Analisi delle statistiche dei centrocampisti della MLS

Alessandro La Farciola

9 dicembre 2021

1 Introduzione

1.1 Contesto e prospettiva

L'obiettivo della nostra analisi è quello di indagare le statistiche principali dei centrocampisti della MLS (Major League Soccer), il massimo campionato di calcio statunitense, al fine di suddividere i vari giocatori in cluster, assimilandoli in base alle caratteristiche di gioco. In questo senso, i risultati che si possono ottenere risulterebbero utili, ad esempio, ad una società che desideri acquistare un nuovo giocatore e potrebbe usufruire della distinzione in cluster per cercare il centrocampista giusto in base alle sue prestazioni e al suo modo di giocare.

1.2 Presentazione del *dataset*

La tabella di dati da analizzare ha come osservazioni i giocatori della MLS classificati come centrocampisti, mentre come fattori presenta le statistiche principali e più rilevanti in tale ruolo relativi all'anno 2019. In particolare, come giocatori sono stati selezionati tutti quelli che durante la stagione hanno giocato almeno 9 partite. Questa decisione è stata presa per evitare speculazioni su giocatori che, evidentemente, giocando troppo poco, non hanno generato un numero soddisfacente di statistiche, rendendo meno efficace l'analisi da svolgere. I fattori dati presi in considerazione, invece, sono:

- *Gls*: numero di goal segnati;
- *Ast*: numero di assist effettuati;
- *CrdY*: numero di ammonizioni;
- *Sh*: numero di tiri effettuati;
- *Touches*: numero totale di tocchi effettuati (stop, controllo e passaggio contano per uno);
- *Drib*: numero di dribbling eseguiti con successo;
- *Fls*: numero di falli compiuti;
- *Off*: numero di fuorigioco;
- *Crs*: numero di cross;
- *Tklw*: numero di tackle vinti;
- *Recov*: numero di recuperi palle;
- *Pass*: numero totale di passaggi realizzati.

1.3 Link alle tabelle di dati

I dati selezionati per la nostra analisi sono stati interamente reperiti sul sito di `www.fbref.com`. In particolare, è possibile accedere direttamente ad essi attraverso i seguenti link:

- `https://fbref.com/en/comps/22/2798/stats/2019-Major-League-Soccer-Stats`;
- `https://fbref.com/en/comps/22/2798/misc/2019-Major-League-Soccer-Stats`;

- `https://fbref.com/en/comps/22/2798/shooting/2019-Major-League-Soccer-Stats`;
- `https://fbref.com/en/comps/22/2798/possession/2019-Major-League-Soccer-Stats`;
- `https://fbref.com/en/comps/22/2798/passing/2019-Major-League-Soccer-Stats`.

2 Analisi

2.1 PCA preliminare

In primo luogo, è opportuno analizzare la matrice delle correlazioni, al fine di incominciare a capire come si legano tra loro i vari fattori della nostra tabella di dati.

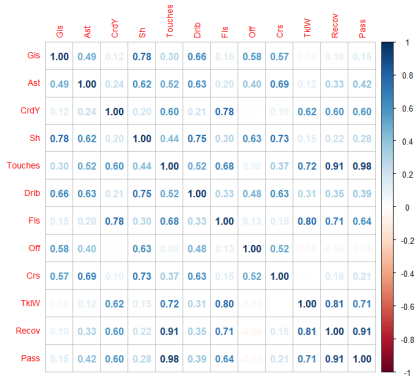
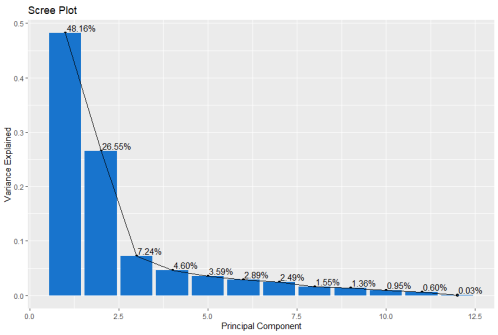
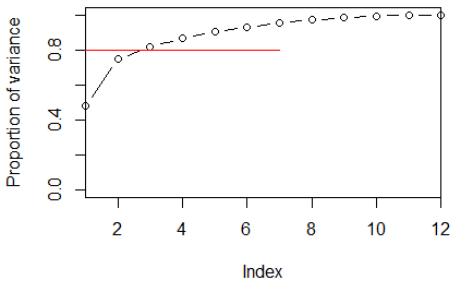


Figura 1: Matrice delle correlazioni

Osservando la figura, emerge fin da subito la presenza di forti correlazioni tra alcuni tipi di fattori. In particolare, si può notare il significativo legame tra *Touches* e molti degli altri fattori e questo risulta coerente con il significato che assume, in quanto la maggior parte delle statistiche considerate necessitano il contatto con la palla. Questo ci suggerisce una buona possibilità di ridurre efficacemente la dimensione del problema. Nonostante ciò, il legame esistente tra i fattori permette di ritenere ciascuno di essi determinante nel fornire informazioni al fine dell'analisi da effettuare.

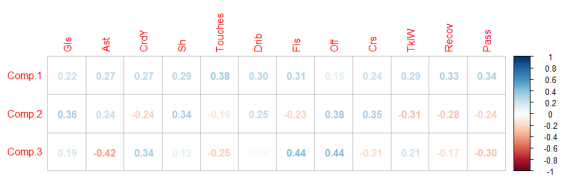
Possiamo, quindi, procedere con una rapida analisi delle componenti principali sulla tabella di dati opportunamente standardizzata. I grafici successivi mostrano l'andamento della varianza spiegata da cui si è ritenuto sufficiente considerare rilevanti le prime tre componenti principali, le quali raggiungono circa l'82% di varianza spiegata.



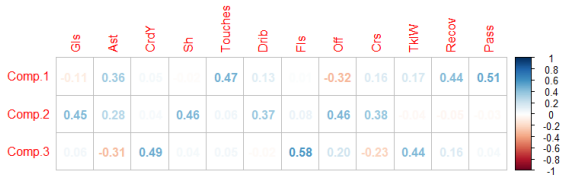


Fin da subito, emerge il fondamentale significato assunto dal piano principale, il quale cattura il 75% della varianza del nostro problema. Da qui segue che lo studio delle osservazioni sul piano principale, come la suddivisione in cluster, che rimane il nostro obiettivo primario, assume un ruolo soddisfacente, catturando gran parte del fenomeno che emerge dai dati.

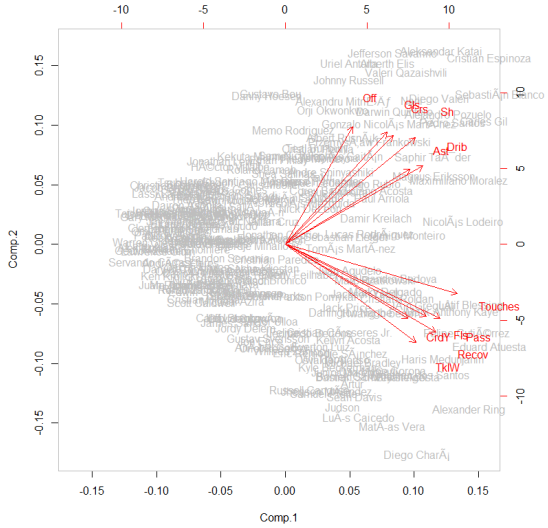
Tornando alla PCA, uno sguardo alla matrice dei loadings (e una rotazione) per le prime tre componenti e a due piani principali ci permette di interpretare le componenti.

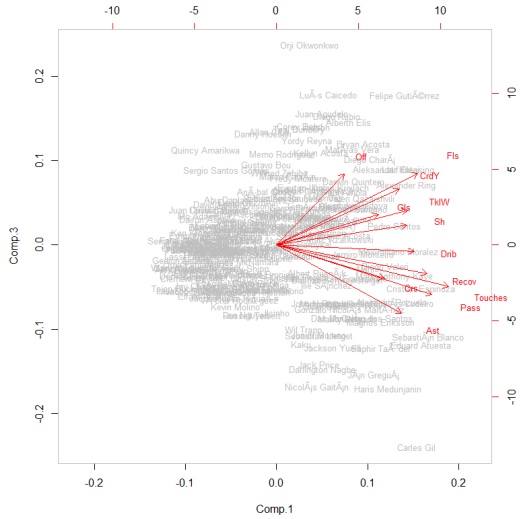


Loadings



Rotazione





Alla luce dei precedenti grafici, è possibile fare le seguenti osservazioni:

- I valori dei loadings relativi alla prima componente, tutti positivi e tutti abbastanza rilevanti, ci suggeriscono un significato riassuntivo assunto da tale componente. Guardando anche il piano principale, si può osservare come tutti i fattori puntino lungo la stessa direzione. Questo significa che la prima componente di un'osservazione tende ad essere maggiore quante più statistiche quel giocatore ha generato durante l'intera stagione. Pertanto, sul piano principale si dispongono sull'ascissa positiva quei giocatori che si sono distinti maggiormente, mentre su quella negativa si dispongono i centrocampisti che hanno generato meno statistiche.

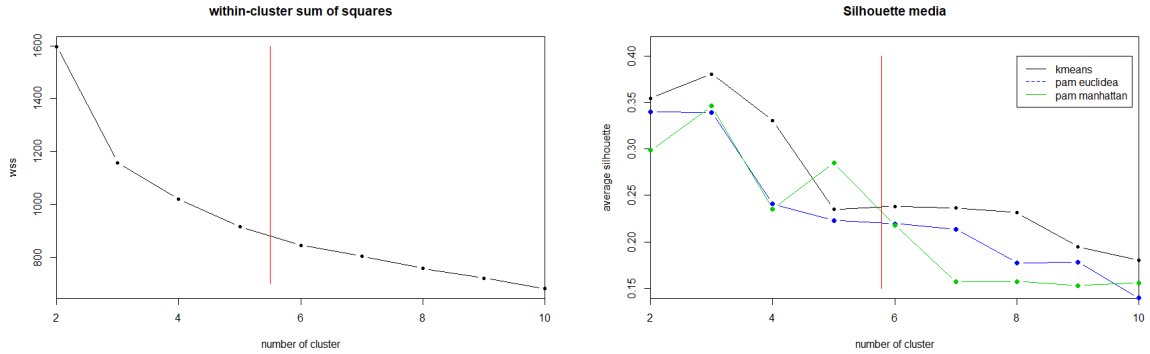
- Il significato assunto dalla seconda componente è determinante ai fini della nostra analisi e risulta evidente osservando il piano principale. Essa infatti distingue nettamente i fattori relativi alle statistiche offensive, quali *Goal*, *Assit*, *Tiri*, *Dribbling*, *Offside*, *Cross*, da quelle più difensive, di contatto e di regia, quali *Tackle*, *Passaggi*, *Tocchi*, *Falli*, *Recuperi*, *Ammonizioni*.

- Il ruolo della terza componente, infine, è determinato, nella direzione positiva, principalmente dai fattori *Ammonizioni*, *Offside*, *Falli*, *Tackle* (altamente correlati al numero di falli), che indicano

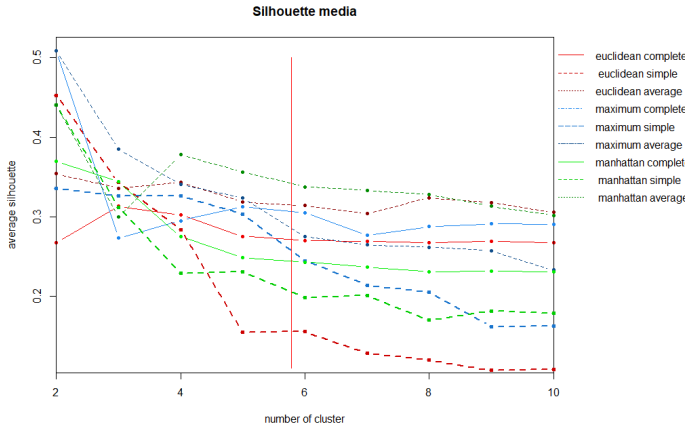
una maggiore aggressività agonistica del giocatore, ma allo stesso tempo una maggiore tendenza a provocare irregolarità di gioco. Infatti, nella direzione opposta, si dispongono i fattori come *Assist*, *Passaggi*, *Cross* e *Recuperi palla* (regolari), che indicano le maggiori qualità tecniche del giocatore.

2.2 Clustering

Sfruttando l'interpretazione delle componenti principali appena svolta, siamo pronti per procedere con il clustering al fine di suddividere le nostre osservazioni. Innanzitutto, si vuole indagare il numero di cluster opportuno in cui suddividere i giocatori. Per farlo, possiamo sfruttare i seguenti grafici che mostrano la *within-cluster sum of squares* per il metodo *k-means* e la *silhouette* media per i metodi *k-means*, *pam* con la distanza euclidea e *pam* con la distanza *manhattan*.



Entrambi i grafici, che si basano sui metodi a prototipo, ci suggeriscono di indagare la divisione fino a 5 cluster. Una conferma di questo si ottiene anche dai metodi gerarchici, come mostra il seguente grafico.

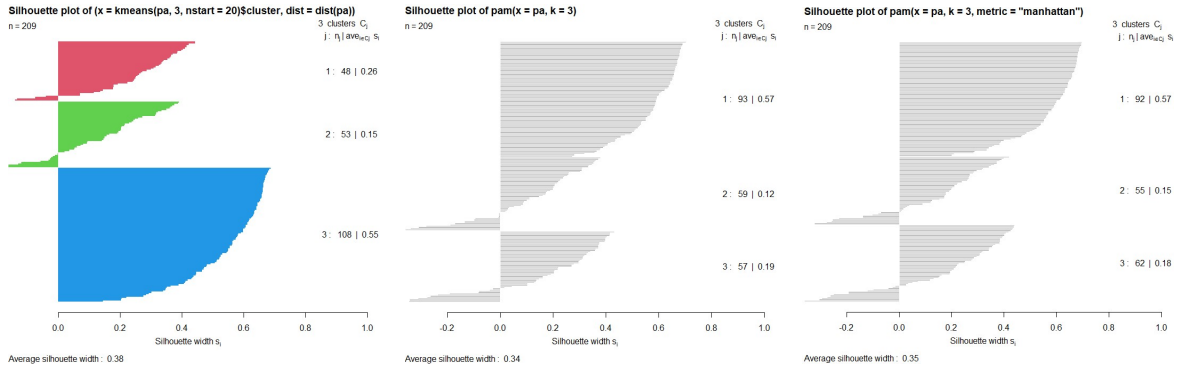


Il nostro obiettivo, ora, è quello di capire quali metodi realizzano una migliore e più efficace divisione in cluster e procedere, quindi, con una loro interpretazione. In particolare, possiamo ricondurci ad

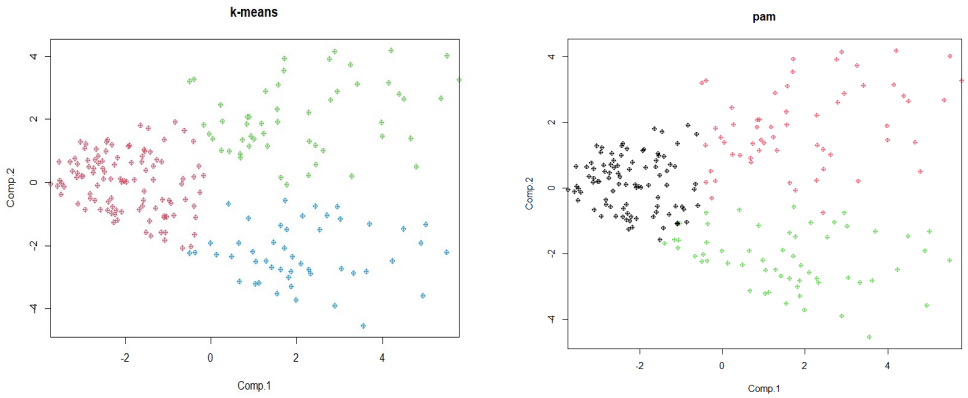
analizzare un numero di cluster pari a $3, 4, 5$, scartando il caso di due cluster che risulterebbe poco significativo per il nostro obiettivo.

2.2.1 3 cluster

Proviamo ad analizzare una possibile divisione in tre cluster attraverso vari metodi e confrontiamoli. I seguenti grafici, in ordine, mostrano le *silhouette* per i metodi *k-means*, *pam* con la distanza euclidea e *pam* con la distanza *manhattan*.



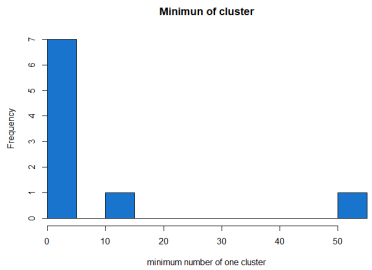
Osservando le *silhouette*, i tre metodi sembrano lavorare abbastanza efficacemente e in maniera piuttosto paragonabile. Si può notare come in tutti vi è la presenza di un cluster molto più numeroso che possiede una *silhouette* media molto alta il che ci suggerisce un'ottima assegnazione in quel gruppo. La bontà di tale assegnazione bilancia le altre due che presentano dei valori di *silhouette* più bassa, con alcuni valori anche negativi, specialmente nei casi del metodo *pam*. Per questo motivo, dei tre sembrerebbe più opportuno considerare il metodo *k-means*. Proviamo a vedere sul piano principale quali cluster si distinguono, limitandoci al caso *k-means* e *pam* con la distanza euclidea (*pam* con la distanza *manhattan* genera quasi gli stessi cluster).



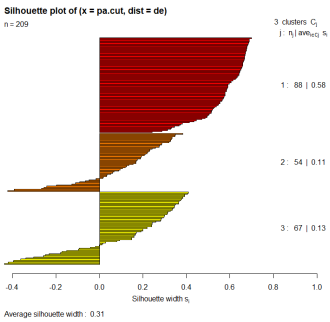
Dai precedenti grafici risulta evidente come le tipologie di cluster individuate dai due metodi si equivalgono. Si può notare come il metodo *k-means* agglomeri nel cluster più grande un numero di

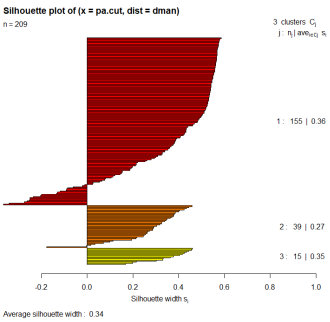
osservazioni maggiore rispetto a *pam*. Mentre sugli altri due si possono evidenziare alcune differenze di assegnazione lungo il confine e questo è sintomo di una certa difficoltà di assegnazione. Tale aspetto suggerisce che i giocatori in quell'area presentano caratteristiche in comune ad entrambi i cluster.

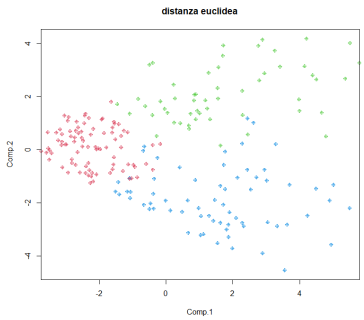
Un risultato di questo tipo è dato anche dai metodi gerarchici che, però, si comportano in maniera peggiore rispetto ai precedenti metodi per l'assegnazione delle osservazioni in cluster. Per scegliere quali distanze utilizzare per applicare il metodo gerarchico, si è visto principalmente la numerosità minima di un cluster dopo l'assegnazione. Molti metodi catturano un unico grande cluster lasciando fuori singole osservazioni che fanno gruppo a sé, come mostra il seguente istogramma.

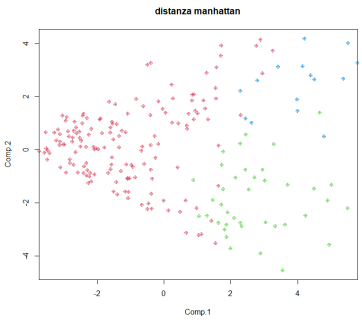


Tale assegnazione, però, risulta errata in quanto dal piano principale, seppur alcuni giocatori spiccano in una determinata direzione, non emerge alcun dato particolarmente anomalo e per questo tali metodi sono stati scartati. I seguenti grafici mostrano le *silhouette* e i cluster sul piano principale ottenute con i migliori metodi gerarchici, ovvero nel caso della distanza euclidea e della *manhattan* con la *complete average linkage*.



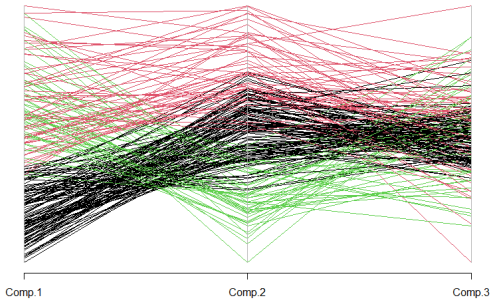






I cluster realizzati ricordano quelli precedenti e questo avvalora il risultato ottenuto. Sono paragonabili anche le *silhouette*, anche se possiamo notare alcune differenze. Per la distanza euclidea, come prima, un cluster numeroso è ben assegnato e bilancia gli altri due (dove, però, emerge un numero significante di *silhouette* negative); per la distanza *manhattan* le silhouette dei singoli cluster sono più uniformi ma al costo di avere un cluster troppo numeroso che, come si vede anche dalla suddivisione sul piano principale, non fornisce un risultato più efficace.

A questo punto siamo pronti per interpretare i cluster ottenuti, sfruttando il significato precedentemente discusso delle componenti principali. In più, grazie alle discussione sui metodi appena svolte, possiamo concentrarci maggiormente sul metodo *k-means*, analizzando quali valori assumono i cluster sulle componenti principali.



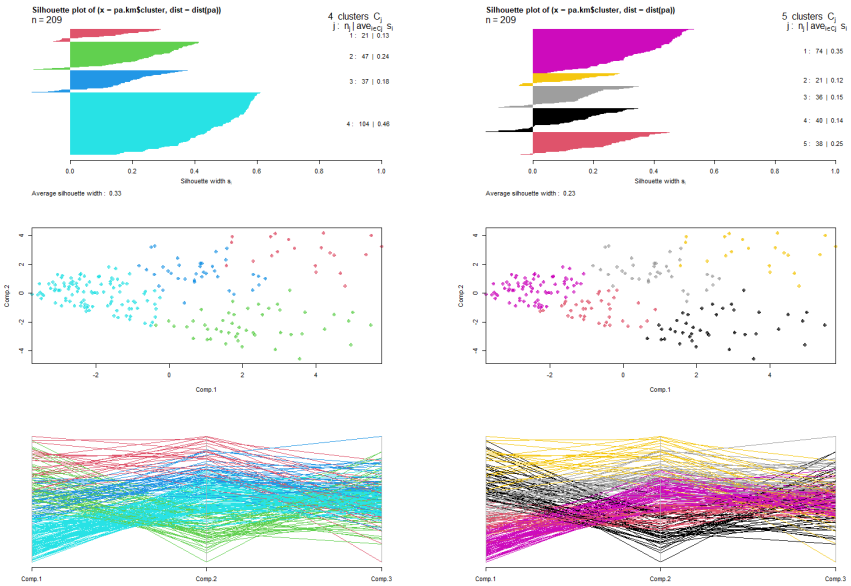
Alla luce del precedente grafico, è evidente come il cluster più numeroso (nero nella figura di sopra) ha un basso valore lungo la prima componente e un valore medio lungo la seconda e la terza componente. Dato il significato riassuntivo della componente principale, ciò vuol dire che i giocatori in questo cluster sono quelli che hanno generato meno statistiche nel corso della stagione, giocando meno partite e non distinguendosi particolarmente nei numeri analizzati (e di conseguenza assumono un valore mediano sulle altre due componenti). Gli altri due cluster, invece, sono composti da giocatori che hanno un valore elevato sulla prima componente e quindi hanno avuto un ruolo più determinante durante le partite. Tali cluster, però, si distinguono nettamente sulla seconda componente: uno è composto da giocatori più offensivi, il cui modo di giocare si concentra più sull'attacco della profondità, i goal, i tiri e così via. Dall'altra parte, invece, vi sono i giocatori che si differenziano maggiormente per le loro statistiche più difensive e quindi si tratta di centrocampisti che giocano più lontano dalla porta e più vicino alla difesa, che toccano molte volte la palla e che vanno più spesso a contrasto. Infine, notiamo come non ci riesce a molto bene distinguere i due cluster precedenti lungo la terza componente principale. Pertanto, può essere interessante andare ad indagare una possibile divisione in più cluster con l'obiettivo di creare dei sottogruppi che si differenzino ulteriormente su questa componente.

2.2.2 4 o 5 cluster

In questa sezione si vuole indagare una possibile suddivisione in 4 o 5 cluster e si procederà esattamente come fatto in precedenza[1]. I risultati ottenuti con i vari metodi, seppur con alcune differenze, sono paragonabili, in particolare per quanto riguarda la disposizione dei cluster realizzati sul piano principale. Ci concentriamo, però, sul metodo che sembra fornire i risultati più soddisfacenti, ovvero

[1]Per motivi di spazio si presenteranno soltanto i risultati rilevanti, dato che la strategia e il metodo con cui si è proceduto è del tutto analogo al caso di 3 cluster.

il *k-means*. I seguenti grafici mostrano le *silhouette* in entrambe le classificazione, la loro distribuzione sul piano principale e il significato assunto sulle componenti principali[2].



Una suddivisione in più cluster può fornirci maggiori informazioni sui giocatori analizzati e, infatti, otteniamo come risultato quello di ridurre leggermente il cluster più numeroso e dividere in due gli altri due cluster.

- Con una suddivisione in 4 cluster (colonna di sinistra) viene diviso il gruppo dei centrocampisti più offensivi (blu e rosso) e guardando il *parcoord* notiamo come il cluster rosso ha un valore maggiore sia sulla componente 1 che sulla 2 e questo vuol dire che quei giocatori si distinguono ancora di più per le qualità offensive. In più si nota una leggera tendenza del cluster rosso ad avere un valore minore sulla componente 3, rispetto al blu, e, dato il significato assunto dalla componente 3, questo significa maggiori qualità tecniche e meno propensione ai falli e al contatto.

- La suddivisione in 5 cluster (colonna destra) genera un'ulteriore separazione, in particolare quella dei centrocampisti più difensivi (nero e rosso). In questo caso, è meno netta la distinzione e, infatti, il cluster rosso è composto più da giocatori del vecchio gruppo numeroso piuttosto che da quelli che si volevano separare. Questa difficoltà di assegnazione si evince anche da una *silhouette* media più bassa. Nonostante questo, possiamo notare come il cluster rosso abbia sì un valore minore sulla prima componente, ma anche un valore maggiore rispetto al nero sulla seconda componente e questo significa una minore attitudine difensiva da parte di tali centrocampisti. Infine, non riusciamo a notare una soddisfacente separazione lungo la terza componente.

3 Conclusioni

Tutti i risultati presentati mostrano una migliore assegnazione (*silhouette* alta) sul cluster più numeroso che si posiziona sulla sinistra del piano principale rispetto agli altri cluster. Ciò si spiega

[2]I colori identificano i cluster e sono coerenti per le figure disposte in verticale.

teoricamente perché, osservando il piano principale, le osservazioni risultano essere molto più vicine e il valore della *silhouette* di un'osservazione dipende fortemente dalla distanza dagli elementi del proprio cluster. D'altro canto, ciò è coerente con le nostre interpretazioni, dato che è sicuramente più facile accomunare quei giocatori che sono stati meno incisivi, rispetto a capire la tipologia di gioco a partire da semplici statistiche. Inoltre, è chiaro come le osservazioni presenti sui bordi di due cluster (che coincidono con quelle osservazioni dalla *silhouette* negativa) a seconda del metodo possano essere assegnati all'uno o all'altro. La difficoltà di assegnazione, però, non determina in negativo la nostra analisi: i giocatori che sono a cavallo tra il cluster più a sinistra e quelli adiacenti sono coloro che, sebbene non abbiano generato un numero molto elevato di statistiche, ne possiedono a sufficienza per capire la tipologia di giocatore (più offensivo, più difensivo). D'altro canto i giocatori che sono a cavallo tra il cluster dei centrocampisti più offensivi e quello dei più difensivi (per capirci, le osservazioni che hanno la componente 2 in un intorno di 0) sono coloro che non spiccano particolarmente in una delle due caratteristiche ma che, data la loro versatilità, possono ricoprire entrambe le posizioni efficacemente (avendo generato un numero importante di statistiche in entrambe le tipologie).

Infine, è doveroso sottolineare che le statistiche registrate non riescono a cogliere a pieno tutti gli aspetti del gioco del calcio. Ciò nonostante, la nostra analisi aiuta in modo sufficiente ad individuare a grandi linee la tipologia di un centrocampista in base alle proprie caratteristiche tecnico-tattiche, offrendo una linea guida sul problema iniziale che ci eravamo proposti di indagare.