UNIVERSITÀ
DI SIENA
1240

DIPARTIMENTO DI ECONOMIA POLITICA E STATISTICA

CORSO DI LAUREA MAGISTRALE IN SCIENZE STATISTICHE PER LE INDAGINI CAMPIONARIE (SSIC)

Modelli statistici per riprodurre caratteristiche di stile di gioco per le squadre di calcio

Relatore: Chiar.mo Prof.
Gian Piero Cervellera

Correlatore: Chiar.mo Prof.
Gianni Betti

Candidato
Leonardo Mori

Anno accademico 2017/18

A nonno Fernando e nonna Loriana.

Abstract

The aim of this document is to analyze the FIFA World Cup 2018 using a mathematical-statistical approach. In the first part, the sixty-four matches of the tournament are grouped in clusters, according to their characteristics. Then, for each of the created groups, the system of play and the playing style are identified. To do this, four indices have been used: two statistical indices related to the graph theory (closeness centrality and betweenness centrality) and two indices concerning the players' position in the field (median position on the long side and median position on the short side). In the end, the document presents a comparison between the data obtained from the analysis described above and the predictions made by the article *Predicting FIFA World Cup 2018 key role and playing style features* (2018), in order to check the goodness of the predictions and verify if the teams given as favorites have played as expected.

Indice

1. Introduzione

Sette parole hanno a lungo dominato il calcio: *questo è il modo in cui è sempre stato fatto*.

Il bel gioco è intriso di tradizioni. Il bel gioco si aggrappa alle sue convinzioni e ai suoi dogmi. Il bel gioco è gestito da uomini che non vogliono vedere il proprio potere sfidato da estranei, che sanno che il loro modo di vedere il calcio è il vero modo di veder il calcio; non vogliono sentirsi dire che, da più di un secolo, a loro manca qualcosa.

Il bel gioco, però, è pronto per il cambiamento. E al centro di quel cambiamento ci sono i numeri. Sono i numeri che cambieranno le convenzioni ed invertiranno le norme. Sono i numeri che ci permetteranno di vedere il calcio come non lo abbiamo mai visto prima. Ma non si tratta solo di raccogliere dati, è importante sapere cosa fare con loro, cogliere la loro verità interiore. Questo è il futuro. Questa è la nuova frontiera del calcio. Ciò non significa, certamente, che tutte le tradizioni del calcio siano sbagliate; i dati che oggi abbiamo a disposizione confermano, infatti, che alcune cose che abbiamo sempre pensato fossero vere siano effettivamente vere. Oltre a ciò, tuttavia, i numeri e la loro analisi ci offrono ulteriori novità e mostrano aspetti del calcio che non potremmo conoscere intuitivamente.

La matematica oggi è estremamente avanzata nel risolvere la grande maggioranza dei problemi, tuttavia, ogni appassionato di sport sa come lo sport possa essere casuale; è quindi possibile utilizzare la matematica per prevedere cosa accadrà? In caso affermativo, quanto spesso il risultato sarà corretto? Ma soprattutto, quali sono i dati da selezionare e i modelli da utilizzare per effettuare un'analisi corretta e statisticamente significativa? Negli ultimi anni il numero di dati disponibili per ogni singola partita è cresciuto in modo esponenziale e la progressiva apertura alla cultura statistica da parte degli operatori calcistici ha portato allo sviluppo di numerosi studi in questo campo e all'elaborazione di altrettanti modelli statistico-matematici.

La natura del gioco del calcio implica che statistiche semplici, come il numero di goal segnati o il numero di tiri effettuati durante una partita, non consentono una misurazione dettagliata delle prestazioni della squadra. Quindi, se tradizionalmente molta attenzione è stata dedicata ai goal e alla loro distribuzione nell'arco dei novanta minuti di gioco, oggi si concentra principalmente sui passaggi, un evento molto più appropriato da analizzare quando si cerca di descrivere le caratteristiche dello stile di gioco di una squadra. Perciò, lo stile di gioco (o tattica), definito come il modo in cui una squadra muove la palla in campo, è certamente diventato un elemento di interesse primario in questo sport. Di recente alcuni studi hanno analizzato questo tema: per

esempio, Hughes e Franks (2005) hanno effettuato un confronto tra le lunghezze delle sequenze di passaggi delle squadre oggetto di studio, Gyarmati, Kwak e Rodriguez (2014) hanno analizzato le strutture di passaggio per trovare similitudini e differenze tra stili di gioco, Clemente, Martins e Mendes (2014) hanno analizzato la varianza di alcuni indicatori relativi ai passaggi per capirne le distribuzioni, Pena (2014) ha studiato il possesso palla delle squadre utilizzando un processo di Markov e Cintia, Giannotti, Pappalardo, Pedreschi e Malvaldi (2015) hanno proposto un indicatore di performance basato sui passaggi per prevedere l'esito delle partite.

In termini statistici, lo stile di gioco di una squadra può essere adeguatamente rappresentato tramite un grafo, una struttura costituita da un insieme di punti (nodi), corrispondenti ai giocatori, e un insieme di linee (archi) che uniscono coppie di nodi, corrispondenti ai passaggi. I grafi (o reti) sono uno strumento utilizzato nello studio di una varietà di tematiche, che vanno dalle questioni tecnologiche e dei trasporti ai fenomeni sociali fino alle discipline biologiche. La versatilità dei grafi è tale che una ricca teoria matematica è stata sviluppata intorno a loro, in particolare da Eulero, in relazione al problema dei sette ponti di Königsberg, Erdòs e molti altri. Il calcio, come gli altri sport di squadra, prevede una fase di possesso, detta anche fase offensiva, che si concretizza quando un giocatore rispetta almeno una delle seguenti condizioni: i) effettua almeno due tocchi consecutivi della palla, ii) esegue un passaggio positivo (permettendo il mantenimento del possesso palla), iii) effettua un tiro verso la porta. Le partite di calcio, per questo motivo, rappresentano un interessante esempio di rete.

Tutte le più famose squadre di calcio della storia hanno avuto uno stile di gioco fortemente rappresentativo (ad esempio il *tiki-taka* del Barcellona di Guardiola), come un'impronta riconoscibile del loro gioco, che è sempre stato pensato come qualcosa di osservabile agli occhi degli esperti di calcio piuttosto che descritto da determinate statistiche. Per identificare questa impronta specifica, si utilizza la distribuzione dei passaggi di una squadra per costruire un grafo, con nodi corrispondenti a giocatori e frecce pesate al numero di passaggi riusciti tra giocatori, ottenendo quindi un'immagine immediata dello stile di gioco della squadra, che può essere proficuamente utilizzato da allenatori o dirigenti per osservare le aree di campo sfruttate o sottoutilizzate e per rilevare il contributo di ciascuno giocatore.

Nella mia analisi utilizzerò la teoria matematica dei grafi per esaminare le informazioni statistiche ricavate dalle partite dei Mondiali di calcio di Russia 2018 e misurare le prestazioni di squadra e dei suoi giocatori. In particolare, farò riferimento all'articolo *Predicting FIFA World Cup 2018 key role and playing style features* di Campagnolo G., Duncan A., Diquigiovanni J.,

Papastathopoulos I. e Zygalakis K. (2018), che analizza le cinquecentotrentotto partite di qualificazione ai Mondiali tramite un modello spaziale per indagare come le caratteristiche di una rete di passaggi quali le misure di centralità e le posizioni degli eventi con la palla determinino diversi stili di gioco. L'obiettivo degli autori è quello di raggruppare le *n* prestazioni di squadra per confrontare i diversi stili di gioco ed individuare quali squadre potrebbero essere le favorite per la vittoria finale della competizione.
Io replicherò la suddetta analisi sulle sessantaquattro partite disputate ai Mondiali di calcio di Russia 2018, per cercare di identificare lo stile di gioco delle squadre che hanno partecipato alla competizione e verificare se le previsioni fatte dall'articolo citato precedentemente si sono effettivamente realizzate. Dato che ogni partita prevede due squadre che si contrappongono, dalle sessantaquattro partite giocate, ottengo centoventotto reti, su ognuna delle quali ho calcolato, per gli undici giocatori titolari, quattro indici:

- *closeness centrality*;
- *betweenness centrality*;
- posizione mediana sul lato x (lato lungo del campo di gioco);
- posizione mediana sul lato y (lato corto del campo di gioco).

Sul dataset così ottenuto, effettuerò un raggruppamento delle prestazioni di squadra tramite un processo di *clustering* con l'obiettivo di identificare, in base ai valori assunti dagli indici considerati, gruppi con caratteristiche simili. Una volta definiti tali gruppi (*cluster*), ne analizzerò peculiarità, differenze e similitudini, fino ad identificare un certo stile di gioco per ognuno di essi.

L'analisi è così strutturata: nel Capitolo 2 faccio un approfondimento sulla teoria dei grafi, le sue proprietà e gli indici che da essi si possono calcolare; nel Capitolo 3 analizzo il caso di studio, calcolando gli indici elencati in precedenza e sviluppando il processo di *clustering*; nel Capitolo 4 definisco lo stile di gioco dei *cluster* individuati; nella sezione 5 traggo le conclusioni, in base ai dati ottenuti; infine, in Appendice, riporto lo script che ho utilizzato in ambito di elaborazione dei dati.
La parte computazionale sui dati, forniti da FootballIntelligence, è eseguita tramite il programma statistico R, mentre i grafici sono creati con Excel.
Ogni elaborazione è supportata dal codice R, per cui si dà la possibilità al lettore di poter replicare fedelmente il lavoro fatto nella mia tesi.

2. Teoria dei grafi

2.1 Storia della teoria dei grafi

In letteratura sono disponibili numerosi metodi per analizzare la tattica delle partite di calcio (Anderson e Sally, 2013). Tra questi, la rappresentazione dello stile di gioco di una squadra come un grafo (o rete) è particolarmente utile, dal momento che la teoria dei grafi offre una traduzione immediata e semplice da interpretare dei dati raccolti e contenuti nelle matrici in concetti e teoremi che possono essere direttamente rapportati ai caratteri essenziali delle reti sociali.

La prima volta che emerge un vero problema riguardante le reti risale al XVIII secolo, più precisamente al 1735, anno in cui il matematico svizzero Leonhard Euler (Eulero) riuscì a risolvere l'enigma adesso conosciuto come il "problema dei sette ponti di Königsberg". L'antica città prussiana di Königsberg (oggi conosciuta come Kaliningrad) è attraversata dal fiume Pregel, al cui interno si trovano due isole, connesse fra di loro e con il resto della città da sette ponti. Per secoli, ci si è chiesto se fosse possibile con una sola passeggiata attraversare tutti e sette i ponti percorrendone ciascuno una ed una sola volta. I cittadini, incapaci di trovare una soluzione, presentarono il problema ad Eulero, il quale capì che la chiave stava nel riuscire a riconoscere che quella che si presentava era in realtà una domanda che riguardava la rete alla base del problema; era innanzitutto necessario identificare e disegnare quella rete. Per prima cosa, eliminò tutti gli elementi contingenti, fatta eccezione per le aree urbane delimitate dalle acque del fiume e dai sette ponti (Figura 1a); in seguito, sostituì le aree urbane con dei punti, detti nodi, e i ponti con delle linee, detti archi (Figura 1b).

Figura 1a.

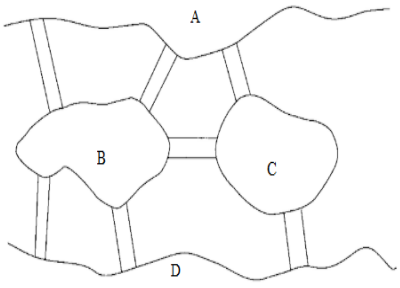
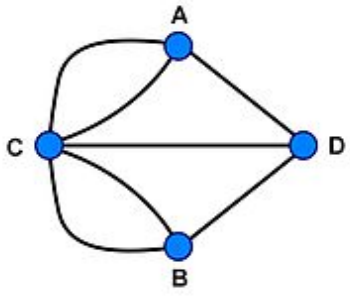


Figura 1b.



Fonte: https://it.wikipedia.org/wiki/Problema_dei_ponti_di_Königsberg

Come osservabile dalla Figura 1b, dai nodi A, B e D partono e arrivano tre ponti, mentre dal nodo C cinque ponti. Questi sono i gradi dei quattro nodi: in ordine 3, 3, 5, 3. Si dice "nodo pari" un nodo di ordine pari, al contrario, si dice "nodo dispari" un nodo di ordine dispari. Eulero giunse alla conclusione che un grafo composto solamente da nodi pari è sempre percorribile e che si può ritornare al punto di partenza senza sovrapposizioni di percorso; se un grafo è composto da nodi pari e soltanto due nodi dispari, è ancora percorribile, a condizione che si parta da uno dei nodi dispari per arrivare all'altro, ma non si può più ritornare al punto di partenza. Se, invece, contiene più di due nodi dispari, non è percorribile senza sovrapposizioni di percorso. Analizzando il grafo dei sette ponti di Königsberg in Figura 1b, è facile osservare come esso presenta ben quattro nodi di ordine dispari: non risulta, quindi, percorribile senza sovrapposizioni.

Nel corso del XIX secolo, è stato enunciato il "problema dei quattro colori", che pone l'interrogativo se, data una superficie piana suddivisa in regioni connesse, siano sufficienti quattro colori per colorare ogni regione in modo che regioni adiacenti non abbiano lo stesso colore. Il problema può essere posto anche in termini di grafi, come segue: è sempre possibile colorare tutti i vertici di un grafo (sia esso finito o infinito) con soli quattro colori, in modo che vertici adiacenti abbiano colori diversi? Nel corso degli anni sono state formulate varie ipotesi per risolvere la questione, ma nessuno è riuscito a trovare la giusta soluzione fino al 1977, quando Kenneth Appel e Wolfgang Haken, due matematici dell'Illinois, riuscirono a formulare la definitiva dimostrazione grazie all'utilizzo di un complesso algoritmo informatico, che si basava sul concetto di grafo planare.

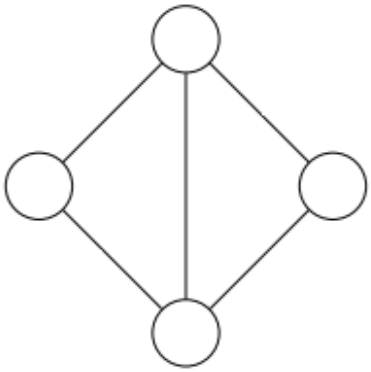
Poche altre teorie sui grafi sono state proposte fino alla seconda metà del XX secolo, quando, invece, sono iniziati numerosi studi, in linea con il forte sviluppo della combinatoria e del calcolo automatico tramite computer. Nel giro di cinquant'anni la teoria dei grafi è oggi un capitolo della matematica molto sviluppato, pieno di risultati e applicazioni interessanti ed estremamente utili in una grande varietà di discipline.

2.2 Definizione di grafo

In termini matematici, è definito grafo una configurazione formata da un insieme di punti (detti nodi) e di linee (detti archi), che uniscono coppie di nodi. Più formalmente, un grafo è descritto come una coppia ordinata $G = (V, E)$ di insiemi, dove V è l'insieme dei nodi (*vertex* in inglese)

ed E è l'insieme degli archi (*edges* in inglese), tali che gli elementi che compongono E siano coppie di elementi di V, ovvero tali che $E \subseteq V \times V$.

Figura 2. *Esempio di grafo.*



Fonte: ns. elaborazione.

2.3 Grafi orientati e grafi non orientati

Esistono varie tipologie di grafi; la principale differenza è quella che li distingue in orientati e non orientati.

È definito grafo orientato un grafo i cui archi hanno un orientamento, ovvero hanno un senso di percorrenza; in tal caso tutti gli archi saranno contrassegnati da frecce, che indicano il senso di percorrenza dal nodo di partenza al nodo di arrivo. Perciò, la coppia di archi (i,j) e (j,i) individua due archi distinti.

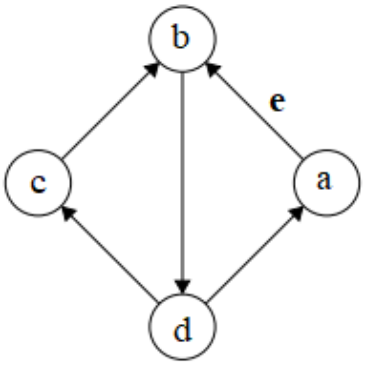
Dato un arco orientato $e = (a,b)$, come raffigurato in Figura 3, sono definiti:

- e arco uscente da a (nodo coda);
- e arco entrante in b (nodo testa);
- a predecessore diretto di b;
- b successore diretto di a.

Dato un nodo *a* di un grafo orientato, come raffigurato in Figura 3, sono definiti:

- δ^+ *(a)* insieme degli archi uscenti da *a*;
- δ^- *(a)* insieme degli archi entranti in *a*.

Figura 3. *Esempio di grafo orientato.*



Fonte: ns. elaborazione.

Al contrario, è definito grafo non orientato un grafo i cui archi non hanno un orientamento, ovvero tali che l'insieme *E* sia composto da coppie non ordinate. In tal caso, la coppia di archi (i,j) e (j,i) identifica il medesimo arco.

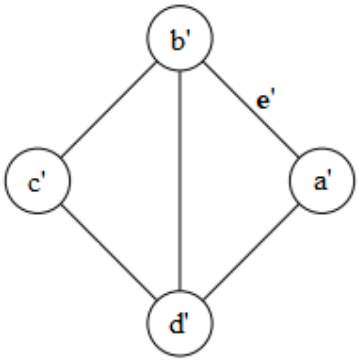
Dato un arco *e′ = (a′,b′)*, come raffigurato in Figura 4, sono definiti:

- *a′*, *b′* estremi dell'arco *e′*;
- *a′*, *b′* nodi adiacenti;
- *e′* arco incidente in *a′* e *b′*.

Dato un nodo *a′* di un grafo non orientato, come raffigurato in Figura 4, sono definiti:

- *N(a′)* intorno di *a′*, ovvero l'insieme di tutti i nodi adiacenti ad *a′*;
- *δ(a′)* stella di *a′*, ovvero l'insieme di tutti gli archi incidenti in *a′*.

Figura 4. *Esempio di grafo non orientato.*

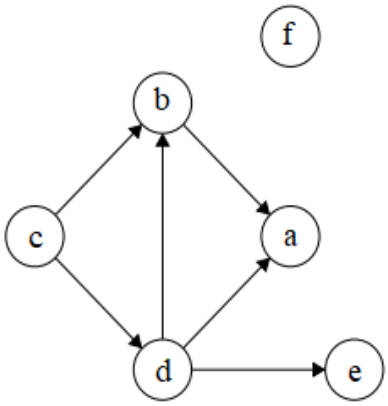


Fonte: ns. elaborazione.

2.4 Ordine di un nodo

È definito ordine di un nodo il numero di archi che terminano sul nodo stesso. Un nodo di ordine zero è detto nodo isolato.

Figura 5. *I nodi a e c sono di ordine 2, il nodo b è di ordine 3, il nodo d è di ordine 4, il nodo e è di ordine 1, mentre il nodo f, essendo di ordine 0, è un nodo isolato.*



Fonte: ns. elaborazione.

2.5 Cammini

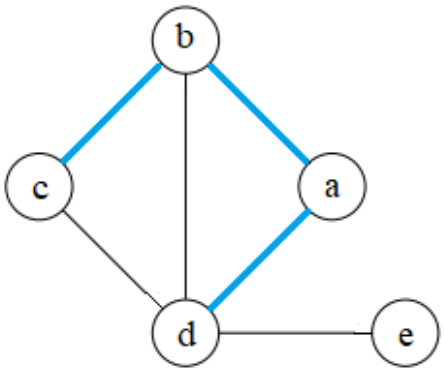
Dato un grafo, è definito cammino una sequenza di archi tali che ogni due archi consecutivi siano adiacenti. Più formalmente, dato un grafo $G = (V, E)$, è detto cammino nel grafo una sequenza di $m+1$ nodi:

$$v_0 \rightarrow v_1 \rightarrow v_2 \rightarrow \ldots \rightarrow v_m$$

tali che per ogni $i = 1, \ldots, m$, si ha che:

$$(v_{i-1}, v_i) \in E.$$

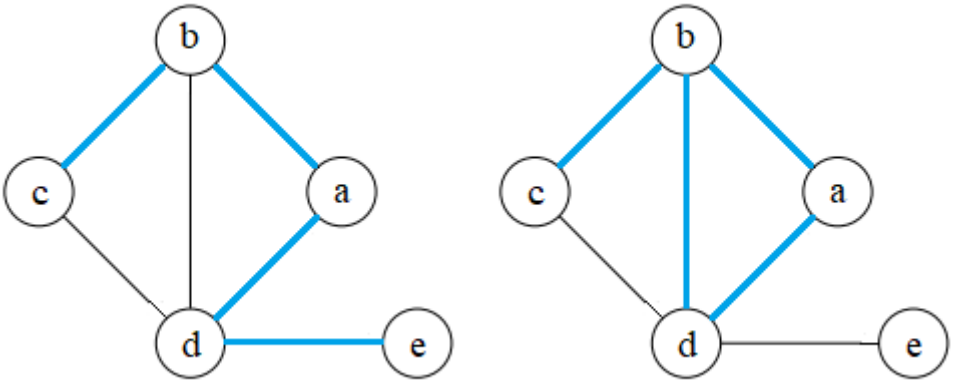
Figura 6. *Esempio di un cammino in un grafo (in azzurro).*



Fonte: ns. elaborazione.

Un cammino è detto semplice se i nodi e gli archi del cammino sono tutti distinti, altrimenti è detto non semplice.

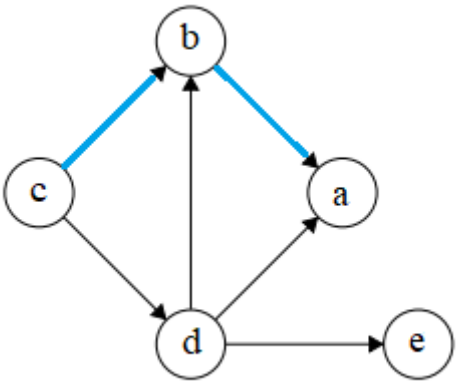
Figura 7. *Esempio di cammino semplice (a sinistra) e di cammino non semplice (a destra).*



Fonte: ns. elaborazione.

Un cammino è detto orientato se per ogni arco $e = (i,j)$ del cammino stesso, il nodo i e il nodo j sono rispettivamente la coda e la testa dell'arco e.

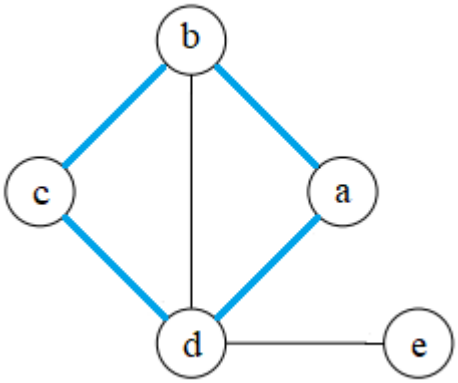
Figura 8. *Esempio di cammino orientato (in azzurro). Il senso di percorrenza è* $c \rightarrow b \rightarrow a$.



Fonte: ns. elaborazione.

Un cammino è detto chiuso se i nodi estremi del cammino stesso coincidono, ovvero se il nodo di partenza e il nodo di arrivo del cammino sono il medesimo.

Figura 9. *Esempio di cammino chiuso.*



Fonte: ns. elaborazione.

2.6 Cicli

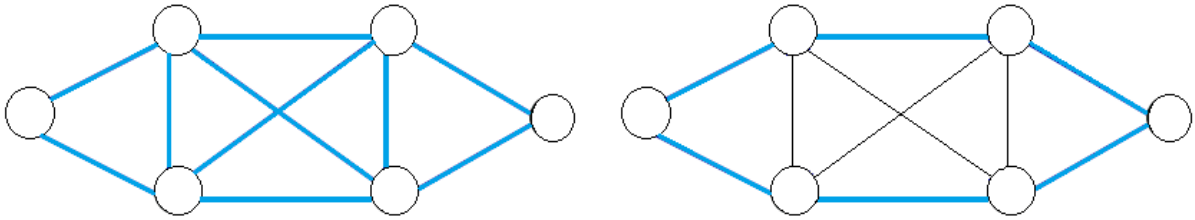
Dato un grafo, è definito ciclo un cammino semplice e chiuso.

Esistono due casi particolari di ciclo, molto significativi all'interno della teoria dei grafi sia per la loro importanza teorica quanto per la loro applicazione pratica: il ciclo euleriano e il ciclo hamiltoniano.

Un ciclo euleriano, che prende il nome dal matematico svizzero Leonhard Euler, è definito come un cammino chiuso e semplice che attraversa ciascun arco del grafo una ed una sola volta. Non sempre è possibile individuare un ciclo euleriano in un grafo; talvolta, infatti, esso può non esistere.

Un ciclo hamiltoniano, che prende il nome dal matematico irlandese William Rowan Hamilton, è definito come un ciclo passante per tutti i nodi del grafo una ed una sola volta. Come nel caso precedente, non è sempre possibile determinare un ciclo hamiltoniano in un grafo.

Figura 10. *Esempio di ciclo euleriano (a sinistra) e di ciclo hamiltoniano (a destra).*



Fonte: ns. elaborazione.

2.7 Network Analysis

La Social Network Analysis (SNA) è un insieme di strumenti finalizzati a descrivere le principali caratteristiche di un grafo. In particolare, considera le relazioni sociali in ottica di network theory, utilizzando algoritmi e metodologie di analisi tipici della teoria dei grafi. Per fare ciò, in ambito di SNA sono impiegati alcuni indici in grado di analizzare una rete sia nel suo complesso come struttura unica (vedi paragrafo 2.8), sia a livello di singoli nodi (vedi paragrafo 2.9).

2.8 Densità

La densità è una delle più importanti statistiche descrittive in ambito di Social Network Analysis, utilizzata come indicatore del livello generale di coesione di un grafo. Perciò, più alto è il valore assunto da tale indice, più un grafo risulta connesso e coeso.

Da un punto di vista matematico, la densità di un grafo indica la proporzione dei legami presenti tra nodi su tutti i legami possibili. In particolare, dato il grafo $G = (V, E)$, siano:

- n la cardinalità dell'insieme V (numero di nodi del grafo G);
- L la cardinalità dell'insieme E (numero di archi del grafo G).

La densità di un grafo non orientato è calcolata come:

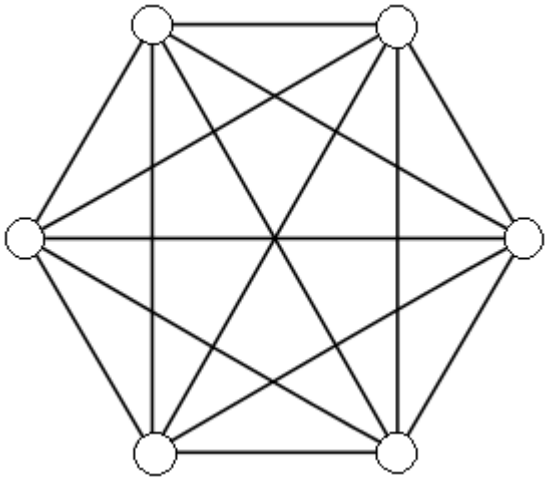
$$\Delta = 2L / n(n-1)$$

La densità di un grafo orientato, invece, è calcolata come:

$$\Delta = L / n(n-1)$$

La densità di un grafo assume valori tra 0 e 1, così che valori prossimi a 0 indicano un basso livello di coesione tra i nodi, mentre valori prossimi a 1 indicano un alto livello di coesione. Un grafo con densità pari a 1, ovvero un grafo in cui tutti i nodi sono collegati agli altri, è definito completo.

Figura 11. *Esempio di grafo completo.*



Fonte: ns. elaborazione.

2.9 Misure di centralità

Fino a questo momento, l'analisi dei grafi è stata effettuata da un punto di vista "macro", con l'obiettivo di descrivere le caratteristiche generali di una rete. Esistono, tuttavia, anche alcune misure di tipo "micro", che permettono di effettuare confronti tra i singoli nodi di una rete. Tra queste, assumono una notevole importanza le misure di centralità.

Il sociologo americano Linton Clarke Freeman propose nel 1979 tre nozioni di centralità :

- *degree centrality*, legata al grado di un nodo;
- *closeness centrality*, legata alla distanza tra nodi;
- *betweenness centrality*, legata ai percorsi che collegano i nodi.

2.9.1 Degree centrality

La prima ad essere stata proposta e allo stesso tempo la più semplice ed intuitiva da utilizzare fra le misure di centralità è la *degree centrality*, che conta il numero di collegamenti che si verificano su un nodo. Nel caso in cui il grafo sia orientato (ovvero i collegamenti fra nodi abbiano una direzione), esistono due misure distinte: *in-degree centrality*, che conteggia il numero di archi entranti in un nodo, e *out-degree centrality*, che conteggia il numero di archi uscenti da un nodo. Formalmente, dato un grafo $G = (V, E)$, con V insieme dei nodi ed E insieme degli archi, la *degree centrality* di un dato nodo i è calcolata come:

$$C_{Di} = d_i / (n-1),$$

dove d_i è il numero di collegamenti del nodo i, ovvero il numero di archi che hanno il nodo i come estremo, e n è il numero totale di nodi presenti nel grafo. Valori bassi indicano una scarsa centralità del nodo all'interno del grafo, mentre valori elevati indicano un'elevata centralità del nodo.

2.9.2 Closeness centrality

La *closeness centrality* misura la prossimità di un nodo agli altri nodi di un grafo. È calcolata come l'inverso della somma della lunghezza dei percorsi più brevi tra un nodo e tutti gli altri. Più precisamente, dato un grafo $G = (V, E)$, con V insieme dei nodi ed E insieme degli archi, la *closeness centrality* di un dato nodo i è calcolata come:

$$C_{Ci} = 1 / \Sigma_{i \neq j} d_{ij},$$

dove d_{ij} è la distanza tra il nodo *i* e il nodo *j*. Normalmente, in ambito scientifico, si fa riferimento alla forma normalizzata, data dalla formula precedente moltiplicata per il fattore *n*-1, dove *n* é il numero totale di nodi presenti nel grafo; per grafi molto estesi, inoltre, è possibile rimuovere il termine -1 dal momento che diventa inconsistente, ottenendo la formula finale:

$C_{Ci} = n / \Sigma_{i \neq j} d_{ij}$.

La *closeness centrality* può essere vista anche come l'efficienza di ogni nodo nella diffusione di "informazioni" a tutti gli altri nodi. Così maggiore è il valore dell'indice assunto da un nodo, più breve è la distanza media dal nodo considerato a qualsiasi altro nodo nel grafo, e quindi migliore la posizione del nodo nel diffondere "informazioni". In conclusione, valori bassi di C_C indicano che un nodo si trova in una posizione periferica del grafo ed è separato dagli altri nodi da distanze rilevanti, mentre valori elevati di C_C indicano che un nodo si trova in una posizione centrale del grafo, a breve distanza dagli altri nodi.

2.9.3 Betweenness centrality

La *betweenness centrality* misura il numero di percorsi che passano per ciascun nodo. Più precisamente, è una misura basata sulla frequenza con cui ogni singolo nodo del grafo si trova nel percorso più breve (detto percorso geodetico) che collega ogni altra coppia di nodi. Per ogni coppia di nodi in un grafo, esiste, infatti, almeno un percorso geodetico tra nodi tale che il numero di archi attraversati dal percorso sia minimizzato. E la *betweenness centrality* per ciascun nodo è proprio il numero di questi percorsi geodetici che attraversano il nodo. Dal punto di vista matematico, dato un grafo $G = (V, E)$, con *V* insieme dei nodi ed *E* insieme degli archi, la *betweenness centrality* di un dato nodo *i* è calcolata come:

$C_{Bi} = \Sigma_{i \neq j \neq k} n^i_{jk} / g_{jk}$,

dove n^i_{jk} è il numero di percorsi geodetici dal nodo *j* al nodo *k* che passano attraverso il nodo *i* e g_{jk}, invece, è il numero totale di percorsi geodetici da *j* a *k*. Dato che il valore assunto dall'indice varia a seconda del numero di coppie di nodi presenti nel grafo, la formula precedente può essere scalata dividendo per il numero di coppie di nodi che non includono il nodo *i*, in modo che $C_B \in [0,1]$. Il termine con cui effettuare questo ridimensionamento è (*n*-1)(*n*-2) per i grafi orientati e (*n*-1)(*n*-2)/2 per i grafi non orientati, dove *n* è il numero di totali di nodi presenti nel grafo.

La *betweenness centrality* può essere interpretata come una misura di quanto un'entità (nodo) sia "intermediaria" tra le altre entità del grafo. Quindi, valori prossimi a 0 indicano che il nodo ricopre un ruolo marginale nei trasferimenti/scambi all'interno del grafo, mentre valori prossimi a 1 indicano che il nodo è uno "snodo" importante all'interno del grafo.

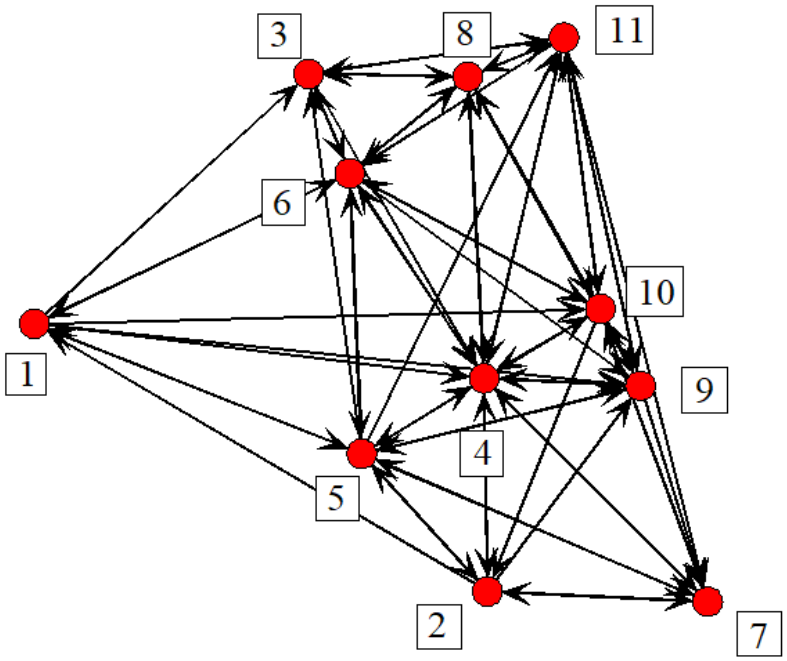
3. Caso di studio

3.1 Premessa

L'oggetto di studio della mia analisi sono i Mondiali di calcio di Russia 2018 e in particolare le sessantaquattro partite che sono state disputate. Dal momento che ogni partita prevede due squadre che si affrontano in campo, è possibile ricavare per ogni incontro due prestazioni di squadra; perciò l'intera competizione disputata in Russia genera un totale di centoventotto prestazioni di squadra, ognuna delle quali può essere rappresentata tramite un grafo ed analizzata secondo i criteri di *network analysis* visti in precedenza.

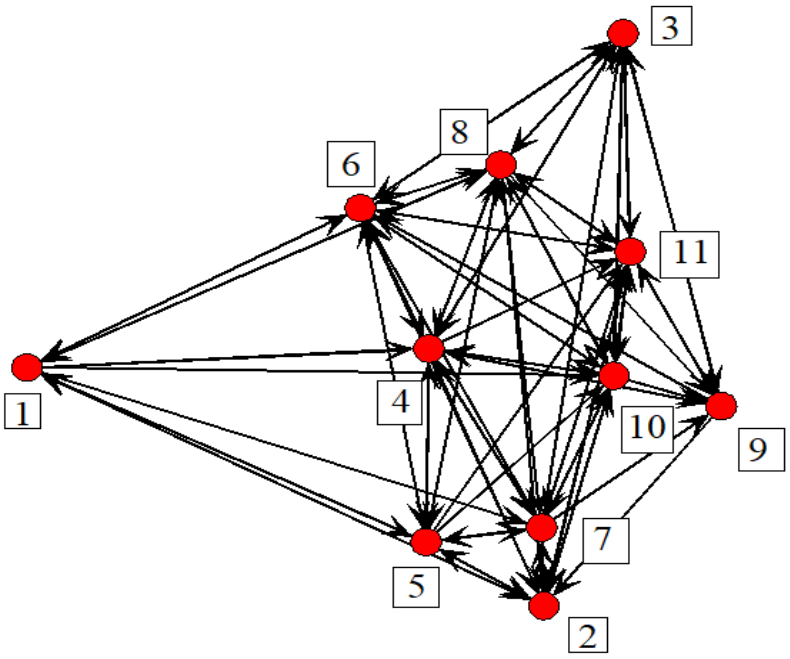
Partendo dalla distribuzione dei passaggi, ogni prestazione di squadra può essere rappresentata con un grafo (detto *passing network*), in cui i nodi corrispondono agli undici giocatori in campo e gli archi ai passaggi che sono stati effettuati con successo fra giocatori. Dato che i passaggi hanno una direzione, il grafo sarà orientato, con archi contrassegnati da frecce indicanti la direzione in cui il passaggio è stato effettuato. Inoltre, posizionando ciascun nodo nella posizione mediana che il giocatore rappresentato ha ricoperto durante la partita, si ottiene un'immagine immediata del posizionamento in campo della squadra e dello stile di gioco che ha utilizzato. Di seguito riporto due esempi di *passing networks*, il primo raffigurante la Francia e il secondo la Croazia, nella partita valevole per la finale della competizione.

Figura 12. *Passing network della Nazionale francese per la partita Francia – Croazia.*



Fonte: ns. elaborazione.

Figura 13. *Passing network della Nazionale croata per la partita Francia – Croazia.*



Fonte: ns. elaborazione.

Dalle centoventotto prestazioni di squadra è possibile, quindi, ricavare centoventotto *passing networks*, su cui effettuerò una *network analysis*. In particolare, per ogni grafo, calcolerò la *closeness centrality* e la *betweenness centrality* per gli undici giocatori titolari, oltre alla loro posizione mediana sul lato lungo del campo (lato x) e la posizione mediana sul lato corto del campo (lato y).

3.2 Indici

L'analisi, in linea con quanto fatto nell'articolo *Predicting FIFA World Cup 2018 key role and playing style features* (2018), prevede l'utilizzo dei seguenti quattro indici, calcolati per gli undici giocatori titolari, per ognuna delle centoventotto prestazioni di squadra, disposte secondo l'ordine cronologico delle partite giocate durante la competizione:

- *closeness centrality*;
- *betweenness centrality*;
- posizione mediana sul lato lungo del campo di gioco (lato x);
- posizione mediana sul lato corto del campo di gioco (lato y).

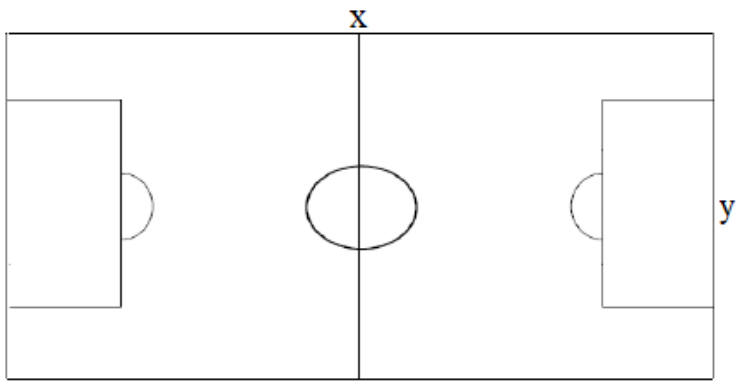
In un contesto sportivo, i due indici di centralità assumono un significato ben definito: la *closeness centrality* fornisce una misura di quanto sia facile raggiungere con la palla un giocatore all'interno di una squadra; valori bassi indicano un'elevata distanza media, ovvero che il giocatore si trova in una posizione del campo isolata rispetto ai propri compagni, mentre valori elevati indicano, al contrario, una piccola distanza media, ovvero che il giocatore è ben collegato all'interno della squadra e facilmente raggiungibile dai propri compagni con un passaggio. In particolare, punteggi elevati di *closeness centrality* sono determinati da passaggi di tipo A-B-A, ovvero da due passaggi consecutivi tra due giocatori (il cosiddetto uno-due), con il giocatore A che passa la palla al giocatore B, il quale, a sua volta, la ripassa al giocatore A.

La *betweenness centrality* in ambito calcistico misura, invece, come il flusso della palla tra gli altri giocatori dipende da un particolare giocatore, fornendo, quindi, una misura di quanto egli sia coinvolto nel gioco della propria squadra. Valori prossimi a 0 indicano che il giocatore non viene coinvolto nel gioco di squadra e che può essere sostituito senza troppe perdite, mentre valori prossimi a 1 indicano che il giocatore è altamente coinvolto e rappresenta un riferimento per il

gioco della squadra, per cui una sua uscita dalla partita (per esempio a causa di un'espulsione o di un infortunio) avrebbe effetti negativi. In particolare, punteggi elevati di *betweenness centrality* sono determinati da passaggi di tipo A-B-C, ovvero da due passaggi consecutivi tra tre giocatori, con il giocatore A che passa la palla al giocatore B, il quale, a sua volta, la passa al giocatore C.

I rimanenti due indici sono calcolati come mediane delle posizioni di tutti gli eventi con la palla nell'arco della partita, ovvero delle posizioni in cui il giocatore si trova ogni volta che ha toccato la palla. Per il terzo indice, in particolare, si fa riferimento alle posizioni sul lato lungo del campo (lato x), mentre per il quarto indice si fa riferimento alle posizioni sul lato corto del campo (lato y).

Figura 14. *Campo da calcio con indicati lato x e lato y.*



Fonte: ns. elaborazione.

3.3 Football Data

I dati che ho a disposizione, forniti da FootballIntelligence, prevedono per ognuna delle sessantaquattro partite disputate ai Mondiali di Russia 2018 quattro dataset. Supponendo una generica partita in cui si affrontano la squadra A e la squadra B, i quattro dataset disponibili per l'incontro A vs B sono:

- matrice dei passaggi della squadra A;
- tabella delle posizioni dei giocatori della squadra A;

- matrice dei passaggi della squadra B;
- tabella delle posizioni dei giocatori della squadra B.

Per matrice dei passaggi si intende una matrice quadrata, sulle cui righe e colonne sono inseriti i nomi dei giocatori e le cui caselle sono completate con il numero di passaggi che questi si sono scambiati. Ragionevolmente, la diagonale sarà composta da caselle tutte pari a 0, dal momento che un passaggio, per essere tale, presuppone un trasferimento della palla tra due giocatori diversi. Inoltre, essendo un passaggio orientato, il valore della generica casella (*i*, *j*) non sarà necessariamente uguale al valore della casella (*j*, *i*): in particolare, le righe della matrice fanno riferimento ai giocatori che effettuano il passaggio, mentre le colonne ai giocatori che ricevono il passaggio. Di seguito un esempio di matrice dei passaggi.

Tabella 1. *Matrice dei passaggi della Nazionale francese per la partita Francia – Croazia nella sua configurazione finale (in seguito spiegherò i passaggi necessari per ottenerla). I giocatori considerati sono gli undici titolari più le tre riserve che sono subentrate, in quanto la matrice conteggia tutti i passaggi che sono stati effettuati nella durata della partita. Da notare la diagonale della matrice composta da tutti 0 e i diversi valori assunti dalle caselle (i, j) e (j, i).*

	GRIEZMANN	VARANE	LLORIS	UMTITI	MATUIDI	GIROUD	KANTE	HERNANDEZ	POGBA	PAVARD	MBAPPE	NZONZI	TOLISSO	FEKIR
GRIEZMANN	0	2	0	1	2	1	0	2	2	1	2	3	3	1
VARANE	2	0	3	4	0	2	1	1	2	3	2	1	0	0
LLORIS	2	1	0	1	0	7	1	0	2	0	0	1	0	2
UMTITI	1	5	2	0	1	1	3	0	3	0	0	1	1	0
MATUIDI	1	0	0	3	0	1	3	3	3	0	0	1	0	0
GIROUD	5	2	0	0	1	0	0	1	2	0	2	1	0	0
KANTE	0	0	0	2	3	0	0	1	1	0	0	0	0	0
HERNANDEZ	1	0	0	1	1	1	2	0	1	0	1	1	1	0
POGBA	2	0	0	2	2	1	0	3	0	5	5	3	0	2
PAVARD	0	3	1	0	0	1	0	0	2	0	3	1	0	0
MBAPPE	1	2	0	0	0	0	0	0	2	2	0	0	0	0
NZONZI	4	1	0	1	1	1	0	2	3	0	0	0	0	1
TOLISSO	2	0	0	0	0	1	0	1	0	0	0	1	0	0
FEKIR	1	0	0	0	0	0	0	0	0	0	0	1	0	0

Fonte: ns. elaborazione.

Per tabella delle posizioni si intende una tabella in cui sono contenute varie informazioni circa la posizione dei giocatori in campo, ogni volta che hanno toccato la palla. Più precisamente, sono presenti cinque colonne (riferite ad altrettante variabili), così disposte:

- Fase: indica se la partita è nei tempi regolamentari (Regular) o nei tempi supplementari (Extra);

- Tempo: indica se la partita è nel primo tempo (1) o nel secondo tempo (2);
- Giocatore: riporta il nome del giocatore che ha toccato la palla;
- X: specifica la posizione del giocatore indicato sul lato lungo del campo (lato x);
- Y: specifica la posizione del giocatore indicato sul lato corto del campo (lato y).

Le variabili X ed Y fanno riferimento ad un campo di dimensioni standardizzate, con valori appartenenti all'intervallo [-1;1]; in seguito spiegherò come scalare questi valori in termini di dimensioni di un campo da gioco di calcio.

La combinazione di queste cinque variabili, rende disponibile per ogni riga della tabella la posizione di un giocatore ogni volta che ha toccato la palla nella durata della sua permanenza in campo in una certa partita. Di seguito un esempio di tabella delle posizioni.

Tabella 2. *Estratto della tabella dei passaggi della Nazionale francese per la partita Francia – Croazia.*

	Fase	Tempo	Giocatore	x	Y
1	Regular	1	MBAPPE	0.060190	-0.948529
2	Regular	1	LLORIS	-0.868381	-0.085294
3	Regular	1	UMTITI	-0.679238	0.305588
4	Regular	1	VARANE	-0.592762	-0.409118
5	Regular	1	UMTITI	-0.495619	0.158529
6	Regular	1	MATUIDI	-0.061524	0.380882
7	Regular	1	UMTITI	-0.350667	0.468235
8	Regular	1	GIROUD	0.181905	0.381176

Fonte: ns. elaborazione.

3.4 Elaborazione dei dati

L'elaborazione dei dati è stata effettuata tramite il software statistico R ed in particolare tramite l'interfaccia RStudio. Di seguito riporterò le varie righe di codice utilizzate e il relativo commento, in modo che chiunque, con i dati necessari disponibili, possa replicare l'analisi da me svolta.

Il primo passo è stato organizzare i vari dataset disponibili in sessantaquattro cartelle, numerate dalla 1 alla 64 e riferite alle altrettante partite giocate durante l'intera competizione, ordinate in

base al calendario ufficiale di gara, così che la cartella 1) indichi la prima partita disputata (Arabia Saudita – Russia) e la cartella 64) l'ultima partita disputata, la finale (Francia – Croazia). Ognuna di queste sessantaquattro cartelle contiene i quattro dataset descritti precedentemente.

3.4.1 Matrice dei passaggi

In questa sezione analizzerò i comandi che ho utilizzato per aprire in R la matrice dei passaggi e trasformarla nella sua forma finale.

```
> ##### GIORNATA 1 #####
> # 1) KSA-RUS
> # 1_A) KSA
```

Le prime tre righe di codice sono un commento, come si può notare dal cancelletto (#) iniziale, e servono per organizzare il lavoro in R, indicando rispettivamente la fase del torneo, la partita e la squadra analizzata.

```
> library(readr)
```

Carica la library *readr*, necessaria per leggere in R dataset rettangolari, come la matrice dei passaggi.

```
> X1_KSA <- read_csv("D:/TESI/DATASET/Partite_Mori/1)KSA-RUS/matriceKSA.csv")
```

Apre in R la matrice dei passaggi, file in formato csv. Quello contenuto fra virgolette è il path (percorso) del file nel mio computer.

```
> nomi_1_A=X1_KSA[,1]
> nomi_1_A=unlist(nomi_1_A)
```

Il primo codice estrae la prima colonna della matrice dei passaggi, contenente i nomi di tutti i giocatori che hanno giocato almeno un minuto nella partita analizzata, creando un nuovo oggetto in R. Il secondo codice trasforma questo oggetto in un vettore.

```
> rownames(X1_KSA)<-nomi_1_A
```

Rinomina le righe della matrice dei passaggi con gli elementi del vettore creato in precedenza, contente i nomi dei giocatori.

```
> X1_KSA[,1]=NULL
```

Elimina la prima colonna dalla matrice dei passaggi. Così facendo, la colonna contenente i nomi dei giocatori non farà più parte della matrice stessa, ma risulterà esterna come etichetta delle unità di studio contenute nelle righe.

```
> X1_KSA=as.matrix(X1_KSA)
```

Trasforma la matrice dei passaggi, che fino a questo punto è stata letta in R come un data frame, in una matrice, rendendo possibile successivamente il calcolo degli indici.

3.4.2 Closeness centrality e betweenness centrality

In R esistono delle funzioni che permettono di effettuare la *network analysis* e calcolare i relativi indici direttamente dalla matrice dei passaggi, senza dover necessariamente trasformare i dati in un grafo. Questi indici fanno comunque riferimento ai grafi e sono calcolati da R come se i dati fossero già stati trasformati in grafi, ma tali funzioni rendono il procedimento molto più snello ed immediato.

I comandi seguenti sono utilizzati per il calcolo della *closeness centrality* e della *betweenness centrality* e per creare una tabella contenente i nomi dei giocatori, in ordine alfabetico, e i valori dei due indici.

```
> library(igraph)
> library(sna)
```

Questi due comandi servono per caricare due library, *igraph* e *sna*, necessarie per effettuare la *network analysis* e calcolare gli indici desiderati.

```
> clos_1_A=closeness(X1_KSA)
```

Calcola la *closeness centrality*, creando un vettore contenente i valori dell'indice per tutti i giocatori della squadra.

```
> betw_1_A=betweenness(X1_KSA)
> betw_rel_1_A=betweenness(X1_KSA)/((nrow(X1_KSA)-1)*(nrow(X1_KSA)-2))
```

Il primo comando calcola la *betweenness centrality* nella sua forma base, creando un vettore contenente i valori dell'indice per tutti i giocatori della squadra. Il secondo calcola la *betweenness centrality* riscalata, ottenuta dividendo l'indice calcolato precedentemente per (*n*-1)(*n*-2), creando un nuovo vettore contenente i valori dell'indice per tutti i giocatori della squadra. Da qui in poi, ogni volta che farò riferimento alla *betweenness centrality*, considererò quest'ultima forma riscalata.

```
> TAB_1_A=cbind(nomi_1_A,clos_1_A,betw_rel_1_A)
```

Crea una tabella con tre colonne, contenenti rispettivamente i nomi dei giocatori, i valori della *closeness centrality* e i valori della *betweenness centrality*.

```
> TAB_1_A=as.data.frame(TAB_1_A)
> TAB_1_A<-TAB_1_A[order(TAB_1_A$nomi_1_A),]
```

Il primo comando trasforma la tabella appena creata in un data frame, per poter poi, con il secondo comando, ordinarlo in base ai nomi dei giocatori in ordine alfabetico.

```
> nomi_ord_1_A=TAB_1_A[,1]
> nomi_ord_1_A=unlist(nomi_ord_1_A)
> rownames(TAB_1_A)<-nomi_ord_1_A
> TAB_1_A[,1]=NULL
```

Questi comandi servono per effettuare il procedimento visto in precedenza per la matrice dei passaggi, ovvero:

- estrarre la prima colonna del data frame contenente i nomi dei giocatori, ordinati alfabeticamente, creando un nuovo oggetto in R;
- trasformare questo oggetto contenente i nomi in un vettore;
- rinominare le righe del data frame con i nomi dei giocatori, contenuti nel vettore creato;
- eliminare la prima colonna del data frame.

Di seguito il data frame ottenuto dopo aver applicato questi comandi in R.

Tabella 3. *Esempio di data frame contenente i valori della closeness centrality e della betweenness centrality dei giocatori (disposti in ordine alfabetico). Da notare come i nomi dei giocatori non siano una colonna del data frame, ma siano i nomi delle unità statistiche su cui sono calcolati i due indici.*

	clos_1_A	betw_rel_1_A
ABDULLAH ALMUAIOUF	0	0.0127289377289377
ABDULLAH OTAYF	0	0.01128663003663
FAHAD ALMUWALLAD	0	0
HATAN BAHBRI	0	0
MOHAMMED ALBURAYK	0	0.0469551282051282
MOHAMMED ALSAHLAWI	0	0.00657051282051282
MUHANNAD ASIRI	0	0
OMAR HAWSAWI	0	0.02372557997558
OSAMA HAWSAWI	0	0.0260836385836386
SALEM ALDAWSARI	0	0.0617673992673993
SALMAN ALFARAJ	0	0.12912851037851
TAISEER ALJASSAM	0	0.0906669719169719
YAHIA ALSHEHRI	0	0.00567002442002442
YASSER ALSHAHRANI	0	0.0149038461538462

Fonte: ns. elaborazione.

3.4.3 Tabella delle posizioni

In questa sezione analizzerò i comandi che ho utilizzato per aprire in R la tabella delle posizioni dei giocatori, calcolare la loro mediana per ogni unità statistica e creare una nuova tabella contenente i nomi dei giocatori e la loro posizione mediana nella partita sul lato x e sul lato y.

```
> Y1_KSA <- read_csv("D:/TESI/DATASET/Partite_Mori/1)KSA-RUS/posKSA.csv")
> attach(Y1_KSA)
```

Apre in R la tabella delle posizioni, file in formato csv. Quello contenuto fra virgolette è il path (percorso) del file nel mio computer. Il comando `attach` in R serve per accedere agli oggetti presenti in un certo database semplicemente richiamando i loro nomi.

```
> Y1_KSA$pos_x_1_A=ifelse(Tempo==1,x,-x)
> Y1_KSA$pos_y_1_A=ifelse(Tempo==1,Y,-Y)
> attach(Y1_KSA)
```

Le posizioni in campo dei giocatori fanno riferimento all'intera durata della partita, senza tener conto del fatto che alla fine di ogni tempo le squadre invertono la metà campo in cui si dispongono. I primi due comandi servono proprio per far sì che le posizioni dei giocatori (sia sul lato x che sul lato y) siano sempre riferite ad una sola metà campo, come se una squadra attaccasse sempre verso la stessa porta per tutta la partita. Allo stesso tempo, però, le posizioni dei giocatori della squadra avversaria (squadra B) nella partita di riferimento risulteranno invertiti, dal momento che questa attacca sempre verso la porta opposta a quella in cui attacca la squadra A. È, quindi, necessario, per la squadra B, non solo effettuare il procedimento usato per la squadra A visto in precedenza, ma anche invertire i dati relativi alle posizioni, in modo da rendere uniformi tutte le posizioni dei giocatori delle due squadre. Il codice in R sottostante serve proprio per ottenere ciò:

```
> Y1_RUS$pos_x_1_B=ifelse(V2==1,-V4,V4)
> Y1_RUS$pos_y_1_B=ifelse(V2==1,-V5,V5)
> attach(Y1_RUS)
```

In questo caso la variabile ‘V2’ corrisponde alla precedente variabile ‘Tempo’, ‘V4’ a ‘X’ e ‘V5’ a ‘Y’. Nel caso della squadra A, quando la variabile ‘Tempo’ è uguale a 1, ovvero quando la partita si trova nel primo tempo, le posizioni sul lato x e sul lato y rimangono invariate, mentre quando la variabile ‘Tempo’ è uguale a 2, le posizioni sul lato x e sul lato y vengono invertite, cambiando il loro segno. Nel caso della squadra B, invece, quando la variabile ‘V2’ (‘Tempo’) è uguale a 1, le posizioni sul lato x e sul lato y vengono invertite, cambiando il loro segno, mentre quando la variabile ‘V2’ (‘Tempo’) è uguale a 2, le posizioni sul lato x e sul lato y rimangono invariate.

```
> POS_1_A=cbind(Giocatore,pos_x_1_A,pos_y_1_A)
```

Crea una tabella con tre colonne, corrispondenti ad altrettante variabili: la prima contiene i nomi dei giocatori ogni volta che hanno toccato la palla, la seconda e la terza le posizioni sul lato x e sul lato y, corrette con il procedimento visto in precedenza. Su questa tabella sarà ora possibile calcolare le mediane per ogni giocatore.

```
> pos_med_x_1_A=tapply(pos_x_1_A,Giocatore,median)
> pos_med_y_1_A=tapply(pos_y_1_A,Giocatore,median)
```

Il primo comando calcola la mediana delle posizioni sul lato lungo del campo per ogni giocatore che ha partecipato alla partita; il secondo comando ripete l’operazione per le posizioni sul lato corto del campo.

```
> pos_med_1_A=cbind(pos_med_x_1_A,pos_med_y_1_A)
```

Crea un nuovo data frame con i valori delle posizioni mediane sul lato x e sul lato y per ogni giocatore. I nomi dei giocatori, in ordine alfabetico, sono anche in questo caso esterni al data frame, in funzione di nomi delle unità statistiche su cui sono calcolate le due mediane. Di seguito il data frame ottenuto.

Tabella 4. *Esempio di data frame contenente i valori delle posizioni mediane sul lato x e sul lato y dei giocatori (in ordine alfabetico). Da notare come i nomi dei giocatori non siano una colonna del data frame, ma siano i nomi delle unità statistiche su cui sono calcolati i due indici.*

	pos_med_x_1_Â	pos_med_y_1_Â
ABDULLAH ALMUAIOUF	-0.800762	-0.0347060
ABDULLAH OTAYF	-0.115619	0.1135290
FAHAD ALMUWALLAD	0.002095	-0.0173530
HATAN BAHBRI	0.250667	-0.7983825
MOHAMMED ALBURAYK	0.116000	-0.9363235
MOHAMMED ALSAHLAWI	-0.032952	-0.2276470
MUHANNAD ASIRI	0.006095	-0.0158820
OMAR HAWSAWI	-0.336571	0.3850000
OSAMA HAWSAWI	-0.287238	-0.4900000
SALEM ALDAWSARI	0.159619	0.6654410
SALMAN ALFARAJ	-0.064286	0.2964710
TAISEER ALJASSAM	-0.101143	-0.3417650
YAHIA ALSHEHRI	0.134286	-0.7626470
YASSER ALSHAHRANI	0.028762	0.8814710

Fonte: ns. elaborazione.

3.4.4 Vettore finale

Obiettivo di questa sezione è spiegare quali comandi utilizzare per creare il vettore finale, contenente, per i soli undici giocatori titolari (ordinati per ruolo), i valori dei quattro indici. Per ognuna delle centoventotto prestazioni di squadra, otterrò, quindi, un totale di undici osservazioni per ognuno dei quattro indici, per un totale di quarantaquattro osservazioni per prestazione di squadra.

```
> DATI_1_A=cbind(TAB_1_A,pos_med_1_A)
```

Crea un data frame che associa ad ognuno dei giocatori, che ha giocato almeno un minuto nella partita, i valori dei quattro indici: *closeness centrality*, *betweenness centrality*, posizione mediana sul lato x e posizione mediana sul lato y. Le posizioni, fino a questo momento, fanno ancora riferimento al campo di dimensioni standardizzate, con valori, sia per il lato x che per il lato y, appartenenti all'intervallo [-1;1].

```
> for (i in 1:dim(DATI_1_A)[1]){
+   DATI_1_A[i,3]=(DATI_1_A[i,3]+1)*108/2
+   DATI_1_A[i,4]=(DATI_1_A[i,4]+1)*68/2
+ }
```

È un ciclo *for*, ovvero un'iterazione programmata di una porzione di codice, utilizzata per modificare le dimensioni del campo a cui sono riferite le posizioni dei giocatori. Più precisamente, tramite questo comando, viene considerato un campo di dimensioni regolamentari, con lato lungo (lato x) di lunghezza 108 metri e lato corto (lato y) di lunghezza 68 metri. Così , l'indice relativo alla posizione mediana del giocatore sul lato x assumerà valori nell'intervallo [0;108], mentre l'indice relativo alla posizione mediana del giocatore sul lato y assumerà valori nell'intervallo [0;68]. Il data frame creato al passaggio precedente assume adesso, dopo aver utilizzato il ciclo *for*, questa configurazione.

Tabella 5. *Esempio di data frame contenente i valori dei quattro indici per i giocatori (in ordine alfabetico) che hanno giocato la partita.*

	clos_1_A	betw_rel_1_A	pos_med_x_1_A	pos_med_y_1_A
ABDULLAH ALMUAIOUF	0	0.0127289377289377	10.75885	32.819996
ABDULLAH OTAYF	0	0.01128663003663	47.75657	37.859986
FAHAD ALMUWALLAD	0	0	54.11313	33.409998
HATAN BAHBRI	0	0	67.53602	6.854995
MOHAMMED ALBURAYK	0	0.0469551282051282	60.26400	2.165001
MOHAMMED ALSAHLAWI	0	0.00657051282051282	52.22059	26.260002
MUHANNAD ASIRI	0	0	54.32913	33.460012
OMAR HAWSAWI	0	0.02372557997558	35.82517	47.090000
OSAMA HAWSAWI	0	0.0260836385836386	38.48915	17.340000
SALEM ALDAWSARI	0	0.0617673992673993	62.61943	56.624994
SALMAN ALFARAJ	0	0.12912851037851	50.52856	44.080014
TAISEER ALJASSAM	0	0.0906669719169719	48.53828	22.379990
YAHIA ALSHEHRI	0	0.00567002442002442	61.25144	8.070002
YASSER ALSHAHRANI	0	0.0149038461538462	55.55315	63.970014

Fonte: ns. elaborazione.

Fino ad adesso, ho lavorato e calcolato gli indici per tutti i giocatori che hanno giocato almeno un minuto nella partita in analisi, ovvero gli undici giocatori titolari più le riserve che sono subentrate a gara in corso. Nel corso dell'intera competizione, le squadre hanno utilizzato un

numero diverso di giocatori, da un minimo di undici (senza effettuare nessuna sostituzione), fino ad un massimo di quindici (è prevista una quarta sostituzione nei tempi supplementari, oltre le tre concesse nei tempi regolamentari). Questo rende impossibile effettuare un confronto tra le varie prestazioni di squadra: è, quindi, necessario considerare un numero uguale di giocatori per ogni prestazione, così da rendere possibile il confronto. Nella mia analisi, ho considerato gli undici giocatori titolari in ognuna delle centoventotto prestazioni di squadra, dal momento che questi sono i giocatori che hanno avuto un maggior minutaggio nell'arco della partita e, di conseguenza, maggiori probabilità di incidere sul risultato e sullo stile di gioco della propria squadra. L'eliminazione dalla tabella precedente dei giocatori che sono subentrati (vedi Tabella 5) è stata fatta manualmente, andando ad escludere dal data frame le osservazioni relative ai giocatori subentrati.

```
> DATI_1_A=DATI_1_A[-c(3,4,7),]
```

Elimina dal data frame creato in precedenza le osservazioni relative ai giocatori che sono entrati in campo a partita in corso, creando un nuovo data frame con i valori dei quattro indici per i soli undici giocatori titolari. In questa prestazione di squadra, le osservazioni relative ai giocatori subentrati a partita in corso, e quindi da eliminare, sono la terza, la quarta e la settima; ovviamente, questo non sarà valido per tutte le altre centoventisette prestazioni di squadra, dove le osservazioni da eliminare potranno essere disposte in ordine diverso.

Per rendere le centoventotto prestazioni di squadra fra loro confrontabili, oltre a svolgere l'analisi su uno stesso numero di giocatori, è necessario, inoltre, identificare i ruoli che i vari giocatori in campo ricoprono in una certa partita. Infatti, solo confrontando un giocatore con il suo pari ruolo, l'analisi risulterà corretta. Confrontare, per esempio, la posizione mediana del portiere di una squadra con la posizione mediana dell'attaccante di un'altra squadra ha, naturalmente, poca utilità e porterebbe a risultati difficili da analizzare.

È importante, quindi, riuscire a capire con quale modulo gioca una squadra ed identificare i ruoli degli undici giocatori in campo. Per fare ciò, ho assegnato manualmente ad ogni ruolo (in base alla posizione ricoperta in campo) un numero, con il seguente ordine:

- 1: portiere;
- 2: terzino destro;
- 3: terzino sinistro;

- 4: centrocampista centrale o difensore centrale (in una difesa a tre);
- 5: difensore destro;
- 6: difensore sinistro;
- 7: ala destra;
- 8: centrocampista sinistro;
- 9: attaccante;
- 10: centrocampista destro o trequartista;
- 11: ala sinistra.

```
> num_1_A=c(1,4,2,9,6,5,11,8,10,7,3)
```

Crea un vettore numerico, assegnando un numero da 1 a 11 ai giocatori titolari, disposti in ordine alfabetico.

```
> DATI_1_A=cbind(num_1_A,DATI_1_A)
```

Aggiunge al data frame precedente, contenente i valori dei quattro indici, il vettore numerico appena creato.

```
> DATI_1_A<-DATI_1_A[order(DATI_1_A$num_1_A),]
```

Ordina le osservazioni del data frame precedente in base al vettore numerico con ordine crescente, ovvero dal numero 1 al numero 11. Di seguito, il data frame così ottenuto.

Tabella 6. *Esempio di data frame contenente i numeri identificativi dei ruoli ordinati e i valori dei quattro indici per i giocatori.*

	num_1_A	clos_1_A	betw_rel_1_A	pos_med_x_1_A	pos_med_y_1_A
ABDULLAH ALMUAIOUF	1	0	0.0127289377289377	10.75885	32.819996
MOHAMMED ALBURAYK	2	0	0.0469551282051282	60.26400	2.165001
YASSER ALSHAHRANI	3	0	0.0149038461538462	55.55315	63.970014
ABDULLAH OTAYF	4	0	0.01128663003663	47.75657	37.859986
OSAMA HAWSAWI	5	0	0.0260836385836386	38.48915	17.340000
OMAR HAWSAWI	6	0	0.02372557997558	35.82517	47.090000
YAHIA ALSHEHRI	7	0	0.00567002442002442	61.25144	8.070002
SALMAN ALFARAJ	8	0	0.12912851037851	50.52856	44.080014
MOHAMMED ALSAHLAWI	9	0	0.00657051282051282	52.22059	26.260002
TAISEER ALJASSAM	10	0	0.0906669719169719	48.53828	22.379990
SALEM ALDAWSARI	11	0	0.0617673992673993	62.61943	56.624994

Fonte: ns. elaborazione.

```
> KSA1=DATI_1_A[,c(2,3,4,5)]
```

Crea un nuovo data frame, rinominato con la sigla della squadra a cui fa riferimento e il turno della competizione in cui la partita in analisi è stata giocata, selezionando dal precedente le sole quattro colonne relative agli indici e tralasciando quella relativa ai numeri, essendo le osservazioni sui giocatori già ordinate dall'1 all'11.

Per effettuare, successivamente, il *clustering* delle prestazioni di squadra, è necessario avere i dati disposti in una singola matrice di dimensioni 128 x 44, dove centoventotto sono le unità statistiche, ovvero le prestazioni di squadra dell'intera competizione, e quarantaquattro sono i valori dei quattro indici per gli undici giocatori titolari. È importante in questa fase, quindi, riuscire a disporre tutte e quarantaquattro le osservazioni di una singola prestazione in un vettore di dimensioni 1 x 44, in modo che, unendo verticalmente i centoventotto vettori così creati, si ottenga la matrice desiderata.

```
> KSA1=as.matrix(KSA1)
> KSA1=as.vector(KSA1)
```

Questi due comandi trasformano il data frame precedente prima in una matrice (passaggio necessario) e poi in un vettore. Il vettore così creato avrà dimensione 1 x 44, proprio come desiderato, e conterrà i valori dei quattro indici per gli undici giocatori titolari in quella partita, così disposti: *closeness centrality* dal numero 1 al numero 11 (C1, …, C11), *betweenness centrality* dal numero 1 al numero 11 (B1, …, B11), posizione mediana sul lato x dal numero 1 al numero 11 (X1, …, X11) e posizione mediana sul lato y dal numero 1 al numero 11 (Y1, …, Y11).

Tutta l'analisi appena descritta viene replicata in modo identico sulle rimanenti centoventisette unità di studio (prestazioni di squadra), con l'unica differenza, vista in precedenza, in fase di conversione delle posizioni in campo, con un procedimento distinto fra squadra A e squadra B. Alla fine, si avranno centoventotto vettori, contenenti, per gli undici giocatori titolari, i valori dei quattro indici: *closeness centrality*, *betweenness centrality*, posizione mediana sul lato x e posizione mediana sul lato y. Ogni vettore è identificato con la sigla della squadra a cui fa riferimento e il turno della competizione in cui la partita è stata giocata. Di seguito, la lista delle sigle delle squadre e i numeri che identificano i vari turni della competizione.

Sigle delle trentadue squadre partecipanti ai Mondiali di calcio di Russia 2018:

- ARG: Argentina
- AUS: Australia
- BEL: Belgio
- BRA: Brasile
- COL: Colombia
- CRC: Costarica
- CRO: Croazia
- DEN: Danimarca
- EGY: Egitto
- ENG: Inghilterra
- ESP: Spagna
- FRA: Francia
- GER: Germania
- JPN: Giappone
- KOR: Corea del sud
- KSA: Arabia Saudita
- ISL: Islanda
- IRN: Iran
- MAR: Marocco
- MEX: Messico
- NGA: Nigeria
- PAN: Panama
- PER: Perù
- POL: Polonia
- POR: Portogallo
- RUS: Russia
- SEN: Senegal
- SRB: Serbia
- SUI: Svizzera
- SWE: Svezia
- TUN: Tunisia
- URU: Uruguay

Numeri che identificano i turni dei Mondiali di calcio di Russia 2018:

- 1: Giornata 1
- 2: Giornata 2
- 3: Giornata 3
- 4: Ottavi di finale
- 5: Quarti di finale
- 6: Semifinali
- 7: Finali(1°/2° posto e 3°/4° posto)

3.4.5 Data frame finale

Questa sezione è finalizzata a spiegare i comandi utilizzati per ottenere la matrice di dimensioni 128 x 44 descritta in precedenza e renderla ottimizzata per poi effettuare il *clustering*. Prima di questo, effettuerò un'approssimazione alla terza cifra decimale dei valori assunti dagli indici.

```
> KSA1=as.numeric(KSA1)
> KSA1=round(KSA1, digits=3)
```

Approssima i quarantaquattro termini del vettore alla terza cifra decimale; per fare ciò, è prima necessario convertire la natura del vettore in numerica e questa operazione è effettuata tramite il primo comando.

Questo procedimento è ripetuto in modo identico anche per gli altri centoventisette vettori, così da avere tutti i dati numerici approssimati allo stesso modo e renderli, quindi, più facilmente confrontabili.

```
> DATI=rbind(KSA1,RUS1,EGY1,URU1,MAR1,IRN1,ESP1,POR1,AUS1,FRA1,
+            ARG1,ISL1,PER1,DEN1,CRO1,NGA1,SRB1,CRC1,GER1,MEX1,
+            BRA1,SUI1,SWE1,KOR1,BEL1,PAN1,TUN1,ENG1,JPN1,COL1,
+            POL1,SEN1,RUS2,EGY2,MAR2,POR2,URU2,KSA2,IRN2,ESP2,
+            DEN2,AUS2,FRA2,PER2,ARG2,CRO2,BRA2,CRC2,NGA2,ISL2,
+            SRB2,SUI2,BEL2,TUN2,KOR2,MEX2,GER2,SWE2,PAN2,ENG2,
+            SEN2,JPN2,COL2,POL2,KSA3,EGY3,URU3,RUS3,IRN3,POR3,
+            ESP3,MAR3,AUS3,PER3,DEN3,FRA3,NGA3,ARG3,ISL3,CRO3,
+            SWE3,MEX3,KOR3,GER3,CRC3,SUI3,SRB3,BRA3,COL3,SEN3,
+            JPN3,POL3,ENG3,BEL3,PAN3,TUN3,FRA4,ARG4,URU4,POR4,
+            ESP4,RUS4,CRO4,DEN4,BRA4,MEX4,BEL4,JPN4,SWE4,SUI4,
```

```
+             COL4,ENG4,URU5,FRA5,BRA5,BEL5,SWE5,ENG5,RUS5,CRO5,
+             FRA6,BEL6,CRO6,ENG6,BEL7,ENG7,FRA7,CRO7)
```

Crea una matrice, unendo, uno dopo l'altro, i centoventotto vettori ottenuti precedentemente. La funzione `rbind` è utilizzata in R per unire verticalmente due o più oggetti: per fare ciò, è necessario che questi abbiano le stesse variabili.

```
> nomi_var=c("C1","C2","C3","C4","C5","C6","C7","C8","C9","C10","C11",
+            "B1","B2","B3","B4","B5","B6","B7","B8","B9","B10","B11",
+            "X1","X2","X3","X4","X5","X6","X7","X8","X9","X10","X11",
+            "Y1","Y2","Y3","Y4","Y5","Y6","Y7","Y8","Y9","Y10","Y11")
> colnames(DATI)<-nomi_var
```

Il primo comando crea un oggetto in R contenente quarantaquattro codici, riferiti ai quattro indici per gli undici giocatori titolari. Il secondo comando rinomina le quarantaquattro variabili della matrice appena creata con i codici contenuti fra virgolette nel vettore.

```
> DATI=as.data.frame(DATI)
```

Trasforma la matrice in un data frame, passaggio necessario per poter effettuare successivamente il *clustering*. Di seguito, il data frame finale ottenuto.

Tabella 7. *Estratto del data frame finale, su cui effettuare il clustering.*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
KSA1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RUS1	0.722	0.722	0.765	0.812	0.722	0.765	0.684	0.812	0.684	0.619	0.765
EGY1	0.591	0.765	0.812	0.812	0.684	0.812	0.684	0.684	0.520	0.929	0.722
URU1	0.722	0.867	0.812	0.929	0.650	0.765	0.650	0.929	0.722	0.765	0.722
MAR1	0.722	0.684	0.650	0.765	0.929	0.765	0.867	0.722	0.448	0.765	0.684
IRN1	0.565	0.591	0.565	0.765	0.722	0.591	0.684	0.565	0.500	0.684	0.565
ESP1	0.650	0.684	0.765	0.812	1.000	0.867	0.867	0.765	0.619	0.765	0.812
POR1	0.684	0.722	0.867	0.812	0.765	0.765	0.684	0.867	0.765	0.684	0.650
AUS1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
FRA1	0.591	0.722	0.812	0.765	0.765	0.929	0.765	0.929	0.684	0.812	0.722
ARG1	0.591	0.765	0.765	0.929	0.812	0.867	0.812	0.812	0.812	0.929	0.684
ISL1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Fonte: ns. elaborazione.

3.4.6 Clustering

In questa ultima sezione analizzerò il processo di *clustering*, ovvero un insieme di tecniche di analisi multivariata dei dati finalizzate al raggruppamento di elementi omogenei in due o più gruppi. L'obiettivo è quello di individuare *cluster* (gruppi), che raccolgano prestazioni di squadra con caratteristiche simili e che identifichino uno stile di gioco ben preciso.

La tecnica da me utilizzata, in linea con quanto fatto nell'articolo *Predicting FIFA World Cup 2018 key role and playing style features* (2018), consiste in una procedura di *clustering model-based* per la riduzione della dimensionalità, che va ad identificare un insieme di combinazioni lineari delle caratteristiche originali, dette direzioni, che catturano la maggior parte della struttura di *clustering* contenuta nei dati. Di default, le informazioni sulla riduzione di dimensionalità sono fornite sia dalle variazioni delle medie del *cluster* sia, in base al modello stimato, dalle variazioni delle covarianze del *cluster*. In tal senso, è importante il valore assegnato al parametro *lambda*, che può variare nell'intervallo [0;1]: di default è impostato un valore pari a 0.5, che dà uguale importanza alle differenze nelle medie e nelle covarianze tra *cluster*. Per riconoscere quali direzioni distinguono maggiormente un *cluster* dall'altro, si imposta *lambda* = 1: questo valore del parametro fa sì che sia utilizzata solo l'informazione sulle medie dei *cluster* per la stima delle direzioni.

Le osservazioni, ovvero le centoventotto prestazioni di squadra nella mia analisi, possono quindi essere proiettate su un sottospazio così ridotto, fornendo grafici riassuntivi che aiutano a visualizzare, in modo estremamente immediato ed intuitivo, la suddivisione in gruppi. Tale sottospazio avrà dimensione pari a $d = min(p, G-1)$, dove p è il numero di variabili presenti nella matrice di riferimento su cui applicare il processo di *clustering* e G il numero di *cluster* generati. Le direzioni che attraversano questo sottospazio sono tali che riescono a catturare la maggior parte delle informazioni di *clustering* disponibili nei dati.

```
> library(mclust)
```

Carica in R la library *mclust*, contenente vari modelli di mistura gaussiana per la stima del clustering, della classificazione e della densità basata su modelli, inclusa la regolarizzazione bayesiana e la riduzione della dimensionalità.

```
> BIC <- mclustBIC(DATI, prior=priorControl(shrinkage=0.01))
> mod1 <- Mclust((DATI),G=3:9, x=BIC)
> mod1dr <- MclustDR(mod1, lambda=1)
```

Sono i comandi necessari per effettuare il *clustering* con la tecnica descritta in precedenza, ovvero quella della riduzione della dimensionalità. In particolare, ho imposto che siano individuati minimo tre diversi *cluster*, in modo da avere una differenziazione fra i diversi stili di gioco utilizzati durante la competizione. In questo caso, $p = 44$ variabili e $G = 3$ *cluster*, così la dimensione del sottospazio è $d = 2$, ovvero un grafico in due dimensioni.

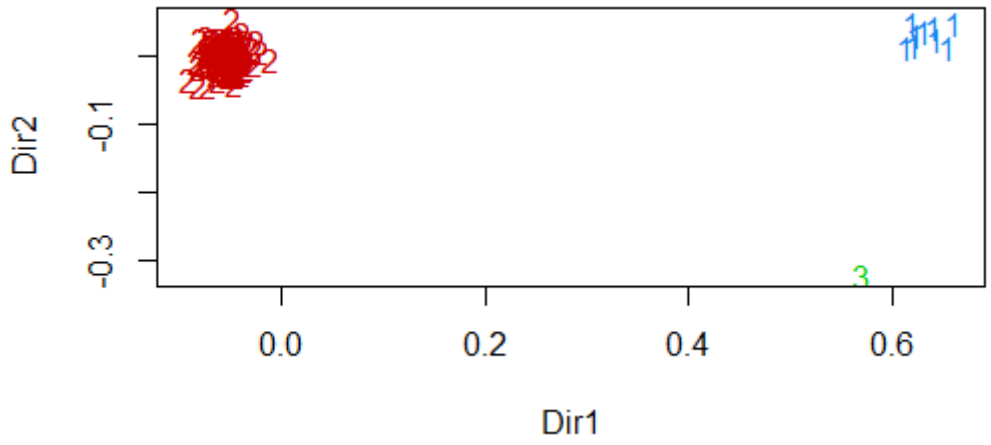
```
> plot(mod1dr, what = "scatterplot", symbols=c("1","2","3"))
```

Crea il grafico con i vari *cluster* ricavati dai comandi precedenti. Come tipologia di grafico, ho scelto uno *scatterplot*, in modo da avere un'idea immediata dei *cluster* già a prima vista; come simboli dei tre *cluster* ottenuti, invece, ho scelto i numeri 1, 2 e 3.

```
> summary(mod1dr)
> mod1dr$class
```

Il primo comando fornisce un sommario generale del processo di *clustering* effettuato, mentre il secondo comando indica a quale *cluster* appartiene ognuna delle centoventotto osservazioni. Di seguito la rappresentazione del sottospazio bidimensionale (di dimensione $d = 2$) ottenuto, con i tre *cluster*.

Figura 15. *Grafico che mostra i tre cluster ottenuti. Le variabili sui due assi (Dir1 e Dir2) sono combinazioni lineari delle caratteristiche originali che catturano la maggior parte della struttura di clustering contenuta nei dati.*



Fonte: ns. elaborazione.

Il processo di *clustering* appena descritto, applicato ai miei dati, genera i seguenti tre *cluster*:

- Cluster 1: contenente nove osservazioni;
- Cluster 2: contenente centodiciotto osservazioni;
- Cluster 3: contenente una sola osservazione.

Più precisamente, i tre *cluster* sono così composti:

- Cluster 1: KSA1, AUS1, DEN1, JPN1, SEN1, MAR3, CRC3, COL3, SEN3;
- Cluster 2: RUS1, EGY1, URU1, MAR1, IRN1, ESP1, POR1, FRA1, ARG1, PER1, CRO1, NGA1, SRB1, CRC1, GER1, MEX1, BRA1, SUI1, SWE1, KOR1, BEL1, PAN1, TUN1, ENG1, COL1, POL1, RUS2, EGY2, MAR2, POR2, URU2, KSA2, IRN2, ESP2, DEN2, AUS2, FRA2, PER2, ARG2, CRO2, BRA2, CRC2, NGA2, ISL2, SRB2, SUI2, BEL2, TUN2, KOR2, MEX2, GER2, SWE2, PAN2, ENG2, SEN2, JPN2, COL2, POL2, KSA3, EGY3, URU3, RUS3, IRN3, POR3, ESP3, AUS3, PER3, DEN3, FRA3, NGA3, ARG3, ISL3, CRO3, SWE3, MEX3, KOR3, GER3, SUI3, SRB3, BRA3, JPN3, POL3, ENG3, BEL3, PAN3, TUN3, FRA4, ARG4, URU4, POR4, ESP4, RUS4, CRO4, DEN4, BRA4, MEX4, BEL4, JPN4, SWE4, SUI4, COL4, ENG4, URU5, FRA5, BRA5, BEL5, SWE5, ENG5, RUS5, CRO5, FRA6, BEL6, CRO6, ENG6, BEL7, ENG7, FRA7, CRO7;
- Cluster 3: ISL1.

```
> CLUSTER1=rbind(KSA1,AUS1,DEN1,JPN1,SEN1,MAR3,CRC3,COL3,SEN3)
> colnames(CLUSTER1)<-nomi_var
> CLUSTER1=as.data.frame(CLUSTER1)
```

Il primo comando crea in R un nuovo oggetto (una matrice), chiamato CLUSTER1, andando ad unire verticalmente i vettori relativi alle prestazioni di squadra appartenenti a questo *cluster*; il secondo comando rinomina le colonne di tale matrice con i codici creati in precedenza mentre il terzo comando trasforma la matrice in un data frame. Lo stesso procedimento è eseguito anche per il Cluster 2, andando a creare l'oggetto CLUSTER2, come segue:

```
> CLUSTER2=rbind(RUS1,EGY1,URU1,MAR1,IRN1,ESP1,POR1,FRA1,ARG1,
+                PER1,CRO1,NGA1,SRB1,CRC1,GER1,MEX1,BRA1,SUI1,
+                SWE1,KOR1,BEL1,PAN1,TUN1,ENG1,COL1,POL1,RUS2,
+                EGY2,MAR2,POR2,URU2,KSA2,IRN2,ESP2,DEN2,AUS2,
+                FRA2,PER2,ARG2,CRO2,BRA2,CRC2,NGA2,ISL2,SRB2,
+                SUI2,BEL2,TUN2,KOR2,MEX2,GER2,SWE2,PAN2,ENG2,
+                SEN2,JPN2,COL2,POL2,KSA3,EGY3,URU3,RUS3,IRN3,
+                POR3,ESP3,AUS3,PER3,DEN3,FRA3,NGA3,ARG3,ISL3,
+                CRO3,SWE3,MEX3,KOR3,GER3,SUI3,SRB3,BRA3,JPN3,
+                POL3,ENG3,BEL3,PAN3,TUN3,FRA4,ARG4,URU4,POR4,
+                ESP4,RUS4,CRO4,DEN4,BRA4,MEX4,BEL4,JPN4,SWE4,
+                SUI4,COL4,ENG4,URU5,FRA5,BRA5,BEL5,SWE5,ENG5,
+                RUS5,CRO5,FRA6,BEL6,CRO6,ENG6,BEL7,ENG7,FRA7,
+                CRO7)
> colnames(CLUSTER2)<-nomi_var
> CLUSTER2=as.data.frame(CLUSTER2)
```

Il procedimento per il Cluster 3 è leggermente diverso, in quanto questo è formato da una sola osservazione, quella relativa alla partita della Nazionale islandese nella prima giornata della competizione, rappresentata dal vettore ISL1. Di seguito i comandi utilizzati:

```
> CLUSTER3=as.data.frame(ISL1)
> CLUSTER3=t(CLUSTER3)
> colnames(CLUSTER3)<-nomi_var
> CLUSTER3=as.data.frame(CLUSTER3)
```

Il primo comando trasforma il vettore ISL1 in un data frame di dimensioni 44 x 1, con i quarantaquattro valori disposti, uno dopo l'altro, in una singola colonna; il secondo comando effettua la trasposizione del data frame precedente (trasformandolo in automatico in una matrice), con i valori adesso disposti su una singola riga in quarantaquattro colonne: la stessa disposizione dei primi due *cluster*. Il terzo comando rinomina le colonne di tale matrice con i codici creati in precedenza mentre il quarto comando trasforma la matrice in un data frame.

4. Stili di gioco

4.1 Premessa

Il processo di *clustering* applicato ha creato tre gruppi di prestazioni di squadra, che identificano tre moduli e tre stili di gioco ben definiti. In questo capitolo, analizzerò proprio questi aspetti, cercando di riconoscere quale sia il modulo collegato ad ognuno dei tre *cluster*, in base alla posizione degli undici giocatori in campo, e quale sia lo stile di gioco adottato, in base ai valori assunti dagli indici di *closeness centrality* e *betweenness centrality*.

4.2 Calcolo delle mediane

Concluso il processo di *clustering* e le necessarie implementazioni in R, ho adesso tre *cluster* di osservazioni con numerosità molto diverse. In particolare il Cluster 2 ha un numero di osservazioni molto più elevato rispetto agli altri due gruppi: questo potrebbe essere spiegato dal fatto che le squadre partecipanti ai Mondiali sono le migliori fra tutte quelle che hanno giocato le qualificazioni e, quindi, il loro livello è piuttosto omogeneo e le varie prestazioni di squadra presentano caratteristiche simili.

Per poter effettuare un confronto tra i tre gruppi, è necessario individuare un criterio che calcoli, per ogni *cluster*, un valore unico per ognuna delle quarantaquattro variabili considerate. Un possibile criterio è quello di utilizzare un indice di posizione. Fra questi, dato che la media è maggiormente influenzata dai valori estremi (*outliers*) della distribuzione, ho deciso di utilizzare la mediana, un indice più robusto.

Il procedimento da seguire adesso, quindi, consiste nel calcolare, per il Cluster 1 e per il Cluster 2, i valori mediani delle quarantaquattro variabili; il Cluster 3, invece, essendo formato da una sola osservazione, non necessita di questo passaggio: i valori assunti dalle variabili, in questo caso, sono da intendersi già come valori mediani del gruppo.

```
> CLUSTER1_med=apply(CLUSTER1,2,median)
> CLUSTER2_med=apply(CLUSTER2,2,median)
```

La funzione `apply` in R applica una funzione alle righe (inserendo 1 al secondo termine del comando) o alle colonne (inserendo 2 al secondo termine del comando) di un data frame. In questo caso, ho applicato la funzione *median*, ovvero il calcolo della mediana, alle colonne del Cluster 1 e del Cluster 2: così viene calcolato il valore mediano per ognuna delle quarantaquattro variabili nei due *cluster*. Gli oggetti così creati sono vettori numerici di dimensioni 44 x 1, con i quarantaquattro valori disposti, uno dopo l'altro, in una singola colonna.

```
> CLUSTER1_med=t(CLUSTER1_med)
> CLUSTER2_med=t(CLUSTER2_med)
```

Questi due comandi effettuano, rispettivamente per il Cluster 1 e per il Cluster 2, la trasposizione del vettore precedente (trasformandolo in automatico in una matrice), con i valori adesso disposti su una singola riga in quarantaquattro colonne.

```
> CLUSTER_MEDIANE=rbind(CLUSTER1_med,CLUSTER2_med,CLUSTER3)
> rownames (CLUSTER_MEDIANE)=c("CLUSTER1","CLUSTER2","CLUSTER3")
```

Crea un nuovo data frame unendo verticalmente i valori mediani, calcolati precedentemente, per i tre *cluster*. Il secondo comando rinomina le righe del data frame con i nomi dei tre *cluster*. Di seguito, il data frame ottenuto, i cui dati saranno utilizzati per effettuare un'analisi descrittiva dei tre *cluster* ed individuare modulo e stile di gioco utilizzati.

Tabella 8. *Data frame contenente i valori mediani delle quarantaquattro variabili per i tre cluster.*

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
CLUSTER1	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.00	0.000
CLUSTER2	0.667	0.722	0.722	0.812	0.765	0.8	0.706	0.812	0.667	0.75	0.722
CLUSTER3	0.000	0.000	0.000	0.000	0.000	0.0	0.000	0.000	0.000	0.00	0.000

B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11
0.005	0.0400	0.0560	0.044	0.0420	0.0380	0.0260	0.088	0.016	0.0510	0.011
0.007	0.0405	0.0385	0.054	0.0355	0.0425	0.0275	0.064	0.016	0.0305	0.021
0.081	0.0890	0.0250	0.021	0.0280	0.0570	0.0360	0.176	0.039	0.0740	0.055

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
10.7590	51.8400	55.1730	47.7570	38.489	38.5920	61.7550	48.960	73.2450	59.174	62.619
8.6475	55.6765	55.8565	49.1865	38.137	37.6255	67.7775	53.766	70.1695	62.856	67.238
6.0740	50.1220	60.7580	25.9820	30.682	18.8230	62.1670	52.365	71.8350	60.439	60.902

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11
32.720	5.330	62.735	33.885	20.17	47.090	8.070	43.355	27.44	22.3800	54.55
33.525	6.125	62.355	31.855	20.33	48.355	13.825	42.320	36.91	28.7075	53.00
38.535	29.950	59.230	33.610	17.38	54.410	3.190	36.940	33.91	30.9050	55.71

Fonte: ns. elaborazione.

4.3 Moduli e stili di gioco

In questa sezione analizzerò i valori mediani calcolati per i tre *cluster* e li metterò a confronto, con l'obiettivo di individuare a quale modulo le prestazioni di squadra di ogni *cluster* possono essere collegate e quale stile di gioco esse hanno adottato. L'analisi si basa sull'interpretazione dei valori assunti dai quattro indici considerati e, in particolare, sul significato che questi hanno in ambito sportivo, come visto nel capitolo precedente. I risultati ottenuti sono una mia personale elaborazione dei dati e non rappresentano, necessariamente, l'unico modo di leggerli.

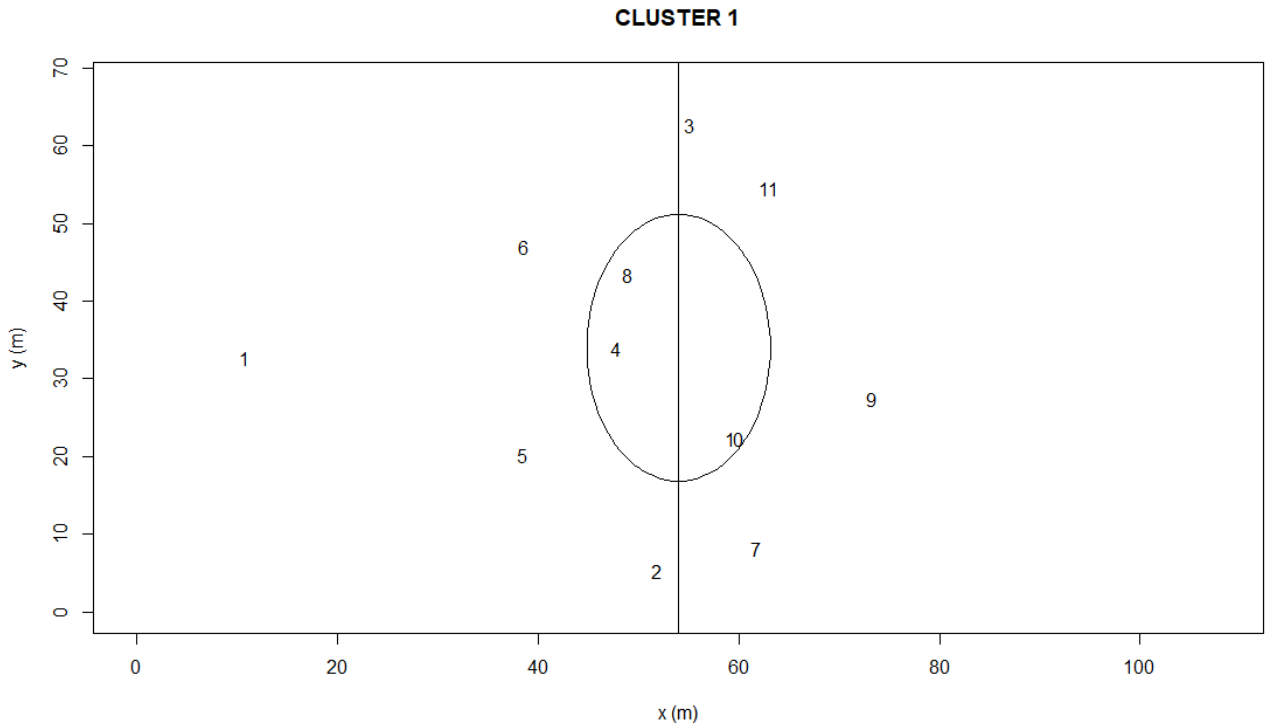
Di seguito, prenderò in analisi i tre *cluster* singolarmente, aiutandomi con l'uso di *barplot* (elaborati tramite Excel) e rappresentazioni grafiche delle posizioni dei giocatori in campo (elaborate tramite R con comandi che in seguito mostrerò).

4.3.1 Cluster 1

Il Cluster 1 è costituito da nove osservazioni, ovvero da nove prestazioni di squadra, che adottano un modulo approssimativamente riconducibile ad un 4-3-3, ovvero un modulo che prevede un portiere, quattro difensori, tre centrocampisti e tre giocatori offensivi (un attaccante e due ali). Di seguito, la rappresentazione grafica delle posizioni mediane in campo degli undici giocatori titolari, ottenuta tramite i seguenti comandi:

```
> numeri=c("1","2","3","4","5","6","7","8","9","10","11")
> library(plotrix)
> lato_x_cluster1=CLUSTER_MEDIANE[1,23:33]
> lato_x_cluster1=t(lato_x_cluster1)
> lato_y_cluster1=CLUSTER_MEDIANE[1,34:44]
> lato_y_cluster1=t(lato_y_cluster1)
> plot(lato_x_cluster1,lato_y_cluster1,xlim=c(0,108),xlab="x
(m)",ylim=c(0,68),ylab="y (m)",main="CLUSTER 1",pch=numeri)
> abline(v=54)
> draw.circle(54,34,9.15)
```

Figura 16. ***R****appresentazione grafica delle posizioni mediane in campo degli undici giocatori titolari del Cluster 1.*



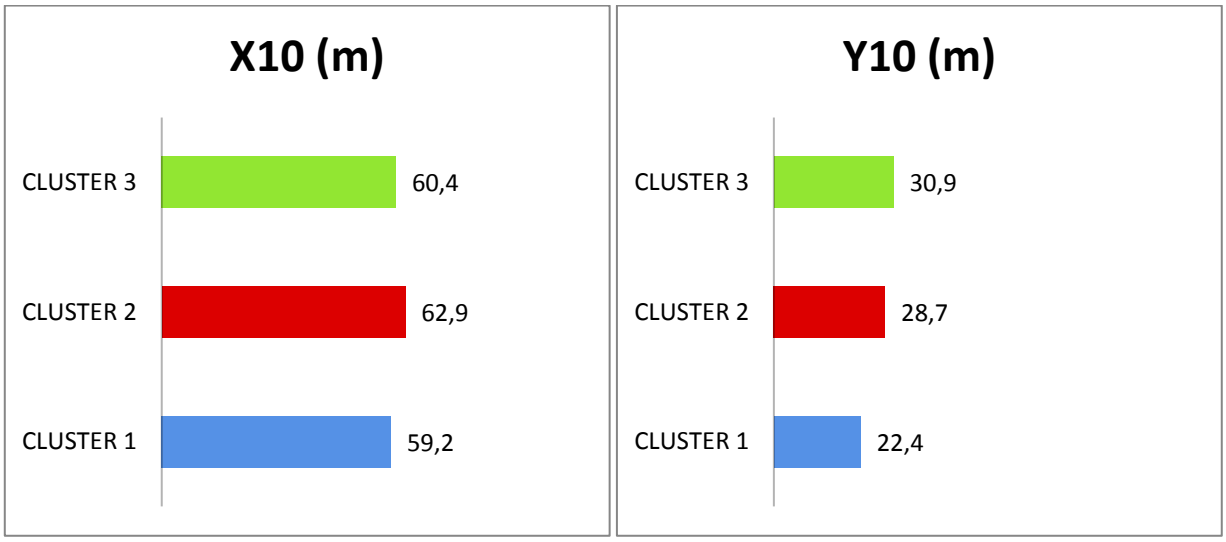
Fonte: ns. elaborazione.

Nella Figura 16 si può facilmente riconoscere la linea difensiva a quattro, con i due terzini (2 e 3) e i due difensori centrali (5 e 6), il centrocampo a tre (4 nel ruolo di centrocampista centrale e 8 e 10 nei ruoli, rispettivamente, di centrocampista sinistro e centrocampista destro), e l'attacco a tre, con un attaccante (9) e due ali (7 a destra e 11 a sinistra).

Osservando con maggiore attenzione i dati relativi alle posizioni mediane dei giocatori in campo, si può notare come la posizione del numero 10, a differenza dei pari ruolo negli altri due *cluster*, sia più arretrata e più decentrata verso destra: in tal caso, la sua posizione, quindi, è da intendersi non come trequartista, ma come centrocampista di destra, a formare un centrocampo a tre,

insieme ai numeri 4 e 8. Di seguito la rappresentazione dei valori relativi alla posizione mediana sul lato lungo e sul lato corto del campo del numero 10 nei tre *cluster*.

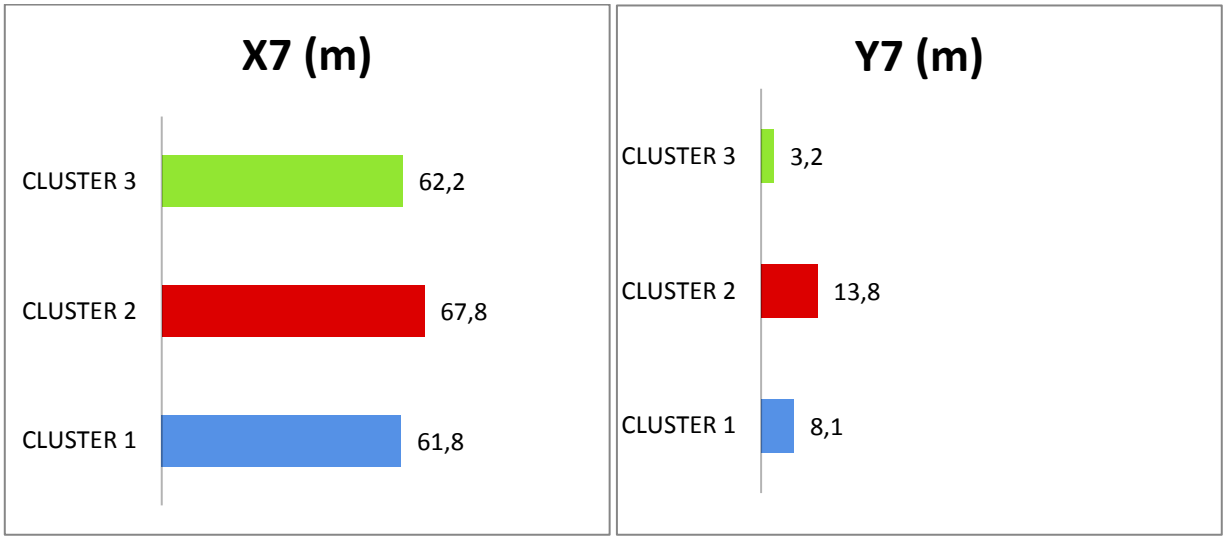
Figura 17. *Posizione mediana sul lato x (a sinistra) e sul lato y (a destra) del giocatore numero 10 nei tre cluster.*



Fonte: ns. elaborazione.

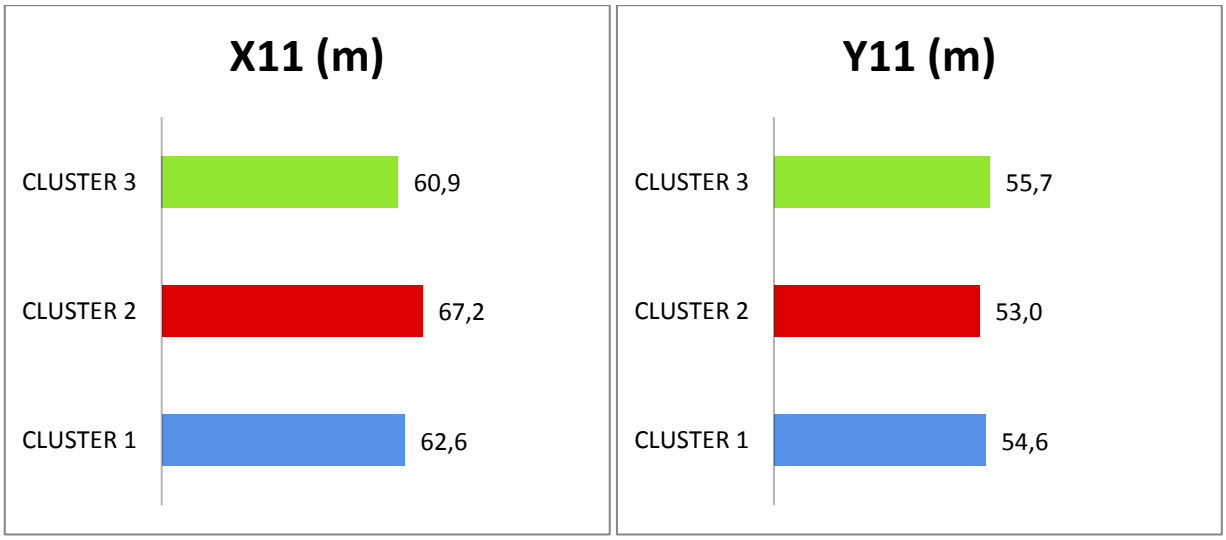
Si può notare, inoltre, come le due ali occupino, rispetto ai pari ruoli degli altri due *cluster*, una posizione più arretrata, con valori di X7 e X11 inferiori, soprattutto rispetto al Cluster 2. Allo stesso tempo, per quanto riguarda il lato corto del campo (lato y), i due giocatori occupano posizioni molto ampie, vicine alle linee laterali, con Y7 basso e Y11 elevato. Questo, da un punto di vista calcistico, può indicare un atteggiamento prudente e maggiormente difensivo della squadra, con le due ali che spesso si abbassano sulla linea di centrocampo a formare un centrocampo a cinque giocatori, insieme ai numeri 4, 8 e 10. A conferma di ciò, riporto i grafici relativi al confronto tra i tre *cluster* per le variabili appena descritte.

Figura 18. *Posizione mediana sul lato x (a sinistra) e sul lato y (a destra) del giocatore numero 7 nei tre cluster.*



Fonte: ns. elaborazione.

Figura 19. *Posizione mediana sul lato x (a sinistra) e sul lato y (a destra) del giocatore numero 11 nei tre cluster.*

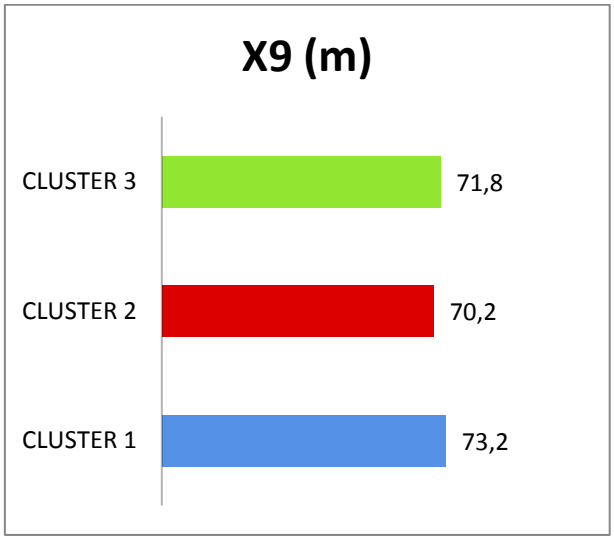


Fonte: ns. elaborazione.

Da segnalare, infine, la posizione del numero 9, che, rispetto a quella dell'attaccante degli altri due *cluster*, risulta essere più avanzata di qualche metro: questa potrebbe essere spiegata col fatto che, essendo le due ali più difensive e di conseguenza in posizione più arretrata, l'attaccante deve sviluppare l'azione offensiva da solo e, di conseguenza, occupa una posizione più avanzata. Questa idea può trovare conferma nel fatto che le prestazioni di squadra del Cluster 1 presentano valori elevati per l'indice di *betweenness centrality*: ciò indica una tendenza ad effettuare

principalmente azioni in contropiede con passaggi del tipo A-B-C, ovvero passaggi lunghi (cross o lanci) verso l'attaccante.

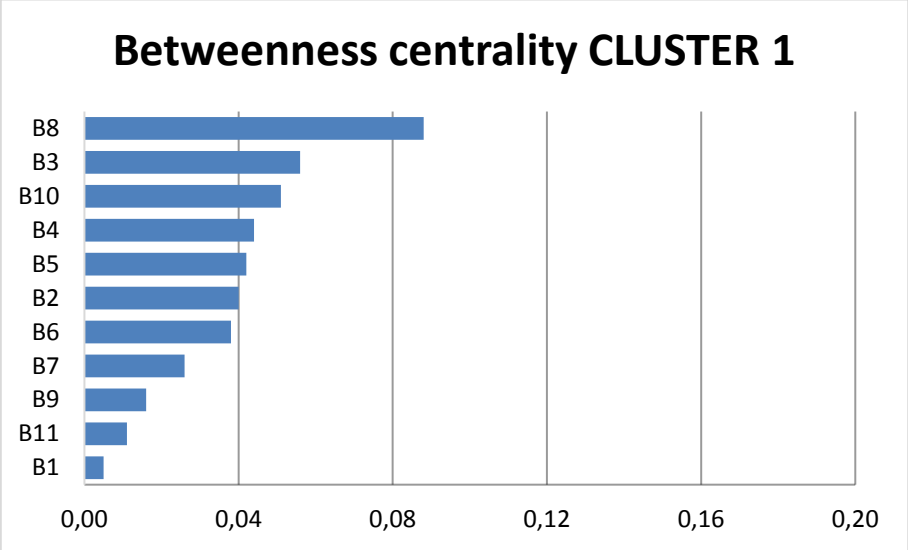
Figura 20. *Posizione mediana sul lato x del giocatore numero 9 nei tre cluster.*



Fonte: ns. elaborazione.

Per quanto riguarda lo stile di gioco utilizzato nelle nove prestazioni di squadra del Cluster 1, i valori della *closeness centrality* sono tutti uguali a zero, mentre la *betweenness centrality* assume valori leggermente più elevati per i numeri 8, 10 e 3, ovvero per i due centrocampisti laterali e per il terzino sinistro. Il fatto che sia il centrocampista sinistro (8) sia il centrocampista destro (10) presentino valori di *betweenness centrality* più alti rispetto a quello del numero 4, come visibile nella Figura 21, può indicare che l'azione offensiva è sviluppata principalmente dai due centrocampisti laterali, mentre il numero 4 svolge il ruolo di mediano, incaricato del recupero palla.

Figura 21. *Barplot con i punteggi di betweenness centrality per gli undici giocatori titolari del Cluster 1.*



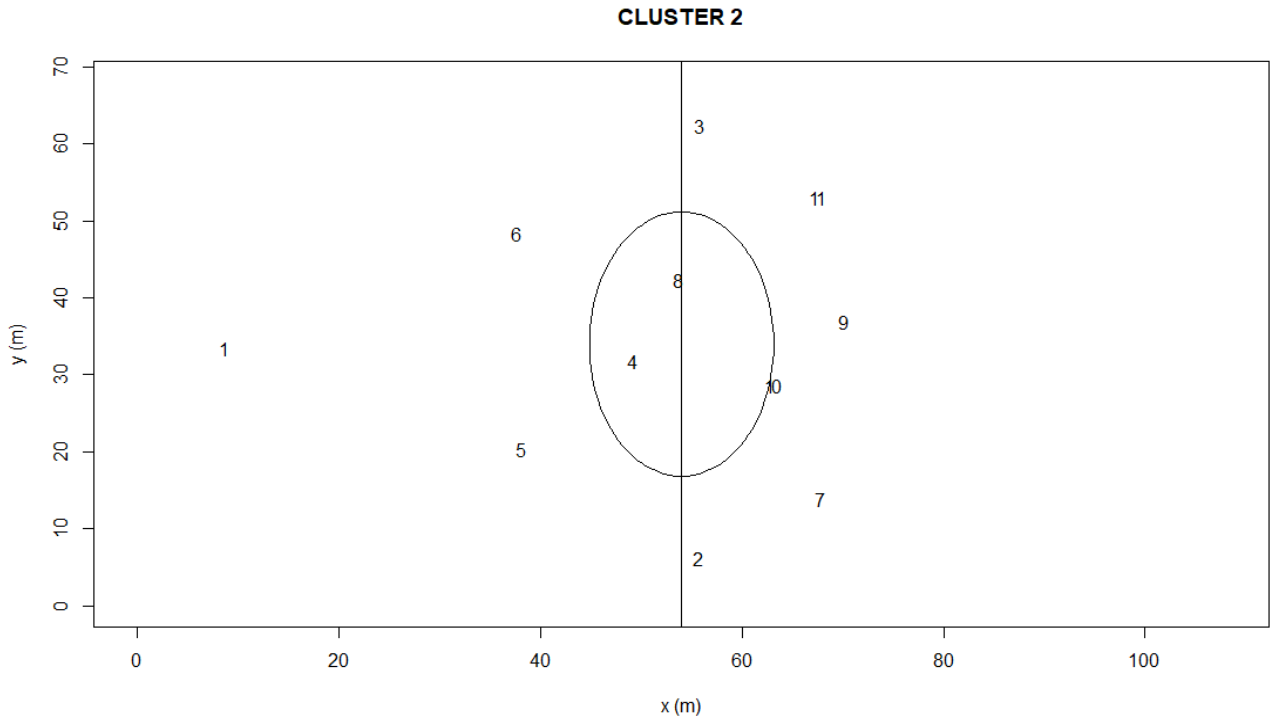
Fonte: ns. elaborazione.

4.3.2 Cluster 2

Il Cluster 2 è costituito da centodiciotto osservazioni, ovvero da centodiciotto prestazioni di squadra, che utilizzano un modulo approssimativamente riconducibile ad un 4-2-3-1, che prevede un portiere, quattro difensori, due centrocampisti e quattro giocatori offensivi (un trequartista, due ali e un attaccante). Di seguito, la rappresentazione grafica delle posizioni mediane in campo degli undici giocatori titolari, ottenuta tramite i seguenti comandi:

```
> lato_x_cluster2=CLUSTER_MEDIANE[2,23:33]
> lato_x_cluster2=t(lato_x_cluster2)
> lato_y_cluster2=CLUSTER_MEDIANE[2,34:44]
> lato_y_cluster2=t(lato_y_cluster2)
> plot(lato_x_cluster2,lato_y_cluster2,xlim=c(0,108),xlab="x
(m)",ylim=c(0,68),ylab="y (m)",main="CLUSTER 2",pch=numeri)
> abline(v=54)
> draw.circle(54,34,9.15)
```

Figura 22. *Rappresentazione grafica delle posizioni mediane in campo degli undici giocatori titolari del Cluster 2.*



Fonte: ns. elaborazione.

Dalla Figura 22 è facile identificare il modulo adottato nelle prestazioni di squadra che compongono il Cluster 2: davanti al portiere, la linea difensiva a quattro, con i due terzini (2 e 3) e i due difensori centrali (5 e 6), il centrocampo a due (4 e 8), il trequartista con il numero 10 e l'attacco a tre, con un attaccante (9) e due ali (7 a destra e 11 a sinistra).

Analizzando con attenzione i dati relativi alle posizioni mediane dei giocatori in campo, emerge come i due terzini siano posizionati entrambi oltre alla linea di metà campo, con X2 e X3 entrambi superiori a 54 (metri): ciò indica, presumibilmente, un atteggiamento delle squadre maggiormente offensivo, con i terzini coinvolti nelle azioni di attacco.

Importanti differenze rispetto al caso precedente si evidenziano nelle posizioni delle due ali: questi giocatori nel Cluster 2 occupano posizioni più avanzate e più interne al campo rispetto ai pari ruolo negli altri due *cluster*. Come mostrato dalla Figura 18 e dalla Figura 19, infatti, X7 assume nel Cluster 2 il valore più elevato, così come X11: le due ali sono, cioè, più avanzate. Allo stesso tempo, Y7 assume il valore più elevato e Y11 quello più basso tra i tre *cluster*: le due ali sono posizionate più internamente, ovvero più lontane dalle linee laterali. Ciò potrebbe essere spiegato con il fatto che in queste prestazioni, le squadre hanno adottato una tattica offensiva,

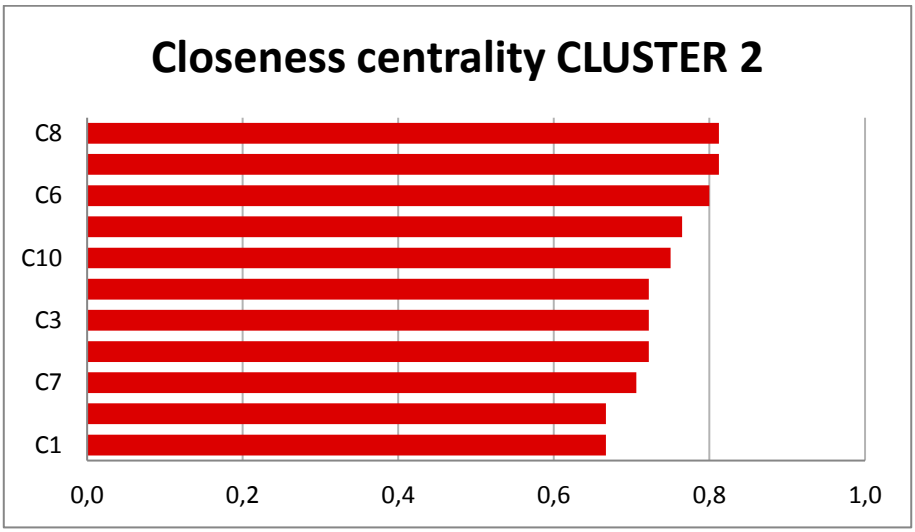
votata all'attacco (in linea con quanto visto per i due terzini), con le due ali posizionate a formare un tridente d'attacco insieme al numero 9.

Diverso è anche il ruolo del numero 10, che nel Cluster 2 agisce da trequartista, svolgendo un ruolo di raccordo tra centrocampo e attacco: la sua posizione, come evidenziato dalla Figura 17, è infatti nella zona centrale del campo, proprio dietro l'attaccante, con Y10 circa uguale a 34 (la metà esatta della larghezza del campo); X10, invece, assume il valore più alto tra i tre *cluster*, in quanto, in questo caso, il numero 10 non è un centrocampista, ma un trequartista e di conseguenza un giocatore più offensivo ed avanzato.

L'attaccante (il numero 9) è, in questo caso, posizionato qualche metro più indietro rispetto ai pari ruolo degli altri due *cluster*, come conseguenza di quanto detto fino ad ora: essendo, infatti, supportato da ben tre giocatori (7, 10 e 11), può giocare in una posizione leggermente più bassa ed essere coinvolto in azioni palla a terra con passaggi corti del tipo A-B-A. Questa idea può essere confermata dal fatto che la *closeness centrality*, che è determinata da passaggi proprio di questa tipologia, assume, nel Cluster 2, valori molto elevati, soprattutto per i giocatori offensivi.

Passando adesso ad analizzare il possibile stile di gioco adottato nelle prestazioni di squadra del Cluster 2, emerge prima di tutto come, a differenza degli altri due *cluster*, i punteggi di *closeness centrality* per gli undici giocatori siano diversi da zero ed, anzi, assumano valori elevati, tutti superiori a 0.5. In particolare sono determinati ruoli ad avere punteggi molto elevati: i due difensori centrali (C5 e C6), i due centrocampisti centrali (C4 e C8) e il trequartista (C10), tutti giocatori che agiscono nella zona centrale del campo. Questo implica che l'azione di attacco parte dai centrali di difesa, passa attraverso i due centrocampisti fino ad arrivare al numero 10: un'azione fatta di passaggi brevi che possono essere restituiti (del tipo A-B-A). Di seguito il *barplot* con i valori di *closeness centrality* nel Cluster 2.

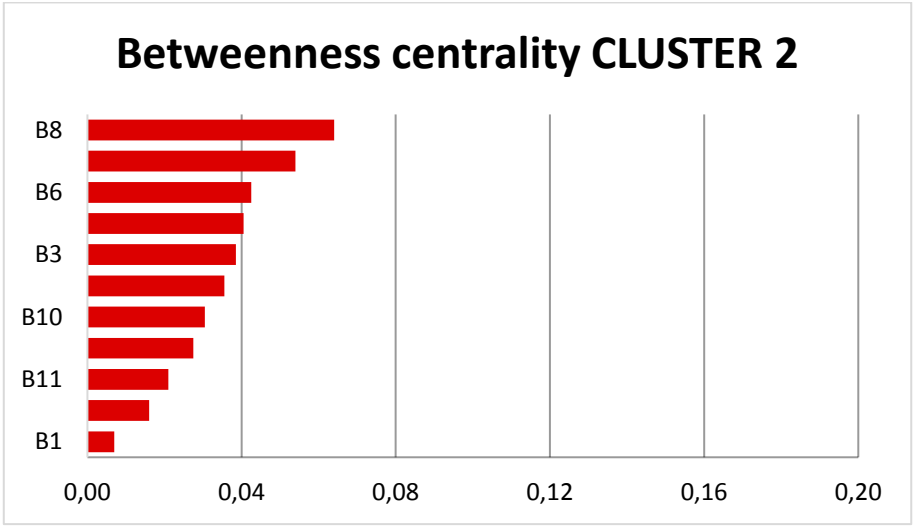
Figura 23. *Barplot con i punteggi di closeness centrality per gli undici giocatori titolari del Cluster 2.*



Fonte: ns. elaborazione.

Per quanto riguarda, invece, la *betweenness centrality*, i valori assunti dagli undici giocatori (B1, …, B11) sono molto bassi, i più bassi fra i tre *cluster*: passaggi lunghi che non permettono di effettuare scambi ravvicinati fra compagni di squadra, ovvero passaggi del tipo A-B-C, sono evitati o, comunque, molto ridotti. Questi dati mostrano una tendenza, nelle centodiciotto prestazioni del Cluster 2, di utilizzare uno stile di gioco basato su passaggi brevi che possono essere restituiti, con passaggi del tipo A-B-A, come visto in precedenza. Di seguito il *barplot* con i valori di *betweenness centrality* nel Cluster 2.

Figura 24. *Barplot con i punteggi di betweenness centrality per gli undici giocatori titolari del Cluster 2.*



Fonte: ns. elaborazione.

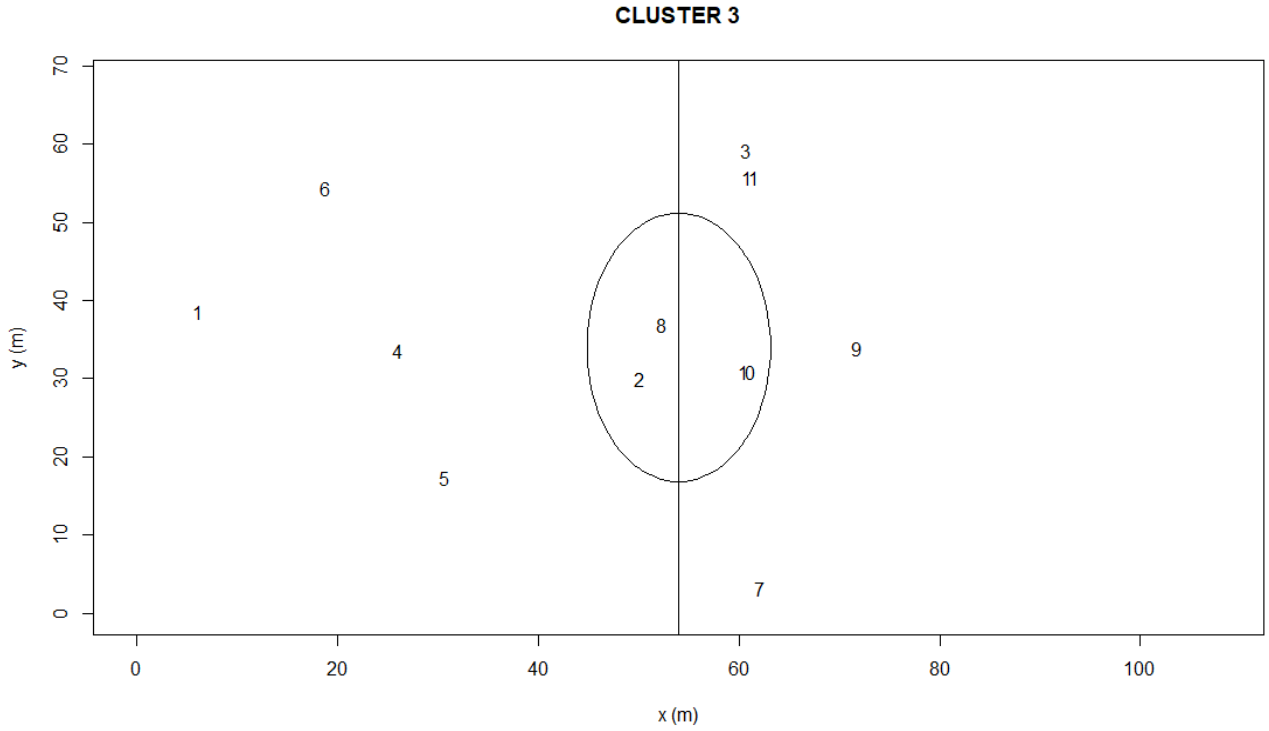
In conclusione, le prestazioni di squadra del Cluster 2 riflettono un modulo puramente offensivo con uno stile di gioco votato all'attacco e caratterizzato da passaggi brevi e numerosi scambi fra giocatori vicini. In particolare, è molto utilizzata la linea centrale, con i due difensori centrali, che assumono un ruolo decisivo nell'impostare l'azione, giocando palla a terra verso i compagni più avanzati, e con i due centrocampisti, che, come mostrato dai punteggi più alti sia per la *closeness centrality* sia per la *betweenness centrality*, sono i due giocatori maggiormente coinvolti nel gioco della propria squadra.

4.3.3 Cluster 3

Il Cluster 3 conta una sola prestazione di squadra, quella della Nazionale islandese nella prima giornata della competizione, in cui ha utilizzato un modulo approssimativamente riconducibile ad un 3-5-1-1, ovvero un modulo che prevede un portiere, tre difensori, cinque centrocampisti e un solo attaccante, supportato da un trequartista. Di seguito, la rappresentazione grafica delle posizioni in campo degli undici giocatori titolari, ottenuta tramite i seguenti comandi:

```
> lato_x_cluster3=CLUSTER_MEDIANE[3,23:33]
> lato_x_cluster3=t(lato_x_cluster3)
> lato_y_cluster3=CLUSTER_MEDIANE[3,34:44]
> lato_y_cluster3=t(lato_y_cluster3)
> plot(lato_x_cluster3,lato_y_cluster3,xlim=c(0,108),xlab="x
(m)",ylim=c(0,68),ylab="y (m)",main="CLUSTER 3",pch=numeri)
> abline(v=54)
> draw.circle(54,34,9.15)
```

Figura 25. *Rappresentazione grafica delle posizioni in campo degli undici giocatori titolari del Cluster 3.*

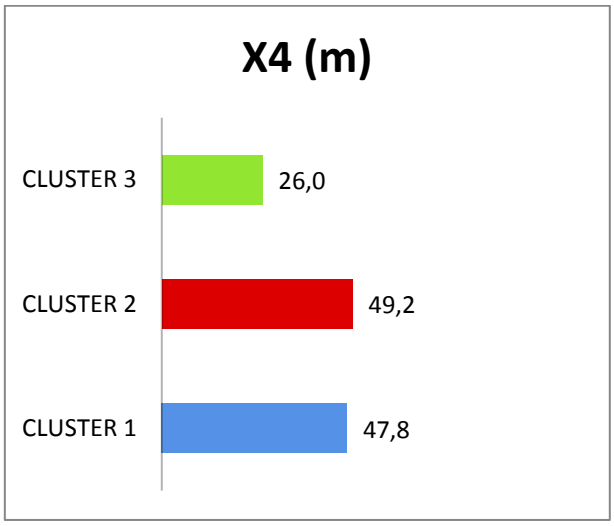


Fonte: ns. elaborazione.

Dalla Figura 25 è immediato riconoscere la difesa a tre, con il numero 4 nel ruolo di difensore centrale e i numeri 5 e 6 nei ruoli, rispettivamente, di difensore destro e sinistro; davanti a loro una linea di centrocampo formata da cinque giocatori e il trequartista (10), a supporto dell'unico attaccante (9): un modulo estremamente difensivo, dovuto, molto probabilmente, al fatto che in quella partita l'Islanda ha affrontato l'Argentina, una delle Nazionali più forti ed offensive fra quelle partecipanti ai Mondiali di calcio di Russia 2018.

Analizzando i dati relativi alle posizioni degli undici giocatori titolari, si può notare come la posizione del numero 4 sia molto più arretrata rispetto ai primi due *cluster*: in questo caso, il numero 4, infatti, non ricopre un ruolo da centrocampista, ma da difensore centrale e, per questo, è posizionato molto più vicino alla propria porta. Questa differenza emerge molto chiaramente osservando la seguente Figura 26.

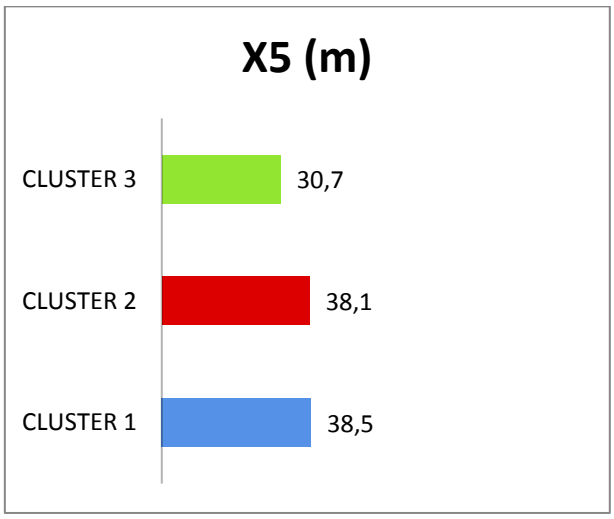
Figura 26. *Posizione mediana sul lato x del giocatore numero 4 nei tre cluster.*



Fonte: ns. elaborazione.

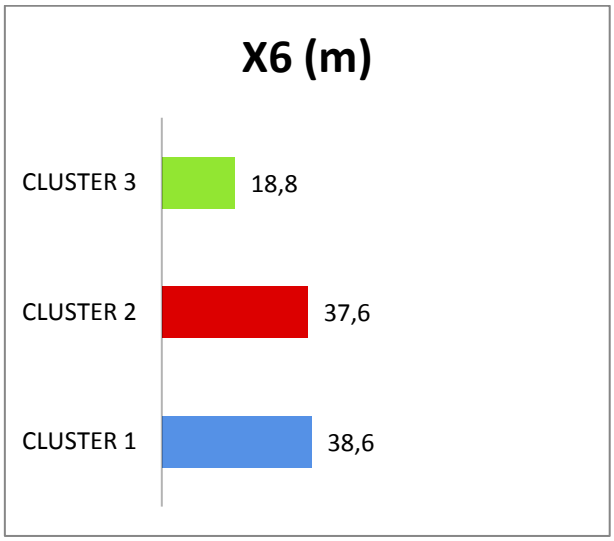
Ugualmente, anche le posizioni del numero 5 e del numero 6, nel Cluster 3, sono molto più arretrate rispetto agli altri due *cluster*, come visibile nelle seguenti Figura 27 e Figura 28: i due giocatori, infatti, pur svolgendo il ruolo, rispettivamente, di difensore destro e difensore sinistro (come fatto dai numeri 5 e 6 sia nel Cluster 1 che nel Cluster 2), sono posizionati molto più vicini alla propria porta e al proprio portiere. Questo atteggiamento fortemente difensivo è in linea con quanto detto precedentemente per il numero 4: il commissario tecnico ha preparato questa partita contro la Nazionale argentina in modo da avere un numero maggiori di giocatori difensivi per poter contrastare il gioco offensivo degli avversari.

Figura 27. *Posizione mediana sul lato x del giocatore numero 5 nei tre cluster.*



Fonte: ns. elaborazione.

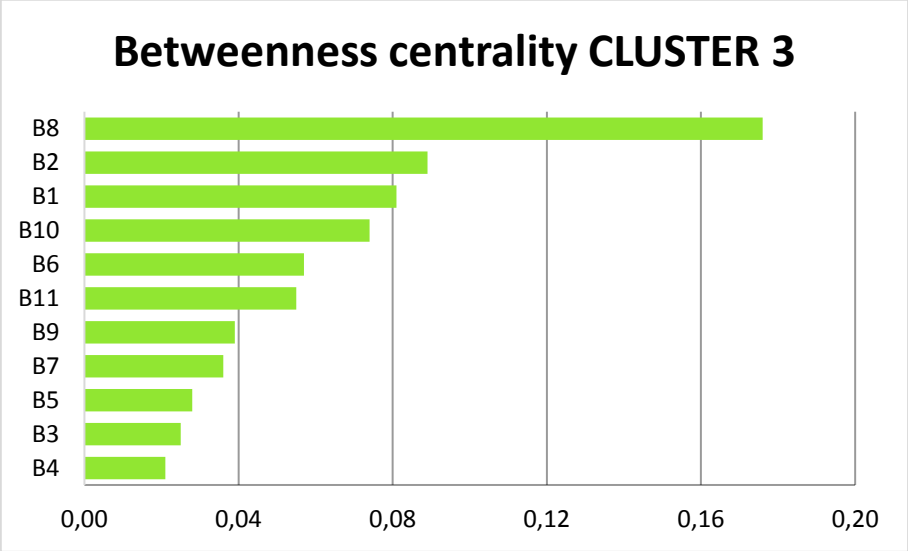
Figura 28. *Posizione mediana sul lato x del giocatore numero 6 nei tre cluster.*



Fonte: ns. elaborazione.

Spostando adesso l'analisi sullo stile di gioco utilizzato dalla Nazionale islandese nella partita contro la Nazionale argentina, i dati mostrano come i valori della *closeness centrality* siano tutti uguali a zero, mentre la *betweenness centrality* assume punteggi più elevati rispetto al Cluster 1 e al Cluster 2. In particolare, i valori più alti sono in corrispondenza dei numeri 1, 2, e 8, ovvero del portiere e dei due giocatori che, in questa partita, hanno svolto il ruolo di centrocampisti. Il fatto che il portiere abbia un valore elevato di *betweenness centrality* potrebbe indicare che questo è stato spesso chiamato ad intervenire e coinvolto nel rilancio della palla: una situazione plausibile dal momento che l'Argentina ha effettuato ben ventisette tiri verso la sua porta durante la partita. I valori elevati per i due centrocampisti, invece, potrebbero indicare che questi sono stati i due giocatori incaricati dello sviluppo dell'azione d'attacco e, quindi, coloro che hanno toccato più volte la palla. Essendo la *betweenness centrality* determinata principalmente da passaggi del tipo A-B-C, questi punteggi mostrano come la dinamica di gioco sia caratterizzata da una manovra veloce con passaggi diretti verso la porta avversaria. Di seguito il *barplot* con i valori di *betweenness centrality* nel Cluster 3.

Figura 29. *Barplot con i punteggi di betweenness centrality per gli undici giocatori titolari del Cluster 3.*



Fonte: ns. elaborazione.

In conclusione, l'analisi del Cluster 3 sembra evidenziare come la Nazionale islandese abbia adottato, nella partita contro la Nazionale argentina, un modulo estremamente difensivo, con una difesa a tre ed una linea di cinque centrocampisti, sviluppando azione veloci con passaggi lunghi (lanci e cross) diretti verso l'attaccante: una scelta che si è dimostrata giusta visto il risultato finale, un pareggio per 1 a 1.

4.4 Heatmap

In questa ultima sezione effettuerò una breve analisi dei due indici statistici, *closeness centrality* e *betweenness centrality*, e dei valori che questi assumono nelle centoventotto prestazioni di squadra realizzate durante i Mondiali di calcio di Russia 2018. In particolare, l'analisi sarà condotta da un punto di vista generale, senza considerare la divisione in *cluster*, ma considerando le osservazioni *in toto*, in modo da avere una visione globale dei punteggi assunti dai due indici. Per fare ciò, mi servirò di particolari grafici, detti *heatmap*, che assegnano ai punteggi degli undici giocatori colori diversi in base al valore assunto.

Per creare in R tali grafici, è prima necessario caricare alcune library; i comandi utilizzati sono i seguenti:

```
> if (!require("gplots")) {
+   install.packages("gplots", dependencies = TRUE)
+   library(gplots)
+ }


> if (!require("RColorBrewer")) {
+   install.packages("RColorBrewer", dependencies = TRUE)
+   library(RColorBrewer)
+ }


> library(lattice)
```

Grazie a queste library, sono adesso in grado di creare le *heatmap* per i due indici: prima analizzerò la *closeness centrality* e dopo la *betweenness centrality*. Le *heatmap* sono costruite in modo che le osservazioni siano disposte nello stesso ordine utilizzato nell'analisi precedente, ovvero secondo l'ordine cronologico delle partite giocate durante la competizione: dalla prima prestazione di squadra nella prima partita (in basso) fino all'ultima prestazione di squadra nella partita finale (in alto).

4.4.1 Closeness centrality

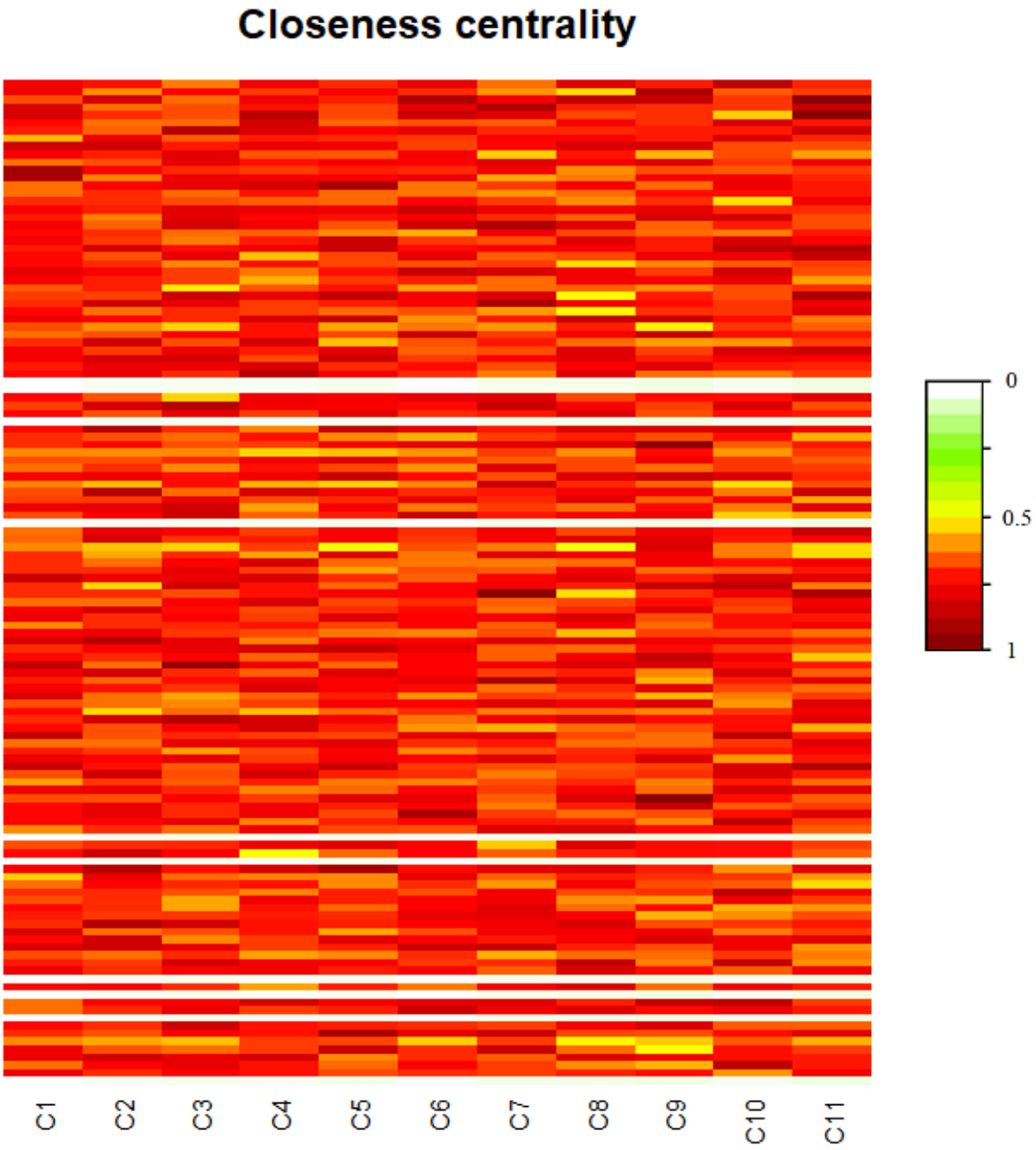
Partendo dal data frame DATI, contenente i valori dei quarantaquattro indici per le centoventotto osservazioni totali, il procedimento per creare la *heatmap* della *closeness centrality* è il seguente:

```
> DATI_CLOSENESS=DATI[,1:11]
> DATI_CLOSENESS <- data.matrix(DATI_CLOSENESS[,1:ncol(DATI_CLOSENESS)])
> my_palette <- colorRampPalette(c("white", "lawngreen", "yellow", "red",
"red4"))(n = 299)
> col_breaks = c(seq(0,0.20,length=100),
+                seq(0.20,0.40,length=100),
+                seq(0.40,0.60,length=100),
+                seq(0.60,0.80,length=100),
+                seq(0.80,1,length=100))
> heatmap <- heatmap(DATI_CLOSENESS, Rowv=NA, Colv=NA, col = my_palette,
+                    scale="column", margins=c(5,10), main="Closeness
centrality")
```

Il primo comando seleziona le prime undici colonne del data frame DATI, quelle contenenti i valori della *closeness centrality* per gli undici giocatori titolari, creando un nuovo data frame,

mentre il secondo comando trasforma il data frame creato in una matrice: passaggio necessario per poter costruire il grafico descritto. I due comandi successivi creano, rispettivamente, una scala di colori (dal bianco al rosso scuro) e le classi di valori per l'indice. L'ultimo comando, infine, genera la *heatmap*, associando la scala di colori alle classi create in precedenza. In particolare, i colori sono così assegnati: il colore bianco si trova in corrispondenza di valori di *closeness centrality* uguali a 0 (Cluster 1 e Cluster 3), mentre più il colore diventa scuro più alto è il valore assunto dall'indice.

Figura 30. *Heatmap relativa alla closeness centrality con relativa legenda dei colori.*



Fonte: ns. elaborazione.

Osservando la Figura 30, si nota come le partite della fase iniziale della competizione, rappresentate nella parte inferiore della *heatmap*, presentano molte celle di colore chiaro (arancione, giallo e addirittura bianco), ovvero punteggi di *closeness centrality* piuttosto bassi. Questo potrebbe essere spiegato con il fatto che in quella particolare fase dei Mondiali, l'obiettivo primario delle squadre è stato quello di non perdere per riuscire a superare la fase dei gironi e qualificarsi alla fase successiva, senza ricercare eccessivamente il bel gioco: le squadre, quindi, potrebbero aver utilizzato tattiche molo difensive, senza ricercare azioni di attacco troppo elaborate, ma ricorrendo a passaggi del tipo A-B-C. Al contrario, le partite della fase finale della competizione, rappresentate nella parte superiore della *heatmap*, presentano molte celle di colore scuro, ovvero valori di *closeness centrality* più alti. Nella fase successiva dei Mondiali (ottavi di finale, quarti di finale, semifinali e finali), invece, l'obiettivo da ricercare è la vittoria e questo richiede alle squadre di esprimere un gioco più offensivo, con azioni palla a terra e passaggi brevi che possono essere restituiti, del tipo A-B-A.

Spostando l'analisi sui singoli giocatori, la *heatmap* mostra come i due difensori (5 e 6) e i due centrocampisti (4 e 8) siano i giocatori con punteggi di *closeness centrality* più elevati, con numerose celle di colore rosso e solo poche celle gialle e bianche: questo risultato riflette quanto visto precedentemente per il Cluster 2, l'unico *cluster* con valori dell'indice diversi da zero.

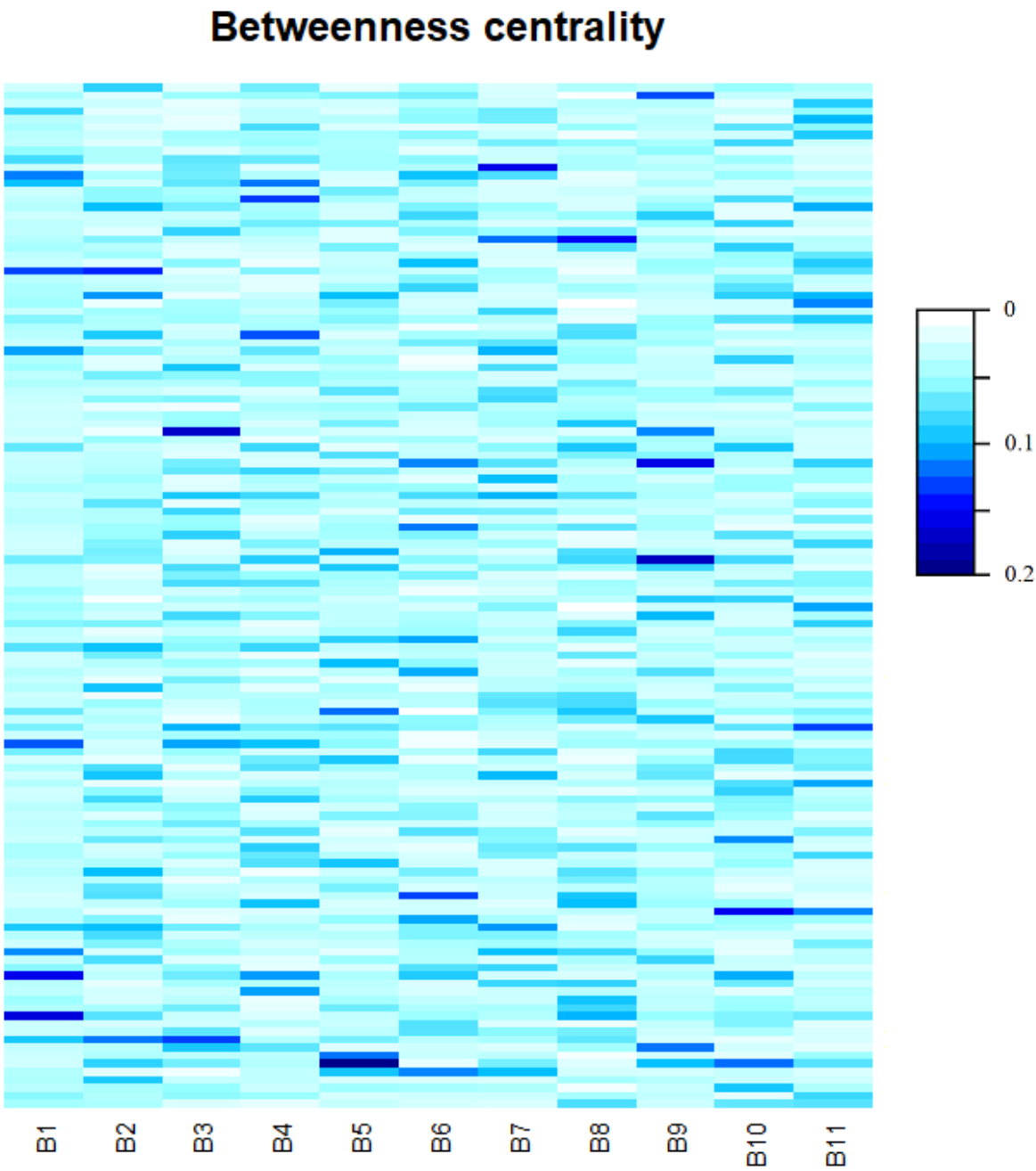
4.4.2 Betweenness centrality

Passando, adesso, ad analizzare la *betweenness centrality*, i comandi utilizzati per creare la *heatmap* sono i seguenti:

```
> DATI_BETWEENNESS=DATI[,12:22]
> DATI_BETWEENNESS <-
data.matrix(DATI_BETWEENNESS[,1:ncol(DATI_BETWEENNESS)])
> my_palette2 <- colorRampPalette(c("white","darkslategray1" ,"deepskyblue",
"blue", "blue4"))(n = 299)
> col_breaks2 = c(seq(0,0.05,length=100),
+                 seq(0.05,0.1,length=100),
+                 seq(0.1,0.15,length=100),
+                 seq(0.15,0.2,length=100))
> heatmap2 <- heatmap(DATI_BETWEENNESS, Rowv=NA, Colv=NA, col = my_palette2,
+                         scale="column", margins=c(5,10), main="Betweenness
centrality")
```

Il procedimento è uguale al caso precedente, con alcune logiche differenze: le colonne selezionate sono, in questo caso, quelle dalla dodici alla ventidue, contenenti i dati relativi alla *betweenness centrality*, la scala di colori assume una colorazione diversa (dal bianco al blu scuro) e anche le classi di valori sono diverse, variando da 0 a 0.2. I colori sono così assegnati: il colore bianco si trova in corrispondenza di valori di *betweenness centrality* uguali a 0, mentre più il colore diventa scuro più alto è il valore assunto dall'indice.

Figura 31. *Heatmap relativa alla betweenness centrality con relativa legenda dei colori.*



Fonte: ns. elaborazione.

La *heatmap* in Figura 31 è costituita principalmente da caselle di colore chiaro, conseguenza di valori di *betweenness centrality* per la maggior parte bassi. Questo dato non sorprende; infatti il Cluster 2, con una numerosità di centodiciotto osservazioni sul totale di centoventotto, presenta, come visto precedentemente, valori bassi per questo indice. Valori elevati, rappresentati in figura da caselle di colore blu scuro, sono molto rari e mai presenti nella stessa prestazione di squadra per più di un giocatore.

Concentrandosi sui singoli giocatori, la *heatmap* mostra come i numeri 4 e 8, ovvero i centrocampisti, siano i giocatori con punteggi di *betweenness centrality* più elevati, con un numero limitato di celle di colore celeste chiaro o bianco: questo risultato indica che, nonostante un livello generale basso per l'indice, i due centrocampisti sono i due giocatori maggiormente coinvolti nel gioco e nella circolazione della palla delle proprie squadre.

5. Conclusioni

L'articolo a cui ho fatto fino ad adesso riferimento, *Predicting FIFA World Cup 2018 key role and playing style features* (2018), nella sua analisi sulle partite di qualificazione ai Mondiali di calcio di Russia 2018, ha suddiviso le trentadue squadre qualificate in tre *cluster*, individuando per ognuno di essi lo stile di gioco utilizzato. I tre *cluster* individuati sono:

- *front-runners*;
- *contenders*;
- *others*.

I *front-runners* comprendono dieci squadre: Argentina, Australia, Brasile, Croazia, Francia, Germania, Inghilterra, Polonia, Spagna e Svizzera. Questo *cluster* ha i punteggi di *closeness centrality* più alti rispetto agli altri due *cluster*, specialmente per i due difensori centrali, che svolgono un ruolo determinante nel tipico gioco posizionale di queste squadre, che prevede molti passaggi palla a terra prima di arrivare al tiro. Allo stesso tempo, i punteggi di *betweenness centrality* sono bassi, soprattutto per i giocatori offensivi: ciò potrebbe significare una tendenza a non effettuare cross o giocare passaggi che non possono essere restituiti come nel tipo di azione di contrattacco. Queste squadre, inoltre, utilizzano spesso, come giocatore d'attacco, il cosiddetto "falso nove", ovvero non un attaccante forte fisicamente ed in grado di raccogliere cross o lanci lunghi, ma bensì un giocatore più mobile ed in grado di abbassare la propria posizione in campo

e partecipare al giro palla. Per le caratteristiche appena descritte, queste squadre sono indicate come le favorite per la vittoria finale della competizione.

I *contenders*, invece, sono formati da sole tre squadre: Belgio, Costa Rica e Serbia. Queste squadre, durante le partite di qualificazione ai Mondiali, hanno utilizzato un centrocampo a tre; in particolare, questo *cluster* mostra l'evidenza di un centrocampista (numero 4) che gioca in profondità e che si muove nella zona centrale del campo. Punteggi elevati di *betweenness centrality* sono misurati per i due centrocampisti laterali (numeri 8 e 10), che svolgono un ruolo di collegamento tra difesa e attacco e risultano, quindi, i giocatori maggiormente coinvolti nel gioco di squadra, insieme al numero 4. Queste tre squadre, nell'analisi svolta dall'articolo sopra citato, sono indicate come le principali contendenti per la vittoria finale, nel caso in cui nessuna delle squadre dei *front-runners* ci riesca.

Il terzo *cluster*, infine, comprende le rimanenti diciannove squadre qualificate ai Mondiali non citate precedentemente: Arabia Saudita, Colombia, Corea, Danimarca, Egitto, Giappone, Iran, Islanda, Marocco, Messico, Nigeria, Panama, Perù, Portogallo, Russia, Senegal, Svezia, Tunisia e Uruguay. Questo *cluster* è caratterizzato da valori elevati di *betweenness centrality* per i ruoli difensivi e valori bassi di *closeness centrality* per i centrocampisti: ciò indica che le squadre in questo gruppo effettuano probabilmente lanci lunghi per portare la palla in avanti, evitando di passare attraverso la zona centrale del campo. Queste squadre, inoltre, utilizzano una linea difensiva a quattro, che difende molto in profondità (con valori di X bassi), mentre l'attaccante è molto isolato e "costretto" a gestire il peso della manovra offensiva da solo.

Confrontando i tre *cluster* appena descritti con i tre generati dalla mia analisi sui Mondiali, emergono alcune importanti similitudini. In particolare, il *cluster* dei *front-runners* ha caratteristiche molto simili a quelle del Cluster 2: un atteggiamento offensivo che si concretizza in posizioni dei laterali (terzini ed ali) molto avanzate e azioni palla a terra con passaggi frequenti e brevi. In entrambi i *cluster*, svolgono un ruolo determinante i due difensori centrali, incaricati di far par partire l'azione e dare un'impronta ben riconoscibile al gioco delle proprie squadre; inoltre, in entrambi i casi, il numero 9 abbassa la propria posizione in campo, partecipando al possesso palla con i compagni: il gol viene ricercato non attraverso cross ma tramite il giro palla.

Il *cluster* dei *contenders*, invece, riflette le caratteristiche del Cluster 1: entrambi utilizzano un centrocampo a tre giocatori, con il numero 4 che gioca in posizione centrale e i numeri 8 e 10 ai suoi lati, con funzioni di raccordo tra difesa ed attacco. E sono proprio questi ultimi che

presentano punteggi di *betweenness centrality* più alti, a sottolineare come essi siano i giocatori più coinvolti nel gioco di squadra.

Il *cluster* rimanente, quello contenente le altre diciannove squadre qualificate, ha delle caratteristiche particolari che non trovano un corrispettivo nei tre *cluster* creati nel corso della mia analisi: uno stile di gioco fortemente difensivo e finalizzato quasi esclusivamente a non prendere goal, che, in fase di qualificazione, può essere utile soprattutto per le squadre sfavorite, ma che viene quasi del tutto abbandonato durante i Mondiali, dove il livello generale delle squadre è più elevato e richiede un atteggiamento meno remissivo.

Tornando all'analisi sui Mondiali, adesso è necessario individuare un criterio che associ ognuna delle trentadue squadre ad uno dei tre *cluster* (Cluster 1, Cluster 2 o Cluster 3), per effettuare un confronto con i *cluster* precedenti e verificare se le squadre che erano date come favorite hanno effettivamente giocato da tali. In tal senso, una squadra è da considerarsi come un'unità singola e non come somma delle varie prestazioni: nel *cluster* viene, cioè, inserita la squadra e non la singola prestazione fatta in una certa fase della competizione. All'interno di uno dei tre *cluster*, quindi, ogni squadra comparirà una sola volta, indipendentemente dal numero di partite giocate. I possibili criteri da utilizzare per tale scopo sono molteplici; qui presenterò quello che io reputo possa essere il più adatto per i dati disponibili. Quella che do è una mia personale interpretazione dei dati, che non risulta essere, ancora una volta, l'unica possibile.

Dato un livello generale molto omogeneo per la competizione, con la quasi totalità delle prestazioni di squadra appartenenti al Cluster 2, il fatto che una squadra abbia fatto una prestazione che rientra nel Cluster 1 o nel Cluster 3 è un dato abbastanza significativo del suo atteggiamento e del suo stile di gioco. Il criterio da me scelto è, quindi, il seguente: inserisco una squadra nel Cluster 2 solo se tutte le prestazioni che ha fatto durante i Mondiali appartengono al Cluster 2, nel Cluster 1 se presenta almeno una prestazione appartenente al Cluster 1, nel Cluster 3 se presenta almeno una prestazione appartenente al Cluster 3. In base a questo criterio, le squadre sono così assegnate:

- Cluster 1: Arabia Saudita, Australia, Colombia, Costa Rica, Danimarca, Giappone, Marocco, Senegal;
- Cluster 2: Argentina, Belgio, Brasile, Corea, Croazia, Egitto, Francia, Germania, Inghilterra, Iran, Messico, Nigeria, Panama, Perù, Polonia, Portogallo, Russia, Serbia, Spagna, Svezia, Svizzera, Tunisia, Uruguay;
- Cluster 3: Islanda.

Tutte le squadre che erano indicate come favorite dall'articolo *Predicting FIFA World Cup 2018 key role and playing style features* (2018), ad eccezione dell'Australia, hanno rispettato il pronostico, utilizzando durante tutte le partite giocate ai Mondiali di Russia 2018 uno stile di gioco in linea con quanto mostrato durante le qualificazioni. E sia la Francia che la Croazia, che facevano parte del *cluster* dei *front-runners* e indicate tra le favorite per la vittoria finale, sono state infatti le due migliori squadre della competizione, aggiudicandosi, rispettivamente, primo e secondo posto. L'Australia, pronosticata come una delle squadre favorite grazie ad uno stile di gioco offensivo durante le qualificazioni, è inserita, invece, nel Cluster 1: questo potrebbe essere spiegato con il fatto che la Nazionale australiana, in fase di qualificazioni, ha dovuto affrontare squadre del raggruppamento dell'Oceania con un livello generale piuttosto basso, che le ha permesso di esprimere uno stile di gioco fortemente offensivo, mentre ai Mondiali ha affrontato squadre di un livello più alto, che le hanno impedito di ripetere quanto fatto fino ad allora.

Belgio e Serbia, facenti parte del *cluster* dei *contenders* per lo stile di gioco utilizzato in fase di qualificazione, hanno dimostrato, invece, durante i Mondiali uno stile di gioco più offensivo, tanto da essere inserite nel Cluster 2: in particolare il Belgio è andato oltre i pronostici, arrivando terzo e dimostrandosi una delle squadre migliori dell'intera competizione. Al contrario, la Costa Rica, inserita fra i *contenders*, per quanto fatto ai Mondiali è stata assegnata al Cluster 1, mostrando grande continuità nello stile di gioco utilizzato.

Infine, le squadre inserite dall'articolo *Predicting FIFA World Cup 2018 key role and playing style features* (2018) nel *cluster* degli *others*, ovvero il *cluster* contenente le squadre considerate come le meno probabili per la vittoria finale a causa dello stile di gioco poco creativo ed eccessivamente difensivo mostrato durante le qualificazioni, si sono diversificate durante i Mondiali: alcune squadre, come ad esempio Russia ed Uruguay, hanno utilizzato un modulo ed uno stile di gioco più offensivo, che ha permesso loro di arrivare fino ai quarti di finale e di entrare nel Cluster 2, mentre altre squadre, come ad esempio Arabia Saudita, Marocco e Senegal, hanno mantenuto uno stile di gioco troppo difensivo, che non ha permesso loro di superare la fase a gironi e sono inseriti nel Cluster 1.

Bibliografia

Anderson C. & Sally D., *The numbers game: Why everything you know about soccer is wrong*, United Kingdom, Penguin UK, 2013.

Campagnolo G., Duncan A., Diquigiovanni J., Papastathopoulos I. & Zygalakis K., *Predicting FIFA World Cup 2018 key role and playing style feauteres,* University of Edinburgh, 2018.

Cintia P., Giannotti F., Pappalardo L., Pedreschi D. & Malvaldi M., *The harsh rule of the goals: Data-driven performance indicators for football teams*, in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA) , Paris, 2015, pagine 1-10.

Clemente F. M., Martins F. M. & Mendes R., *Applying networks and graph theory to match analysis: identifying the general properties of a graph*, in VIII Congreso Internacional de la Asociación Española de Ciencias del Deporte, Caceres, 2014.

Freeman L. C., *Centrality in social networks: conceptual clarification*, in "Social Networks", 1979, pagine 215-239.

Gyarmati L., Kwak H. & Rodriguez P., *Searching for a unique style in soccer*, arXiv preprint arXiv:1409.0308, 2014.

Higgins P. M., *Nets, puzzles and postmen: an exploration of mathematical connections*, Oxford University Press, 2008.

Hughes M. & Franks I., *Analysis of passing sequences, shots and goals in soccer*. Journal of sports sciences 23(5), 2005, pagine 509-514.

Pena J. L. & Touchette H., *A network theory analysis of football strategies*. arXiv preprint arXiv:1206.6904, 2012.

Pena J. L. & Navarro R. S., *Who can replace xavi? a passing motif analysis of football players*. arXiv preprint arXiv:1506.07768, 2015.

Scrucca L., *Dimension reduction for model-based clustering*, arXiv preprint arXiv:1508.01713, 2015.

Sitografia

https://cran.r-project.org/

https://www.rdocumentation.org/

https://www.wikipedia.org/