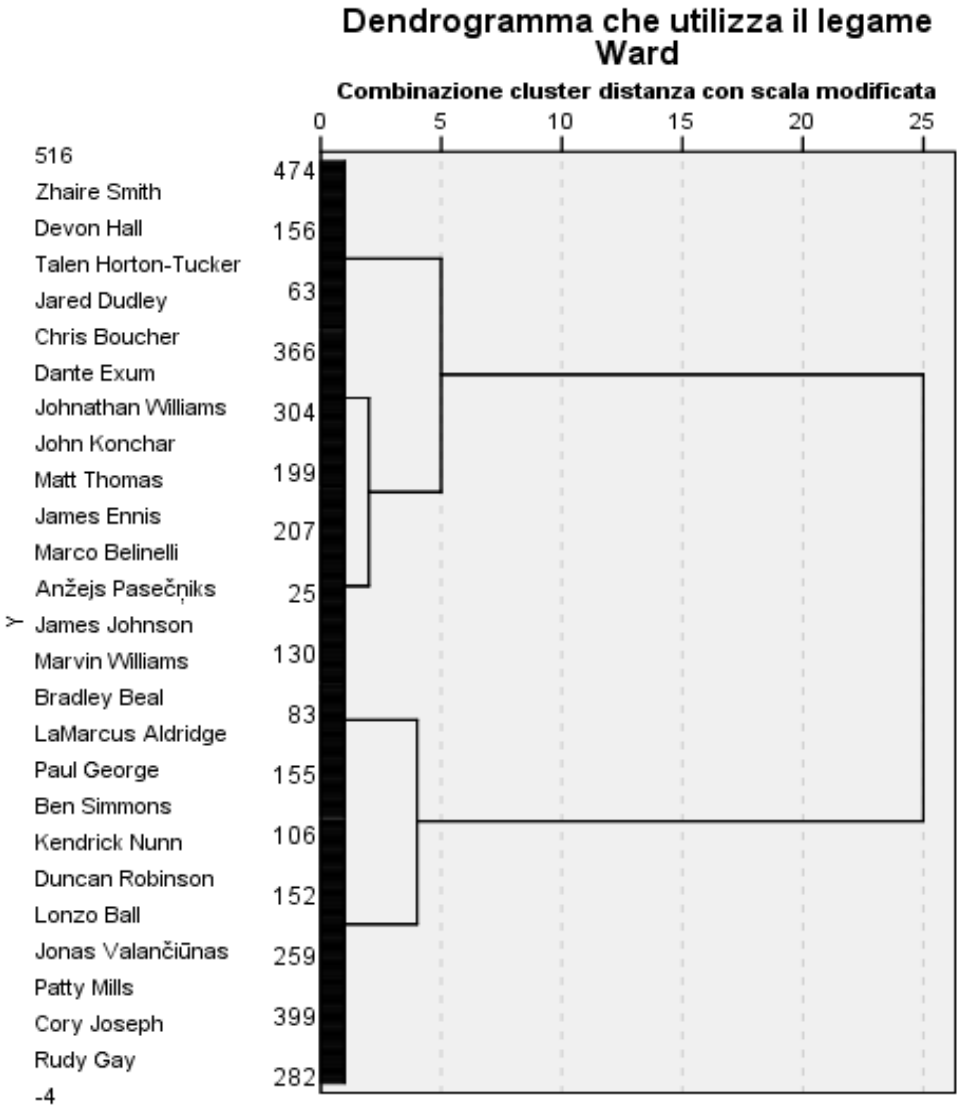
Pianificazione di agglomerazione

Stadio	Combinato in cluster		Coefficienti	Stadio prima apparizione cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	55	0,010	0	0	97
2	281	506	0,030	0	0	23
3	353	385	0,070	0	0	233
4	156	246	0,115	0	0	94
5	179	218	0,160	0	0	196
6	267	485	0,215	0	0	43
7	51	414	0,270	0	0	135
8	32	372	0,325	0	0	93
9	97	292	0,380	0	0	264
10	215	283	0,435	0	0	58
11	373	379	0,505	0	0	87
12	29	326	0,585	0	0	82
13	249	464	0,670	0	0	97
14	253	343	0,755	0	0	50
15	238	313	0,845	0	0	151
…	…	…	…	…	…	…
499	4	58	4436,977	476	477	507
500	5	8	4728,643	490	493	507
501	16	22	5051,655	497	482	506
502	1	19	5406,609	492	496	506
503	9	11	5763,351	488	459	509
504	2	15	6219,626	489	498	505
505	2	12	7287,902	504	495	508
506	1	16	8367,844	502	501	508
507	4	5	10372,039	499	500	509
508	1	2	17721,997	506	505	510

509	4	9	25605,082	507	503	510
510	1	4	74483,986	508	509	0

La prima e l'ultima colonna indicano i livelli di agglomerazione, mentre le colonne denominate "combinato in cluster" e "stadio prima apparizione cluster" specificano la composizione del cluster e a che livello sono stati creati tali cluster. Infine, la quarta colonna specifica la distanza di agglomerazione. Quest'ultima è particolarmente rilevante, siccome dalla sua verifica è possibile determinare il numero ottimale di gruppi.

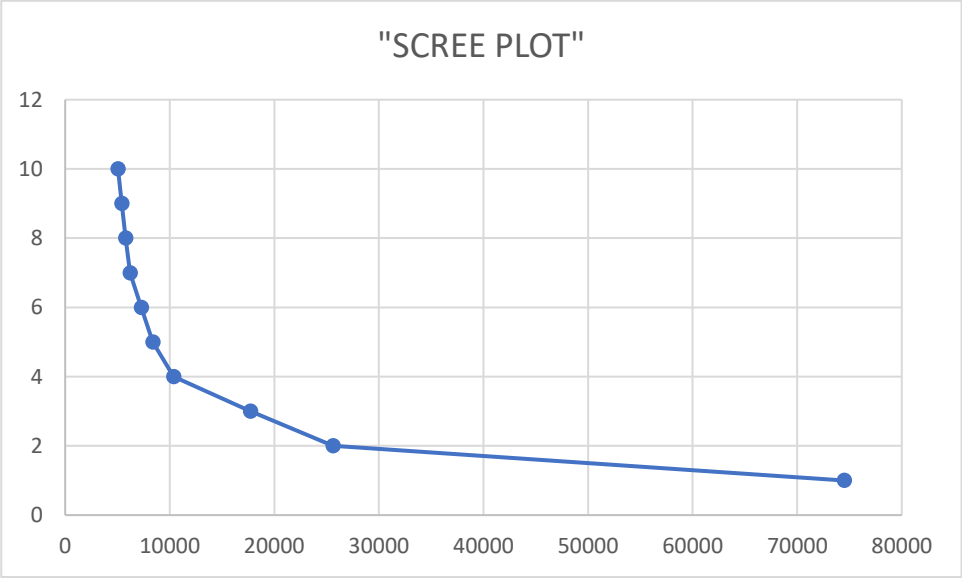
Dalla tabella di pianificazione di agglomerazione, che rappresenta una sintesi numerica del risultato, è possibile ottenere il dendogramma che fornisce una sintesi grafica del risultato ottenuto.



Per determinare il punto in cui tagliare il dendogramma, possiamo affidarci al seguente criterio: se nel passaggio da g gruppi a g+1 gruppi si registra un forte incremento della distanza di fusione, si deve "tagliare" a g gruppi. Pertanto, ho costruito una tabella utilizzando gli ultimi 10 dati della quarta colonna della pianificazione di agglomerazione, da cui ho ricavato l'incremento della distanza di agglomerazione da uno stadio all'altro e il relativo valore percentuale.

stadio	distanza	incremento	incremento %	N°Cluster
501	5051,655	323,012	6,831	10
502	5406,609	354,954	7,026	9
503	5763,351	356,742	6,598	8
504	6219,626	456,275	7,917	7
505	7287,902	1068,276	17,176	6
506	8367,844	1079,942	14,818	5
507	10372,039	2004,195	23,951	4
508	17721,997	7349,958	70,863	3
509	25605,082	7883,085	44,482	2
510	74483,986	48878,904	190,895	1

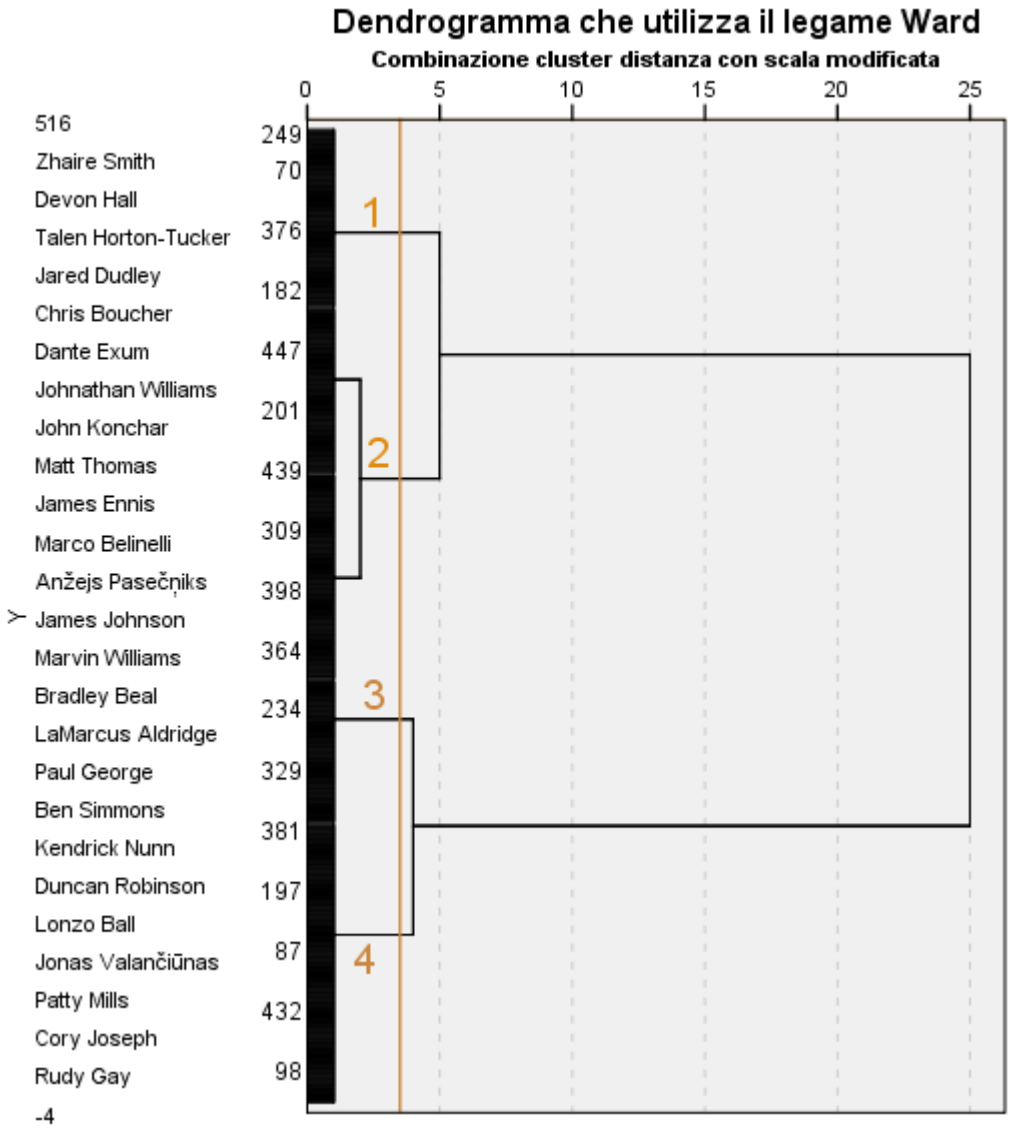
Successivamente, utilizzando il programma Excel, ho costruito un grafico inserendo sull'asse delle ascisse le distanze di agglomerazione, mentre sull'asse delle ordinate ho inserito il numero di cluster. Questo grafico è anche chiamato con il nome "Scree plot".



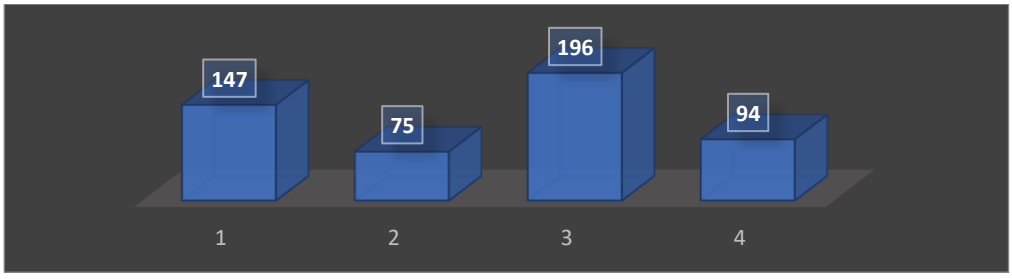
Dal grafico si nota che nel punto con ordinata pari a 4 gruppi, la curva riduce decisamente la sua pendenza diventando quasi-piatta. In quel punto identifichiamo la soluzione "a gomito". Infatti, dalla tabella precedente possiamo osservare che dal passo 507, a cui corrispondono 4 gruppi, al

passo 508, l'incremento della distanza tra gruppi è maggiore rispetto al passo precedente. Quindi, il numero ottimale di cluster da prendere in considerazione è 4.

Tagliando il dendogramma possiamo individuare i cluster trovati.



Infine, ecco la composizione di ogni gruppo.



DESCRIZIONE DEI SINGOLI CLUSTER INDIVIDUATI

Cluster 1

	Statistiche Descrittive							
	N	Minimo	Massimo	Media	Errore Std.	Deviazione Std.	Varianza	Asimmetria
PTS	147	2.0	18.6	11.143	.2100	2.546	6.483	.171
TRB	147	1.6	10.5	4.510	.1635	1.982	3.928	.946
AST	147	.0	8.7	2.388	.1159	1.405	1.975	1.682
STL	147	.0	2.0	.837	.0282	.342	.117	.515
BLK	147	.0	2.5	.486	.0357	.433	.188	2.194
MP	147	19.1	34.4	26.487	.2589	3.139	9.854	-.035

Percentili							
	5	10	25	50	75	90	95
MEDIA	7.400	8.400	9.400	10.800	12.600	15.200	15.620
	2.000	2.300	3.000	4.200	5.800	6.920	9.100
	1.000	1.200	1.500	1.900	3.000	4.560	5.340
	.400	.400	.600	.800	1.100	1.300	1.560
	.100	.100	.200	.400	.600	.920	1.360
	21.440	22.000	24.100	26.600	29.000	30.320	31.520

il cluster 1 contiene 147 giocatori, è il secondo raggruppamento più grande uscito dall'analisi.

Il giocatore medio del primo cluster ha una media punti per partita pari a 11.143, cattura 4.510 rimbalzi a partita, effettua 2.388 assist a partita, ruba 0.837 palloni a partita, stoppa 0.486 tiri a partita e gioca 26.287 minuti a partita.

Si tratta di un gruppo costituito da giocatori che registrano un ottimo rendimento, in relazione ad un minutaggio non eccessivamente alto, il quale però si limita, in alcuni casi, ad una particolare categoria. Per esempio: Deandre Jordan (PTS 8.4, TRB 10.0, AST 1.9, STL 0.4, BLK 0.9, MP 22.0) è uno specialista nei rimbalzi; Jordan Clarkson (PTS 15.3, TRB 2.6, AST 2.0, STL 0.6, BLK 0.3, MP 24.2) offre un maggiore contributo a livello di punti; Elfrid Payton (PTS 9.8, TRB 4.7, AST 7.1, STL 1.6, BLK 0.4, MP 27.7) e Ricky Rubio (PTS 12.7, TRB 4.5, AST 8.7, STL 1.6, BLK 0.2, MP 31.6) hanno medie assist a partita, che sono di parecchio superiori alla media del gruppo. Oppure, il cluster contiene giocatori

con un rendimento globale o semi-globale, nel senso che hanno medie abbastanza buone in due o più campi di osservazione. Per esempio: Lonzo Ball (PTS 11.7, TRB 6.1, AST 6.8, STL 1.3, BLK 0.5, MP 32.4) ha buone medie in termini di punti, rimbalzi totali, assist e palle rubate; Tristan Thompson (PTS 12.2, TRB 10.3, AST 2.1, STL 0.6, BLK 0.9, MP 30.2) e Jonas Valančiūnas (PTS 14.5, TRB 10.5, AST 1.8, STL 0.4, BLK 1.1, MP 26.3) registrano ottime medie sia nei punti che nei rimbalzi; Jonathan Isaac (PTS 12.0, TRB 6.9, AST 1.4, STL 1.6, BLK 2.4, MP 29.7) produce discrete medie a livello di punti, rimbalzi tot, palle rubate e stoppate.

Tuttavia, nel gruppo ritroviamo un giocatore che con le medie del gruppo centra poco nulla. Stiamo parlando di Mychal Mulder (PTS 2.0, TRB 4.0, AST 0.0, STL 0.0, BLK 0.0, MP 29.1), il quale dovrebbe rientrare nel quarto cluster, ma siccome presenta una media minuti per partita abbastanza alta, il software lo ha inserito in questo gruppo.

Quindi, in definitiva possiamo identificare questo gruppo come il cluster dei giocatori da quintetto iniziale o sesti uomini di livello.

Cluster 2

	Statistiche Descrittive							
	N	Minimo	Massimo	Media	Errore Std.	Deviazione Std.	Varianza	Asimmetria
PTS	75	13.9	35.2	20.872	.506	4.382	19.199	.936
TRB	75	2.4	15.3	6.723	.364	3.155	9.957	1.015
AST	75	1.2	10.6	4.532	.255	2.205	4.864	.409
STL	75	.4	2.1	1.037	.043	.373	.139	.829
BLK	75	.1	3.1	.648	.067	.583	.339	1.982
MP	75	26.0	36.9	33.104	.247	2.139	4.576	-.655

Percentili							
	5	10	25	50	75	90	95
MEDIA	14.960	15.860	17.600	19.400	23.300	27.340	29.780
	3.100	3.500	4.300	6.300	7.900	11.920	13.880
	1.400	1.500	2.900	4.200	6.500	7.420	8.300
	.600	.660	.700	1.000	1.300	1.640	1.820
	.100	.200	.300	.500	.800	1.600	1.900
	29.260	30.160	31.800	33.100	34.600	35.840	36.200

Il cluster 2 contiene 75 giocatori, ed è il cluster meno popolato.

Il giocatore medio del secondo cluster ha 20.872 punti di media a partita, 6.723 rimbalzi totali di media a partita, 4.532 assist di media a partita, 1.037 palle rubate di media a partita, 0.583 stoppate di media a partita e gioca 33.104 minuti di media a partita.

In questo gruppo ritroviamo i migliori giocatori della lega di basket statunitense, cioè coloro che hanno un ruolo di prima punta nelle rispettive squadre e che costantemente registrano il miglior rendimento, sia con riferimento ad un'unica variabile sia a livello complessivo. Per esempio: Lebron James (PTS 25.5, TRB 7.7, AST 10.6, STL 1.2, BLK 0.5, MP 34.9), Giannis Antetokounmpo (PTS 29.7, TRB 13.7, AST 5.8, STL 1.1, BLK 1.1, MP 30.9), James Harden (PTS 35.2, TRB 6.4, AST 7.3, STL 1.7, BLK 0.9, MP 36.7).

Nonostante, il cluster sia abbastanza identificabile e poco misto al suo interno, ho comunque individuato alcuni giocatori, le cui loro medie hanno poco a che fare con la tendenza del cluster. Tra questi abbiamo: Carmelo Anthony (PTS 15.5, TRB 6.4, AST 1.5, STL 0.8, BLK 0.5, MP 33.2), Harrison Barnes (PTS 15.0, TRB 4.7, AST 2.3, STL 0.6, BLK 0.2, MP 34.9), Luke Kennard (PTS 15.8, TRB 3.5, AST

4.1, STL 0.4, BLK 0.2, MP 32.9). Si tratta di giocatori che hanno un rendimento buono ma non da giocatore di prima classe, il semplice motivo per cui il software gli ha inseriti in questo raggruppamento è legato al fattore minuti per partita, in caso contrario sarebbero ricaduti nel primo cluster.

Cluster 3

	Statistiche Descrittive							
	N	Minimo	Massimo	Media	Errore Std.	Deviazione Std.	Varianza	Asimmetria
PTS	195	1.0	10.2	5.384	.124	1.729	2.992	.139
TRB	195	.8	8.0	2.990	.105	1.464	2.142	.857
AST	195	.0	5.2	1.232	.059	.835	.698	2.078
STL	195	.0	1.4	.478	.020	.282	.080	.865
BLK	195	.0	1.5	.368	.024	.332	.110	1.531
MP	195	8.7	24.5	15.623	.250	3.493	12.203	.053

Percentili							
	5	10	25	50	75	90	95
MEDIA	2.600	3.200	4.300	5.200	6.500	7.800	8.220
	1.100	1.300	1.900	2.700	3.900	5.100	5.800
	.400	.500	.700	1.000	1.500	2.300	3.000
	.000	.100	.100	.300	.500	.900	1.100
	.100	.200	.300	.400	.600	.900	1.020
	10.280	10.900	12.600	15.800	18.400	20.040	21.020

Il cluster 3 accoglie 195 giocatori, ed è il gruppo più numeroso della nostra analisi.

Il giocatore medio del cluster 3 registra 5.384 punti di media a partita, 2.990 rimbalzi totali di media a partita, 1.232 assist di media a partita, 0.478 palle rubate di media a partita, 0.368 stoppate di media e gioca 15.623 minuti di media a partita.

I giocatori di questo cluster hanno nelle loro rispettive squadre un ruolo marginale, il quale si sostanzia in un minutaggio medio basso e in prestazioni non molto esaltanti. Difatti, rientrano giocatori al primo anno nella lega, per esempio: Nickeil Alexander-Walker (PTS 5.1, TRB 2.0, AST 1.8, STL 0.3, BLK 0.2, MP 12.2), Ty Jerome (PTS 3.8, TRB 1.6, AST 1.6, STL 0.6, BLK 0.3, MP 11.3), Nicolò Melli (PTS 6.7, TRB 3.0, AST 1.1, STL 0.5, BLK 0.3, MP 17.1). Oltre a questa categoria, in questo insieme ritroviamo anche i cosiddetti “veterani”, cioè coloro che anni che militano nel campionato statunitense e che adesso hanno un ruolo parecchio ridimensionato, come: J.J. Barea (PTS 8.9, TRB 2.2, AST 3.9, STL 0.2, BLK 0.1, MP 15.0), Marco Belinelli (PTS 5.8, TRB 1.8, AST 1.3, STL 0.2, BLK 0.0, MP 15.0), Vince Carter (PTS 4.9, TRB 2.1, AST 0.8, STL 0.4, BLK 0.5, MP 14.6).

Nonostante sia il cluster più consistente, non presenta enormi errori di fusione. Sono presenti solo alcuni casi anomali, quali: Zylan Cheatham (PTS 1.3, TRB 2.0, AST 0.7, STL 0.0, BLK 0.3, MP 10.3), Josh Gray (PTS 1.0, TRB 1.0, AST 1.0, STL 0.0, BLK 0.0, MP 11.5), Keldon Johnson (PTS 2.6, TRB 1.2, AST 0.6, STL 0.0, BLK 0.0, MP 10.2). Questi si caratterizzano per un rendimento al di sotto dell'andamento complessivo del gruppo, soprattutto per quanto riguarda le palle rubate e le stoppate e per questi motivi dovrebbero rientrare nel cluster 4. All'opposto, il giocatore Doug McDermott (PTS 10.2, TRB 2.5, AST 1.1, STL 0.1, BLK 0.1, MP 20.0), presenta un rendimento, dal punto di vista della media punti a partita, superiore ma dato che nelle altre categorie presenta valori simili a quelli delle medie del gruppo, è corretto che sia stato inserito in questo gruppo.

Cluster 4

	Statistiche Descrittive							
	N	Minimo	Massimo	Media	Errore Std.	Deviazione Std.	Varianza	Asimmetria
PTS	94	.0	6.0	1.562	.112	1.086	1.181	.627
TRB	94	.0	3.5	1.001	.083	.808	.653	.994
AST	94	.0	1.3	.350	.031	.299	.090	.920
STL	94	.0	1.5	.173	.024	.228	.052	2.836
BLK	94	.0	1.5	.128	.022	.213	.045	3.581
MP	94	.5	9.5	5.386	.231	2.237	5.001	.076

Percentili							
	5	10	25	50	75	90	95
MEDIA	.000	.000	1.000	1.450	2.350	3.000	3.125
	.000	.000	.400	.800	1.400	2.100	2.725
	.000	.000	.000	.100	.300	.400	.500
	.000	.000	.100	.300	.500	.800	1.000
	.000	.000	.000	.100	.200	.300	.525
	1.950	2.650	3.650	5.200	7.500	8.600	9.000

Il cluster 4 accoglie 94 giocatori, ed è il terzo gruppo più capiente.

Il giocatore medio di questo cluster ha 1.562 punti di media a partita, 1.001 rimbalzi totali a partita, 0.350 assist di media a partita, 0.173 palle rubate a partita, 0.128 stoppate di media a partita e gioca 5.386 minuti di media a partita.

Questo cluster è quello più identificabile tra i 4, dal momento che in esso rientrano solo giocatori che hanno un impatto sulla partita quasi nullo, questo è dato da diverse motivazioni: scelte tecniche degli allenatori, alta competizione tra giocatori di una stessa squadra, infortuni, fine carriera ecc.

Per dare un'idea, mostro un breve elenco di alcuni giocatori che fanno parte di questo gruppo: Michael Frazier (PTS 1.8, TRB 0.4, AST 0.1, STL 0.3, BLK 0.0, MP 8.3), Udonis Haslem (PTS 1.7, TRB 2.7, AST 0.3, STL 0.0, BLK 0.0, MP 7.0), Mfiondu Kabengele (PTS 3.5, TRB 0.9, AST 0.2, STL 0.2, BLK 0.2, MP 5.3).

Quindi possiamo identificarli con il termine "riserve".

Conclusioni

L'analisi tramite il software SPSS, dopo diversi tentativi sul tipo di procedura da utilizzare, è stata comunque eseguita senza grossi problemi.

Per quanto concerne l'affidabilità del risultato ottenuto, è necessario ricordare che il dataset conteneva 511 elementi, e questo sicuramente ha reso più complesso il processo di classificazione da parte del software. Inoltre, nella matrice di correlazione si evidenziava una certa connessione fra la variabile punti per partita e la variabile minuti per partita, per la quale ho ritenuto non intervenire con ulteriori modifiche nel dataset e nelle procedure di analisi, siccome verificando la procedura alternativa di aggregazione, i risultati ottenuti mostravano maggiori casi di anomalia nei gruppi. Quindi, nonostante ciò la procedura è stata realizzata con successo. Infatti, in ogni gruppo è stato possibile identificare in maniera abbastanza precisa una determinata macro-classe di giocatori:

- Il cluster 1 è rappresentato da giocatori di quintetto o sesti uomini di livello
- Il cluster 2 è rappresentato dai fuoriclasse o giocatori di prima gamma
- Il cluster 3 è rappresentato da giocatori di rotazione
- Il cluster 4 è rappresentato dalle riserve

Ovviamente si tratta di una classificazione abbastanza generale, siccome ho utilizzato come variabili le principali voci di valutazione del rendimento di un giocatore (punti, rimbalzi, assist, palle rubate, stoppate e minuti giocati), senza usufruire di ulteriori dati, i quali sicuramente mi avrebbero permesso di ottenere una soluzione più specifica, ma che sarebbe andata oltre l'obiettivo di questa analisi.

Sitografia

Capitolo V ANALISI DEI GRUPPI di Andrea Cerioli e Sergio Zani

https://people.unica.it/lucafrigau/files/2012/04/Cap-V-_ANALISI-DEI-GRUPPI.pdf

Introduzione alla Cluster Analysis

http://www.federica.unina.it/economia/analisi-statistica-sociologica/introduzione-cluster-analysis/

Sito da cui ho ricavato il Dataset

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

Cluster Analysis (S. Terzi)

http://host.uniroma3.it/facolta/economia/db/materiali/insegnamenti/185_903.pdf

Manuale SPSS Statistics Base 25, in

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/it/client/Manuals/IBM_SPSS_Statistics_Base.pdf

esempi di applicazioni della Cluster Analysis nella pagina e-learning del corso “Metodi statistici per il management”

Appendice Formule

Indice di correlazione di Pearson:

$$\rho XY = \frac{\sigma XY}{\sigma X \sigma Y}$$

Distanza euclidea quadratica:

$$d\,(\mathcal{A}, \mathcal{B}) = \sum_{i=1}^{k} |X_{ai} - X_{bi}|^2$$

Metodo di raggruppamento di Ward:

$$d_{(i,j)k} = \frac{1}{ni\ +\ nj + nk}\left[(n\mathrm{i} + nk)d_{jk}^2 +\ (nj + nk)d_{jk}^2 -\ n_k d_{ij}^2\ \right]$$