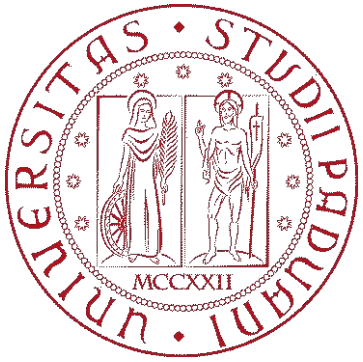


UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

Corso di Laurea in Statistica per
l'Economia e l'Impresa

Analisi statistica delle squadre maschili e femminili del campionato italiano di A1 di pallavolo

Relatore:

Antonio Canale

Candidato:

Irene Veghin
Matricola 2005393

ANNO ACCADEMICO 2022/2023

Abstract

Il presente lavoro mira ad analizzare le principali caratteristiche dei massimi campionati femminili e maschili di pallavolo in Italia degli ultimi 15 anni, con un focus particolare sulle dinamiche interne alle squadre tramite modelli per dati di rete. Infatti, ci si propone di studiare le relazioni tra i giocatori e tra le società, con l'intento di sottolineare l'appetibilità e la dinamicità dei campionati.

I dati utilizzati sono stati ricavati dai siti di *Lega Pallavolo Serie A* e *Lega Pallavolo Serie A Femminile* con un processo di scraping.

Le analisi sono state svolte utilizzando il linguaggio R, mentre per la visualizzazione delle reti è stato sfruttato il software Gephi.

Indice

1 Introduzione **1**
1.1 La pallavolo in Italia . . . 1
1.1.1 Serie A1 in Italia . . . 1
1.1.2 La struttura del campionato . . . 2
1.2 Obiettivi e dati . . . 2
1.2.1 Scraping dei giocatori . . . 2
1.2.2 Dataset e pulizia . . . 3
1.2.3 Scraping delle società . . . 4

2 Metodi **5**
2.1 Network analysis . . . 5
2.1.1 Statistiche descrittive di rete . . . 5
2.1.2 Rappresentazione della rete . . . 6
2.2 Modelli lineari (GLM) . . . 7

3 Analisi **8**
3.1 Analisi di rete dei giocatori . . . 8
3.1.1 Creazione delle reti . . . 8
3.1.2 Struttura delle reti . . . 9
3.1.3 Relazione tra statistiche di rete e covariate . . . 12
3.1.4 Visualizzazione delle reti . . . 18
3.1.5 Sviluppo temporale delle reti . . . 18
3.1.6 Permanenza nel campionato . . . 24
3.2 Analisi delle società sportive . . . 25
3.2.1 Permanenza delle società nel campionato . . . 25
3.2.2 Permanenza e spostamenti dei giocatori nei club . . . 26
3.2.3 Relazione tra risultati e struttura della squadra . . . 29

4 Conclusioni **32**

A Codice R: Scraping e pulizia **34**
A.1 Giocatori campionato maschile . . . 34
A.2 Giocatrici campionato femminile . . . 36
A.3 Pulizia dei giocatori . . . 40
A.4 Pulizia delle squadre . . . 40
A.5 Scraping delle società maschili . . . 42
A.6 Scraping delle società femminili . . . 45

B Codice R: Analisi delle reti dei giocatori **48**
B.1 Creazione delle reti 48
B.2 Struttura delle reti 50
B.3 Modelli per grado e betweenness 52
B.4 Reti con finestra di 5 anni 55
B.5 Permanenza degli atleti nel campionato 58

C Codice R: Analisi dei club **60**
C.1 Permanenza dei club 60
C.2 Creazione delle reti dei club 61
C.3 Modello per risultato in Regular Season 62

D Altre visualizzazioni di reti **65**
D.1 Campionato maschile (finestra di 5 anni) 65
D.2 Campionato femminile (finestra di 5 anni) 67
D.3 Società del campionato maschile 69
D.4 Società del campionato femminile 71

Capitolo 1

Introduzione

1.1 La pallavolo in Italia

La pallavolo è uno degli sport più diffusi e conosciuti in Italia, molti sono gli appassionati e chi la pratica anche in modo non agonistico [1]. In particolare negli ultimi vent'anni, secondo quanto riportato nei rapporti del CONI [2], la pallavolo è sempre stata sul podio degli sport con più tesserati in Italia e seconda solo al calcio per numero di società sportive.

Un movimento così importante e ricco, con oltre 300.000 atleti (ad eccezione del difficile periodo COVID, in cui questi numeri si sono temporaneamente ridotti [3]), ha permesso all'Italia di collocarsi tra i paesi migliori al mondo in questo sport.

Secondo la classifica stilata dalla Federazione Internazionale di Pallavolo (FIVB), che tiene in considerazione i risultati di tutte le partite delle squadre nazionali nelle competizioni ufficiali dal 1 gennaio 2019, sia la nazionale maschile italiana [4] che quella femminile [5] compaiono al terzo posto nel ranking mondiale. Basti pensare agli ultimi risultati ottenuti: la nazionale maschile ha vinto l'oro sia nel campionato europeo del 2021 che nel campionato mondiale del 2022, mentre la femminile rispettivamente oro e bronzo. Tali successi, mettono ancora più in luce questo sport e avvicinano altri al gioco, con aumento di tesserati, creando così un circolo virtuoso.

1.1.1 Serie A1 in Italia

Il massimo capionato italiano è uno dei migliori al mondo e ha una lunga tradizione. Rispetto alla classifica che ordina tutti i club di massima serie nel mondo secondo un punteggio che dipende dal posizionamento nei tornei e dall'importanza delle competizioni stesse, nei primi 100 club maschili [6] se ne trovano 11 italiani, nei primi 100 femminili [7] quelli italiani sono 9. Questo si traduce nel fatto che l'Italia sia il primo paese per numero di club nella top 100 maschile e il secondo, dopo la Russia, nella top 100 femminile. Tale considerazione, mostra l'effettiva qualità delle squadre del campionato italiano, che lo rendono molto attrattivo anche per giocatori di livello provenienti dall'estero.

1.1.2 La struttura del campionato

Sia il campionato di A1 femminile che quello maschile presentano prima una fase di **Regular Season**, seguita poi dai **Playoff Scudetto**, che determinano il club vincitore del titolo di Campione d'Italia per quella stagione.

La Regular Season si struttura come un girone all'italiana — in cui ogni squadra si scontra con tutte le altre — con gare di andata e ritorno. I risultati di queste partite determinano una classifica[1]che, al termine della Regular Season, decreta chi potrà partecipare ai Playoff ed eventuali retrocessioni in A2.

Tipicamente le prime otto classificate in Regular Season accedono ai Playoff, che si strutturano in quarti di finale, semifinali e finali, giocate non come gara unica ma al meglio delle tre o cinque partite. Non sempre questa formula viene replicata tutte le stagioni, anche se è la più standard: in certi casi l'accesso ai Playoff è permesso a più squadre, introducendo però una fase di ottavi di finale, oppure in alcune stagioni è stata sperimentata la finale a gara unica.

Inoltre, è bene specificare, che nel corso di una stagione sportiva, le squadre non si scontrano solamente in campionato ma possono partecipare anche ad altre coppe italiane ed europee, oltre che al campionato mondiale per club. Tuttavia, in seguito non considereremo queste altre competizioni,concentrandoci unicamente sui risultati ottenuti in campionato.

1.2 Obiettivi e dati

L'intenzione è quella di analizzare e descrivere l'appetibilità del campionato.In particolare si è scelto di guardare al fenomeno con l'approccio delle reti, per evidenziare non tanto caratteristiche e risultati individuali, quanto più la forza delle connessioni e dei legami, che rappresentano — tra l'altro — l'essenza dello sport di squadra.

I dati che si utilizzeranno sono stati tutti ricavati dai siti ufficiali dei campionati, maschile [8] e femminile [9], con un processo di scraping. Si è valutato di fare riferimento ad un periodo di 15 annate sportive, in modo da avere uno spettro sufficientemente ampio per rappresentare un numero adeguato di situazioni ma non eccessivamente grande da non permettere confronti a parità di condizioni.

Il codice R utilizzato si può trovare in Appendice A.

1.2.1 Scraping dei giocatori

Anzitutto si sono ricavate le informazioni relative agli atleti nel periodo considerato, ovvero le stagioni sportive dal 2008/2009 al 2022/2023. In particolare, per ogni giocatore si è tenuto traccia del suo nome e cognome, l'anno di nascita, la nazionalità sportiva, l'altezza in centimetri e il ruolo in campo. In certi casi, specie nel sito del campionato femminile, alcuni di questi campi risultavano mancanti o

[1]La classifica è stilata considerando in primis i punti ottenuti dalle squadre: ad ogni partita vengono assegnati 3 punti, nel caso di risultato di 3-0 o 3-1 questi sono assegnati tutti al vincitore e 0 allo sconfitto, invece nel caso di vittoria 3-2 il vincente ottiene 2 punti mentre la squadra perdente 1. Nell'eventualità di parità a livello di punti si considerano, nell'ordine, il numero di partite vinte, il rapporto tra set vinti e persi (quoziente set) e il rapporto tra punti realizzati e subiti (quoziente punti).

con informazioni contraddittorie riguardo lo stesso atleta: in queste occasioni si è fatto riferimento alla pagina personale del giocatore sul sito utilizzato, considerando l'informazione ivi riportata, se presente, come quella definitiva. Ciò ha permesso di ridurre il numero di osservazioni con dati mancanti a due sia per gli uomini (entrambe riguardanti l'altezza) che per le donne (una riguardo altezza e nazionalità, l'altra rispetto alla sola nazionalità).

A questo punto, si è proseguito con lo scraping della carriera dei giocatori nelle ultime 15 annate sportive. In particolare, le stagioni in cui gli atleti non hanno giocato, hanno giocato all'estero oppure in Italia ma in serie minori, sono stati codificati come NA. Per le stagioni in cui, invece, gli atleti hanno giocato nel massimo campionato italiano, si è tenuto traccia del nome della squadra. In certi casi si è rilevato di giocatori che hanno partecipato nello stesso anno a due squadre del campionato di A1 italiano: in queste situazioni si è conservata unicamente l'informazione sulla seconda di queste squadre. Infatti, per regolamento un atleta non può entrare in campo nello stesso campionato con due squadre diverse. Ciò significa che il giocatore in questione non era mai stato schierato in campo dalla squadra che compare per prima in quell'anno della sua carriera e dunque si è ritenuto di poter considerare quest'informazione come trascurabile.

1.2.2 Dataset e pulizia

Con le informazioni sin qui ricavate, si sono creati due dataset simmetrici, uno per gli atleti uomini e uno per le donne, strutturati in questo modo:

- cognome e nome del giocatore
- anno di nascita
- nazionalità sportiva
- altezza in centimetri
- ruolo giocato
- url della pagina personale del giocatore sul sito del campionato
- quindici variabili denominate 2008, 2009,... 2022 con indicazione del nome della squadra se in quell'anno il giocatore ha giocato in A1 in Italia, altrimenti NA

Il risultato è un dataset con 1020 osservazioni per gli uomini e uno con 985 per le donne, entrambi con 21 variabili.

A questo punto si sono applicate delle verifiche di correttezza, come ad esempio il controllo che il numero di squadre rilevato sui dati per ogni anno corrispondesse con l'effettivo numero di squadre presenti nel campionato (questo numero non è costante, varia di anno in anno, tipicamente tra 12 e 14). Inoltre, si è notato della presenza ripetuta di uno stesso atleta in quanto, contrariamente a tutti gli altri, dispone di due pagine a lui dedicate sul sito del campionato maschile: si sono accorpate le informazioni e ricondotto il dataset a 1019 unità osservate. In seguito si è proceduto a riportare alle stesse modalità la variabili corrispondenti al ruolo: nel dataset per

le atlete donne i ruoli erano declinati al femminile e presentavano anche la modalità "opposto", ovvero un ruolo che nel dataset degli uomini era ricompreso in quello più generale di "schiacciatore".

Successivamente si è ritenuto opportuno condurre una pulizia sui nomi delle squadre: questi comprendevano anche lo sponsor e potevano riferirsi alla stessa società anche non univocamente. Si è conservato, quindi, solo il nome della città di riferimento del club, con eventualmente qualche sigla per distinguere società distinte ma che, nel corso degli anni, si sono sviluppate nella stessa città.

L'intenzione legata a quest'ultima pulizia è di permettere il confronto non solo entro la stessa stagione, ma anche tra stagioni diverse a parità di club di appartenenza. Infatti, successivamente, si è proceduto anche a raccogliere alcune informazioni sulle società sportive e i risultati conseguiti nel campionato nel periodo di riferimento.

1.2.3 Scraping delle società

Utilizzando come punto di partenza i nomi dei club precedentemente puliti, si è associato ad ognuno il posizionamento in classifica al termine della stagione regolare e l'indicazione se avesse vinto poi il campionato, per tutti gli anni che la società ha partecipato al campionato di A1 tra le stagioni sportive 2008/2009 e 2022/2023.

I dataset così ottenuti presentano 31 variabili:

- nome del club
- quindici variabili denominate rs.2008, rs.2009,... rs.2022 contenti il posizionamento in Regular Season (NA se in quell'anno la società non ha partecipato al campionato)
- quindici variabili denominate primo.2008, primo.2009,... primo.2022 con indicazione se il club ha vinto o meno il campionato nell'anno (NA se la società non ha nemmeno partecipato)

Il numero di osservazioni nel dataset delle società maschili è 28, mentre per quelle femminili è 48.

È rilevante sottolineare che nel campionato femminile nelle stagioni 2011/2012 e 2012/2013, rispettivamente una e due società si sono ritirate a campionato iniziato: i risultati degli incontri disputati da queste sono stati annullati, dunque il loro posizionamento in classifica è stato salvato nel dataset come 99. Inoltre, in entrambi i campionati, la stagione 2019/2020 è stata interrotta per via dell'emergenza pandemica da COVID-19: non è stato assegnato il titolo e la classifica riportata è parziale, con un numero di partite giocate non esattamente uguale tra tutte le squadre (differiscono di una o due).

Capitolo 2

Metodi

2.1 Network analysis

Il termine "rete" fa riferimento ad un insieme di elementi e alle loro inter-relazioni [10]. In termini più matematici è possibile riferirsi alle reti come "grafi" $G = (V, E)$, composti da un insieme di nodi(o vertici V) e archi(E) . I nodi rappresentano gli elementi base del grafo, mentre gli archi sono i legami che intercorrono tra i vertici e dunque l'insieme E è caratterizzato come un insieme di coppie di nodi.

Un altro modo per rappresentare la struttura di una rete è, oltre agli insiemi di nodi e archi, quello delle matrici di adiacenza. Queste sono matrici quadrate la cui dimensione è pari al numero dei vertici e ogni elemento della matrice corrisponde al legame tra il nodo in riga e quello in colonna: se l'arco è assente nella matrice sarà riportato il valore 0, altrimenti il valore corrispondente al collegamento tra i nodi.

Infatti, sinora non abbiamo discusso le varie tipologie di rete:

- **Reti binarie:** due nodi sono o non sono connessi, non si definisce una forza del legame. La corrispondente matrice di adiacenza presenta come valori solamente 0 e 1.
- **Reti pesate:** esiste una misura del legame tra due nodi che definisce quanto sono connessi. La matrice può contenere anche valori diversi da 1 per indicare l'esistenza di un arco, mentre l'assenza è comunque rappresentata con lo 0.

Un'altra importante distinzione tra le tipologie di rete riguarda la direzionalità delle relazioni tra nodi.

- **Reti indirette:** non esiste una direzione degli archi, il legame è sempre bidirezionale. In questo caso la corrispondente matrice di adiacenza è simmetrica.
- **Reti dirette:** permettono archi unidirezionali e la matrice di adiacenza non risulta simmetrica.

2.1.1 Statistiche descrittive di rete

Essendo la rete un tipo di dato diverso da quello classico, è necessario introdurre degli strumenti per riuscire a valutarne la struttura e fare delle osservazioni.

Anzitutto esistono delle statistiche che caratterizzano i nodi:

- **Grado:** numero di archi incidenti su un nodo. Nel caso di grafi orientati si distinguono il grado entrante (numero di archi rivolti verso il nodo stesso) e il grado uscente (numero di archi che dal nodo si dirigono verso altri vertici).

- **Betweenness:** misura che rappresenta quanto un vertice sia posizionato tra altre coppie di nodi e dunque la sua centralità nella rete. Infatti sfrutta il concetto di *shortest path*, ovvero il percorso minimo tra due nodi, che viene calcolato tenendo in considerazione non solo il numero di archi necessari ma anche il loro peso. Nello specifico, la betweenness di un nodo i è definita come proporzione dei cammini minimi che passano dal vertice i ($n_{sv|i}$), rispetto al totale degli shortest paths tra due nodi s e v (n_{sv}), considerata per ogni coppia di nodi distinti e diversi da i nella rete.

$$c_B(i) = \sum_{s \neq v \neq i} \frac{n_{sv|i}}{n_{sv}}$$

Inoltre altre statistiche sono calcolabili a livello globale, sull'intera rete:

- **Densità:** esprime il rapporto tra numero di archi esistenti e possibili, nel caso in cui il grafo fosse completo (ovvero con ogni nodo collegato ad ogni altro). Posto $|V|$ il numero di vertici, in caso di una rete non diretta il numero di archi possibili è $|V|(|V|-1)/2$, mentre per una rete diretta è $|V|(|V|-1)$. Il valore della densità è compreso tra 0 (nessun collegamento) e 1 (grafo completo).

- **Diametro:** la lunghezza del percorso minimo più lungo.

- **Average shortest path:** indica, in media, quanto è lungo il percorso minimo tra due nodi.

2.1.2 Rappresentazione della rete

Un aspetto fondamentale nell'analisi di rete è la visualizzazione grafica della stessa. Infatti, semplicemente con una buona rappresentazione è possibile fare molte valutazioni.

In primis, il posizionamento dei nodi è molto importante. È possibile sfruttare layout diversi a seconda dell'esigenza di rappresentazione: un esempio è il layout circolare che, ordinati secondo qualche criterio i nodi, li dispone lungo una circonferenza e permette di vedere all'interno del cerchio tutti i collegamenti. Inoltre esistono algoritmi che determinano il posizionamento dei nodi unicamente in base alla forza delle connessioni nella rete: sono chiamati algoritmi di *Force Directed Placement*, presentano una componente stocastica e possono risultare molto informativi a fini interpretativi.

In seguito, se si dispone di informazioni ulteriori riguardo i vertici o gli archi, può essere utile sfruttarle per agevolare l'interpretabilità. Infatti, con colore, misura, forma dei nodi è possibile rappresentare le loro caratteristiche,. Questo permette di valutare la proprietà di *omofilia*, ovvero la tendenza dei vertici a creare legami con nodi in qualche modo a loro simili. L'utilizzo di spessori o colori diversi può essere applicato anche per la caratterizzazione degli archi.

Inoltre spesso è utile visualizzare anche le misure della rete stessa, come ad esempio il grado o la betweenness dei vertici o il peso degli archi: queste informazioni visive possono permettere, ad esempio, di riconoscere nella rete i nodi *Hub*, ovvero quelli centrali — cioè con una betweenness elevata— e con molte connessioni.

2.2 Modelli lineari (GLM)

I modelli lineari sono una classe di modelli di regressione molto conosciuta e utilizzata per via della sua semplicità di interpretazione. I modelli lineari generalizzati (GLM), in particolare, permettono la trattazione unificata di un ampio insieme di modelli di regressione, per diverse tipologia di variabile risposta, con procedure analoghe di inferenza sui parametri.

La specificazione dei GLM prevede l'assunzione che le osservazioni sulla risposta siano realizzazione di $Y_1, ..., Y_n$ variabili casuali indipendenti con distribuzione proveniente da una famiglia di dispersione esponenziale: questa comprende classi distributive quali la distribuzione normale, di Poisson, binomiale e gamma. Il valore atteso della risposta è modellato attraverso un predittore lineare che, per l'i-esima osservazione, risulta

$$\eta_i = x_i\beta = \sum_{j=1}^{p} x_{ij}\beta_j$$

per $i = 1, ...n$, dove x_i è il vettore riga delle esplicative non stocastiche e β il vettore dei coefficienti di regressione (di dimensione p). La funzione di legame $g(.)$ definisce il collegamento tra $E(Y_i) = \mu_i$ e il predittore lineare:

$$g(\mu_i) = \eta_i$$

Dunque, in base alla funzione di legame scelta, è possibile interpretare i coefficienti stimati dal modello come effetti lineari su $g(\mu_i)$. Per ciascuna specificazione della distribuzione della risposta, esiste una funzione di legame privilegiata detta canonica, in quanto semplifica le procedure inferenziali per la stima dei parametri. In particolare in seguito si farà ricorso ai modelli con risposta normale, dicotomica e di Poisson, la cui funzione di legame canonica è rispettivamente la funzione identità, logit ($f(x) = \log(\frac{x}{1-x})$) e logaritmo. [11]

Capitolo 3

Analisi

3.1 Analisi di rete dei giocatori

Il codice R di riferimento per le seguenti analisi si trova in Appendice B.

3.1.1 Creazione delle reti

Le reti sono state create separatamente per atleti uomini e donne. Si è stabilito che due giocatori fossero legati nel momento in cui avessero giocato almeno un anno nella stessa squadra: il legame così definito è biunivoco e, dunque, la rete prodotta è risultata indiretta. Inoltre si è scelto di lavorare con reti pesate, considerando il numero di stagioni condivise da ogni coppia di atleti. Si è valutato, però, che la maggior parte dei collegamenti avrebbe avuto lo stesso peso (vedi Figura 3.1) e dunque potesse risultare più informativo normalizzare il numero di stagioni passate nella stessa squadra rispetto alla media del numero di stagioni trascorse in Italia dai due giocatori nei 15 anni di riferimento. In questo modo risultano come più significativi i legami tra due atleti che hanno condiviso una buona parte dei loro anni — se non tutti — nel campionato italiano insieme e vengono penalizzati quelli con atleti che permangono nel campionato poco, solo una stagione sportiva. È bene osservare che ovviamente questa trasformazione comporta un cambio di dominio dei pesi: originariamente su scala discreta $\{0, 1, ...15\}$ (anche se il valore massimo osservato è 8), ora variano in modo continuo in $[0, 1]$.

Le connessioni così definite sono state riportate in matrici di adiacenza, aventi righe e colonne associate a degli identificativi dei giocatori. In seguito, attraverso il software Gephi, si è proceduto a visualizzare tali reti. L'algoritmo di layout utilizzato è "Force Atlas 2", un layout di tipo force-directed che tiene in considerazione anche il peso degli archi [12]. Come è possibile vedere in Figura 3.2, le reti con pesi non normalizzati risultano molto più concentrate: la dimensione dei nodi è in realtà la stessa in tutte le quattro rappresentazioni, semplicemente nei primi due casi i vertici sono più vicini e, occupando una porzione di spazio più piccola, risultano ingranditi. D'ora in avanti si porrà l'attenzione unicamente sulle reti con pesi normalizzati.

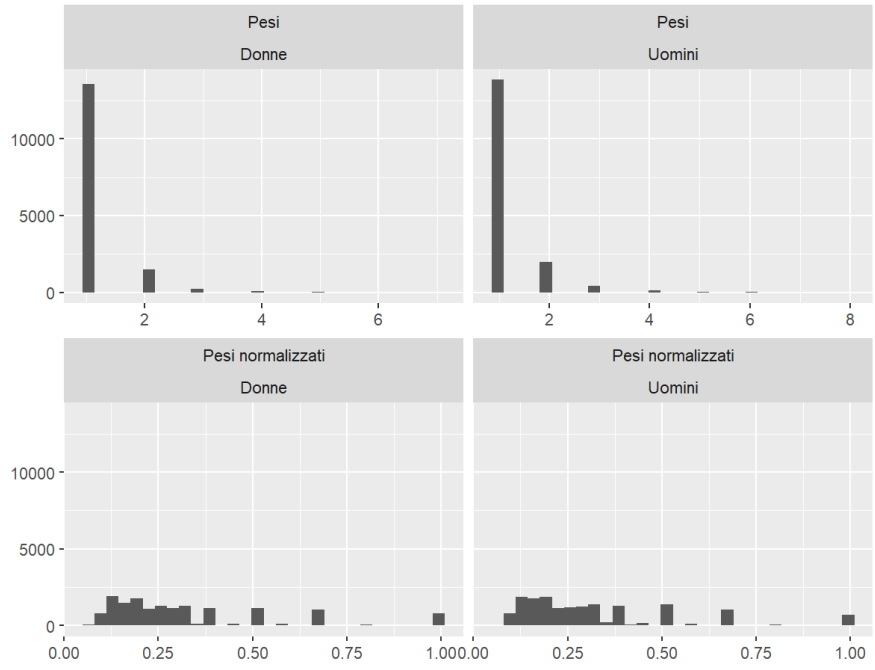


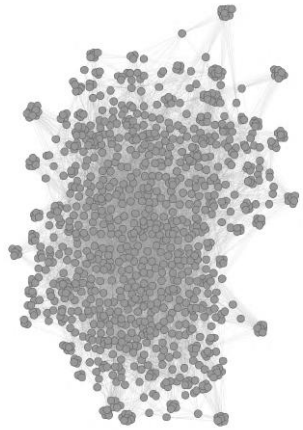
Figura 3.1: Distribuzione di frequenza dei pesi degli archi nella rete degli atleti uomini e in quella delle donne, utilizzando o meno la normalizzazione. Il peso nullo, ovvero l'assenza dell'arco, non è qui rappresentato.

3.1.2 Struttura delle reti

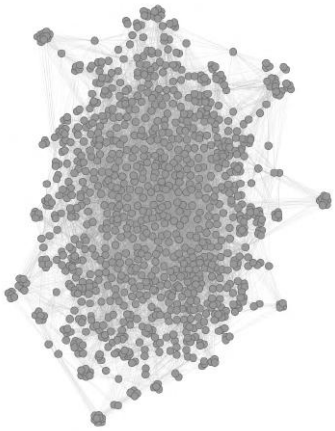
Per valutare la struttura delle reti ottenute, si procederà sfruttando le statistiche descrittive precedentemente introdotte alla Sezione 2.1.1. In primis si può notare (vedi Figura 3.3) la forte asimmetria che caratterizza il grado e la betweenness dei nodi. In particolare quest'ultima ha una proporzione di zeri sul totale che è pari all'82.7% nella rete degli atleti uomini e all'85.5% in quella delle donne. Ciò significa che la stragrande maggioranza dei vertici sono nodi non centrali rispetto alla rete, a dispetto del loro grado non necessariamente piccolo (vedi Figura 3.4). Infatti l'andamento di queste due statistiche non è lo stesso: sono molto pochi i nodi con betweenness elevata mentre molti di più quelli con grado più elevato. Inoltre queste due misure sono tendenzialmente concordi nell'assegnare importanza ad un nodo, ma non corrispondono esattamente: la correlazione tra grado e betweenness è 0.65 nella rete degli atleti uomini e 0.61 in quella delle donne.

Passando a considerare la densità, il valore osservato è molto basso (0.0317 in entrambe le reti) e ciò configura i grafi come sparsi. Tuttavia, la lunghezza del diametro e del percorso minimo medio hanno valori piccoli (vedi Tabella 3.1): anche se ogni nodo ha un numero limitato di connessioni ed è ben distante dall'essere connesso a tutti gli altri, con un breve serie di connessioni intermedie può raggiungere ogni altro vertice della rete. Questa condizione rispecchia una proprietà tipica delle reti detta *Small World*.

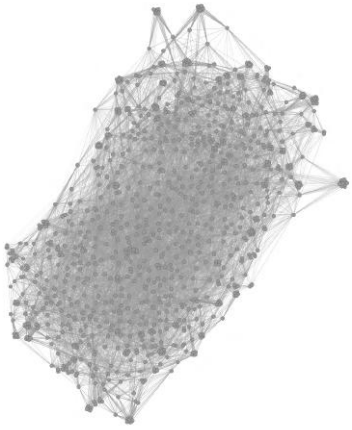
Dunque siamo in presenza di reti poco dense ma con alcuni importarti nodi centrali che per la loro caratteristica di fungere da "ponte" di legame tra nodi distanti sono detti *hub*.



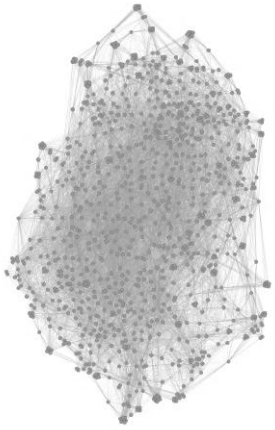
(a) Donne.



(b) Uomini.



(c) Donne con pesi normalizzati.



(d) Uomini con pesi normalizzati.

Figura 3.2: Reti dei due campionati con layout Force Atlas 2, sia con pesi normalizzati che non.

Tabella 3.1: Statistiche di connessione delle reti dei due campionati.

	Uomini	**Donne**
Densità	0.0317427	0.0317259
Diametro	0.8196673	0.8380952
Av. shortest path	0.3820421	0.3847095

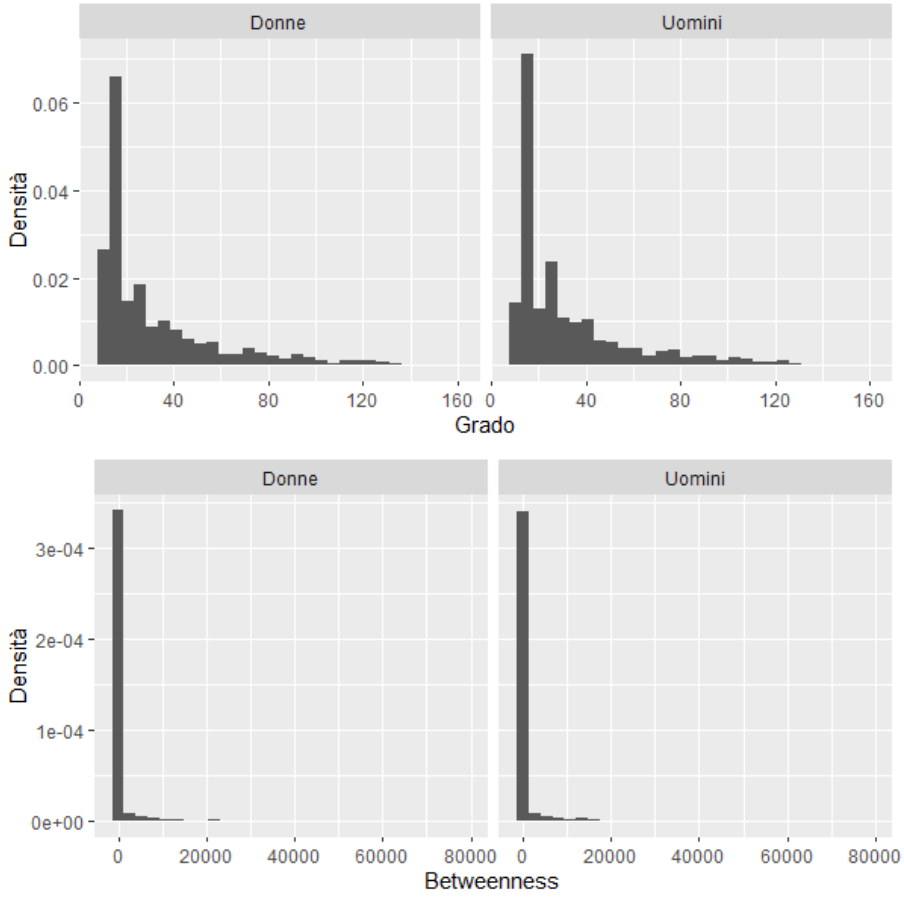


Figura 3.3: Distribuzione di grado e betweenness dei vertici nelle reti di giocatori e giocatrici.

(a) Donne. (b) Uomini.

Figura 3.4: Reti dei due campionati: la dimensione dei nodi rappresenta il grado, il colore la betweenness (il valore massimo è visualizzato in blu).

3.1.3 Relazione tra statistiche di rete e covariate

Con l'intenzione di interpretare la struttura della rete e chiarire quali sono le caratteristiche che descrivono al meglio la centralità o meno dei nodi, si sono sviluppati dei modelli di regressione per mettere in relazione le statistiche di rete — in particolare grado e betweenness — con le informazioni sui nodi. Come covariate nei modelli sono state inserite le seguenti variabili riferite ai giocatori:

- genere: *gen*
- ruolo : *ruolo*
- nazionalità (dicotomizzata in italiana o no): *ita*
- numero di stagioni nel campionato: *nstag*

La scelta di adattare un modello unico per le due reti nasce dall'intenzione di distinguere l'effetto marginale delle variabili sulle statistiche di rete da un effetto specifico e potenzialmente diverso nei due campionati. Per realizzare ciò, sono state considerate nel modello tutte le interazioni a due tra il genere del giocatore e le altre covariate.

La decisione di non inserire nei modelli, invece, le altre variabili a disposizione (anno di nascita e altezza) è motivata da ragioni distinte per le due. Nel primo caso, da una evidente relazione (vedi Figura 3.5) con il numero di stagioni in Italia: infatti, per questioni legate all'utilizzo di una finestra temporale per queste analisi, risulta che gli atleti che permangono per un numero elevato di stagioni si trovano tutti in fasce d'età centrali e non sono, invece, coloro che hanno più anni di esperienza, come sarebbe più logico. Essendo questo un chiaro bias e non ritenendo in ogni caso la variabile "anno di nascita" particolarmente importante allo scopo prefisso, si è optato per l'omissione di questa dai modelli. Per quanto riguarda l'altezza, invece, non si può considerare da sola come una proxy della bravura o della qualità dell'atleta, nemmeno a parità di ruolo e genere, dunque sul piano logico non ha senso considerarla. Inoltre, non sembra esserci nemmeno una relazione (lineare o quadratica) da un punto di vista statistico né con grado né con betweenness (vedi Figura 3.6), dunque si ritiene di poter escludere dai modelli anche questa covariata.

Modello per il grado

Si è assunto che le osservazioni sulla variabile *grado* siano realizzazioni di variabili aleatorie $Y_1, ...Y_{2004}$ indipendenti con distribuzione di Poisson con media $\mu_i, i = 1, ..., 2004$. Utilizzando la funzione di legame canonica per la regressione di Poisson, il modello è specificato come segue:

$$\begin{aligned}\log(\mu_i) = \beta_0 + \beta_1\ gen_i + \beta_2\ ruolo_i + \beta_3\ ita_i + \beta_4\ nstag_i + \\ +\beta_5\ gen_i\ ruolo_i + \beta_6\ gen_i\ ita_i + \beta_7\ gen_i\ nstag_i\end{aligned}$$

Si è proceduto alla selezione del modello sulla base del criterio di Akaike (AIC), rimuovendo così l'interazione di *gen* con *ita*. L'analisi dei residui di questo modello, però, non risulta soddisfacente: i residui mostrano un andamento quadratico, con

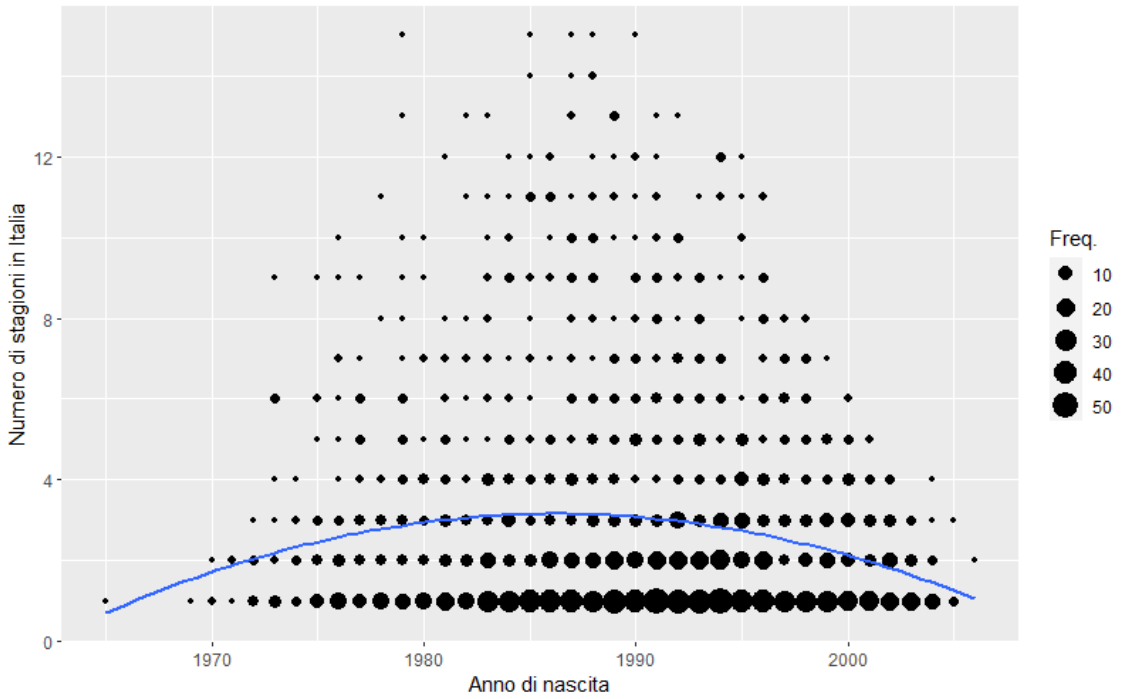


Figura 3.5: Grafico di dispersione dell'anno di nascita di giocatori e giocatrici rispetto al numero di stagioni di partecipazione al campionato italiano. La curva è stata stimata con una regressione polinomiale di grado 2. Il diametro dei punti rappresentati dipende dalla frequenza delle osservazioni coincidenti.

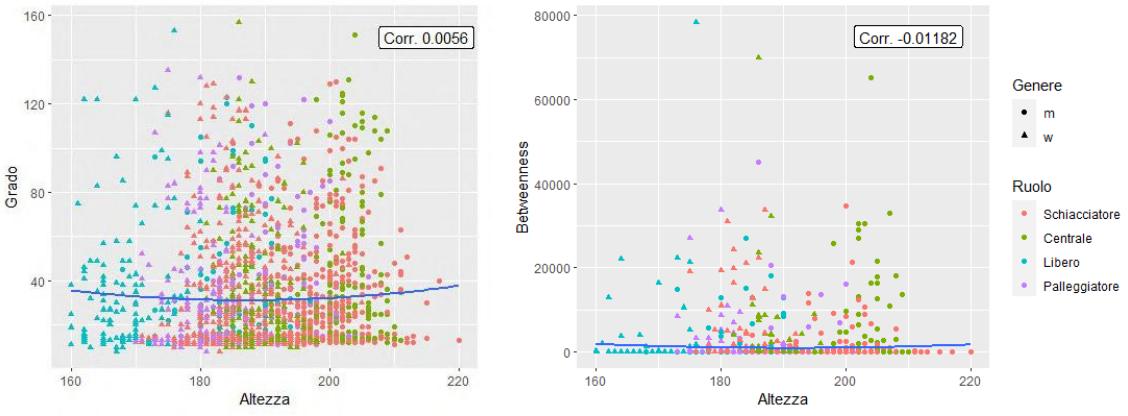


Figura 3.6: Grafico di dispersione dell'altezza del giocatore rispetto a corrispettivi grado (a sinistra) e betweenness (a destra). Le curve sono stimate con una regressione polinomiale di grado 2, mentre nel riquadro in alto a destra dei due grafici sono riportate le correlazioni tra le due coppie di variabili.

valori predetti che prima sovrastimano quelli osservati, poi li sottostimano, per poi sovrastimarli ancora. La ragione di questo andamento è che la funzione di legame utilizzata non rispecchia il legame esistente tra la risposta e le covariate, in particolare la variabile *nstag*. Infatti, questa relazione è sostanzialmente lineare, con una correlazione di 0.980, mentre il modello specificato prevede una relazione lineare tra la media della risposta e e^{nstag} (vedi Figura 3.7). Dunque si è proceduto ad una specificazione del modello che rispettasse questa linearità, utilizzando la funzione di legame identità.

$$\mu_i = \beta_0 + \beta_1\ gen_i + \beta_2\ ruolo_i + \beta_3\ ita_i + \beta_4\ nstag_i + \\ + \beta_5\ gen_i\ ruolo_i + \beta_6\ gen_i\ ita_i + \beta_7\ gen_i\ nstag_i$$

In questo caso, il rispetto della coerenza tra supporto del predittore lineare e spazio delle medie della risposta ($\mathbb{R}^+$) è garantito in quanto nessuna delle covariate assume valori negativi. Si è proceduto con una selezione delle variabili secondo AIC, giungendo al seguente modello finale:

$$\mu_i = \beta_0 + \beta_1\ gen_i + \beta_2\ nstag_i + \beta_3\ gen_i\ nstag_i$$

L'analisi dei residui mostra un adattamento ai dati abbastanza buono. Inoltre, l'utilizzo della funzione di legame identità rende questo modello di semplice interpretazione. I coefficienti stimati (vedi Tabella 3.2) mostrano che, all'aumento unitario del numero di stagioni, corrisponde un incremento di circa 10 sul grado: ciò significa che, in media, ogni anno in più che un atleta trascorre nel campionato, ha la possibilità di giocare con 10 persone nuove (per la precisione, 10.298 per il campionato femminile e 9.753 per quello maschile). La differenza evidenziata per atleti uomini e donne è dovuta alla significatività dell'interazione, che porta infatti a stimare rette distinte nei due campionati, non solo per quanto riguarda l'intercetta ma anche per il coefficiente angolare (vedi Figura 3.8). La variabile dicotomica *gen* ha come livello di riferimento il campionato maschile, dunque la retta stimata per questo risulta $grado = \beta_0 + \beta_2\ nstag = 4.56 + 9.75\ nstag$, mentre per il campionato femminile è $grado = (\beta_0 + \beta_1) + (\beta_2 + \beta_4)\ nstag = 3.39 + 10.30\ nstag$.

Modello per la betweenness

Nel caso della betweenness,come già mostrato in Figura 3.3, il numero di zeri è preponderante. Per questa ragione, si è preferito concentrare l'attenzione nel distinguere le caratteristiche di chi ha betweenness nulla rispetto a chi l'ha diversa da zero, non differenziando così i valori piccoli da quelli molto grandi: infatti, sebbene questo

Tabella 3.2: Coefficienti stimati del modello finale per il grado.

	Stima	Std. Error	z value	p value
β_0	4.54738	0.21734	20.923	< 2e-16
β_1	-1.17023	0.30735	-3.807	0.00014
β_2	9.75326	0.09295	104.932	< 2e-16
β_3	0.54013	0.13607	3.970	7.2e-05

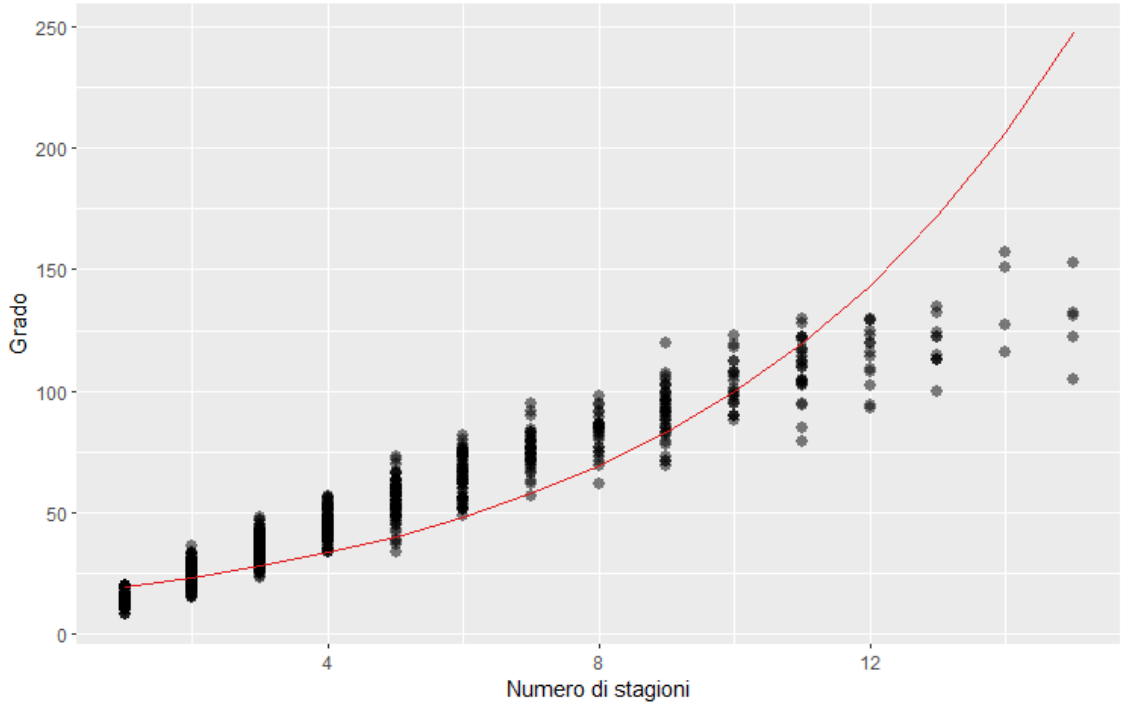


Figura 3.7: Grafico di dispersione del numero di stagioni di partecipazione al campionato italiano rispetto al grado del nodo della rete del corrispondente campionato. La curva in rosso è quella stimata dal modello di Poisson con legame canonico.

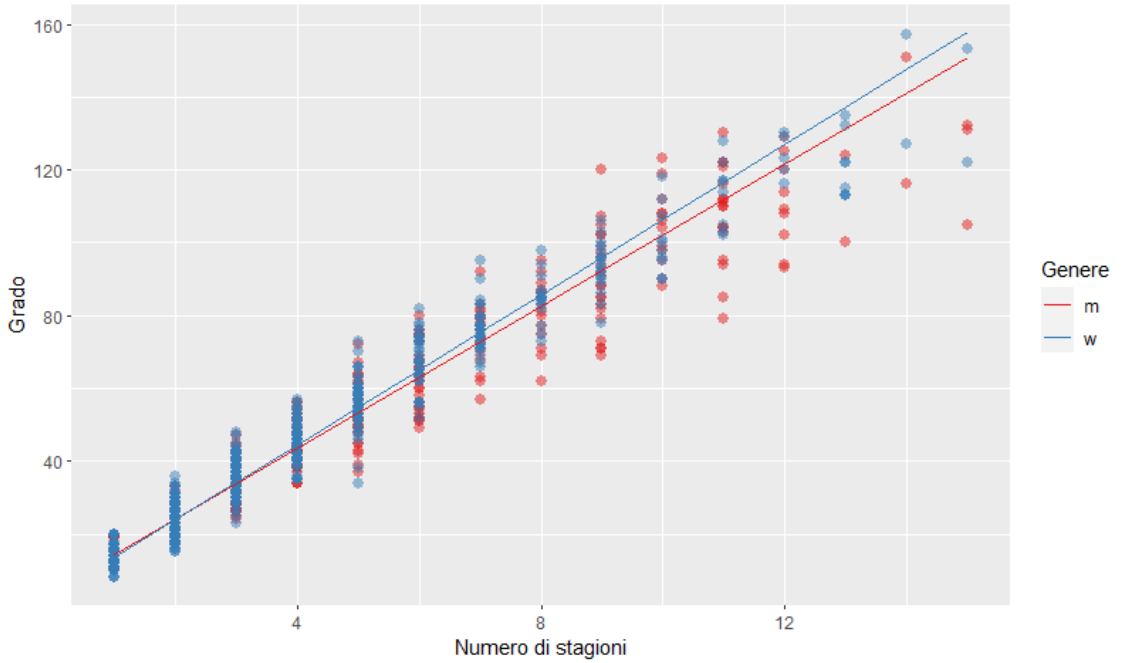


Figura 3.8: Curve stimate con modello di Poisson con legame identità per i due campionati sul grafico di dispersione dei dati.

approccio appiattisca il fenomeno ad una binarietà che non gli è propria e permetta di interpretarne solo una parte, si è preferito apportare questa semplificazione piuttosto che modellare l'intero andamento basando le stime su poche osservazioni.

Si è impiegato un modello di regressione logistica per valutare le caratteristiche degli atleti con betweenness positiva. Come risposta si è utilizzato la variabile $betw01$, ovvero la betweenness dicotomizzata in valori nulli (0) e valori positivi (1). Si sono assunte le osservazioni per $betw01$ come realizzazioni di variabili casuali indipendenti $Y_i \sim Bin(\pi_i), i = 1, ..., 2004$. La specificazione del modello completo è la seguente:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ gen_i + \beta_2\ ruolo_i + \beta_3\ ita_i + \beta_4\ nstag_i + \\ + \beta_5\ gen_i\ ruolo_i + \beta_6\ gen_i\ ita_i + \beta_7\ gen_i\ nstag_i$$

Dopo una selezione delle variabili sulla base dell'AIC, il modello finale risulta:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ gen_i + \beta_2\ nstag_i + \beta_3\ gen_i\ nstag_i$$

Il modello finale sembra avere un buon adattamento ai dati: la matrice di confusione, fissato $\pi_0 = 0.5$ il valore di soglia delle previsioni basate sul modello per distinguere le due classi, risulta quella in Tabella 3.4. L'accuratezza, ovvero la proporzione di osservazioni correttamente classificate, risulta 0.96657.

Da un punto di vista interpretativo, invece, il modello appena adattato sembra indicare che l'effetto principale sulla probabilità per un atleta di avere betweenness positiva e, dunque, acquisire un ruolo di centralità nella rete, sia dettato dal numero di stagioni di partecipazione al campionato. Ciò sembra sensato, anche in relazione a quanto visto con il grado: più annate sportive si passano nel campionato, più legami si creano, più realtà diverse si conoscono e quindi più è facile essere in grado di fungere da ponte tra gruppi distinti. È interessante, però, evidenziare il ruolo svolto dalla variabile *gen*: indica delle differenze di struttura nelle reti dei due campionati, non solo dal punto di vista marginale, ma anche in relazione al numero di stagioni, vista la significatività dell'interazione con *nstag* (vedi Figura 3.9). Sembra, infatti, che nel campionato femminile una betweenness pari a 0 sia più frequente e che quindi siano meno i nodi centrali nella rete. In compenso, si stima che la quota — ovvero il rapporto tra la probabilità che la betweenness sia maggiore di zero rispetto alla probabilità che sia nulla — sia moltiplicata per 16 ad ogni incremento unitario del numero di stagioni, mentre nel campionato maschile questo incremento è di un fattore 7. Dunque si può dire che nel campionato femminile è ancor più vero che chi ha betweenness positiva, ha transitato nel campionato per un buon numero di stagioni.

Tabella 3.3: Coefficienti stimati del modello finale per la betweenness dicotomizzata.

	Stima	Std. Error	z value	p value
β_0	-9.4134	0.8436	-11.159	$<$ 2e-16
β_1	-4.7498	2.0874	-2.275	0.0229
β_2	1.9238	0.1821	10.564	$<$ 2e-16
β_3	0.8509	0.4245	2.005	0.0450

Tabella 3.4: Matrice di confusione del modello finale per la betweenness dicotomizzata.

	$\pi_0 <= 0.5$	$\pi_0 > 0.5$
$Betw. = 0$	1657	28
$Betw. > 0$	39	280

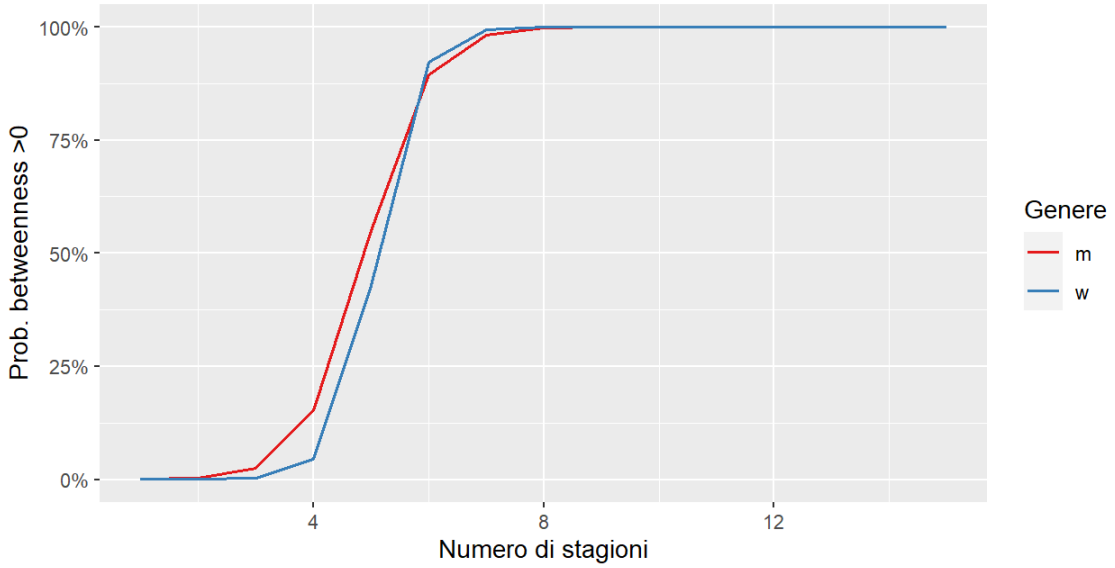


Figura 3.9: Curve di probabilità stimate dal modello logistico per la betweenness per i due campionati.

3.1.4 Visualizzazione delle reti

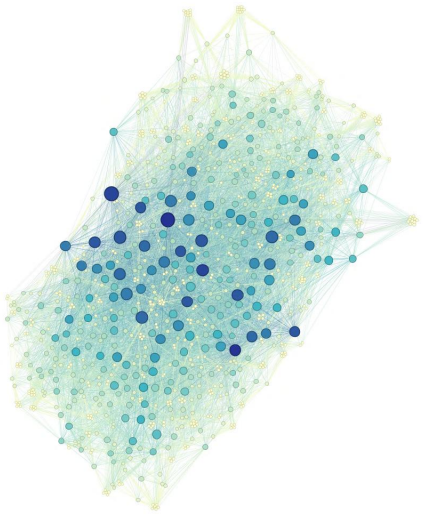
Date le considerazioni fatte con i modelli per grado e betweenness, si intende ora valutare anche visivamente le caratteristiche dei nodi centrali. Anzitutto si nota che, effettivamente, il numero di stagioni è una misura strettamente collegata alla centralità nella rete, come si vede in Figura 3.10: i nodi che presentano elevato grado e/o betweenness hanno una dimensione maggiore nella rappresentazione e sono sostanzialmente tutti di un blu intenso — che sta a significare un alto numero di stagioni di partecipazione al campionato corrispondente — mentre tutti i nodi gialli sono anche di piccole dimensioni.

Visualizzando, successivamente, la rete con le altre covariate, si nota che molte non hanno un andamento chiaro: ad esempio, il ruolo del giocatore (vedi Figura 3.11) non definisce gruppi distinti nella rete — d'altronde non avrebbe senso, i componenti di una squadra necessariamente hanno caratteristiche e quindi ruoli diversi — né risulta una particolare preponderanza di un ruolo nei nodi centrali, con numero di anni di partecipazione al campionato più elevato. Tuttavia, una caratteristica prevalente nei nodi centrali è la nazionalità italiana, come si vede in Figura 3.12. Ciò significa che, anche se il numero di atleti stranieri nel campionato è notevole (la proporzione di stranieri a stagione, in media nei 15 anni considerati, è 40.06% nel campionato maschile e 35.50% in quello femminile), di questi sono pochissimi quelli che decidono di trascorrervi un numero di anni elevato, come si può osservare nel diagramma a barre in Figura 3.13.

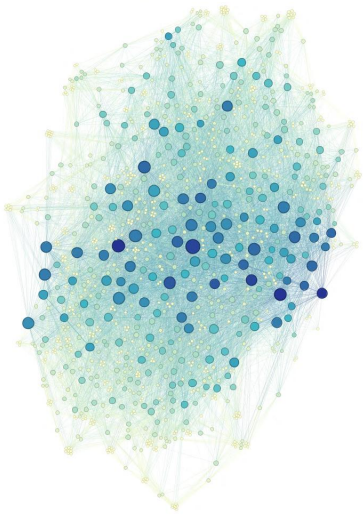
Dalla visualizzazione delle reti si nota, inoltre, che entrambe assumono una forma allungata, ad indicare il tentativo dell'algoritmo di posizionamento dei nodi di distanziare il più possibile i vertici alle estremità opposte, pur mantenendoli legati alla rete. Questa forma suggerisce quindi un andamento che, in questo caso, è dettato dallo scorrere del tempo: gli atleti che hanno giocato nel massimo campionato italiano nei primi anni del periodo considerato avranno meno collegamenti con coloro che vi hanno partecipato negli ultimi anni. Ciò è stato verificato anche suddividendo le quindici stagioni considerate in due periodi di otto annate sportive (le stagioni 2008/09 - 2015/16 e 2015/16 - 2022/23) e assegnando i giocatori al periodo in cui hanno prevalentemente giocato (i pochi che hanno partecipato esattamente allo stesso numero di stagioni nei due periodi sono stati assegnati ad una terza categoria). Osservando le reti con questa suddivisione (vedi Figura 3.14) si nota una forte omofilia, con gruppi quasi nettamente distinti e che dividono a metà l'ellisse in lunghezza, confermando l'interpretazione temporale della forma della rete. Un andamento simile, anche se meno netto, si nota visualizzando inoltre l'anno di nascita degli atleti (vedi Figura 3.15), che con il procedere delle annate sportive è mediamente aumentato. La distinzione meno decisa è dovuta alla fisiologica presenza contemporanea di atleti di età diverse entro la stessa squadra: si possono notare, infatti, sia atleti più giovani all'interno del gruppo di nodi degli atleti più esperti, che il contrario.

3.1.5 Sviluppo temporale delle reti

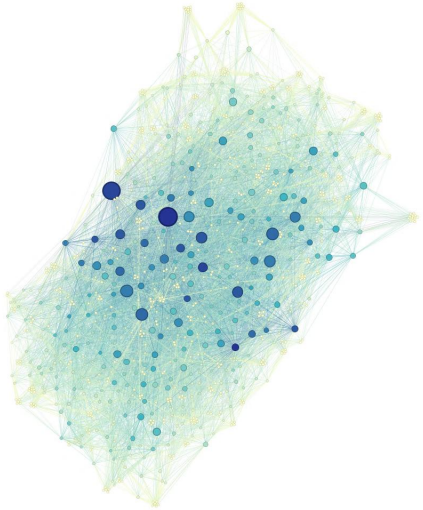
Data la dimensione temporale legata alle reti dei giocatori, si è pensato di creare anche delle reti delle connessioni per sotto-periodi dei 15 anni considerati e valutar-



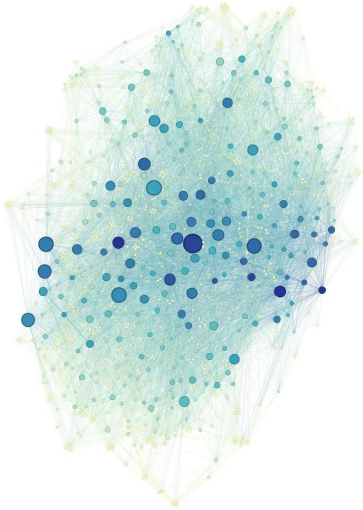
(a) Donne (grado).



(b) Uomini (grado).



(c) Donne (betweenness).



(d) Uomini (betweenness).

Figura 3.10: Reti dei due campionati: il colore dei nodi rappresenta il numero di stagioni di partecipazione per ciascun giocatore (il valore massimo è rappresentato in blu), la dimensione è proporzionale alternativamente al grado (in alto) e alla betweenness (in basso).

(a) Donne. (b) Uomini.

Figura 3.11: Reti dei due campionati: il colore dei nodi rappresenta il ruolo, la dimensione è proporzionale al numero di stagioni di partecipazione al campionato.

(a) Donne. (b) Uomini.

Figura 3.12: Reti dei due campionati: gli atleti con nazionalità italiana sono rappresentati in verde e gli altri in rosa, mentre la dimensione è proporzionale al numero di stagioni di partecipazione al campionato.

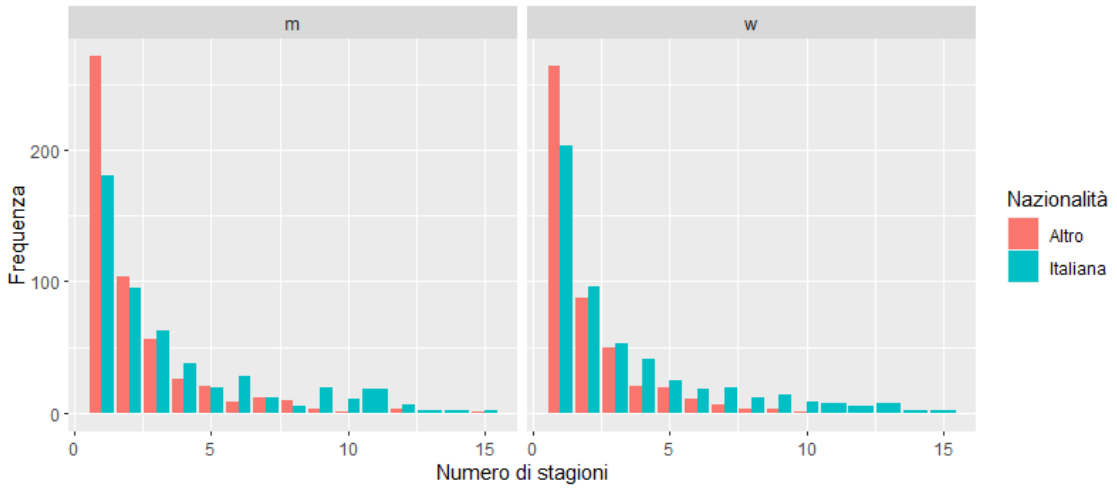


Figura 3.13: Diagramma a barre di frequenza affiancate del numero di stagioni per nazionalità — italiana o meno — degli atleti. A sinistra il campionato maschile, a destra quello femminile.

(a) Donne. (b) Uomini.

Figura 3.14: Reti dei due campionati: in azzurro il periodo 2008/09 - 2015/16 e in rosa 2015/16 - 2022/23, in viola il gruppo intermedio. La dimensione dei nodi è proporzionale al numero di anni nel rispettivo campionato.

(a) Donne. (b) Uomini.

Figura 3.15: Reti dei due campionati: il colore rappresenta l'anno di nascita dei giocatori (in rosso i più vecchi, in blu i più giovani). La dimensione dei nodi è proporzionale al numero di anni nel rispettivo campionato.

ne l'evoluzione. In particolare, si è utilizzata una finestra temporale di 5 anni che, traslata di 1 anno ad ogni passo rispetto alle 15 stagioni considerate, ha definito 11 sotto-periodi in sequenza, ovvero 2008/09 — 2012/13, 2009/10 — 2013/14, ..., 2018-19 — 2022/23. Per ognuno di questi, si sono create le corrispondenti reti dei collegamenti tra i giocatori per i due campionati. La definizione di legame utilizzata è del tutto analoga alle reti trattate sinora, con la differenza che il peso degli archi qui corrisponde, sì, al rapporto tra il numero di stagioni condivise dai due giocatori e la media del numero di stagioni di partecipazione al campionato dei due, ma entro la finestra temporale piuttosto che sull'intero periodo di 15 anni. Inoltre, per permettere la confrontabilità, nelle reti appena definite sono stati mantenuti tutti i giocatori delle reti principali, anche se non presentano collegamenti in quanto in quella finestra temporale non hanno partecipato al campionato. Ciò va ovviamente a incidere sulle statistiche di rete, con una sovrabbondanza di nodi con grado e betweenness nulli. Infatti, si può notare che tutte le reti per le varie finestre temporali presentano una distribuzione di grado e betweenness con mediana 0 (vedi Figura 3.16), con l'unica eccezione del grado dell'ultima finestra temporale per il campionato femminile: questa eccezione è dovuta ad una minor proporzione di punti isolati nella rete rispetto alle finestre precedenti, per via di un aumento del numero di squadre nel campionato e dunque un maggiore numero di atlete. La distribuzione della betweenness presenta in tutti i casi, addirittura, terzo quartile nullo: infatti, come già visto per le reti principali, questa statistica ha una distribuzione di per sé molto asimmetrica e con un'elevata proporzione di zeri che, in questo modo, è ulteriormente incrementata. In generale, però, non si nota un particolare andamento di queste statistiche nelle diverse reti. L'unica differenza che si può effettivamente notare è che nelle reti del campionato maschile sono molti di più i nodi con betweenness alta — ad esempio maggiore di 6000 — rispetto a quello femminile.

Figura 3.16: Boxplot di grado (in alto) e betweenness (in basso) per le reti con finestra temporale di 5 anni, suddivise per campionato.

Passando ora alla visualizzazione di queste reti, si è pensato di sfruttare un layout circolare piuttosto che un layout force-directed. Questa scelta rende più difficile la valutazione di una singola rete, non mettendo in evidenza i nodi centrali, ma favorisce il confronto nel tempo, dal momento che il posizionamento dei nodi è mantenuto costante. In particolare, l'ordine di disposizione dei nodi sul cerchio è stato determinato in base al primo anno di partecipazione al campionato italiano dei giocatori, nelle 15 stagioni prese qui a riferimento. Si possono osservare le reti rispettivamente in Appendice D.3 per il campionato maschile e in Appendice D.4 per quello femminile: i nodi rappresentati hanno dimensione pari al grado e il colore divide in gruppi secondo l'anno di prima partecipazione al campionato. Visivamente si nota subito che, sia per gli uomini che per le donne, il gruppo la cui prima stagione in Italia sarebbe 2008/09 è più cospicuo degli altri, rappresentato da un arco di circonferenza più lungo: ciò è dovuto al bias per cui tutti coloro che hanno partecipato a quella stagione confluiscono in quel gruppo, essendo il primo degli anni considerati. I gruppi successivi presentano, infatti, ampiezze confrontabili, con poche differenze che sembrano determinate dai fluttuamenti del numero di squadre che competono in campionato nelle varie stagioni. Si nota come, per ogni finestra temporale, il grado degli ultimi arrivati sia sostanzialmente lo stesso per tutti: hanno partecipato solo ad una stagione, l'ultima considerata nella finestra di 5 anni, dunque hanno un numero di compagni di squadra pressoché uguale. Diverso è, invece, per coloro che hanno una prima partecipazione precedente, specialmente se addirittura antecedente alla finestra temporale considerata: il grado può variare di più, in funzione principalmente del numero di stagioni nel campionato nella finestra. Si nota, inoltre, con l'evolversi dei periodi, una progressiva diminuzione nel numero di giocatori nei gruppi riferiti alle prime stagioni, tuttavia quelli che permangono sembrano avere tipicamente grado elevato, a rappresentare l'assidua presenza in campionato. Il fatto di considerare 5 anni rende, però, difficile osservare la permanenza continua nel campionato: infatti, periodi di gioco altrove di durata inferiore ai 5 anni comportano solo una riduzione del grado del giocatore rispetto al quello che in media corrisponde a 5 stagioni, ma non ad un grado zero. Infatti sono pochi i nodi che si vedono scomparire e riapparire con l'evolversi del tempo, dal momento che il numero di anni tra le varie stagioni in Italia dovrebbe essere maggiore di 5. Tuttavia, il fenomeno "boomerang" di atleti che hanno giocato nel massimo campionato italiano e, dopo un periodo, vi ritornano è molto comune e d'interessante approfondimento.

3.1.6 Permanenza nel campionato

La permanenza nel campionato sinora è stata descritta con la variabile "Numero di stagioni", a cui si è fatto riferimento specialmente per la sua importanza come predittore della centralità dei nodi nelle reti. Tuttavia questa non descrive interamente il fenomeno, in quanto un atleta potrebbe svolgere quegli anni in maniera continuativa o meno. Anzitutto è necessario sottolineare che il 44.36% degli atleti e il 47.51% delle atlete ha svolto nel massimo campionato italiano solamente una stagione. Tuttavia, considerando coloro che vi hanno transitato per almeno due stagioni, si può stabilire che una buona parte di questi non ha svolto il proprio numero di stagioni nel campionato in modo continuativo: il 41.80% degli uomini e il 52.80% delle donne.

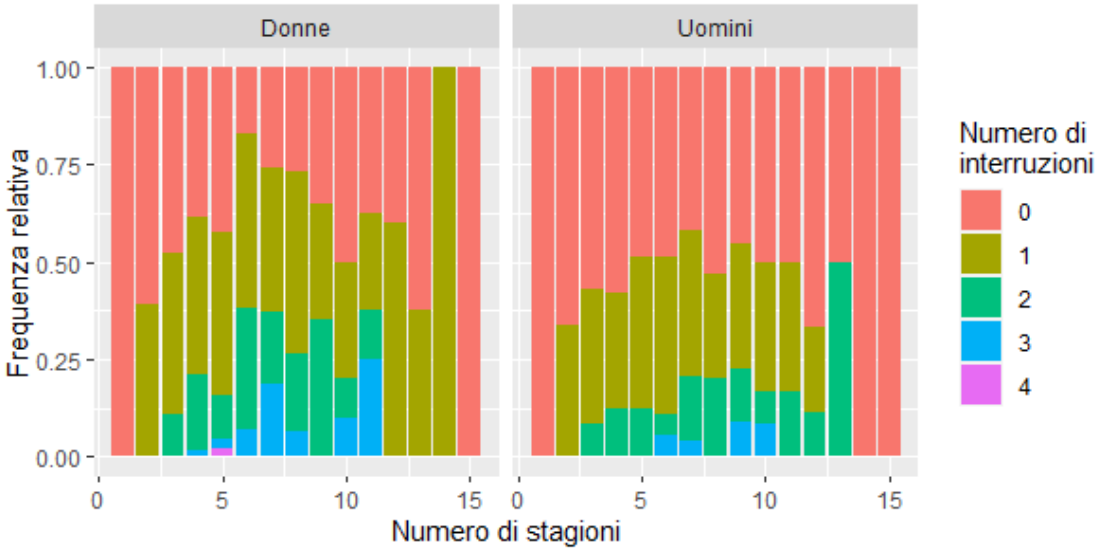


Figura 3.17: Grafico a barre sovrapposte della proporzione del numero delle interruzioni alla partecipazione al campionato di A1 italiano rispetto al numero di stagioni di partecipazione, suddiviso per genere.

Questa proporzione è sicuramente legata al numero complessivo di stagioni svolte (vedi Figura 3.17): sembra avere un andamento prima crescente e poi decrescente in entrambi i campionati, ma nel caso femminile risulta generalmente più alta che in quello maschile. Infatti, sembra emergere che, nonostante una distribuzione analoga del numero di stagioni per uomini e donne, i primi tendano ad una presenza più continuativa mentre le seconde siano un po' più soggette al fenomeno "boomerang". In ogni caso, quest'ultima è una caratterizzazione di entrambe le realtà che, aggiungendosi ad un gran numero di atleti presenti solo per una stagione, determina un campionato in costante rinnovamento. Infatti, in media circa il 40% dei giocatori di una determinata stagione, non erano presenti la stagione precedente, sia per il campionato maschile che femminile.

3.2 Analisi delle società sportive

Il focus passa ora dai giocatori nel campionato alle società sportive nel campionato: la permanenza di queste nel campionato e dei giocatori nelle stesse, i risultati ottenuti e il legame tra risultati, stabilità della squadra e caratteristiche degli atleti.

3.2.1 Permanenza delle società nel campionato

Prima di valutare la permanenza dei giocatori nei club, è d'obbligo soffermarsi sulla permanenza dei club stessi nel campionato, in quanto questo necessariamente comporta delle implicazioni per gli atleti. Ciò è specialmente importante in quanto si evidenzia una forte differenza tra i due campionati, che si può notare anche semplicemente considerando il numero di club distinti che hanno partecipato alle 15 stagioni sportive in esame: in totale 28 per il campionato maschile e 48 per quello femminile.

Infatti, concentrando anzitutto l'attenzione sul numero totale di stagioni in serie A1 per i vari club, si nota che la distribuzione è molto diversa per i due campionati (vedi Figura 3.18): la distribuzione per il numero di stagioni dei club femminili è molto asimmetrica, con molte osservazioni basse e poche elevate, mentre quella dei club maschili presenta anche un buon numero di osservazioni elevate. Ciò emerge anche dal confronto delle statistiche delle due distribuzioni (vedi Tabella 3.5): la media nel caso femminile è nettamente più bassa, ma soprattutto la moda è 1, con circa un terzo delle osservazioni, mentre per gli uomini la moda è 4. Ciò significa che, in generale, sembra esserci un maggiore ricambio delle società sportive femminili, mentre quelle maschili sembrano avere maggior stabilità.

In ogni caso, come precedentemente visto con gli atleti, il numero di stagioni non descrive in toto il fenomeno della permanenza, bisogna anche considerare come queste stagioni sportive in serie A1 sono collocate nel tempo. In Figura 3.19 si è rappresentato in quali anni i vari club hanno partecipato al massimo campionato. Per quanto si possano notare dei casi sia di club maschili che femminili in cui la partecipazione è stata interrotta per qualche anno e poi ripresa, la maggior parte di questi sembra permanere in modo continuativo in serie A1 per alcuni anni e poi non riuscire a ripetere più l'esperienza della massima serie. Ciò implica che, nel valutare la permanenza degli atleti, sarà necessario tener conto solamente del numero di stagioni del club nel rispettivo campionato.

3.2.2 Permanenza e spostamenti dei giocatori nei club

Per rappresentare la consistenza degli spostamenti tra le società sportive del campionato e il numero di atleti che, invece, di volta in volta sono rimasti nello stesso club per più stagioni, si sono considerati i passaggi tra le varie annate sportive — per i 15 anni, dunque, sono 14 — e si sono utilizzate delle matrici di adiacenza per i club per rappresentare il numero di giocatori che passavano da una società all'altra entro i campionati di A1 italiani e, conseguentemente, anche quelli che rimanevano nella stessa squadra.

Con queste matrici si sono costruite delle reti, dirette e pesate: ognuna rappresenta gli scambi di giocatori tra le società tra due stagioni. La visualizzazione delle reti si può osservare in Appendice D.3 per il campionato maschile e in Appendice D.4 per quello femminile. La rappresentazione che si è fatto delle reti utilizza il numero di atleti rimasti nella società come dimensione dei nodi, mentre la classifica al termine della Regular Season della stagione passata come colore (rosso - bassa classifica, giallo - media classifica, blu - alta classifica, grigio - non ha partecipato). Nelle stagioni 2011/12 e 2012/13 nel campionato femminile si sono ritirate prima del

Tabella 3.5: Statistiche descrittive della distribuzione del numero di stagioni in serie A1 dei club per le annate sportive 2008/09 — 2022/23, suddivisi per campionato maschile e femminile.

	Media	**Mediana**	**Moda**	**Varianza**
Uomini	7.07	4	4	23.42
Donne	3.96	3	1	12.12

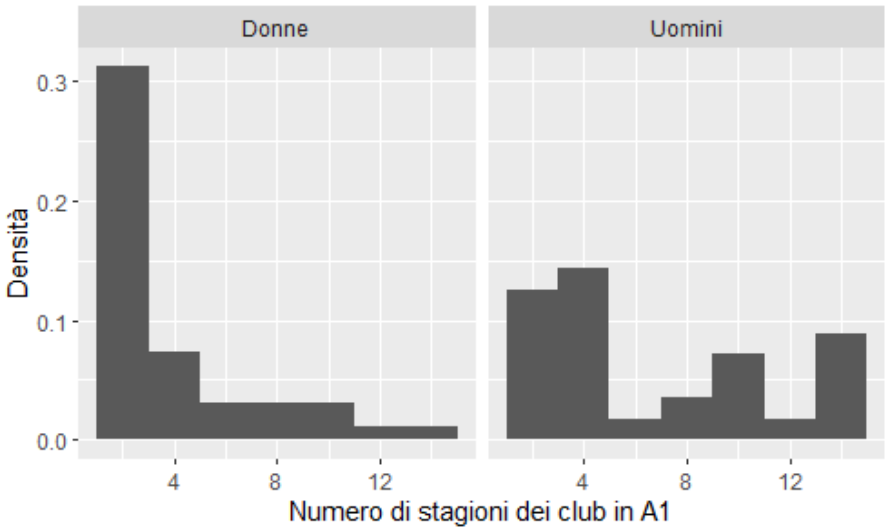


Figura 3.18: Istogramma di densità di frequenza del numero di stagioni delle società sportive in A1 nelle stagioni 2008/09 — 2022/23, distintamente per i due campionati.

termine del campionato rispettivamente una e due società: queste non compaiono nella classifica finale, essendo stati annullati tutti i risultati degli scontri disputati con esse, quindi anche questi casi sono rappresentati in grigio. Inoltre, il nome del vincitore del campionato precedente è in viola. Si può notare che nella stagione 2019/20 nessuno dei nomi dei club dei due campionati è in viola: per l'avvento della pandemia da COVID-19, il campionato è stato interrotto prima del termine della Regular Season (la classifica considerata è, dunque, parziale) e non è stato assegnato il titolo di Campione d'Italia. Il layout utilizzato nella visualizzazione è di tipo force-directed e tiene conto anche del peso dei collegamenti tra nodi. Gli archi sono stati rappresentati curvi per evitare sovrapposizioni di frecce con verso opposto che congiungono coppie di società e, per facilitare l'interpretazione, hanno il medesimo colore del nodo sorgente. Lo spessore delle frecce rappresenta il numero di giocatori traferitisi da una squadra all'altra: tipicamente la maggior parte degli archi hanno peso 1 o 2.

Se si osserva la direzione degli archi, nella maggior parte dei casi vanno da squadre di più bassa classifica a squadre di più alta classifica, infatti sono pochissime quelle di colore blu. In ogni caso qui non sono stati considerati gli acquisti al di fuori del campionato, che rappresentano buona parte dei movimenti in entrata e in uscita per i club.

Da questa rappresentazione si nota anche che tipicamente i nodi più grandi (con più atleti rimasti) sono blu, ovvero riferiti a squadre con un buon posizionamento in classifica, anche se, specialmente nel campionato maschile, in certi casi si può notare una certa stabilità anche dei club in rosso, tanto che nelle reti riferite ai periodi 2018/19 — 2019/20 e 2020/21 — 2021/22 sono proprio queste ad avere dimensione maggiore. Sembra esserci, infatti, una relazione tra risultati delle squadre e il numero di atleti che vi rimangono. Considerando la media dei risultati tra le varie stagioni di partecipazione e il numero medio di atleti rimasti a stagione, si ottiene per entrambi i campionati un diagramma di dispersione sostanzialmente lineare, descritto da una

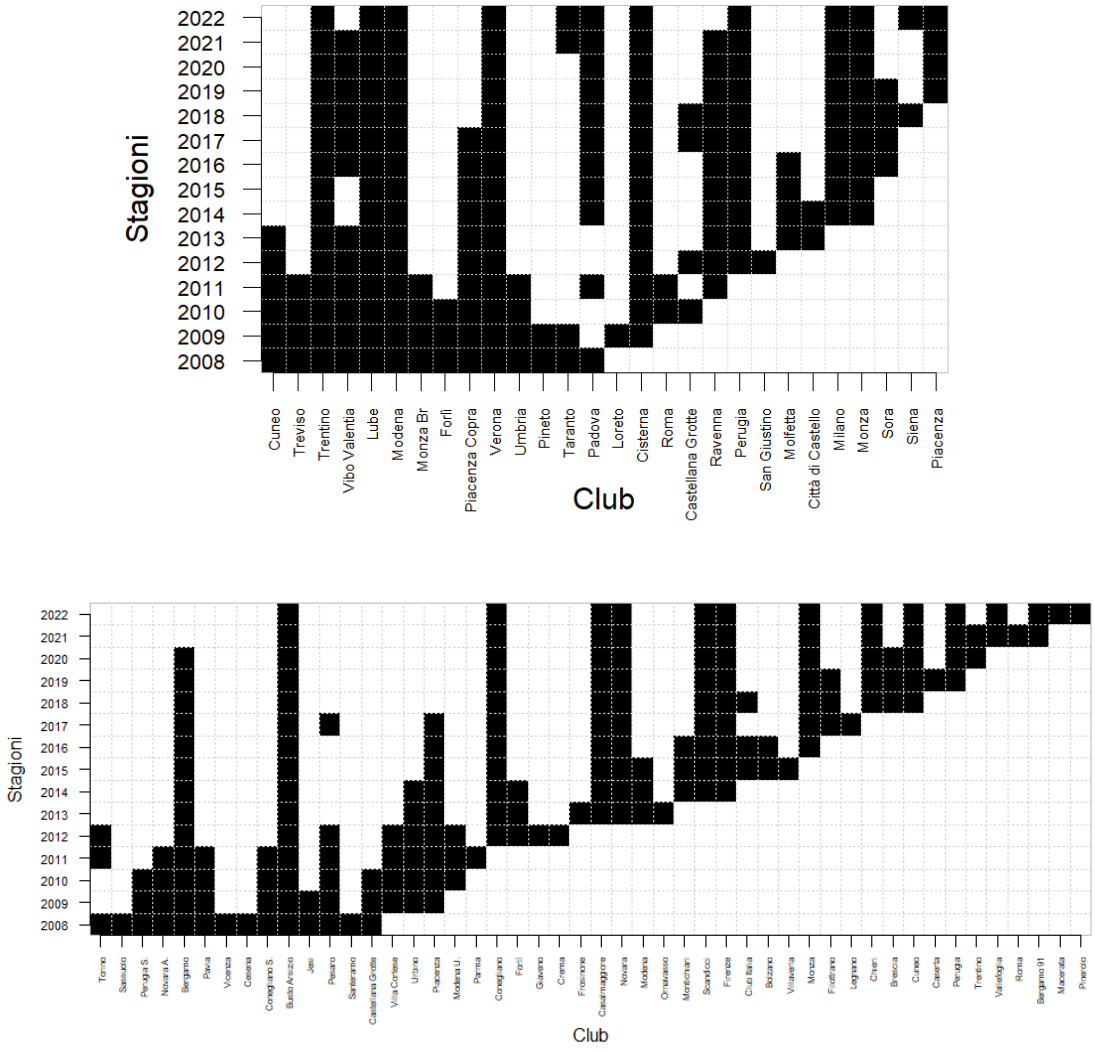


Figura 3.19: Rappresentazione della presenza delle varie società sportive nelle stagioni di riferimento, distintamente per il campionato maschile (in alto) e femminile (in basso).

correlazione di -0.8075 per il campionato maschile e di -0.7372 per quello femminile: passando da un posizionamento migliore ad uno peggiore, diminuisce in media il numero di atleti che rimangono nella società di anno in anno. Queste relazioni tra risultati e composizione delle squadre verranno approfonditi nel paragrafo seguente con la specificazione e l'adattamento di un modello di regressione.

3.2.3 Relazione tra risultati e struttura della squadra

Per esplorare le caratteristiche di composizione della squadra che più influiscono sui risultati in campionato, si è specificato un modello lineare normale. In particolare, si è messo in relazione il risultato in Regular Season (rs) dei club in una stagione sportiva con:

- l'età media della squadra: $m.eta$
- la proporzione di giocatori in squadra con la nazionalità italiana: $p.ita$
- la media del grado dei giocatori della squadra nella rete di 15 annate sportive: $m.grado$
- la media della betweenness dei giocatori della squadra nella rete di 15 annate sportive: $m.betw$
- l'indicazione se il club ha partecipato alla precedente stagione sportiva: $prec01$
- il risultato in Regular Season della stagione precedente (se in A1): $rs.prec$
- il numero di atleti rimasti nel club dalla stagione precedente (se in A1): $rimasti$

In particolare, si sono considerate le stagioni 2009/10 — 2022/23 per riuscire a valutare per ciascuna di esse i risultati dell'annata precedente. Inoltre, in questo caso, si è preferito adattare due modelli distinti per i campionati, seppur con le stesse variabili, per valutarli separatamente.

Si è assunto che le osservazioni sulla variabile rs per i vari anni e per i vari club fossero realizzazioni di $Y_1, ..., Y_n$ variabili casuali indipendenti con $n = 184$ per il campionato maschile e $n = 173$ per quello femminile (le tre osservazioni riferite a club ritirati a campionato in corso sono state rimosse). Come detto, si è assunta anche la normalità della risposta: $Y_i \sim N(\mu_i, \sigma^2), i = 1, ..., n$. Il modello specificato risulta dunque:

$$\begin{aligned}\mu_i = \beta_0 + \beta_1\ m.eta_i + \beta_2\ p.ita_i + \beta_3\ m.grado_i + \beta_4\ m.betw_i + \\ + prec01_i\ (\alpha_0 + \alpha_1\ rs.prec_i + \alpha_2\ rimasti_i)\end{aligned}$$

In particolare, dal punto di vista della notazione, si sono distinti i coefficienti in β e α in quanto i primi fanno riferimento alle caratteristiche di composizione della squadra della stagione considerata, mentre gli altri si riferiscono ai risultati in campionato nella stagione precedente per i club che vi hanno partecipato. Infatti, le variabili $prec$ e $rimasti$ sono presenti nel modello solo in relazione a $prec01$ e non a sé stanti.

I coefficienti stimati si trovano in Tabella 3.6 per il campionato femminile e in Tabella 3.7 per quello maschile. L'adattamento dei due modelli risulta soddisfacente,

Tabella 3.6: Coefficienti stimati del modello per il risultato in Regular Season per il campionato femminile.

	Stima	Std. Error	t value	p value
β_0	18.1795578	3.1232871	5.821	2.98e-08
β_1	-0.1553840	0.0963588	-1.613	0.108752
β_2	1.0660257	1.8668984	0.571	0.568767
β_3	-0.1233206	0.0216719	-5.690	5.66e-08
β_4	0.0001506	0.0000907	1.661	0.098589
α_0	-3.4905028	0.9877224	-3.534	0.000531
α_1	0.4222055	0.0732394	5.765	3.93e-08
α_2	-0.2575732	0.1058724	-2.433	0.016048

Tabella 3.7: Coefficienti stimati del modello per il risultato in Regular Season per il campionato maschile.

	Stima	Std. Error	t value	p value
β_0	12.5527543	3.3313563	3.768	0.000224
β_1	0.1093012	0.1063398	1.028	0.305432
β_2	1.0791317	2.0707525	0.521	0.602931
β_3	-0.1186377	0.0222661	-5.328	3.01e-07
β_4	0.0001366	0.0001027	1.330	0.185379
α_0	-4.8438045	1.1090637	-4.367	2.14e-05
α_1	0.4909581	0.0669054	7.338	7.66e-12
α_2	-0.2370759	0.0958719	-2.473	0.014354

con residui normali e senza andamenti sistematici. Il coefficiente R^2 risulta nel primo caso 0.6176 e nel secondo 0.6676.

In generale i coefficienti stimati per i due modelli risultano simili. L'unico caso in cui un coefficiente presenta addirittura verso opposto nei due modelli è β_1, ovvero quello riferito alla variabile $m.eta$. Tuttavia questa differenza non pare particolarmente rilevante in quanto, in entrambi i casi, il coefficiente sembra essere prossimo a 0, come indica il test di nullità per β_1. Di particolare rilevanza sembra, invece, il grado medio dei giocatori: il coefficiente stimato è negativo, quindi un miglior posizionamento in classifica corrisponde ad un grado medio più elevato. Data la forte relazione che si era riscontrata tra grado del giocatore e numero di stagioni nel campionato, sembra dunque che le squadre beneficino, dal punto di vista dei risultati, di atleti con permanenze nel campionato più lunghe. A conferma di ciò è anche il coefficiente relativo alla variabile *rimasti*: essendo negativo, indica che una maggiore stabilità degli atleti in squadra porta a migliori risultati in campo. L'influenza della stagione passata, però, mostra attraverso la variabile $rs.prec$ anche che i risultati migliori sono tipicamente appannaggio di coloro che già li avevano raggiunti, così come i peggiori. In generale, però la precedente partecipazione al campionato sembra avere un effetto migliorativo sul piazzamento in classifica, specialmente per il campionato femminile in cui, come si è visto alla Sezione 3.2.1, la permanenza dei club in A1 è più breve. Infatti, tipicamente le squadre al primo anno di partecipazione difficilmente ottengono posizionamenti elevanti in campionato.

Capitolo 4

Conclusioni

L'intenzione di questo lavoro è di approcciare l'analisi statistica in ambito sportivo da un punto di vista diverso. L'utilizzo dei dati di rete ha lo scopo di evidenziare una dinamicità insita nella realtà trattata, sottolineando i legami che questa caratteristica porta con sé. I campionati che si sono trattati sono dinamici, ma lo sono solo in relazione ad un insieme di atleti e società che li compongono. Sarebbe stato sicuramente interessante approfondire il tema dei risultati individuali degli atleti, definendo degli indicatori di "qualità" del giocatore sulla base del ruolo, utilizzando statistiche di gioco. In questo caso, però, si è preferito parlare di risultati solo a livello di squadra e, nel farlo, caratterizzare la squadra non tanto come somma di punteggi di "bravura" dei suoi componenti ma dal punto di vista dei legami che li caratterizzano.

Questo approccio sicuramente ha dei difetti, oltre che dei limiti in termini interpretativi e previsivi, ma tenta di tenere in considerazione l'aspetto delle relazioni in un ambito, come quello di squadra, fondato sui legami, un aspetto che non si riesce a rilevare con le statistiche di gioco ma è fondamentale a livello di risultato.

Bibliografia

[1] Italia in Dati. *Il mondo dello sport in Italia*. https://italiaindati.com/sport-in-italia/.

[2] CONI. *I numeri dello sport*. www.coni.it/it/i-numeri-dello-sport.

[3] FIPAV. *Volani i tesseramenti FIPAV, gli atleti superano quota 300.000*. https://www.federvolley.it/news/volano-i-tesseramenti-fipav-gli-atleti-superano-quota-300000.

[4] FIVB. *FIVB Mens Volleyball World Ranking*. https://en.volleyballworld.com/volleyball/world-ranking/men.

[5] FIVB. *FIVB Womens Volleyball World Ranking*. https://en.volleyballworld.com/volleyball/world-ranking/women.

[6] Volleybox. *TOP club di pallavolo maschile*. https://volleybox.net/it/teams/ranking.

[7] Volleybox. *TOP club di pallavolo femminile*. https://women.volleybox.net/it/teams/ranking.

[8] Lega Pallavolo Serie A. www.legavolley.it.

[9] Lega Pallavolo Serie A Femminile. www.legavolleyfemminile.it.

[10] Eric D. Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data with R*. Springer, 2020.

[11] Alessandra Salvan, Nicola Sartori, and Luigi Pace. *Modelli Lineari Generalizzati*. Springer, 2020.

[12] Gephi. *Tutorial layouts*. https://gephi.org/users/tutorial-layouts/.

Appendice A

Codice R: Scraping e pulizia

A.1 Giocatori campionato maschile

```
library(rvest)
library(tidyverse)
library(xml2)

#per lettera dell'alfabeto
urls<-paste0("https://www.legavolley.it/ricerca/?Lettera=",LETTERS,
"&TipoOgg=ATL")
tab_names<-paste0("tab",LETTERS)

for (i in 1:26){
  tab_html <- read_html(urls[i]) %>% html_nodes("table") %>%.[[2]]

  tab<-tab_html%>%html_table%>%select(-1,-7)

  list_href <- tab_html %>% html_nodes("tr") %>% html_nodes("td")%>%
  html_attrs()
  sq<-seq(2,length(list_href),7)
  sel<- list_href[sq] %>% unlist
  link<-sel[which(names(sel)=="onclick")]
  link_clean <- gsub("return window.top.location.href=('",
  "",link,fixed=T) %>%
    gsub("')","",x=.,fixed=T)

  tab$link<-link_clean

  tab<-tab%>%filter(Stagione>2007) #considero gli ultimi 15 anni

  assign(tab_names[i],tab)
}

men<-rbind(tabA,tabB,tabC,tabD,tabE,tabF,tabG,tabH,tabI,tabJ,
           tabK,tabL,tabM,tabN,tabO,tabP,tabQ,tabR,tabS,tabT,
```

```
           tabU,tabV,tabX,tabY,tabZ) #W ha 0 obs
rm(list=setdiff(ls(),c("men")))


add_teams_info<-function(tab){
  df<-matrix(NA,nrow=nrow(tab),ncol=18,
             dimnames = list(NULL,c("Nasc","Naz.sport", "Alt",
                                    as.character(2008:2022))))%>%
    as_tibble()
  tab1<-cbind(tab,df)


  for (i in 1:nrow(tab1)){
    pag<-read_html(tab1[i,6])

    #controllo se ha giocato almeno una stagione in Italia in A1
    A1<-pag%>%html_node("#td-outer-wrap > div.td-main-content-
    wrap.td-container-wrap >div > div > div.lvrow > div.s33 >
    div:nth-child(2) > table")%>%
      html_table()%>%.[1,3]%>%as.character()

    if (A1=="-") next

    #informazioni sul giocatore
    nodi<-pag%>%html_nodes("#td-outer-wrap > div.td-main-content-
    wrap.td-container-wrap >div > div > div.lvrow > div.s41")%>%
      html_children()
    nasc<-nodi%>%.[[3]]%>%html_text()%>%
      substr(start=(nchar(.)-3),stop=nchar(.)) %>% as.numeric()
    naz.sport<-nodi[[5]]%>%html_text%>%
      substr(start=15,stop=17)
    alt<-nodi[[7]]%>%html_text()%>%
      substr(start=(nchar(.)-5),stop=nchar(.)-3) %>% as.numeric()

    tab1[i,7:9]<-c(nasc,naz.sport,alt)

    #informazioni sui campionati
    gioc<-pag %>% html_node("#carriera")%>%html_table()
    gioc$Anno<-gsub("/[[:digit:]]{4}","",gioc$Periodo)

    gioc_sel<-gioc %>% #hanno giocato in Italia, in A1
      filter( (! is.na(Maglia)) &
                Serie=="A1" & as.numeric(Anno)>2007)

    if (dim(gioc_sel)[1]>0){
      gioc_sel <- gioc_sel%>%
        mutate(Squadra2=strsplit(Squadra," \n",fixed=TRUE)%>%
                 unlist%>%.[seq(1,length(.),2)])
```

```
      gioc_sel<-gioc_sel%>%group_by(Anno)%>%
      summarise(Squadra=first(Squadra2))

      tab1[i,which(colnames(tab1) %in% gioc_sel$Anno)]<-
      gioc_sel$Squadra
      }
  }

  #rimuovere chi ha NA per tutte le stagioni
  pieni<-apply(tab1[,10:24],1,function(x) sum(is.na(x))<15)

  return(tab1[pieni,])
}



men_def<-add_teams_info(men[,c(1,2,6)])

men_def<-men_def%>%
  mutate(Alt=as.numeric(Alt))%>%
  mutate(Nasc=as.numeric(Nasc))


rm(list=setdiff(ls(),c("men_def")))
```

A.2 Giocatrici campionato femminile

```
library(rvest)
library(tidyverse)
library(xml2)

#per stagione sportiva
urls<-paste0("https://www.legavolleyfemminile.it/atlete/?stagione=",
             as.character(2008:2022))
tab_names<-paste0("w",as.character(2008:2022))

for (i in 1:15){
  tab_node<-read_html(urls[i])%>%html_nodes("table")%>%.[[1]]

  tab<-tab_node%>%html_table()
  tab[which(tab$Nazionalità==""),]$Nazionalità = NA
  tab[which(tab$Altezza=="cm"),]$Altezza = NA

  tab<-tab%>%mutate(Altezza=Altezza%>%substr(start=1,stop=3)%>%
                    as.numeric)%>%
    mutate(`Profilo personale`=NA)
```

```
  colnames(tab)[6:7]<-c("Nasc","link")

  links<-tab_node%>%html_nodes("tr")%>%html_nodes("td")%>%
    html_nodes("a")%>%html_attrs()%>%unlist
  names(links)<-NULL
  tab$link<-links

  assign(tab_names[i],tab)
}

women<-bind_rows(w2008,w2009,w2010,w2011,w2012,w2013,w2014,w2015,
                 w2016,w2017,w2018,w2019,w2020,w2021,w2022)

rm(list=setdiff(ls(),c("women")))


women[which(women$Nome=="0"),]$Nome<-""

women_unique <- women%>%distinct%>%
  group_by(link)%>%
  summarise(Atleta=paste0(unique(Cognome)," ",unique(Nome)),
            Alt=round(mean(Altezza,na.rm=TRUE)),
            Nasc=round(mean(Nasc,na.rm=TRUE)),
            Naz.sport=list(unique(Nazionalità)),
            Ruolo=list(unique(Ruolo)))

#mettiamo NA per liste e recupereremo poi il valore
women_unique[grep("c(",women_unique$Naz.sport,
                  fixed=T),]$Naz.sport <- list(NA)
women_unique[grep("c(",women_unique$Ruolo,
                  fixed=T),]$Ruolo <- list(NA)

women_unique<-women_unique%>%
  mutate(Naz.sport=unlist(Naz.sport))%>%
  mutate(Ruolo=unlist(Ruolo))%>%
  mutate(Alt=as.character(Alt))%>%
  mutate(Nasc=as.character(Nasc))

women_unique$Alt[which(women_unique$Alt=="NaN")]<-NA
women_unique$Nasc[which(women_unique$Nasc=="NaN")]<-NA

#sul profilo personale i ruoli sono in inglese
trad_ruolo<-function(eng){
  if (eng=="Spiker") return("Schiacciatrice")
  if (eng=="Middle Blocker") return("Centrale")
  if (eng=="Setter") return("Palleggiatrice")
  if (eng=="Libero") return(eng)
  if (eng=="Opposite") return("Opposto")
  if (eng=="Universal") return("Universale")
```

```
  return(NA)
}

get_attr<-function(quali_attr,url){
  pag<-read_html(url)
  out<-c()

  info<-pag%>%
    html_nodes("#content > div:nth-child(3) > div > div >
    div:nth-child(3)")

  if (quali_attr[1]==TRUE){ #altezza
    alt<-info%>%html_nodes("span:nth-child(4)")%>%html_text%>%
      substr(start=nchar(.)-5,stop=nchar(.)-3)
    out<-c(out,alt)
  }

  if (quali_attr[2]==TRUE){ #nascita
    nasc<-info%>%html_nodes("span:nth-child(1)")%>%html_text%>%
      substr(start=nchar(.)-3,stop=nchar(.))
    out<-c(out,nasc)
  }

  if (quali_attr[3]==TRUE){ #nazionalità
    naz<-info%>%html_nodes("span:nth-child(3)")%>%html_text%>%
      substr(start=nchar(.)-2,stop=nchar(.))
    out<-c(out,naz)
  }

  if (quali_attr[4]==TRUE){ #ruolo
    role<-pag%>%html_nodes("#content > div:nth-child(3) > div >
                           div > div:nth-child(2) >
                           span:nth-child(2) > b")%>%
      html_text()
    out<-c(out,trad_ruolo(role))
  }
  return(out)
}



for (i in 1:dim(women_unique)[1]){
  quali<-is.na(women_unique[i,3:6]%>%as.vector())
  if (sum(quali)==0) next

  attr<-get_attr(quali,women_unique[i,1]%>%as.character())
  women_unique[i,c(FALSE,FALSE,quali)]<-
      t(attr)
}
```

```r
apply(women_unique,2,function(x) sum(is.na(x)))
#non ci sono più NA
#ma in certi casi le informazioni sono ancora mancanti
women_unique[which(women_unique$Naz.sport=="a: "),]$Naz.sport<-NA
women_unique[which(women_unique$Alt==": 0"),]$Alt<-NA


women_unique<-women_unique%>%
  mutate(Alt=as.numeric(Alt))%>%
  mutate(Nasc=as.numeric(Nasc))


#ora selezioniamo chi ha giocato in A1 in Italia
add_teams_info_F<-function(tab){
  df<-matrix(NA,nrow=nrow(tab),ncol=15,
             dimnames = list(NULL,as.character(2008:2022)))%>%
    as_tibble()
  tab1<-cbind(tab,df)

  for (i in 1:nrow(tab1)){
    gioc <-read_html(tab1[i,]$link) %>% html_nodes("table") %>%
      .[[1]]%>%html_table()

    gioc$Anno<-gsub("-[[:digit:]]{4}","",gioc$Stagione)
    gioc_sel<-gioc %>% filter(as.numeric(Anno)>2007 &
                              Serie=="A1" &
                              (! is.na(Numero)))

    if (dim(gioc_sel)[1]>0){
      gioc_sel<-gioc_sel%>%group_by(Anno)%>%
        summarise(Squadra=first(Squadra))
      colonne<-which(colnames(tab1) %in% gioc_sel$Anno)
      tab1[i,colonne] <-gioc_sel$Squadra
    }
  }

  #rimuovere chi ha NA per tutte le stagioni
  pieni<-apply(tab1[,7:21],1,function(x) sum(is.na(x))<15)

  return(tab1[pieni,])
}


women_def<-add_teams_info_F(women_unique)

rm(list=setdiff(ls(),c("women_def")))
```

A.3 Pulizia dei giocatori

```
apply(women_def,2,function(x) length(unique(x))) #ok

apply(men_def,2,function(x) length(unique(x)))

which(table(men_def$Atleta)>1)
men_def[which(men_def$Atleta=="Michieletto Alessandro"),]
#Michieletto nel 2019 ha giocato anche con Unitrento (A3)
#riuniamo in una sola osservazione

url_keep<-"http://www.legavolley.it/player/MIC-ALE-01"
url_rm<-"http://www.legavolley.it/player/MIC-ALE-01-SL"
men_def[which(men_def$link==url_keep),]$`2019`<-"Itas Trentino"
men_def<-men_def[-which(men_def$link==url_rm),]

apply(men_def,2,function(x) length(unique(x))) #ok


men_def$Ruolo<-men_def$Ruolo%>%as.factor()
women_def$Ruolo<-women_def$Ruolo%>%
  gsub(pattern="Opposto",repl="Schiacciatore",fixed=T)%>%
  gsub(pattern="Schiacciatrice",repl="Schiacciatore",fixed=T)%>%
  gsub(pattern="Palleggiatrice",repl="Palleggiatore",fixed=T)%>%
  as.factor()
```

A.4 Pulizia delle squadre

```
men_clean_teams<-men_def[7:21]
men_clean_teams<-men_clean_teams%>%
  mutate_all(function(x) gsub(x,pattern=" Volley",
                              replacement="",fixed=T))%>%
  mutate_all(function(x) str_split_i(x,pattern = " ",i=-1))

matr<-men_clean_teams%>%as.matrix()
matr<-matr%>%
  gsub(pattern="Belluno",replacement="Treviso",fixed=T)%>%
  gsub(pattern="Romagna",replacement="Ravenna",fixed=T)%>%
  gsub(pattern="Franca",replacement="Taranto",fixed=T)%>%
  gsub(pattern="Latina",replacement="Cisterna",fixed=T)%>%
  gsub(pattern="Macerata",replacement="Lube",fixed=T)%>%
  gsub(pattern="Civitanova",replacement="Lube",fixed=T)%>%
  gsub(pattern="Treia",replacement="Lube",fixed=T)%>%
  gsub(pattern="Sansepolcro",replacement="Castello",fixed=T)%>%
  gsub(pattern="Castello",replacement="Città di Castello",fixed=T)%>%
  gsub(pattern="Grotte",replacement="Castellana Grotte",fixed=T)%>%
```

```
  gsub(pattern="Valentia",replacement="Vibo Valentia",fixed=T)

matr[,1:10]<-matr[,1:10]%>%
  gsub(pattern="Piacenza",replacement="Piacenza Copra",fixed=T)

matr[,1:4]<-matr[,1:4]%>%
  gsub(pattern="Monza",replacement="Monza Br",fixed=T)%>%
  gsub(pattern="Brianza",replacement="Monza Br",fixed=T)%>%
  gsub(pattern="Montichiari",replacement="Monza Br",fixed=T)

matr[,1:4]<-matr[,1:4]%>%
  gsub(pattern="Giustino",replacement="Umbria",fixed=T)%>%
  gsub(pattern="Perugia",replacement="Umbria",fixed=T)

matr[,5]<-matr[,5]%>%
  gsub(pattern="Giustino",replacement="San Giustino",fixed=T)
men_clean_teams<-matr%>%as_tibble()


#squadre per anno
soc<-matr%>%apply(., 2, unique)%>%unlist%>%unique%>%na.omit
club<-matrix(0,nrow=length(soc),ncol=15,
             dimnames = list(soc,as.character(2008:2022)))
for (i in 1:15){
  q<-which(soc %in% matr[,i])
  club[q,i]<-1
}



women_clean_teams<-women_def[7:21]
women_clean_teams<-women_clean_teams%>%
  mutate_all(function(x) gsub(x,pattern=" Volley",
                              replacement="",fixed=T))%>%
  mutate_all(function(x) gsub(x,pattern=" 1991",
                              replacement="",fixed=T))%>%
  mutate_all(function(x) gsub(x,pattern=" Club",
                              replacement="",fixed=T))%>%
  mutate_all(function(x) gsub(x,pattern=" Crai",
                              replacement="",fixed=T))%>%
  mutate_all(function(x) str_split_i(x,pattern = " ",i=-1))

matr<-women_clean_teams%>%as.matrix()
matr<-matr%>%
  gsub(pattern="Arsizio",replacement="Busto Arsizio",fixed=T)%>%
  gsub(pattern="Cortese",replacement="Villa Cortese",fixed=T)%>%
  gsub(pattern="Grotte",replacement="Castellana Grotte",fixed=T)%>%
  gsub(pattern="Italia",replacement="Club Italia",fixed=T)
matr%>%apply(., 2, unique)%>%unlist%>%unique
```

```
matr%>%apply(., 2, unique)

matr[,15]<-matr[,15]%>%
  gsub(pattern="Milano",replacement="Monza",fixed=T)%>%
  gsub(pattern="S.Bernardo",replacement="Cuneo",fixed=T)

matr[,14:15]<-matr[,14:15]%>%
  gsub(pattern="Bergamo",replacement="Bergamo 91",fixed=T)

matr[,9:10]<-matr[,9:10]%>%
  gsub(pattern="Modena",replacement="Piacenza",fixed=T)

matr[,3:5]<-matr[,3:5]%>%
  gsub(pattern="Modena",replacement="Modena U.",fixed=T)

matr[,5]<-matr[,5]%>%
  gsub(pattern="Bologna",replacement="Forlì",fixed=T)

matr[,1:4]<-matr[,1:4]%>%
  gsub(pattern="Novara",replacement="Novara A.",fixed=T)%>%
  gsub(pattern="Conegliano",replacement="Conegliano S.",fixed=T)%>%
  gsub(pattern="Perugia",replacement="Perugia S.",fixed=T)

matr[,1]<-matr[,1]%>%
  gsub(pattern="Chieri",replacement="Torino",fixed=T)

matr[,8]<-matr[,8]%>%
  gsub(pattern="Vicenza",replacement="Villaverla",fixed=T)

women_clean_teams<-matr%>%as_tibble()


#squadre per anno
soc_w<-matr%>%apply(., 2, unique)%>%unlist%>%unique%>%na.omit
club_w<-matrix(0,nrow=length(soc_w),ncol=15,
               dimnames = list(soc_w,as.character(2008:2022)))
for (i in 1:15){
  q<-which(soc_w %in% matr[,i])
  club_w[q,i]<-1
}
```

A.5 Scraping delle società maschili

```
#scraping per pagina della società sportiva

links<-c("https://www.legavolley.it/team/4830",
```

```
          "https://www.legavolley.it/team/4400",
          "https://www.legavolley.it/team/5493",
          "https://www.legavolley.it/team/4981",
          "https://www.legavolley.it/team/6233",
          "https://www.legavolley.it/team/5949",
          "https://www.legavolley.it/team/4681",
          "https://www.legavolley.it/team/4529",
          "https://www.legavolley.it/team/5356",
          "https://www.legavolley.it/team/6242",
          "https://www.legavolley.it/team/4525",
          "https://www.legavolley.it/team/4258",
          "https://www.legavolley.it/team/4398",
          "https://www.legavolley.it/team/5124",
          "https://www.legavolley.it/team/4416",
          "https://www.legavolley.it/team/6084",
          "https://www.legavolley.it/team/4688",
          "https://www.legavolley.it/team/5627",
          "https://www.legavolley.it/team/5226",
          "https://www.legavolley.it/team/6082",
          "https://www.legavolley.it/team/4849",
          "https://www.legavolley.it/team/4992",
          "https://www.legavolley.it/team/4971",
          "https://www.legavolley.it/team/6394",
          "https://www.legavolley.it/team/5355",
          "https://www.legavolley.it/team/5489",
          "https://www.legavolley.it/team/6278",
          "https://www.legavolley.it/team/6396")

club.m.df<-tibble(nome=rownames(club),url=links)

library(rvest)
library(tidyverse)
library(xml2)

pag<-read_html("https://www.legavolley.it/team/5353")
tab<-pag%>%html_node("table.palmares")%>%html_table()
#piazzamento in regular season

prova<-pag%>%html_node("table.albo")
if (class(prova)!="xml_missing") prova<-prova%>%html_table()
#vittorie campionato
#attenzione che va tolto un anno per l'annata sportiva


matr<-matrix(NA,nrow=28,ncol=(30))
club.m.df<-bind_cols(club.m.df,matr)
rs<-paste0("rs.",2008:2022)
primo<-paste0("primo.",2008:2022)
colnames(club.m.df)[c(-1,-2)]<-c(rs,primo)
```

```r
for (i in 1:dim(club.m.df)[1]){
  pag<-read_html(club.m.df$url[i])
  tab<-pag%>%html_node("table.palmares")%>%html_table()

  A23<-c(tab$Torneo%>%grep("A2",.,fixed=T),
         tab$Torneo%>%grep("A3",.,fixed=T))
  if (length(A23)>0) tab<-tab[-A23,]

  quali<-tab$Torneo%>%grep("Regular Season",.,fixed=T)
  tab<-tab[quali,]
  anni<-tab$Stagione%>%gsub(pattern="/[[:digit:]]{4}",replacement="")
  tab<-tab%>%mutate(anno=anni%>%as.numeric)%>%filter(anno>2007)
  pos<-tab$`Pos.`%>%strsplit(split="ł", fixed=T)%>%unlist%>%
    .[seq(1,length(.),2)]

  colonne<-which(as.character(2008:2022)%in% anni)
  club.m.df[i,(colonne+2)]<-t(rev(pos))
}

club.m.df<-club.m.df%>%mutate_at(3:17,as.numeric)

#stagione 2022/23
classifica<-c("Perugia","Trentino","Modena","Lube","Verona",
              "Piacenza","Monza","Milano","Cisterna","Padova",
              "Taranto","Siena")
for (i in 1:length(classifica)){
  club.m.df$rs.2022[which(club.m.df$nome==classifica[i])]<-i
}


#distinzione tra NA e 0 per chi ha vinto il campionato
#NA: non ha partecipato
#0: ha partecipato ma non ha vinto

for (i in 1:15){
  part<-which(club[,i]==1)
  club.m.df[part,(17+i)]<-0
}



for (i in 1:dim(club.m.df)[1]){
  albo<-read_html(club.m.df$url[i])%>%html_node("table.albo")
  if (class(albo)=="xml_missing") next
  tab<-albo%>%html_table()

  tab<-tab%>%mutate(anno=(Anno-1))
```

```
  tab<-tab%>%filter(anno>2007)
  quali<-which(tab$Torneo=="Campionato Italiano")

  if (length(quali)==0) next
  colonne<-which((2008:2022)%in% tab[quali,]$anno)
  club.m.df[i,(colonne+17)]<-1
}

club.m.df<-club.m.df[,-2]
```

A.6 Scraping delle società femminili

```
library(rvest)
library(tidyverse)
library(xml2)

#scraping per classifica di ogni anno + albo d'oro per i vincitori

classifica<-matrix(NA,ncol=15,nrow=14)%>%
  as_tibble()%>%`colnames<-`((2008:2022)%>%as.character)

urls<-paste("https://www.legavolleyfemminile.it/classifica/?stagione=",
            2008:2022)
for (i in 1:15){
  tab<-read_html(urls[i]) %>% html_nodes("table")%>%
    .[[1]]%>%html_table() %>% .[,2]

  classifica[1:nrow(tab),i]<-tab
}


tab_vincitori<-read_html("https://www.legavolleyfemminile.it/
                         eventi-campionato-serie-a1/") %>%
  html_nodes("table")%>%.[[1]]%>%html_table()
vincitori<-tab_vincitori[15:1,2]%>%t()%>%
  `colnames<-`((2008:2022)%>%as.character)%>%
  `rownames<-`(NULL)%>%
  as_tibble()



#stessa pulizia applicata ai nomi delle società nel df delle atlete
pulizia_nomi<-function(tab){
  women_clean_teams<-tab
  women_clean_teams<-women_clean_teams%>%
    mutate_all(function(x) gsub(x,pattern=" Volley",
```

```
                                replacement="",fixed=T))%>%
    mutate_all(function(x) gsub(x,pattern=" 1991",
                                replacement="",fixed=T))%>%
    mutate_all(function(x) gsub(x,pattern=" Club",
                                replacement="",fixed=T))%>%
    mutate_all(function(x) gsub(x,pattern=" Crai",
                                replacement="",fixed=T))%>%
    mutate_all(function(x) str_split_i(x,pattern = " ",i=-1))

  matr<-women_clean_teams%>%as.matrix()
  matr<-matr%>%
    gsub(pattern="Arsizio",replacement="Busto Arsizio",fixed=T)%>%
    gsub(pattern="Cortese",replacement="Villa Cortese",fixed=T)%>%
    gsub(pattern="Grotte",replacement="Castellana Grotte",fixed=T)%>%
    gsub(pattern="Italia",replacement="Club Italia",fixed=T)
  matr%>%apply(., 2, unique)%>%unlist%>%unique
  matr%>%apply(., 2, unique)

  matr[,15]<-matr[,15]%>%
    gsub(pattern="Milano",replacement="Monza",fixed=T)%>%
    gsub(pattern="S.Bernardo",replacement="Cuneo",fixed=T)

  matr[,14:15]<-matr[,14:15]%>%
    gsub(pattern="Bergamo",replacement="Bergamo 91",fixed=T)

  matr[,9:10]<-matr[,9:10]%>%
    gsub(pattern="Modena",replacement="Piacenza",fixed=T)

  matr[,3:5]<-matr[,3:5]%>%
    gsub(pattern="Modena",replacement="Modena U.",fixed=T)

  matr[,5]<-matr[,5]%>%
    gsub(pattern="Bologna",replacement="Forlì",fixed=T)

  matr[,1:4]<-matr[,1:4]%>%
    gsub(pattern="Novara",replacement="Novara A.",fixed=T)%>%
    gsub(pattern="Conegliano",replacement="Conegliano S.",fixed=T)%>%
    gsub(pattern="Perugia",replacement="Perugia S.",fixed=T)

  matr[,1]<-matr[,1]%>%
    gsub(pattern="Chieri",replacement="Torino",fixed=T)

  matr[,8]<-matr[,8]%>%
    gsub(pattern="Vicenza",replacement="Villaverla",fixed=T)

  women_clean_teams<-matr%>%as_tibble()

  return(women_clean_teams)
}
```

```
classifica_clean<-pulizia_nomi(classifica)
vincitori_clean<-pulizia_nomi(vincitori)


club.w.df<-bind_cols(tibble(nome=rownames(club_w)),
                     matrix(NA,nrow=48,ncol=(30)))
rs<-paste0("rs.",2008:2022)
primo<-paste0("primo.",2008:2022)
colnames(club.w.df)[-1]<-c(rs,primo)


#NA vs 0
for (i in 1:15){
  part<-which(club_w[,i]==1)
  club.w.df[part,(16+i)]<-0
}

#vincitori
for (i in 1:15){
  riga<-which(club.w.df$nome== (vincitori_clean[1,i]%>%
                                  as.character()))
  club.w.df[riga, 16+i]<-1
}

#classifica
for (i in 1:15){
  righe<-which(club_w[,i]==1)
  soc.i<-rownames(club_w)[righe]
  pos<-sapply(soc.i, function(x) which(classifica_clean[,i]==x))
  club.w.df[righe,1+i]<-pos%>%as_tibble_col()
}

#si sono ritirati (e quindi non compaiono nelle classifiche finali)
#- Conegliano S. nel 2011
#- Crema nel 2012
#- Modena U. nel 2012

club.w.df$rs.2011[[which(club.w.df$nome=="Conegliano S.")]]<-99
club.w.df$rs.2012[[which(club.w.df$nome=="Crema")]]<-99
club.w.df$rs.2012[[which(club.w.df$nome=="Modena U.")]]<-99

club.w.df$rs.2011<-sapply(club.w.df$rs.2011,
                          function(x) if(is.null(x)) x<-NA else x<-x)
club.w.df$rs.2012<-sapply(club.w.df$rs.2012,
                          function(x) if(is.null(x)) x<-NA else x<-x)
```

Appendice B

Codice R: Analisi delle reti dei giocatori

B.1 Creazione delle reti

```
#creiamo degli id per i giocatori a partire dagli url
men_ids<-men_def%>%select(link)%>%as.vector()%>%unlist%>%
  gsub(pattern="http://www.legavolley.it/player/",
       replacement='',fixed=T)

women_ids<-women_def%>%select(link)%>%as.vector()%>%unlist%>%
  substr(start=nchar(.)-10,stop=nchar(.)-1)

#matrici di adiacenza per ogni anno
library(Matrix)
create_matr<-function(col){
  m<-Matrix(0,nrow=length(col),ncol=length(col),sparse=TRUE)
  lev=levels(as.factor(col))
  for (i in 1:length(lev)){
    quali<-which(col==lev[i])
    m[quali,quali]<-1
  }
  diag(m)<-rep(0,length(col))
  return(drop0(m))
}


matr_names_m<-paste0("m",2008:2022)
matr_names_w<-paste0("w",2008:2022)

for (i in 1:15){
  assign(matr_names_m[i],
         create_matr(men_def[,(6+i)]))
}
```

```
for (i in 1:15){
  assign(matr_names_w[i],
         create_matr(women_def[,(6+i)]))
}

#matrice complessiva come somma delle matrici dei vari anni
ad_matr_m<-m2008+m2009+m2010+m2011+m2012+m2013+m2014+m2015+m2016+
  m2017+m2018+m2019+m2020+m2021+m2022

ad_matr_w<-w2008+w2009+w2010+w2011+w2012+w2013+w2014+w2015+w2016+
  w2017+w2018+w2019+w2020+w2021+w2022

dimnames(ad_matr_m)<-list(men_ids,men_ids)
dimnames(ad_matr_w)<-list(women_ids,women_ids)

#esportiamo le due matrici di adiacenza per usarle su Gephi
write.csv(as.matrix(ad_matr_m),"men_matr_since2008.csv")
write.csv(as.matrix(ad_matr_w),"women_matr_since2008.csv")

#creo la variabile "numero di stagioni nel campionato"
men_def$nstag<-apply(men_def[7:21],1,
                     function(x) sum(! is.na(x)))
women_def$nstag<-apply(women_def[7:21],1,
                       function(x) sum(! is.na(x)))

#matrici con pesi normalizzati
m<-Matrix(rep(men_def$nstag,1019),ncol=1019)
medie<-(m+t(m))*0.5
weight_matr_m<-ad_matr_m * (1/medie)

write.csv(as.matrix(weight_matr_m),"weight_matr_men2008.csv")


m<-Matrix(rep(women_def$nstag,985),ncol=985)
medie<-(m+t(m))*0.5
weight_matr_w<-ad_matr_w * (1/medie)

write.csv(as.matrix(weight_matr_w),"weight_matr_women2008.csv")


#creazione vertici
vert_men<-cbind(men_ids,
                men_def$Ruolo,
                men_def[4:6])%>%`row.names<-`(NULL)

vert_women<-cbind(women_ids,
                  women_def$Ruolo,
                  women_def[4:5],
                  women_def$Alt)%>%`row.names<-`(NULL)
```

```
colnames(vert_men)<-colnames(vert_women)<-c("id","ruolo","nasc",
                                           "naz","alt")

#creazione di 2 periodi (+ uno intermedio)
periodo1<-apply(men_def[7:14],1,function(x) sum(! is.na(x)))
periodo2<-apply(men_def[14:21],1,function(x) sum(! is.na(x)))
p1<-ifelse(periodo1>0.5*men_def$nstag,1,0)
p2<-ifelse(periodo2>0.5*men_def$nstag,2,0)
p<-p1+p2
p[which(p==0)]<-3
vert_men$p<-p


periodo1<-apply(men_def[7:14],1,function(x) sum(! is.na(x)))
periodo2<-apply(women_def[14:21],1,function(x) sum(! is.na(x)))
p1<-ifelse(periodo1>0.5*women_def$nstag,1,0)
p2<-ifelse(periodo2>0.5*women_def$nstag,2,0)
p<-p1+p2
p[which(p==0)]<-3
vert_women$p<-p

#raggruppiamo il numero di stagioni in 4 categorie
vert_men$nstag<-NA
vert_men$nstag[which(men_def$nstag==1)]<-"1"
vert_men$nstag[which(men_def$nstag %in% c(2,3))]<-"2"
vert_men$nstag[which(men_def$nstag %in% 4:7)]<-"3"
vert_men$nstag[which(men_def$nstag>7)]<-"4"


vert_women$nstag<-NA
vert_women$nstag[which(women_def$nstag==1)]<-"1"
vert_women$nstag[which(women_def$nstag %in% c(2,3))]<-"2"
vert_women$nstag[which(women_def$nstag %in% 4:7)]<-"3"
vert_women$nstag[which(women_def$nstag>7)]<-"4"


write.csv(vert_men,"vert_men_2008_per.csv",row.names = FALSE)
write.csv(vert_women,"vert_women_2008_per.csv",row.names = FALSE)
```

B.2 Struttura delle reti

```
#distribuzione dei pesi normalizzati e non
weight_matr_m<-as(weight_matr_m,"symmetricMatrix")
weight_matr_w<-as(weight_matr_w,"symmetricMatrix")

w<-tibble(pesi=c(ad_matr_m@x,ad_matr_w@x,
```

```
                weight_matr_m@x,weight_matr_w@x),
          tipo=c(rep("Pesi",length(c(ad_matr_m@x,ad_matr_w@x))),
                 rep("Pesi normalizzati",
                     length(c(weight_matr_m@x,weight_matr_w@x)))),
          gend=rep(c(rep("Uomini",length(ad_matr_m@x)),
                     rep("Donne",length(ad_matr_w@x))),2))

ggplot(data=w,aes(pesi))+
  geom_histogram()+
  facet_wrap(~tipo+gend,scales="free_x")+
  labs(y="",x="")


#creazione delle reti
library(igraph)
graph_m<-graph_from_adjacency_matrix(weight_matr_m,
                                     mode="undirected",
                                     weighted=TRUE)

graph_w<-graph_from_adjacency_matrix(weight_matr_w,
                                     mode="undirected",
                                     weighted=TRUE)


deg_m<-degree(graph_m)
deg_w<-degree(graph_w)
btw_m<-betweenness(graph_m)
btw_w<-betweenness(graph_w)

g.stat<-tibble(val=c(deg_m,deg_w,btw_m,btw_w),
               stat=c(rep("Grado",(1019+985)),
                      rep("Betweenness",(1019+985))),
               gend=rep(c(rep("Uomini",1019),
                          rep("Donne",985)),2))

ggplot(g.stat%>%filter(stat=="Grado"),
       aes(val,y=after_stat(density)))+
  geom_histogram()+
  facet_wrap(~gend)+
  labs(y="Densità",x="Grado")

ggplot(g.stat%>%filter(stat=="Betweenness"),
       aes(val,y=after_stat(density)))+
  geom_histogram()+
  facet_wrap(~gend)+
  labs(y="Densità",x="Betweenness")


sum(btw_m==0)/1019
```

```
sum(btw_w==0)/985

cor(btw_m,deg_m)
cor(btw_w,deg_w)


diameter(graph_m)
diameter(graph_w)

edge_density(graph_m)
edge_density(graph_w)

mean_distance(graph_m)
mean_distance(graph_w)
```

B.3 Modelli per grado e betweenness

```
giocatori<-bind_rows(vert_men,vert_women)
giocatori$gen<-c(rep("m",1019),rep("w",985))
giocatori$ita<-ifelse(giocatori$naz=="ITA",1,0)
giocatori$grado<-c(deg_m,deg_w)
giocatori$betw<-c(btw_m,btw_w)
giocatori<-giocatori%>%
  select(ruolo,nasc,ita,alt,nstag,gen,grado,betw)

str(giocatori)
giocatori<-giocatori%>%mutate_at(c(1,3,6),as.factor)%>%
  mutate(ruolo=relevel(ruolo,ref="Schiacciatore"))


#anno di nascita
ggplot(giocatori,aes(x=nasc,y=nstag))+
  geom_count()+
  geom_smooth(method = "lm",formula = y ~ poly(x, 2),
              se=F,fullrange=T)+
  labs(x="Anno di nascita",
       y="Numero di stagioni in Italia",
       size="Freq.")



#altezza
alt_bewt_plot<-ggplot(giocatori,aes(x=alt,y=betw))+
  geom_point(aes(shape=gen,color=ruolo))+
  geom_smooth(method = "lm",formula = y ~ poly(x, 2),
              se=F,fullrange=T)+
```

```
  labs(x="Altezza",y="Betweenness",
       shape="Genere",color="Ruolo")

cor.b<-cor(giocatori$alt,giocatori$betw,
           use="pairwise.complete.obs")%>% round(5)

alt_bewt_plot+
  annotate("label",x=210,y=75000,
           label=paste0("Corr. ",cor.b))


alt_gr_plot<-ggplot(giocatori,aes(x=alt,y=grado))+
  geom_point(aes(shape=gen,color=ruolo))+
  geom_smooth(method = "lm",formula = y ~ poly(x,2),
              se=F,fullrange=F)+
  labs(x="Altezza",y="Grado",
       shape="Genere",color="Ruolo")

cor.g<-cor(giocatori$alt,giocatori$grado,
           use="pairwise.complete.obs")%>% round(5)

alt_gr_plot+
  annotate("label",x=215,y=150,
           label=paste0("Corr. ",cor.g))



giocatori<-giocatori%>%select(-nasc,-alt)
giocatori$betw01<-ifelse(giocatori$betw==0,0,1)







#MODELLO DI POISSON PER IL GRADO
grado1<-glm(grado~gen*.,
            data=giocatori%>%select(-betw,-betw01),
            family = poisson)
step.gr<-step(grado1)

par(mfrow=c(1,4))
plot(step.gr,which=1:4)
par(mfrow=c(1,1))

plot(resid(step.gr)~step.gr$model$nstag)

plot(step.gr$y,fitted(step.gr))
```

```r
abline(a=0,b=1,col=2)


sjPlot::plot_model(step.gr,ci.lvl=NA,type="pred",
                   terms=c("nstag"),show.data = T,
                   axis.title = c("Numero di stagioni","Grado"),
                   title="")
cor(giocatori$grado,giocatori$nstag)


#funzione di legame identità
grado.id1<-glm(grado~gen*(ita+ruolo+nstag),
               data=giocatori%>%na.omit,
               family=poisson(link="identity"))
step.gr.id<-step(grado.id1)
summary(step.gr.id)

par(mfrow=c(1,4))
plot(step.gr.id,which=1:4)
par(mfrow=c(1,1))

plot(step.gr.id$y,step.gr.id$fitted.values)
abline(a=0,b=1,col=2)

sjPlot::plot_model(step.gr.id,ci.lvl=NA,type="pred",
                   terms=c("nstag","gen"),show.data=T,
                   axis.title = c("Numero di stagioni","Grado"),
                   title="",legend.title = "Genere")

coeff<-coef(step.gr.id)
coeff[3]+coeff[4]




#MODELLO PER LA BETWEENNESS
betw.glm<-glm(betw01~gen*(ruolo+ita+nstag),
               family=binomial,data=giocatori%>%na.omit)
step.betw<-step(betw.glm1)
summary(step.betw)


#matrice di confusione
tab<-table(betw01.glm$y,betw01.glm$fitted.values>0.5)

#accuratezza
(tab[1,1]+tab[2,2])/sum(tab)
```

```
#probabilità stimate
sjPlot::plot_model(betw01.glm,type="int",ci.lvl=NA)+
  labs(x="Numero di stagioni",color="Genere",
       y="Prob. betweenness >0",title="")
```

B.4 Reti con finestra di 5 anni

```
m_list<-list(m2008,m2009,m2010,m2011,m2012,m2013,m2014,m2015,
             m2016,m2017,m2018,m2019,m2020,m2021,m2022)
w_list<-list(w2008,w2009,w2010,w2011,w2012,w2013,w2014,w2015,
             w2016,w2017,w2018,w2019,w2020,w2021,w2022)


#lista delle matrici con pesi = nř stagioni insieme nei 5 anni
wind.m.list<-list()
for (i in 3:13){
  wind.m.list[[i-2]]<-m_list[[i-2]]+m_list[[i-1]]+m_list[[i]]+
    m_list[[i+1]]+m_list[[i+2]]
}
wind.m.list<-sapply(wind.m.list,
                    function(x) `dimnames<-`(x,list(men_ids,men_ids)))

wind.w.list<-list()
for (i in 3:13){
  wind.w.list[[i-2]]<-w_list[[i-2]]+w_list[[i-1]]+w_list[[i]]+
    w_list[[i+1]]+w_list[[i+2]]
}
wind.w.list<-sapply(wind.w.list,
                    function(x)`dimnames<-`(x,list(women_ids,women_ids)))


#nř di stagioni per giocatore nei 5 anni
wind.nstag.m<-list()
wind.nstag.w<-list()
for (i in 1:11){
  wind.nstag.m[[i]]<-apply(men_def[(6+i):(10+i)],1,
                           function(x) sum(! is.na(x)))
  wind.nstag.w[[i]]<-apply(women_def[(6+i):(10+i)],1,
                           function(x) sum(! is.na(x)))
}

#matrici con pesi normalizzati
wind.m.norm.list<-list()
for (i in 1:11){
  m<-Matrix(rep(wind.nstag.m[[i]],1019),ncol=1019)
  medie<-(m+t(m))*0.5
```

```
  medie[medie==0]<-1 #per non dividere per zero
  wind.m.norm.list[[i]] <- wind.m.list[[i]] * (1/medie)
}

wind.w.norm.list<-list()
for (i in 1:11){
  m<-Matrix(rep(wind.nstag.w[[i]],985),ncol=985)
  medie<-(m+t(m))*0.5
  medie[medie==0]<-1 #per non dividere per zero
  wind.w.norm.list[[i]] <- wind.w.list[[i]] * (1/medie)
}


print_csv<-function(ls,csv.names){
  for (i in 1:11){
    path<-paste(csv.names[i],".csv", sep="")
    write.csv(as.matrix(ls[[i]]),path)
  }
}

zeros<-c(rep(0,9),rep("",2))
men.names<-paste0("w5_men_",zeros,1:11)
men.names.norm<-paste0("w5_men_norm_",zeros,1:11)
women.names<-paste0("w5_women_",zeros,1:11)
women.names.norm<-paste0("w5_women_norm_",zeros,1:11)

print_csv(wind.m.list,men.names)
print_csv(wind.m.norm.list,men.names.norm)
print_csv(wind.w.list,women.names)
print_csv(wind.w.norm.list,women.names.norm)



#creazione di una variabile "primo anno in Italia"
#per ordinare osservazioni sul cerchio

anno1<-apply(men_def[7:21],1,
             function(x) (!is.na(x))%>%which%>%.[1]) +2007
names(anno1)<-NULL
ordine<-paste0(anno1,men_ids)

anno1w<-apply(women_def[7:21],1,
              function(x) (!is.na(x))%>%which%>%.[1]) +2007
names(anno1w)<-NULL
ordinew<-paste0(anno1w,women_ids)

#da aggiungere ai vertici
v.men<-read.csv("vert_men_2008_per.csv")
v.women<-read.csv("vert_women_2008_per.csv")
```

```
v.men<-cbind(v.men,anno1,ordine)
v.women<-cbind(v.women,anno1w,ordinew)

write.csv(v.men,"vert_men_ORD.csv",row.names = FALSE)
write.csv(v.women,"vert_women_ORD.csv",row.names = FALSE)



#andamento di grado e betweenness
deg_betw_long<-function(ls){
  nomi<-paste0(2008:2018,"-",2012:2022)
  n<-ls[[1]]@Dim[1]
  tab<-tibble(value=NA,
              var=rep(c(rep("Grado",n),
                        rep("Betweenness",n)),11),
              wind=rep(nomi,each=n*2))

  s<-0

  for (i in 1:11){
    g<-graph_from_adjacency_matrix(ls[[i]],
                                   mode="undirected",
                                   weighted=TRUE)
    grado<-degree(g)
    betw<-betweenness(g)

    tab[(s+1):(s+n),]$value<-grado
    tab[(s+n+1):(s+n*2),]$value<-betw

    s<-s+2*n
  }
  return(tab)
}



tab_men<-deg_betw_long(wind.m.norm.list)
tab_men$gen<-rep("Uomini",nrow(tab_men))
tab_women<-deg_betw_long(wind.w.norm.list)
tab_women$gen<-rep("Donne",nrow(tab_women))

box.gr<-tab_compl%>%filter(var=="Grado")%>%
  ggplot(aes(x=wind,y=value,color=gen))+
  geom_boxplot(show.legend = F,outlier.alpha = 0.2)+
  facet_wrap(~gen,nrow=2)+
  labs(x="",y="Grado")
box.gr

box.btw<-tab_compl%>%filter(var=="Betweenness")%>%
```

```
  ggplot(aes(x=wind,y=value,color=gen))+
  geom_boxplot(show.legend = F,outlier.alpha = 0.2)+
  facet_wrap(~gen,nrow=2)+
  labs(x="",y="Betweenness")
box.btw
```

B.5 Permanenza degli atleti nel campionato

```
table(men_def$nstag==1)%>%prop.table()
table(women_def$nstag==1)%>%prop.table()


stag_stop<-function(x){
  nas<-is.na(x)
  first_last<-which(!nas)%>%.[c(1,length(.))]
  x1<-x[first_last[1]:first_last[2]]
  nas1<-nas[first_last[1]:first_last[2]]

  anni_stop<-sum(nas1)
  #numero di anni di stop tra il primo e l'ultimo in italia

  ls<-split(x1[!nas1], cumsum(nas1)[!nas1])
  n_stop<-ls%>%length() -1
  #numero di stop tra le stagioni in italia
  m_nstag_cons<-sapply(ls, length)%>%max
  #massimo numero di anni consecutivi in italia

  return(list(n_stop,anni_stop,m_nstag_cons))
}

stop.m.df<-apply(men_def[,7:21],1,stag_stop)%>%
  do.call(rbind.data.frame,.)%>%
  `colnames<-`(c("n_stop","anni_stop","max_cons"))%>%
  mutate(gen="Uomini")%>%
  mutate(nstag=men_def$nstag)

stop.w.df<-apply(women_def[,7:21],1,stag_stop)%>%
  do.call(rbind.data.frame,.)%>%
  `colnames<-`(c("n_stop","anni_stop","max_cons"))%>%
  mutate(gen="Donne")%>%
  mutate(nstag=women_def$nstag)

stop.df<-bind_rows(stop.m.df,stop.w.df)

#proporzione di stop per chi è stato almeno 2 stagioni
pt<-prop.table(table(stop.m.df$n_stop,men_def$nstag==1),2)
```

```
apply(pt[4:1,],2,cumsum)%>%round(.,digits=4)
pt<-prop.table(table(stop.w.df$n_stop,women_def$nstag==1),2)
apply(pt[4:1,],2,cumsum)%>%round(.,digits=4)

ggplot(stop.df,aes(x=nstag,fill=n_stop%>%as.factor))+
  geom_bar(position = "fill")+
  facet_wrap(~gen)+
  labs(x="Numero di stagioni",y="Frequenza relativa",
       fill="Numero di
interruzioni")



#rinnovamento annuale del campionato
ngiocM<-c()
no_precM<-c(NA)
for (anno in 1:15){
  nas<-men_def%>%.[6+anno]%>%is.na(.)
  ngiocM[anno]<-1019-nas%>%sum

  if (anno==1) next
  no_precM[anno]<-men_def%>%.[!nas,6+anno-1]%>%
    is.na(.)%>%sum
}


(no_precM/ngiocM)%>%mean(na.rm=T)


ngiocW<-c()
no_precW<-c(NA)
for (anno in 1:15){
  nas<-women_def%>%.[6+anno]%>%is.na(.)
  ngiocW[anno]<-1019-nas%>%sum

  if (anno==1) next
  no_precW[anno]<-women_def%>%.[!nas,6+anno-1]%>%
    is.na(.)%>%sum
}

(no_precW/ngiocW)%>%mean(na.rm=T)
```

Appendice C

Codice R: Analisi dei club

C.1 Permanenza dei club

```
nstag.clubM<-apply(club.m.df[,2:16],1,function(x) 15-sum(is.na(x)))
nstag.clubW<-apply(club.w.df[,2:16],1,function(x) 15-sum(is.na(x)))

mean(nstag.clubM)
mean(nstag.clubW)

nstag.plot<-tibble(nstag=c(nstag.clubM,nstag.clubW),
                   gen=c(rep("Uomini",28),rep("Donne",48)))%>%
  ggplot(aes(x=nstag,y=after_stat(density)))+
  geom_histogram(bins=8)+
  facet_wrap(~gen)
nstag.plot+
  labs(x="Numero di stagioni dei club in A1",
       y="Densità")


image(x=1:length(soc),y=2008:2022,z=club,col=c(0,1),xlab="Club",ylab="Stagioni",
      xaxt="n",yaxt="n")
abline(v=1.5:(length(soc)-0.5),col="lightgray",lty="dotted")
abline(h=2008.5:2021.5,col="lightgray",lty="dotted")
box(col="gray")
axis(2, at =2008:2022 ,labels=T,cex.axis=0.7,las=2)
axis(1, at =1:length(soc) ,labels=soc, las=2,cex.axis=0.5)

image(x=1:length(soc_w),y=2008:2022,z=club_w,col=c(0,1),xlab="Club",ylab="Stagioni",
      xaxt="n",yaxt="n")
abline(v=1.5:(length(soc_w)-0.5),col="lightgray",lty="dotted")
abline(h=2008.5:2021.5,col="lightgray",lty="dotted")
box(col="gray")
axis(2, at =2008:2022 ,labels=T,cex.axis=0.7,las=2)
axis(1, at =1:length(soc_w) ,labels=soc_w, las=2,cex.axis=0.5)
```

C.2 Creazione delle reti dei club

```
#campionato maschile
for (anno in 1:14){

  matr<-matrix(0,nrow=length(soc),ncol=length(soc),
               dimnames = list(soc,soc))


  atl.soc<-men_clean_teams%>%
    filter(! is.na(.[anno+1])& ! is.na(.[anno]))%>%
    .[c(anno,anno+1)]

  for (i in 1:nrow(atl.soc)){
    riga<-which(rownames(matr)==atl.soc[i,1]%>%as.character())
    colonna<-which(colnames(matr)==atl.soc[i,2]%>%as.character())
    matr[riga,colonna]<- matr[riga,colonna]+1
  }

  assign(paste0("soc.M.",2008+anno),matr)
}


soc.M.list<-list(soc.M.2009,soc.M.2010,soc.M.2011,soc.M.2012,soc.M.2013,
                 soc.M.2014,soc.M.2015,soc.M.2016,soc.M.2017,soc.M.2018,
                 soc.M.2019,soc.M.2020,soc.M.2021,soc.M.2022)

rm("soc.M.2009","soc.M.2010","soc.M.2011","soc.M.2012","soc.M.2013",
   "soc.M.2014","soc.M.2015","soc.M.2016","soc.M.2017","soc.M.2018",
   "soc.M.2019","soc.M.2020","soc.M.2021","soc.M.2022")


for (i in 1:14){
  write.csv(soc.M.list[[i]],paste0("club_csv/clubM",2008+i,".csv"))

  vert.club.m.i<-tibble(id=soc,
                        rimasti=diag(soc.M.list[[i]]),
                        class=club.m.df[,i+1]%>%pull,
                        vinc=club.m.df[,i+16]%>%pull)%>%
    filter(!is.na(class))

  write.csv(vert.club.m.i,
            paste0("club_csv/clubM",2008+i,"_vert.csv"),row.names = F)
}


#campionato femminile
for (anno in 1:14){
```

```
  matr<-matrix(0,nrow=length(soc_w),ncol=length(soc_w),
               dimnames = list(soc_w,soc_w))


  atl.soc<-women_clean_teams%>%
    filter(! is.na(.[anno+1]) & ! is.na(.[anno]))%>%
    .[c(anno,anno+1)]

  for (i in 1:nrow(atl.soc)){
    riga<-which(rownames(matr)==atl.soc[i,1]%>%as.character())
    colonna<-which(colnames(matr)==atl.soc[i,2]%>%as.character())
    matr[riga,colonna]<- matr[riga,colonna]+1
  }

  assign(paste0("soc.W.",2008+anno),matr)
}


soc.W.list<-list(soc.W.2009,soc.W.2010,soc.W.2011,soc.W.2012,soc.W.2013,
                 soc.W.2014,soc.W.2015,soc.W.2016,soc.W.2017,soc.W.2018,
                 soc.W.2019,soc.W.2020,soc.W.2021,soc.W.2022)

rm("soc.W.2009","soc.W.2010","soc.W.2011","soc.W.2012","soc.W.2013",
   "soc.W.2014","soc.W.2015","soc.W.2016","soc.W.2017","soc.W.2018",
   "soc.W.2019","soc.W.2020","soc.W.2021","soc.W.2022")


for (i in 1:14){
  write.csv(soc.W.list[[i]],paste0("club_csv/clubW",2008+i,".csv"))

  vert.club.w.i<-tibble(id=soc_w,
                        rimasti=diag(soc.W.list[[i]]),
                        class=club.w.df[,i+1]%>%pull,
                        vinc=club.w.df[,i+16]%>%pull)%>%
    filter(!is.na(class))

  write.csv(vert.club.w.i,
            paste0("club_csv/clubW",2008+i,"_vert.csv"),row.names = F)
}
```

C.3 Modello per risultato in Regular Season

```
#CAMPIONATO MASCHILE
men.df<-tibble(nasc=men_def$Nasc%>%as.numeric,
               ita=ifelse(men_def$Naz.sport=="ITA",1,0),
               grado=deg_m,
```

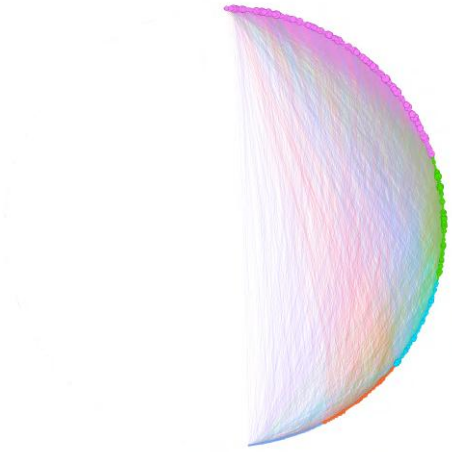
```
                betw=btw_m)

anni.list<-list()
for (i in 2:15){
  sel<-men.df%>%
    mutate(anno = men_clean_teams[,i]%>%pull)%>%
    filter(!is.na(anno))%>%
    group_by(anno)%>%
    summarise(m.eta=2007+i-mean(nasc,na.rm=T),
              p.ita=mean(ita,na.rm=T),
              m.grado=mean(grado),
              m.betw=mean(betw))


  anno.df<-tibble(club=soc,
                  rimasti=diag(soc.M.list[[i-1]]),
                  rs=club.m.df[,1+i]%>%pull,
                  rs.prec=club.m.df[,i]%>%pull)%>%
    inner_join(sel,by=c("club"="anno"))%>%
    mutate(anno = 2007+i)

  anni.list[i-1]<-list(anno.df)
}

men.class<-do.call(bind_rows,anni.list)
men.class<-men.class%>%
  mutate(prec01=ifelse(is.na(men.class$rs.prec),0,1))%>%
  mutate(rs.prec0=ifelse(is.na(men.class$rs.prec),
                         0,men.class$rs.prec))

m.rs.lm<-lm(rs~m.eta+p.ita+m.grado+m.betw+
              prec01+prec01:rs.prec0+prec01:rimasti,
            data=men.class)
summary(m.rs.lm)



#CAMPIONATO FEMMINILE
women.df<-tibble(nasc=women_def$Nasc%>%as.numeric,
                 ita=ifelse(women_def$Naz.sport=="ITA",1,0),
                 grado=deg_w,
                 betw=btw_w)

club.w.df.99<-club.w.df%>%
  replace(club.w.df==99,NA)

anni.list<-list()
for (i in 2:15){
  sel<-women.df%>%
```

```
    mutate(anno = women_clean_teams[,i]%>%pull)%>%
    filter(!is.na(anno))%>%
    group_by(anno)%>%
    summarise(m.eta=2007+i-mean(nasc,na.rm=T),
              p.ita=mean(ita,na.rm=T),
              m.grado=mean(grado),
              m.betw=mean(betw))


  anno.df<-tibble(club=soc_w,
                  rimasti=diag(soc.W.list[[i-1]]),
                  rs=club.w.df[,1+i]%>%pull,
                  rs.prec=club.w.df[,i]%>%pull)%>%
    inner_join(sel,by=c("club"="anno"))%>%
    mutate(anno = 2007+i)

  anni.list[i-1]<-list(anno.df)
}

women.class<-do.call(bind_rows,anni.list)

#rimuoviamo osservazioni su società che si sono ritirate
women.class<-women.class%>%
  filter(rs!=99)
women.class<-women.class%>%
  mutate(prec01=ifelse(is.na(women.class$rs.prec),0,1))%>%
  mutate(rs.prec0=ifelse(is.na(women.class$rs.prec),
                         0,women.class$rs.prec))

w.rs.lm<-lm(rs~m.eta+p.ita+m.grado+m.betw+
              prec01+prec01:rs.prec0+prec01:rimasti,
            data=women.class)
summary(w.rs.lm)
```

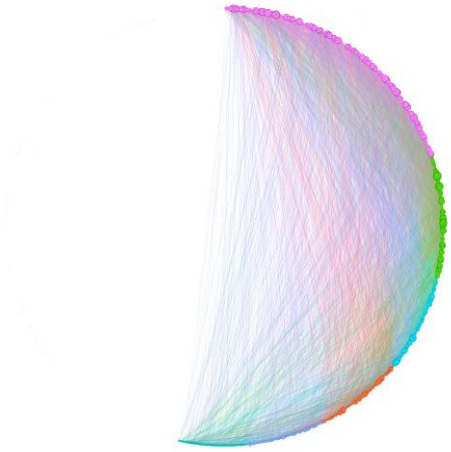
Appendice D

Altre visualizzazioni di reti

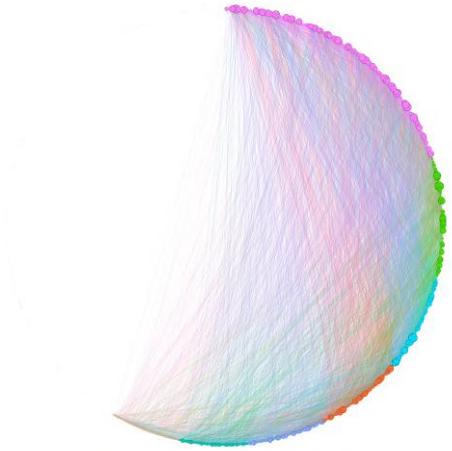
D.1 Campionato maschile (finestra di 5 anni)



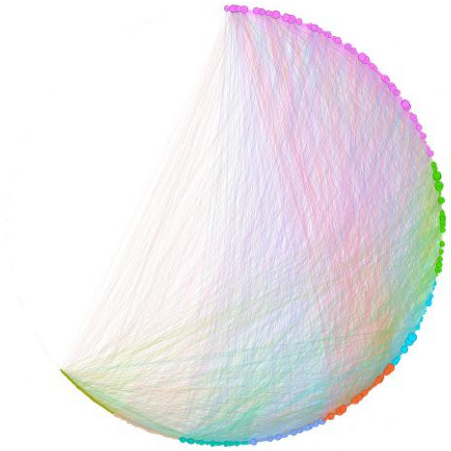
(a) 2008/09 — 2012/13.



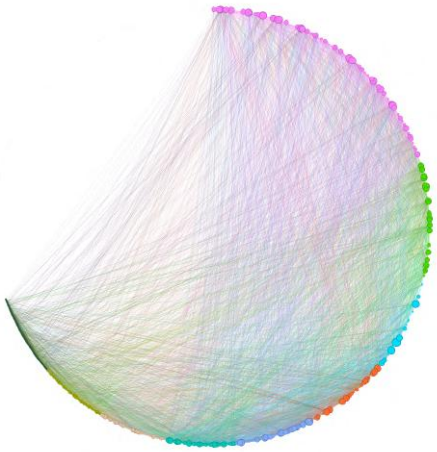
(b) 2009/10 — 2013/14.



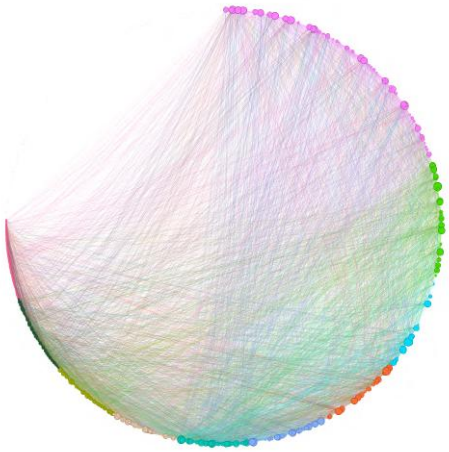
(c) 2010/11 — 2014/15.



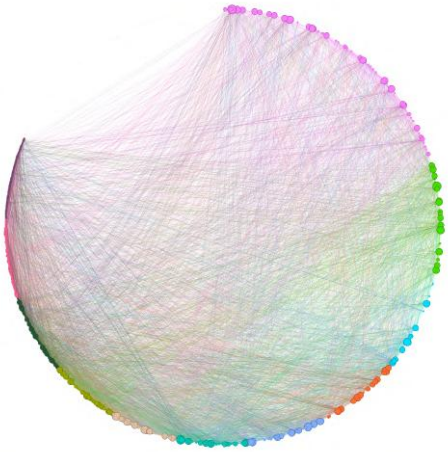
(d) 2011/12 — 2015/16.



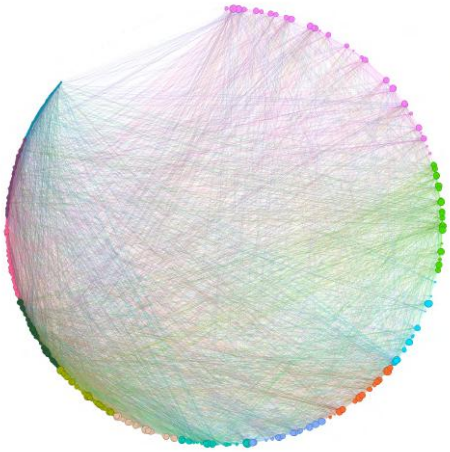
(e) 2012/13 — 2016/17.



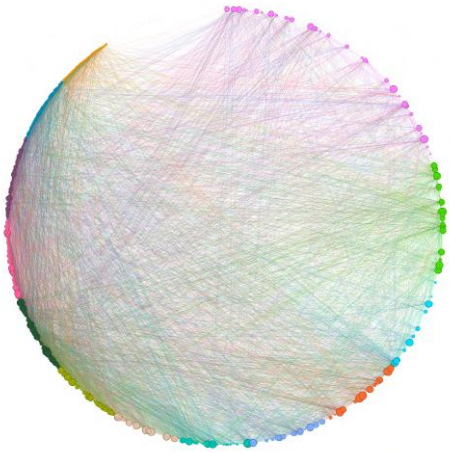
(f) 2013/14 — 2017/18.



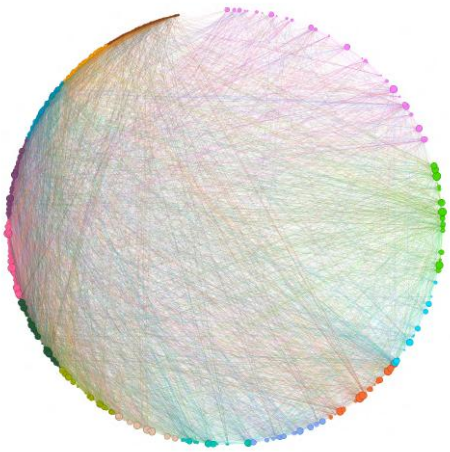
(g) 2014/15 — 2018/19.



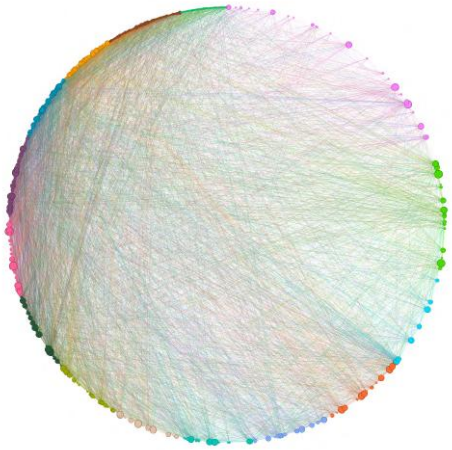
(h) 2015/16 — 2019/20.



(i) 2016/17 — 2020/21.



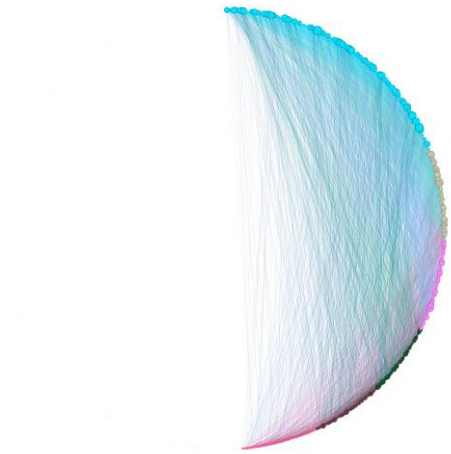
(j) 2017/18 — 2021/22.



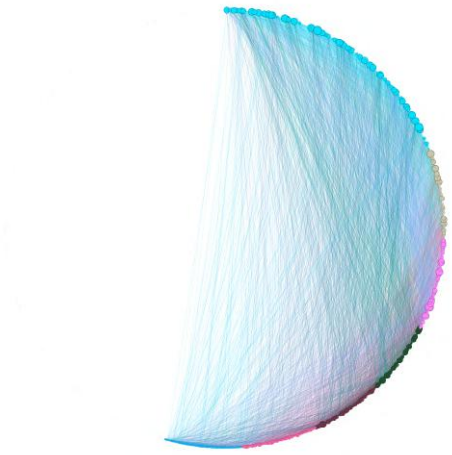
(k) 2018/19 — 2022/23.

D.2 Campionato femminile (finestra di 5 anni)

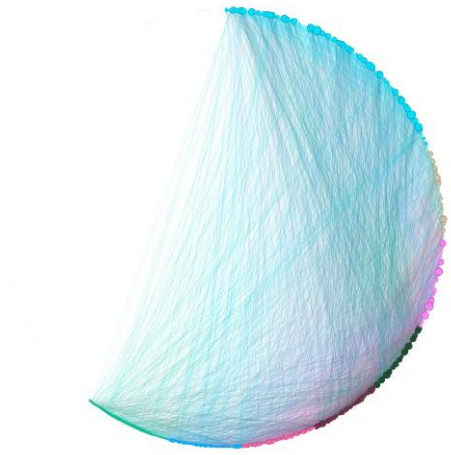
(a) 2008/09 — 2012/13.



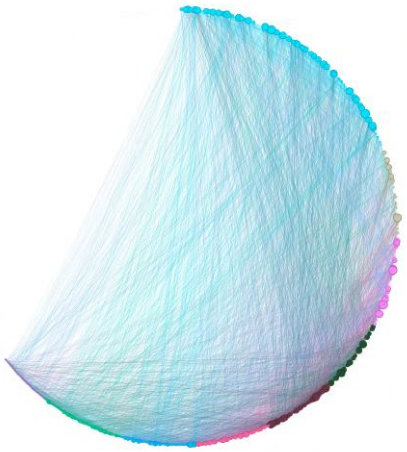
(b) 2009/10 — 2013/14.



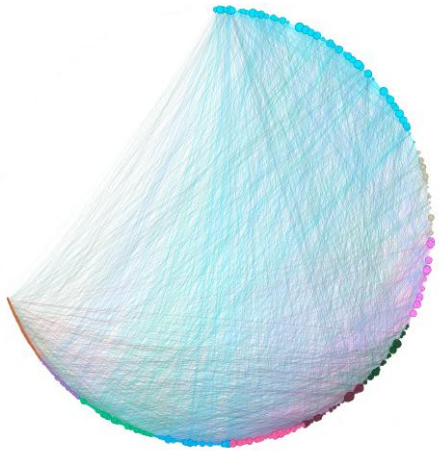
(c) 2010/11 — 2014/15.



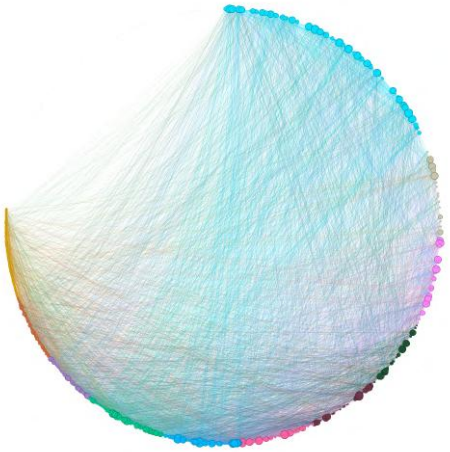
(d) 2011/12 — 2015/16.



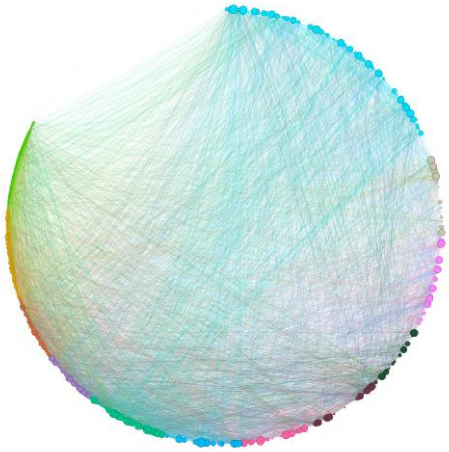
(e) 2012/13 — 2016/17.



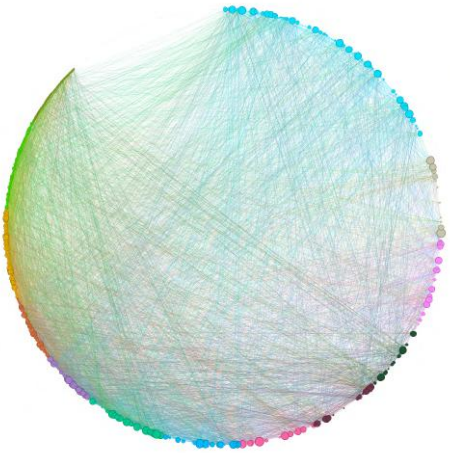
(f) 2013/14 — 2017/18.



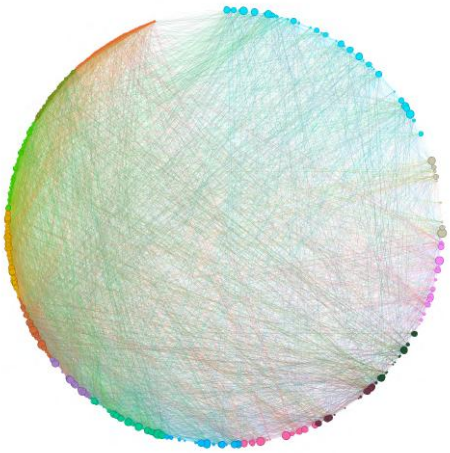
(g) 2014/15 — 2018/19.



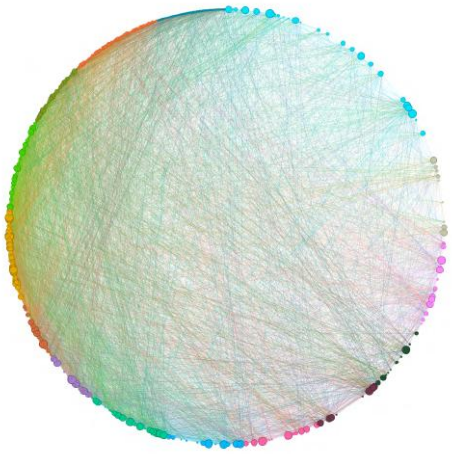
(h) 2015/16 — 2019/20.



(i) 2016/17 — 2020/21.

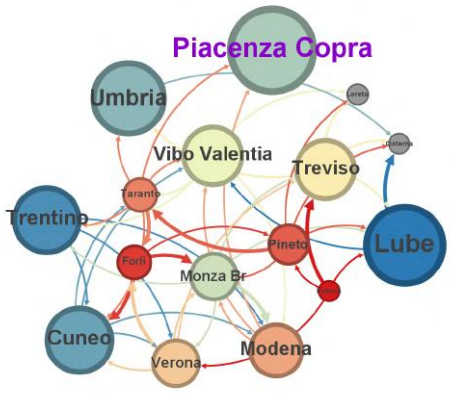


(j) 2017/18 — 2021/22.

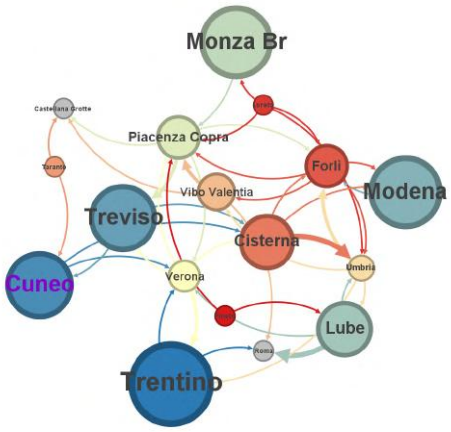


(k) 2018/19 — 2022/23.

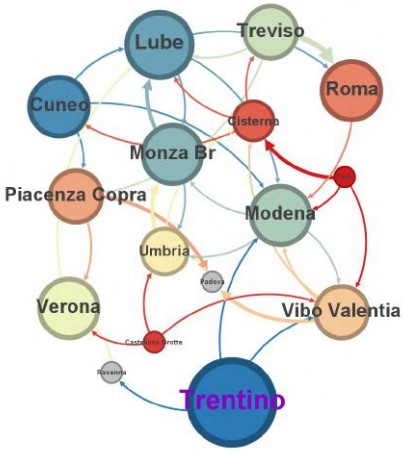
D.3 Società del campionato maschile



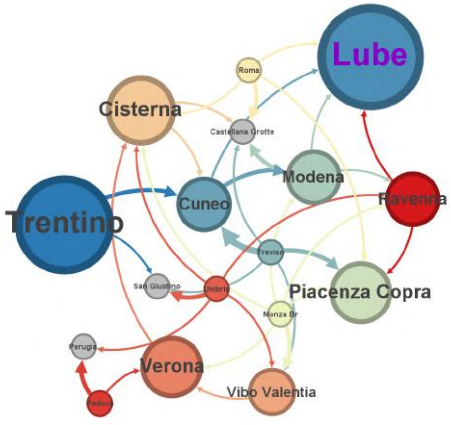
(a) 2008/09 — 2009/10.



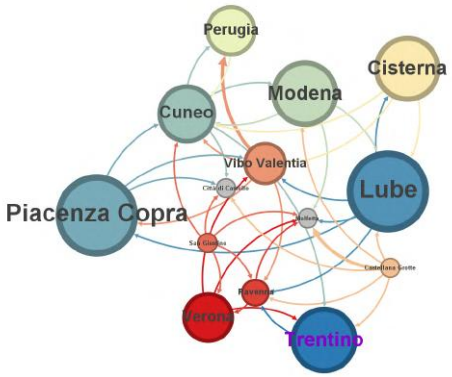
(b) 2009/10 — 2010/11.



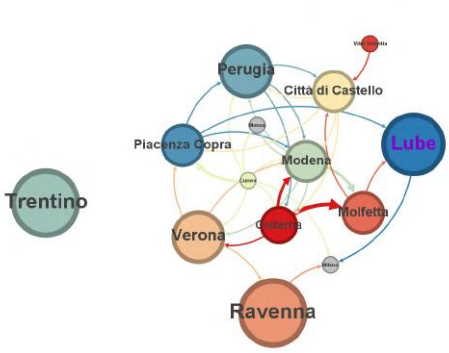
(c) 2010/11 — 2011/12.



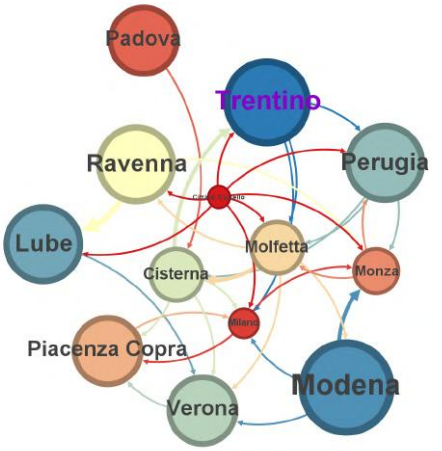
(d) 2011/12 — 2012/13.



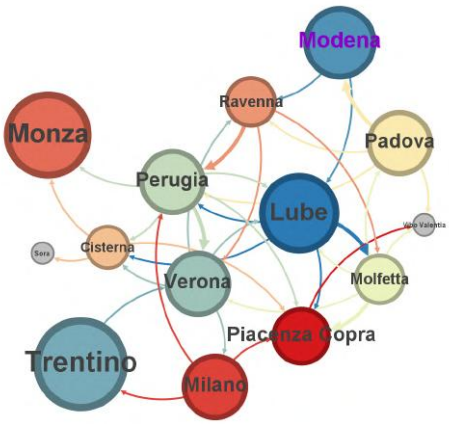
(e) 2012/13 — 2013/14.



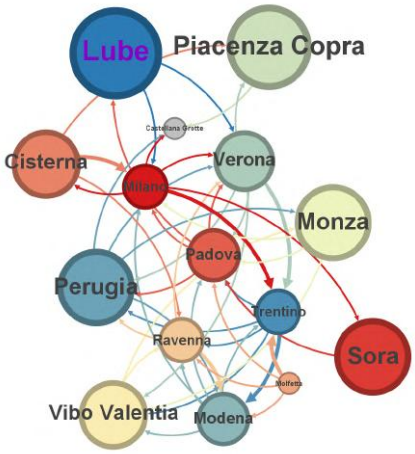
(f) 2013/14 — 2014/15.



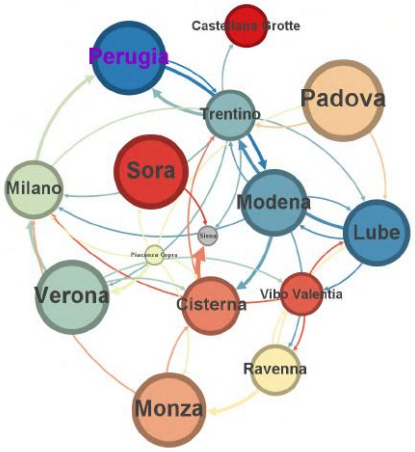
(g) 2014/15 — 2015/16.



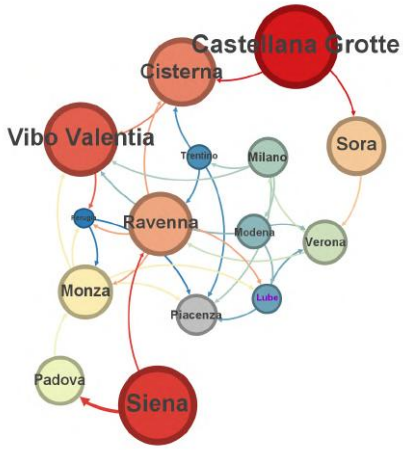
(h) 2015/16 — 2016/17.



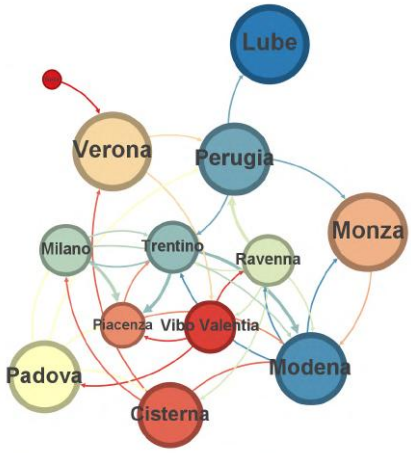
(i) 2016/17 — 2017/18.



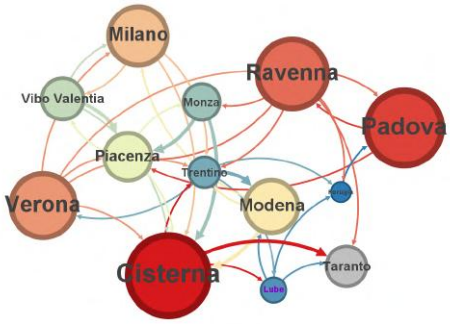
(j) 2017/18 — 2018/19.



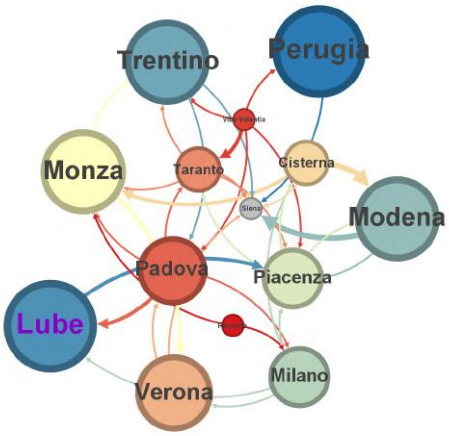
(k) 2018/19 — 2019/20.



(l) 2019/20 — 2020/21.

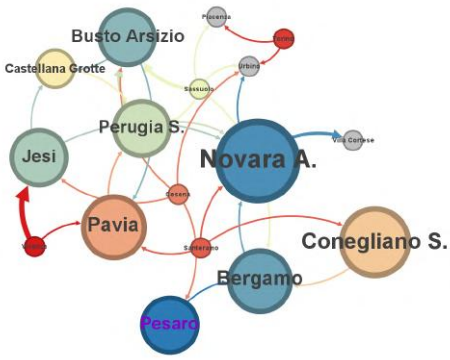


(m) 2020/21 — 2021/22.

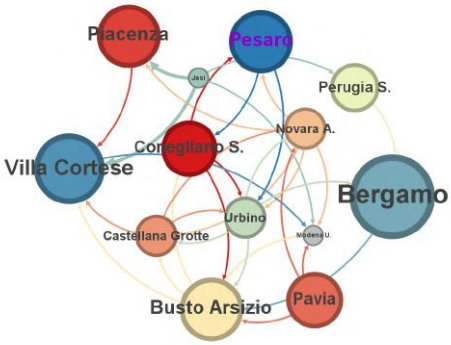


(n) 2021/22 — 2022/23.

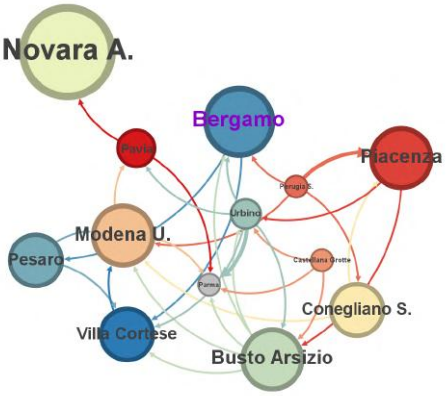
D.4 Società del campionato femminile



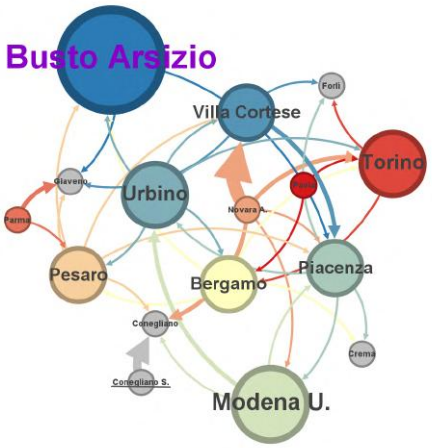
(a) 2008/09 — 2009/10.



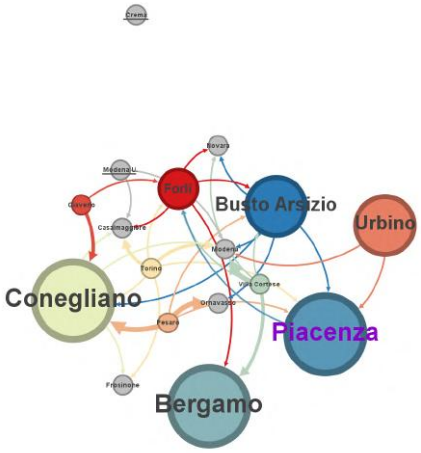
(b) 2009/10 — 2010/11.



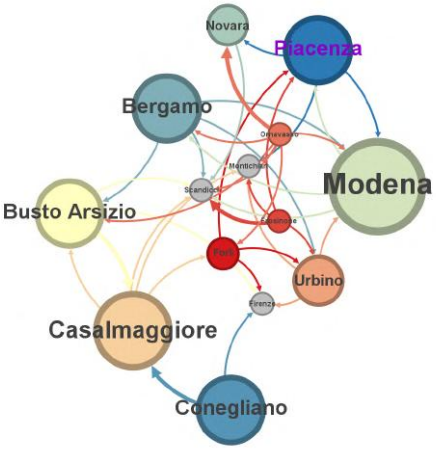
(c) 2010/11 — 2011/12.



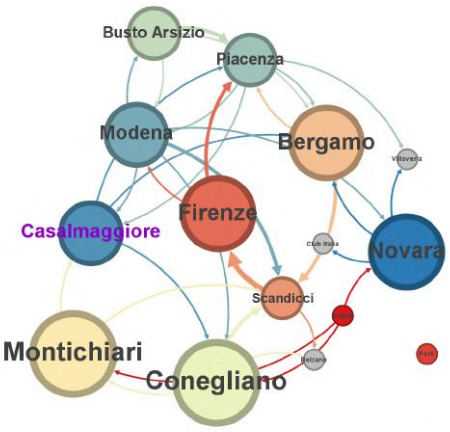
(d) 2011/12 — 2012/13.



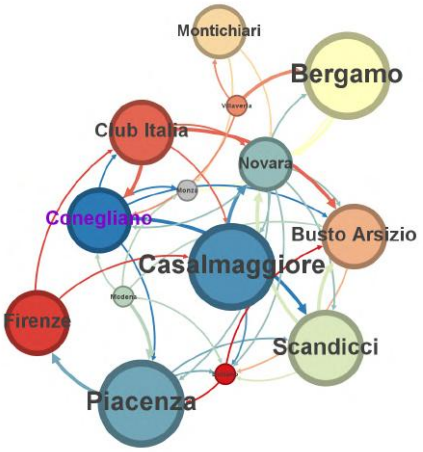
(e) 2012/13 — 2013/14.



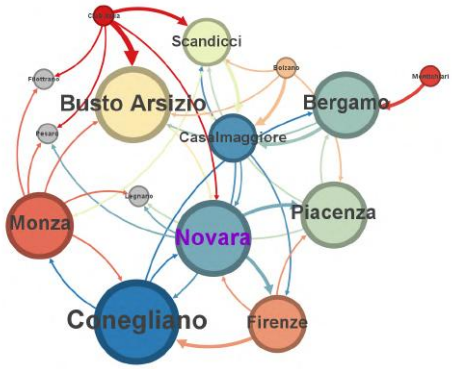
(f) 2013/14 — 2014/15.



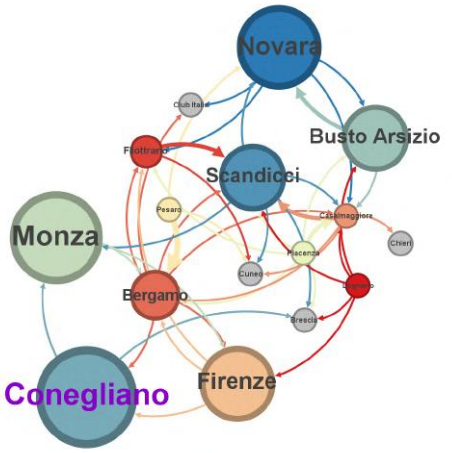
(g) 2014/15 — 2015/16.



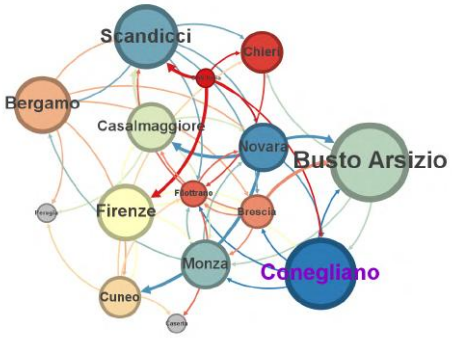
(h) 2015/16 — 2016/17.



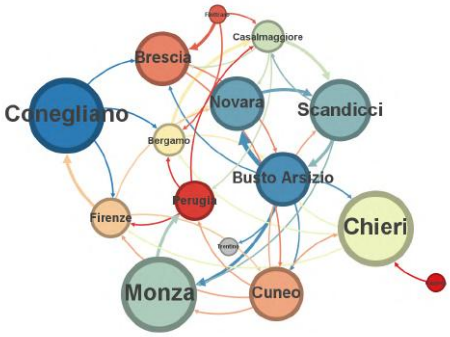
(i) 2016/17 — 2017/18.



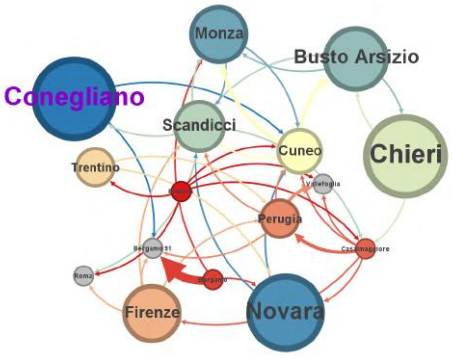
(j) 2017/18 — 2018/19.



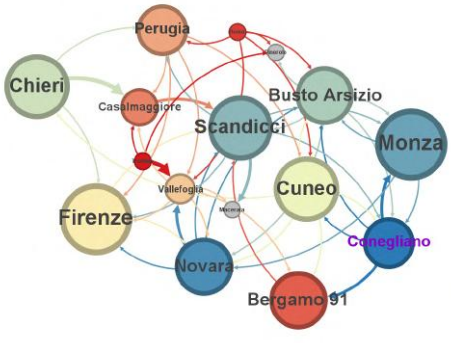
(k) 2018/19 — 2019/20.



(l) 2019/20 — 2020/21.



(m) 2020/21 — 2021/22.



(n) 2021/22 — 2022/23.