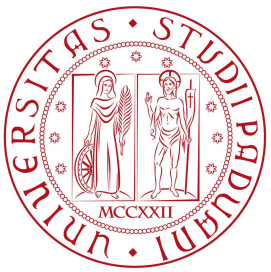
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche



Raggruppamento tra reti: sviluppi metodologici e un'applicazione al gioco del calcio

Relatore: dott. Bruno Scarpa
Dipartimento di Scienze Statistiche

Laureando: Jacopo Diquigiovanni
Matricola n. 1130021

Anno Accademico 2016/2017

Indice

Capitolo 1

Introduzione

I metodi di raggruppamento costituiscono un variegato insieme di tecniche il cui scopo è partizionare le unità statistiche in un certo numero di gruppi. A seconda della tipologia di dato a disposizione, l'approccio utilizzato varia in modo da tenere debitamente in considerazione la natura specifica del problema affrontato.

Il presente lavoro ha come obiettivo l'ideazione di una metodologia apposita in presenza di *dati di rete*: in tale contesto, il fine sarà quello di suddividere le n reti che contraddistinguono il campione statistico in un determinato numero di *clusters* in accordo con un criterio opportuno. Vista la vastità della tematica, l'attenzione sarà posta su reti *pesate indirette* che condividono l'insieme $\mathcal{K}$ dei nodi.

La trattazione si sviluppa come segue.

Nel capitolo 2 viene presentata una formulazione matematica rigorosa della *rete* e il lettore è introdotto al concetto chiave di *popolazione di reti*.

Nel capitolo 3, dopo aver descritto il panorama letterario riguardante la *cluster analysis*, vengono evidenziate le criticità del contesto analizzato e viene illustrato il metodo di raggruppamento nel suo complesso.

Nel capitolo 4 si valuta l'accuratezza della metodologia in alcuni possibili scenari attraverso uno studio di simulazione, confrontando i risultati con quelli ottenuti attraverso un approccio alternativo.

Infine, nel capitolo 5 il metodo trova applicazione nell'analisi degli *schemi di gioco* delle squadre del campionato di calcio di Serie A TIM stagione 2015-2016.

Capitolo 2

I dati di rete

L'aumento esponenziale di informazione disponibile che ha caratterizzato gli ultimi anni ha comportato nuove affascinanti sfide nei più disparati ambiti del sapere, tra cui la statistica. Tale fenomeno richiede profonda innovazione in campo scientifico, dall'archiviazione dei *byte* fruibili all'utilizzo di rinnovate metodologie per l'estrazione dell'informazione dai dati stessi. Illuminanti, in tal senso, sono le parole del celebre statistico Jerome H. Friedman, che afferma:

> "*Ogniqualvolta una tecnologia aumenta il volume di dati di un fattore di dieci, bisogna completamente ripensare a come analizzarli.*"(Friedman, 1998)

Tale complessità non riguarda solamente la quantità d'informazione disponibile, ma anche la sua tipologia: difatti è possibile rilevare e gestire strutture sempre più complesse di dati rispetto ai tipici *dataset* composti da righe (osservazioni) e colonne (variabili).

Uno strumento straordinariamente utile per rappresentare particolari tipologie di dato è la *rete*, esprimibile nella sua forma più elementare come un insieme di *nodi*, noti anche come *attori* soprattutto in ambito sociometrico, collegati tra loro da una qualche forma di associazione, sia essa concreta o totalmente astratta. Nonostante lo sviluppo piuttosto recente, l'analisi di tali strutture, a partire da quella puramente descrittivo-esplorativa per arrivare a modellazioni assai più elaborate, ha avuto una diffusione estesa ai più svariati settori di ricerca grazie alla notevole flessibilità ed adattabilità. Come riportato da Wasserman & Faust (1994), i primi esempi in letteratura di reti riguardano l'ambito prettamente sociologico (Barnes, 1954), ma l'impiego in numerose scienze applicate quali la finanza, la bioinformatica e le neuroscienze è

stato capillare: Górski *et al.* (2008), Jonsson *et al.* (2006) e Bassett & Bullmore (2006) sono solo alcuni dei numerosissimi esempi a proposito.

In ambito statistico i motivi del crescente interesse nei confronti di tale tipologia di dato risiedono nell'intrinseca capacità di cogliere le connessioni esistenti tra gli attori del fenomeno oggetto di studio. A differenza dei tradizionali *dataset* costituiti da osservazioni di un certo numero di variabili rilevate su soggetti indipendenti, il centro dell'analisi delle reti è rappresentato dalle relazioni d'interdipendenza esistenti tra i nodi della rete, aspetto di sempre maggiore interesse in numerose applicazioni.

Nel paragrafo 2.1 vengono presentate le nozioni principali riguardanti le reti e la relativa teoria dei grafi: per non appesantire eccessivamente la trattazione verranno riportati solo i concetti basilari per comprendere le tematiche descritte, evitando di indugiare eccessivamente su aspetti di marginale utilità. Per una disamina completa e dettagliata si rimanda, ad esempio, a Newman (2010).

2.1 Introduzione alle reti

Prima di fornire una definizione rigorosa di rete è necessario evidenziare una distinzione fondamentale a seconda della tipologia delle relazioni che compongono la stessa; tale scelta risulta obbligata in quanto la classificazione influisce sulla notazione da utilizzare.

Si definisce rete *binaria* una rete che ammette al più una sola connessione tra due nodi, delineando dunque una relazione del tipo presente/assente; una rete viene definita invece *pesata* se al collegamento esistente tra due nodi viene associato un peso che ne determina l'intensità: nel corso della trattazione, se non esplicitato diversamente, verrà considerato solo il caso di pesi positivi.

Una formulazione matematica rigorosa del concetto di rete è resa possibile dalla sottostante teoria dei grafi. In accordo con la notazione utilizzata da Boccaletti *et al.* (2006), si definisce *grafo monolivello binario* $G = (\mathcal{K}, \mathcal{L})$ la struttura matematica costituita dai due insiemi $\mathcal{K} \neq \emptyset$ e $\mathcal{L}$, di cui l'ultimo composto da coppie di elementi di $\mathcal{K}$; in particolare, gli elementi di $\mathcal{K} = \{k_1, k_2, \ldots, k_K\}$ rappresentano i K nodi che caratterizzano la rete e gli elementi di $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$ descrivono gli L archi che connettono le coppie di nodi dell'insieme $\mathcal{K}$. Sebbene accettabile da un punto di vista teorico, il caso in cui $\mathcal{L} = \emptyset$ non verrà preso in considerazione in quanto costituisce una rete priva

di connessioni: conseguentemente, considereremo $L > 0$. La definizione di *grafo monolivello pesato* rappresenta la naturale generalizzazione del concetto appena esposto: tale grafo $G = (\mathcal{K}, \mathcal{L}, \mathcal{W})$ presenta, rispetto al precedente, un ulteriore insieme $\mathcal{W} = \{w_1, w_2, \ldots, w_L\}$ contenente i pesi relativi agli archi dell'insieme $\mathcal{L}$.

Un'ulteriore classificazione dipende dalla direzionalità delle connessioni: una rete *diretta* prevede che un determinato arco abbia una direzione, ossia che il collegamento esistente tra due nodi consti di un soggetto che promuove la relazione e di un soggetto che la riceve, mentre una rete *indiretta* implica l'assenza di direzionalità degli archi che compongono il grafo. Da un punto di vista matematico, la rete diretta prevede un ordinamento tra le coppie di nodi che costituiscono l'insieme $\mathcal{L}$: conseguentemente i generici elementi $\{u, v\}$ e $\{v, u\}$, con $u \neq v$, dell'insieme $\mathcal{L}$ rappresentano elementi distinti e indicanti due relazioni differenti. L'elemento $\{u, u\}$ denota la presenza di un arco che collega un nodo con se stesso: noto come *self-loop*, risulta particolarmente interessante soprattutto in presenza di reti pesate. Poiché la procedura che verrà illustrata nel capitolo 3 prevede l'utilizzo di reti pesate indirette, successivamente l'attenzione sarà posta esclusivamente su questa tipologia di rete: le proprietà e le definizioni varranno quindi per tali reti e non necessariamente in generale.

Alternativamente a quanto esposto finora, una rete può essere agevolmente rappresentata tramite una *matrice di pesi* W quadrata di dimensione $K \times K$ il cui generico elemento w_{ij} esprime la relazione esistente tra il nodo i ed il nodo j. Nello specifico $w_{ij} \in \mathbb{R}$ ha valore pari al peso della relazione esistente tra il nodo i ed il nodo j, con $w_{ij} = 0$ qualora i due nodi non siano connessi. La matrice W nel caso in considerazione risulta simmetrica, con gli elementi sulla diagonale ad indicare la presenza, ed eventualmente il peso, di *self-loops* all'interno della rete considerata.

Di seguito vengono riportate alcune proprietà che torneranno utili nello sviluppo della trattazione:

- *grado di un nodo*: peso complessivo degli archi che connettono un determinato nodo ai nodi presenti nella rete. A partire dalla matrice dei pesi è possibile calcolare il grado d_i relativo al nodo i:

$$d_i = \sum_{j=1}^{K} w_{ij} = \sum_{j=1}^{K} w_{ji} \tag{2.1}$$

L'uguaglianza è resa possibile dalla mancanza di direzionalità degli archi caratterizzanti il grafo.

- *media dei gradi*: statistica di sintesi che riporta la media dei gradi dei diversi nodi presenti all'interno di una rete:

$$\bar{d} = \frac{\sum_{i=1}^{K} d_i}{K} \tag{2.2}$$

- *varianza dei gradi*: similmente al valore medio, statistica che riporta la varianza dei gradi dei nodi della rete:

$$s_D^2 = \frac{\sum_{i=1}^{K} (d_i - \bar{d})^2}{K} \tag{2.3}$$

2.2 Popolazioni di reti

A partire dal concetto di grafo monolivello pesato, è possibile definire come *grafo multilivello pesato a nodi comuni* $\mathbf{G} = \{G^1, \ldots, G^n\}$ una collezione di n grafi monolivello pesati, con:

$$G^i = (\mathcal{K}, \mathcal{L}^i, \mathcal{W}^i) \qquad \forall i = 1, 2, \ldots, n \tag{2.4}$$

La rete $\mathbf{G}$ rappresenta un caso particolare di un concetto ben più generale noto come rete *multilayer*, qui non trattato: approfondimenti sono disponibili in Boccaletti *et al.* (2014).

Si noti come i nodi che costituiscono il grafo multilivello considerato siano comuni tra i livelli, mentre gli archi e i relativi pesi siano lasciati liberi di variare.

Una diversa interpretazione della struttura appena presentata permette un utile parallelismo con l'analisi statistica classica: in tale ottica, la rete multilivello può essere interpretata come un *campione statistico* di cui le n reti monolivello costituiscono le singole unità. L'unico vincolo richiesto da tale formulazione è che le unità statistiche condividano l'insieme $\mathcal{K}$ dei nodi. Nel prosieguo, per semplicità, con il termine rete si indicherà una rete monolivello.

Il concetto di campione di reti - e quello di *popolazione* strettamente connesso - apre a nuovi interessanti scenari: a partire da una collezione di reti può risultare d'interesse applicare procedure inferenziali mirate ad ottenere informazioni sulla popolazione di riferimento, analogamente

a quanto si è soliti fare in presenza dei classici *dataset*. Chiaramente la complessità del dato di rete richiede metodologie *ad hoc* che tengano opportunamente in considerazione la struttura di interdipendenza dei nodi e la variabilità delle reti che compongono la popolazione studiata.

In tal senso, un esempio è fornito dai dati esaminati nel capitolo 5: durante un campionato di calcio, in ogni partita due squadre si affrontano e, quando sono in possesso della palla, muovono il pallone lungo il campo definendo il cosiddetto *schema di gioco*. L'esistenza di uno spazio comune a tutte le formazioni - ossia il campo da calcio - permette di inquadrare lo studio di quest'aspetto all'interno del contesto presentato. In questo caso, ogni unità statistica si riferisce alla prestazione di una specifica squadra in una determinata partita, l'insieme dei nodi comune $\mathcal{K}$ sarà costituito dalle fasce di terreno (o *zone*) in cui il campo di gioco viene diviso, e gli archi rappresentano le transizioni del pallone tra le diverse zone.

Capitolo 3

Clustering di reti

3.1 Introduzione ai metodi di raggruppamento

All'interno della vasta gamma di tecniche di apprendimento non supervisionato (ad esempio, Hastie *et al.*, 2009), la classificazione di un insieme di elementi in differenti categorie rappresenta una pratica assai diffusa ed estremamente utile per i più disparati scopi: a tal proposito, si definiscono *metodi di raggruppamento* un eterogeneo insieme di metodologie volte a suddividere i dati in una serie di gruppi attraverso il soddisfacimento di un determinato criterio guida. Nel corso degli anni si sono susseguiti differenti approcci al problema, a seconda delle assunzioni formulate e della complessità del dato analizzato.

In presenza della classica unità statistica costituita da un vettore di variabili $x = (x_1, x_2, \ldots, x_p)$, i più noti metodi di raggruppamento si basano sul calcolo di una matrice di *dissimilarità* capace di esprimere numericamente il grado di *lontananza* tra le coppie di osservazioni che compongono il *dataset* (ad esempio, Azzalini & Scarpa, 2012). A fianco di queste procedure di natura prettamente geometrica si sono sviluppate ulteriori tecniche, come ad esempio i metodi *basati sul modello*, che permettono di trarre conclusioni inferenziali tramite la specificazione di una distribuzione di probabilità alla base del processo generatore dei dati (Fraley & Raftery, 1998). Con il progressivo aumento dell'informazione disponibile, sono state proposte apposite metodologie per l'analisi dei gruppi che, spesso, rappresentano generalizzazioni di quanto appena descritto: ad esempio, Viroli (2011) propone un approccio basato sul modello per raggruppare osservazioni costituite da p variabili rilevate ad r istanti temporali - o luoghi - differenti, estendendo quindi la *cluster*

analysis ad un campione di matrici indipendenti di dimensione $p \times r$.

L'analisi delle relazioni incognite presenti nei dati riveste un ruolo fondamentale nello studio delle reti: come affermato nel capitolo 2, gli archi che collegano i nodi di una rete delineano una struttura fortemente interconnessa che si presta agevolmente, per esempio, alla creazione di gruppi latenti sottostanti. Tale peculiarità trova fondamento nel concetto - anch'esso nato in ambito sociologico - di *omofilia*, ossia la tendenza da parte di un nodo a connettersi maggiormente con altri attori simili per una o più caratteristiche (McPherson *et al.*, 2001). L'attenzione a tale fondamentale tematica è riscontrabile, innanzitutto, nella capillare diffusione di indicatori descrittivi della propensione dei singoli nodi o della rete nel complesso a creare sottografi fortemente connessi; per una trattazione più approfondita si rimanda a Newman (2010).

Il fulcro della *cluster analysis* in presenza di dati di rete è rappresentato da un insieme di procedure noto come *analisi delle comunità*, punto di partenza del presente lavoro. Considerando per il momento un singolo grafo $G = (\mathcal{K}, \mathcal{L}, \mathcal{W})$, in numerose applicazioni risulta particolarmente interessante rilevare insiemi di nodi contraddistinti da un'alta connettività infragruppo e da una più modesta extragruppo. Sviluppati inizialmente per reti binarie, i metodi basati sulla massimizzazione di opportune funzioni obiettivo (Wu & Huberman, 2004), su algoritmi divisivi (Girvan & Newman, 2002) o agglomerativi (Pons & Latapy, 2006) rappresentano solo alcune direzioni verso cui la ricerca in tal senso si è sviluppata.

In siffatto contesto, un approccio particolarmente efficace, che verrà utilizzato all'interno della procedura spiegata nel paragrafo 3.2, è il *metodo di Louvain* (Blondel *et al.*, 2008). La metodologia è incentrata sul concetto di *modularità* di una partizione (Newman & Girvan, 2004): supponendo di disporre di una certa suddivisione dei K nodi in comunità, questa è definita come:

$$Q(c_1, \ldots, c_K; W) = \frac{1}{2m} \sum_{ij} \left[w_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3.1}$$

con W matrice di pesi, w_{ij} pari al peso della relazione esistente tra il nodo i e il nodo j, $k_i = \sum_{j=1}^{K} w_{ij}$ somma dei pesi degli archi connessi al nodo i, c_i comunità di cui fa parte il nodo i, m peso complessivo degli archi che compongono la rete e $\delta(x, y)$ funzione *delta di Kronecker*:

$$\delta(x, y) = \begin{cases} 1 & \text{se } x = y \\ 0 & \text{se } x \neq y \end{cases}$$

La modularità indica quindi la differenza tra la frazione di archi (considerati con il relativo peso w) che connettono nodi appartenenti allo stesso gruppo e il valore atteso della medesima quantità qualora le connessioni all'interno della rete fossero casuali. Pertanto, tale indicatore assume valore positivo se il peso degli archi che connettono i nodi *simili* è maggiore di quello atteso in caso di assegnazione aleatoria, risulta invece negativo in caso contrario; alla luce di ciò, appare chiaro che la modularità può essere considerata come un efficace strumento per valutare la qualità di una data partizione, e dunque una funzione obiettivo da massimizzare (Newman, 2004; Clauset *et al.*, 2004). Sfortunatamente, l'ottimizzazione esatta risulta spesso computazionalmente problematica (Brandes *et al.*, 2006): per questo motivo il metodo di Louvain propone un algoritmo capace di approssimare il valore massimo, rendendo quindi il calcolo possibile anche in presenza di reti contraddistinte da un numero elevato di nodi.

Tale procedimento iterativo inizialmente assegna ogni nodo ad una comunità diversa, per poi alternare due fasi ciclicamente:

- *Prima fase*: A partire dal generico nodo i connesso con $r_i \in \{1, \ldots, K-1\}$ altre comunità $c_1, \ldots, c_{r_i}$, viene calcolato *il guadagno di modularità* $\Delta Q_{ic_1}, \ldots, \Delta Q_{ic_{r_i}}$ ottenuto spostando tale nodo in ciascuna delle altre r_i comunità. Il nodo i viene quindi assegnato alla comunità per cui il guadagno di modularità risulta massimo, ma solamente se maggiore di 0. Qualora qualsiasi spostamento non produca un aumento in termini di modularità, il nodo i rimane assegnato alla comunità di partenza. Il procedimento avviene sequenzialmente per tutti i nodi con almeno un arco (*self-loops* esclusi); la prima fase termina quando nessuno spostamento comporta un ulteriore aumento di modularità.

- *Seconda fase*: Si costruisce una nuova rete i cui nuovi nodi sono costituiti dalle comunità trovate durante la prima fase; i collegamenti tra i nodi originali appartenenti alla medesima comunità costituiranno i *self-loops* dei nuovi nodi, mentre tutte le connessioni tra i nodi originali appartenenti a diverse comunità rappresentano la relazione esistente tra i nodi della rete appena creata.

Definendo come *passo* la combinazione delle due fasi appena descritte (si veda la Figura 3.1 per una schematizzazione grafica), l'algoritmo procede iterativamente finché la prima fase di un determinato passo non propone alcuna modifica nella partizione rilevata allo *step* precedente.

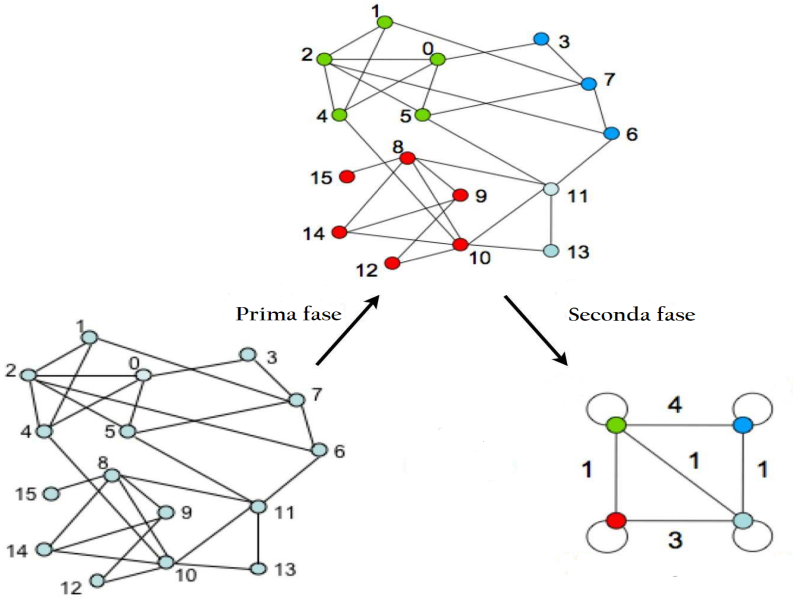


Figura 3.1: Rappresentazione di un passo dell'algoritmo: per semplicità, l'esempio riportato è quello di una rete priva di *self-loops* e con tutti gli archi di peso unitario. Fonte: Blondel *et al.* (2008), Figura 1 (parzialmente rivisitata).

Tale metodologia ha avuto ampio successo grazie alle ottime tempistiche di calcolo, in quanto il numero di comunità evidenziate ad ogni iterazione decresce e gran parte dello sforzo computazionale è concentrato nel primo passo. Sebbene non sia una regola, generalmente il numero di *step* necessari per giungere al valore (approssimativamente) massimo è alquanto contenuto, come messo in luce dagli stessi autori (Blondel *et al.*, 2008).

L'efficienza del metodo è parzialmente dovuta alla semplicità del calcolo della quantità ΔQ_{ic_j}, che risulta:

$$\Delta Q_{ic_j} = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

con $\sum_{in}$ somma dei pesi degli archi che connettono i nodi della comunità c_j tra loro, $k_{i,in}$ somma dei pesi degli archi tra il nodo i e la comunità c_j, $\sum_{tot}$ somma dei pesi degli archi incidenti ai nodi appartenenti alla comunità c_j, k_i e m definiti come nella formula (3.1).

Un limite dell'approccio proposto, proprio di tutti i metodi di raggruppamento basati sull'ottimizzazione della modularità come chiarito da Fortunato & Barthélemy (2007), è la difficoltà nel rilevare comuni-

tà di grandezza inferiore ad una certa soglia, che varia a seconda della dimensione totale della rete e dell'intensità delle connessioni tra le comunità. Tale problematica, però, è in parte superabile in quanto è possibile considerare il raggruppamento proposto dall'algoritmo ad un passo precedente rispetto a quello conclusivo; a discapito di una modularità inferiore, è disponibile una partizione contraddistinta da gruppi di dimensioni ridotte che, in alcuni casi, può avere maggiore valenza interpretativa. Altra peculiarità del metodo è che, almeno potenzialmente, la divisione in comunità proposta ad ogni *step* dipende dall'ordine in cui i nodi vengono considerati durante la prima fase; alcuni test condotti su specifici casi di studio sembrano indicare la mancanza di variazioni significative in tal senso, a differenza delle tempistiche che sembrano essere influenzate da questo aspetto.

3.2 Il metodo

La trattazione che segue ha come obiettivo la definizione e l'analisi di un metodo di raggruppamento specifico per popolazioni di reti. A chi scrive non risulta siano stati sviluppati approcci appositi per rilevare la presenza di *gruppi di reti* all'interno di un campione, sebbene esistano modelli applicabili a tale ambito come quello bayesiano proposto da Durante *et al.* (2016) per reti binarie. Pertanto, il presente lavoro tenta di proporre un primo sviluppo dell'argomento cercando di applicare una metodologia coerente. In tale contesto, il fine non sarà più quello di trovare una partizione per i nodi che compongono una singola rete, ma di suddividere le n reti che contraddistinguono il campione statistico in un certo numero di gruppi in accordo con un apposito criterio. Al fine di evitare possibili equivoci, nel proseguio - se non chiarito diversamente - con il termine *comunità* si indicherà unicamente uno specifico insieme di nodi in una data rete, mentre *gruppo* sarà utilizzato solo per indicare un particolare *cluster* di reti.

La procedura iterativa che verrà descritta presenta una *struttura gerarchica*: a differenza dei *metodi di partizione*, in cui le unità statistiche vengono progressivamente riallocate in un certo numero di categorie decise a priori, i metodi gerarchici propongono una formulazione annidata dei dati e delle partizioni collegate (per questa e altre differenze si veda ad esempio Kaufman & Rousseeuw, 1990). L'approccio è di tipo *agglomerativo*: a partire da una situazione iniziale in cui il numero di gruppi è posto pari al numero di reti, l'algoritmo procede iterativamente fino

alla creazione di un unico gruppo contenente tutte le osservazioni. Il termine gerarchico deriva dal fatto che la partizione proposta al passo h, con $h \in \{1, \ldots, n-1\}$, è direttamente ottenibile da quella precedente tramite l'unione di due gruppi. Inoltre, è possibile associare a questa coppia di gruppi una misura di similarità tramite un opportuno indicatore; l'analisi del *dendrogramma* associato rappresenta una classica regola per determinare il numero di *clusters*, qualora non sia noto a priori.

Infine, la metodologia richiede la specificazione di una *soglia*, illustrata nel paragrafo 3.2.1.

3.2.1 Pre-elaborazione delle reti

Si supponga di disporre di un campione statistico n-variato. Un primo metodo di raggruppamento *naive* consiste, per esempio, nel considerare ogni rete singolarmente, calcolarne la struttura di comunità ed unire, passo dopo passo, le reti che presentano maggiore somiglianza nelle partizioni ottenute. Il limite di questo approccio appare evidente, in quanto ogni unità statistica viene di fatto considerata come un'entità priva di qualsiasi relazione con il resto dei dati. In uno scenario così complesso, una metodologia appropriata deve quindi valutare un duplice aspetto: da un lato l'interpretazione della rete come un insieme di nodi interconnessi tra loro, dall'altro la sua valenza come unità statistica estratta da un campione avente determinate caratteristiche. È necessario, pertanto, tenere in considerazione che, nella sua accezione più generica, la popolazione di riferimento può presentare:

- *Una media dei gradi diversa da rete a rete.* Il numero medio di connessioni $\bar{d}$ (2.2) può variare nettamente all'interno del campione, con alcune reti scarsamente connesse ed altre densamente collegate

- *Grado dei nodi variabile all'interno della rete.* Considerando una singola rete, i vari archi possono presentare pesi molto differenti a seconda della relazione esistente tra i nodi in quella specifica unità statistica

- *Connessione tra la stessa copia di nodi variabile all'interno del campione.* La relazione tra il nodo i e il nodo j può differire significativamente a seconda della rete considerata; di conseguenza, le strutture di comunità ravvisate nei dati possono essere molteplici

Alla luce di ciò, prima di qualsiasi tecnicismo è necessario stabilire quando due reti possano definirsi "simili". Il metodo proposto considera

le strutture di comunità presenti all'interno delle reti - rilevate utilizzando il metodo di Louvain - dopo aver *contestualizzato* quest'ultime nel campione di cui si dispone: in altre parole, due reti R_t, R_s saranno considerate tanto più simili quanto maggiore sarà il grado di somiglianza tra le comunità presenti all'interno delle reti $R_t^{'}, R_s^{'}$, con la generica rete $R^{'}$ opportuna trasformazione della rete di partenza R. Di seguito viene quindi descritta la procedura per la costruzione della rete $R^{'}$.

A partire dal vettore $w_{ij} = (w_{ij}^1, \ldots, w_{ij}^n)$ dei pesi associati agli archi che connettono i nodi i e j con riferimento alle n unità statistiche, è possibile calcolare una vasta gamma di indici di posizione, tra cui:

$$b_{ij} = max(w_{ij}^1, \ldots, w_{ij}^n) \tag{3.2}$$

$$a_{ij} = min(w_{ij}^1, \ldots, w_{ij}^n) \tag{3.3}$$

e di conseguenza procedere alla *normalizzazione* del vettore w_{ij}, il cui generico elemento risulta:

$$u_{ij}^t = \begin{cases} \frac{w_{ij}^t - a_{ij}}{b_{ij} - a_{ij}} \in [0,1] & \text{se } a_{ij} \neq b_{ij}, \quad t = 1, \ldots, n \\ 0.5 & \text{se } a_{ij} = b_{ij}, \quad t = 1, \ldots, n \end{cases} \tag{3.4}$$

Per semplicità, il valore u_{ij}^t, nel caso in cui $a_{ij} = b_{ij}$, è stato fissato pari al valore atteso di una variabile aleatoria con distribuzione uniforme nell'intervallo $[0, 1]$.

Un valore alto della quantità normalizzata indica quindi una relazione tra i due nodi numericamente rilevante rispetto a quella osservata nelle altre unità del campione, mentre un valore basso è espressione di un collegamento esiguo in rapporto alle altre reti. La decisione di far variare gli indicatori statistici a e b a seconda della coppia di nodi considerata permette di cogliere le differenze che intercorrono tra le distribuzioni dei pesi associati ai diversi archi, aspetto fondamentale già evidenziato in precedenza.

La suddetta trasformazione non risulta esauriente in quanto sussiste un'ulteriore problematica irrisolta dalla normalizzazione. Si consideri il seguente esempio: da un campione di n reti con $K = 5$ nodi si estragga una rete R_1 che presenta pesi normalizzati piuttosto bassi $u_{1,b} = 0.05$, eccezion fatta per l'arco che collega il nodo 1 ed il nodo 2, contraddistinto da peso doppio rispetto al precedente $u_{1,a} = 0.10$ (Figura 3.2, in basso a sinistra). Sia R_2 un'altra unità statistica caratterizzata da pesi normalizzati decisamente più elevati $u_{2,b} = 0.40$ e, come nella prima rete, peso doppio nella connessione tra i primi due nodi ($u_{2,a} = 0.8$, Figura 3.2 in alto). La struttura di comunità rilevata nei due casi risulta

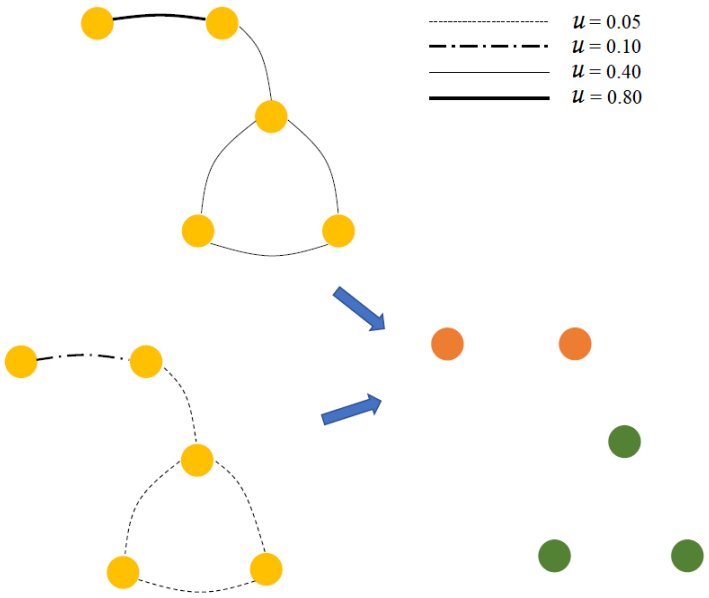


Figura 3.2: Esempio di problematica nell'applicazione del metodo di Louvain.

la medesima (Figura 3.2, in basso a destra): i primi due nodi appartengono ad un *cluster* differente rispetto a quello costituito dai restanti, con conseguente similarità massima tra le due partizioni, supponendo - come è logico attendersi - di utilizzare un indicatore che assegna valore massimo a raggruppamenti identici. Nella prima circostanza, però, la presenza di una comunità costituita dai nodi 1 e 2 è resa unicamente possibile, considerando il peso normalizzato assai contenuto $u_{1,a}$, dalla presenza di connessioni ancora inferiori nel resto della rete; in tale frangente è pertanto sconsigliabile evidenziare una comunità costituita dai primi due nodi visto che la quantità normalizzata $u_{1,a}$ è assai esigua se rapportatata al resto del campione. La motivazione di tale criticità risiede nell'utilizzo del metodo di Louvain che, per costruzione, confronta le relazioni esistenti all'interno di una data rete ignorando il contesto più ampio ed articolato in cui il presente lavoro si pone.

Per questo motivo, un possibile approccio per cercare di risolvere il problema, o quantomeno contenerlo, consiste nell'introduzione di una *soglia* $s \in [0, 1]$. Riprendendo la notazione precedente, a partire dal peso normalizzato u_{ij}^t (3.4) si ottiene il valore $w_{i,j}^{'t}$ definito come:

$$w_{i,j}^{'t} = \begin{cases} u_{ij}^t & \text{se } u_{ij}^t \geq s \qquad t = 1, \ldots, n \\ 0 & \text{se } u_{ij}^t < s \qquad t = 1, \ldots, n \end{cases} \tag{3.5}$$

Quindi, la rete trasformata R' presenterà gli stessi nodi della rete originaria R, ma con archi di peso w'_{ij} al posto del precedente w_{ij}. La scelta della soglia riveste chiaramente un ruolo centrale: a livello numerico, difatti, la trasformazione proposta lascia inalterati i pesi superiori ad s, mentre pone a zero tutte le altre relazioni ravvisate nei dati. Un valore alto della soglia, dunque, da un lato assicura che vengano segnalate solo le comunità caratterizzate da connessioni veramente sorprendenti per la popolazione considerata, dall'altro rende indistinguibili tra loro tutte le relazioni di intensità inferiore alla soglia. Contrariamente, valori prossimi allo zero comportano la problematica messa in luce dall'esempio precedente, ma d'altro canto sfruttano maggiormente l'informazione presente nei dati.

A livello interpretativo, le due modifiche proposte - normalizzazione prima e introduzione della soglia poi - afferiscono a due aspetti per certi versi complementari e, nello specifico, congiuntamente necessari: lo scopo della prima è quello di unificare l'intervallo di valori osservabili in modo da non penalizzare le relazioni tra coppie di nodi che, per caratteristiche intrinseche della popolazione, sono numericamente inferiori ad altre; il fine della seconda, invece, è limitare le problematiche relative alla dimensione puramente *marginale* del metodo di Louvain.

La scelta di s rappresenta perciò l'ultimo fondamentale *step* in questa prima fase di pre-elaborazione del campione statistico. Considerando le possibili differenze in termini distributivi dei pesi associati ai vari archi, sembra ragionevole imporre una soglia che tenga conto della coppia di nodi (i, j) di volta in volta presa in esame. Si definisca pertanto:

$$s_{ij} = q(u_{ij}, \alpha) \qquad\qquad i, j = 1, \ldots, K \tag{3.6}$$

con $u_{ij} = (u^1_{ij}, \ldots, u^n_{ij})$ e $q(x, \alpha)$ quantile di ordine α della distribuzione disaggregata x. A chi scrive la formulazione (3.6) appare come un accettabile compromesso: a fronte di un solo scalare α, la soglia è lasciata libera di variare a seconda delle caratteristiche distributive del vettore u_{ij}. Quindi, la relazione (3.5) viene aggiornata nel seguente modo:

$$w'^t_{i,j} = \begin{cases} u^t_{ij} & \text{se } u^t_{ij} \geq s_{ij} \qquad t = 1, \ldots, n \\ 0 & \text{se } u^t_{ij} < s_{ij} \qquad t = 1, \ldots, n \end{cases} \tag{3.7}$$

Poiché il contesto in cui si pone l'analisi di raggruppamento è tipicamente non supervisionato, nella maggior parte dei casi risulta impossibile selezionare il valore per α ottimale sfruttando un *campione di stima* di cui si conoscono preventivamente le etichette di gruppo (ad esempio,

Azzalini & Scarpa, 2012): pertanto, sebbene qualche indicazione possa essere ottenuta da studi simulativi come mostrato nel capitolo 4, la scelta si basa sulle peculiarità specifiche del problema di volta in volta esaminato.

3.2.2 Indice di similarità

A partire dalle reti trasformate, è possibile quindi rilevare e successivamente confrontare le n strutture di comunità che contraddistinguono il campione. Chiaramente, l'interpretazione dei raggruppamenti muta radicalmente: eventuali comunità di nodi fortemente interconnessi segnalano una relazione particolarmente sorprendente per la popolazione e non necessariamente ravvisabile dall'analisi delle unità statistiche originali.

Lo strumento utilizzato per determinare il grado di similarità tra due partizioni è una variante dell'*Indice di Rand* (Rand, 1971): per non appesantire eccessivamente la relazione, in questo paragrafo si presenta solo l'idea sottostante l'indicatore utilizzato, mentre maggiori dettagli sono disponibili nell'articolo scritto dagli ideatori Hubert & Arabie (1985).

Sia $X = (x_1, \ldots, x_K)$ un insieme composto da K elementi da cui si ottengono due partizioni $U = (u_1, \ldots, u_g)$ e $V = (v_1, \ldots, v_q)$, con g e q numero di gruppi rispettivo ed ogni elemento di U e V composto da un sottoinsieme di valori di X tale che:

$$\cup_{i=1}^{g} u_i = \cup_{j=1}^{q} v_j = X \tag{3.8}$$

$$u_i \cap u_{i'} = v_j \cap v_{j'} = \emptyset \tag{3.9}$$

Per $1 \leq i \neq i' \leq g$, $1 \leq j \neq j' \leq q$. Inoltre siano:

- c_1 il numero di coppie di elementi di X appartenenti allo stesso gruppo in U e in V
- c_2 il numero di coppie di elementi di X appartenenti ad un gruppo differente sia in U che in V
- d_1 il numero di coppie di elementi di X appartenenti allo stesso gruppo in U, ma gruppo diverso in V
- d_2 il numero di coppie di elementi di X appartenenti ad un gruppo differente in U, ma stesso gruppo in V

Quindi, l'Indice di Rand è calcolabile come:

$$R = \frac{c_1 + c_2}{c_1 + c_2 + d_1 + d_2} \tag{3.10}$$

Tale indice assume valore compreso tra 0, riscontrabile quando nessuna coppia di elementi è classificata nella medesima maniera nelle due partizioni, e 1, quando i raggruppamenti risultano identici. Nel corso del tempo, diversi autori hanno evidenziato alcuni limiti dell'indicatore in questione: per esempio, Morey & Agresti (1984) sottolinea come l'Indice di Rand sia fortemente dipendente dal numero di *clusters* ravvisati, mentre Fowlkes & Mallows (1983) mostra che, all'aumentare del numero di gruppi, l'indicatore converge a 1 anche se U e V sono indipendenti.

Per tali motivi, nel presente lavoro verrà utilizzata la variante nota come *Indice di Rand Aggiustato* (o ARI, Hubert & Arabie, 1985), che differisce dal precedente per il fatto di avere valore atteso nullo quando l'accordo tra le due partizioni è unicamente dovuto al caso. Conseguentemente, l'intervallo di valori per l'ARI è $[-1, 1]$, con valori negativi che segnalano un accordo tra i raggruppamenti inferiore a quello ottenuto in caso di etichette di gruppo assegnate casualmente.

Come anticipato nel paragrafo 3.2, il metodo di raggruppamento sviluppato propone un approccio di tipo gerarchico agglomerativo: pertanto, nel corso della procedura è richiesta la specificazione di una misura di similarità capace non solo di confrontare le strutture di comunità delle singole unità statistiche, ma anche di confrontare il grado di accordo tra insiemi di reti sempre più ampi. Per tale motivo, l'accorgimento attuato considera - per ogni gruppo - un'opportuna rete come rappresentativa dell'intero *cluster*. A partire dalla situazione iniziale in cui ogni unità statistica viene assegnata ad un gruppo differente, si costruisce la *rete centroide* delle due reti aventi strutture di comunità maggiormente simili: pertanto i pesi relativi a un dato arco della rete centroide saranno pari alla media aritmetica dei pesi di quell'arco nelle due reti. Questo nuovo costrutto, costituito da una specifica struttura di comunità, permette di operare nuovi confronti con le strutture di comunità delle altre reti tramite l'Indice di Rand Aggiustato. Allo *step* successivo, la rete centroide di fatto sostituisce le due reti unite al passo precedente: la procedura prosegue finché tutte le osservazioni appartengono al medesimo cluster.

La scelta operativa del numero di gruppi rappresenta l'ultimo fondamentale aspetto da considerare: a tal proposito, la valutazione della differenza tra l'indice di similarità massimo in due passi successivi del metodo risulta una prassi alquanto diffusa per i metodi gerarchici (ad

esempio, Azzalini & Scarpa, 2012). Una *caduta di similarità* considerevole, ravvisabile graficamente tramite un dendrogramma, fornisce indicazione del fatto che i gruppi identificati nei dati presentano caratteristiche troppo differenti per poter procedere a successive unioni. Chiaramente, il criterio proposto - su cui non si indugia ulteriormente in quanto assai noto nell'ambito della *cluster analysis* - può essere affiancato da valutazioni *qualitative* da parte dell'analista.

3.2.3 L'algoritmo

Di seguito si presenta sinteticamente la procedura sviluppata in questo capitolo.

La metodologia consta di due fasi distinte: una *prima fase* di preelaborazione dei dati ed una *seconda fase* finalizzata al conseguimento della struttura di gruppo presente negli stessi.

Sia pertanto:

- $R = \{R_1, \ldots, R_n\}$ un campione statistico composto da n reti pesate indirette, con $\mathcal{K} = \{k_1, k_2, \ldots, k_K\}$ l'insieme dei nodi comune a tutte le unità statistiche. Il generico valore w_{ij}^t sarà il peso associato all'arco che collega il nodo i al nodo j nella rete t, con $t = 1, \ldots, n$ e $i, j = 1, \ldots, K$. L'eventuale assenza di collegamento tra due nodi determina una relazione con peso nullo

- α_0 il valore selezionato per la soglia α

Prima fase

Per ogni $t = 1, \ldots, n$, $i, j = 1, \ldots, K$ si calcoli la quantità $w_{ij}^{'t}$ tramite le operazioni di normalizzazione e introduzione della soglia descritte dalle formule (3.2),(3.3),(3.4),(3.6),(3.7). Si ottenga quindi il nuovo campione statistico $R^{'} = \{R_1^{'}, \ldots, R_n^{'}\}$, con la generica rete $R_t^{'}$ costituita dall'insieme di nodi $\mathcal{K}$ e nuovi pesi $w_{ij}^{'t}$.

Seconda fase

Sia $P = (p_1, \ldots, p_n)$ il vettore indicante le etichette di gruppo per le n reti: inizialmente, $P = (1, \ldots, n)$. Si eseguano i seguenti passaggi iterativamente:

1. **Ottenimento della struttura di comunità** Per ogni rete del campione $R^{'}$ si definisca la struttura di comunità utilizzando il metodo di Louvain

2. **Calcolo della similarità tra le partizioni** Per ogni coppia di partizioni si calcoli l'Indice di Rand Aggiustato. Sia (R'_{m_1}, R'_{m_2}), con $m_1 < m_2$, la coppia di reti che presenta ARI più elevato; nell'eventualità non sia unica, viene selezionata casualmente tra quelle che massimizzano l'indicatore.

3. **Unione dei gruppi** Si assegni valore p_{m_1} a tutti gli elementi di P con valore pari a p_{m_2}. Inoltre, si sostituiscano i pesi della rete $R^{'}_{m_1}$ con la media aritmetica dei pesi delle reti $R^{'}_{m_1}, R^{'}_{m_2}$. Si rimuova, quindi, la rete $R^{'}_{m_2}$ dal campione $R^{'}$.

La procedura termina quando tutte le osservazioni appartengono al medesimo gruppo.

3.3 Possibili sviluppi

Vengono di seguito illustrate due modifiche potenzialmente interessanti al metodo presentato.

Innanzitutto la fase di pre-elaborazione del campione statistico ricopre un ruolo fondamentale all'interno della procedura, ma esistono alcune alternative alla normalizzazione (3.4) ugualmente funzionali. Una possibilità è data dalla ben nota *standardizzazione*: riprendendo la notazione utilizzata nel corso del capitolo, i pesi standardizzati risultano:

$$z_{ij}^t = \frac{w_{ij}^t - \bar{w}_{ij}}{\hat{\sigma}_{ij}} \tag{3.11}$$

con:

$$\bar{w}_{ij} = \frac{\sum_{t=1}^{n} w_{ij}^t}{n}$$

$$\hat{\sigma}_{ij} = \sqrt{\frac{\sum_{t=1}^{n} (w_{ij}^t - \bar{w}_{ij})^2}{n}}$$

Tale trasformazione introduce un elemento da considerare attentamente: per costruzione, difatti, i pesi z_{ij}^t assumono segno negativo ogni volta che il valore osservato w_{ij}^t risulta inferiore alla media $\bar{w}_{ij}$. In termini di

analisi delle comunità, le conseguenze sono ben note in letteratura: come evidenziato da Gómez *et al.* (2009), l'introduzione di pesi negativi comporta la mancata interpretazione in termini probabilistici della quantità $\frac{k_i}{2m}$ all'interno della modularità (3.1), con conseguente impossibilità di utilizzare il metodo di Louvain nella forma presentata. Per tale motivo, diversi autori hanno proposto differenti approcci al fine di definire opportunamente la modularità in presenza di *signed networks* (Traag & Bruggeman, 2009; Gómez *et al.*, 2009), rendendo di fatto utilizzabile la standardizzazione (3.11).

Altro aspetto di primario interesse riguarda l'estensione della metodologia alle reti dirette: infatti, Dugué & Perez (2015) ha dimostrato, attraverso studi empirici e dimostrazione teorica, come l'inclusione dell'informazione sulla direzionalità delle relazioni all'interno delle singole reti permetta di trovare, in specifici casi, strutture di comunità maggiormente informative rispetto a quelle individuate ignorando tale indicazione. Nella fattispecie risulta particolarmente utile l'estensione del concetto di modularità proposta da Leicht & Newman (2008), che permette quindi di applicare il metodo di Louvain in presenza di tale tipologia di dato.

Capitolo 4

Simulazioni

L'obiettivo del seguente studio di simulazione è quello di valutare l'appropriatezza del metodo di raggruppamento sviluppato. In particolare si analizzerà l'incidenza del valore scelto per la soglia sull'accuratezza complessiva della metodologia, in modo da avere un'indicazione circa il suo impatto sulle partizioni ottenute. I risultati verranno confrontati con quelli conseguiti da un approccio alternativo.

4.1 Scenari di simulazione

Sia $R = \{R_1, \ldots, R_n\}$ un campione statistico composto da n reti pesate indirette, con $\mathcal{K} = \{k_1, k_2, \ldots, k_K\}$ l'insieme dei nodi comune a tutte le unità statistiche. Siano inoltre presenti due gruppi all'interno del campione statistico, di numerosità rispettivamente n_1 e n_2. In tutte le simulazioni, si impone $n = 100$, $K = 30$, $n_1 = n_2 = 50$. Per semplicità, i nodi saranno etichettati con i numeri da 1 a 30; con nodi *pari* (*dispari*) si intenderanno i nodi la cui etichetta è un numero *pari* (*dispari*). Per facilitare la comprensione, una rappresentazione grafica dello scenario 1 e 2 è proposta in Figura 4.1, dello scenario 3 e 4 in Figura 4.2.

- **Primo scenario**

  Sia X variabile aleatoria discreta che assume valori $\{x_1, x_2, x_3\}$ con probabilità $(\pi_1^x, \pi_2^x, \pi_3^x)$, e Y variabile aleatoria discreta che assume valori $\{y_1, y_2, y_3\}$ con probabilità $(\pi_1^y, \pi_2^y, \pi_3^y)$. Le reti appartenenti al primo gruppo presentano:

  – peso di ogni arco che collega due nodi *pari* generato in maniera indipendente dalla variabile casuale X

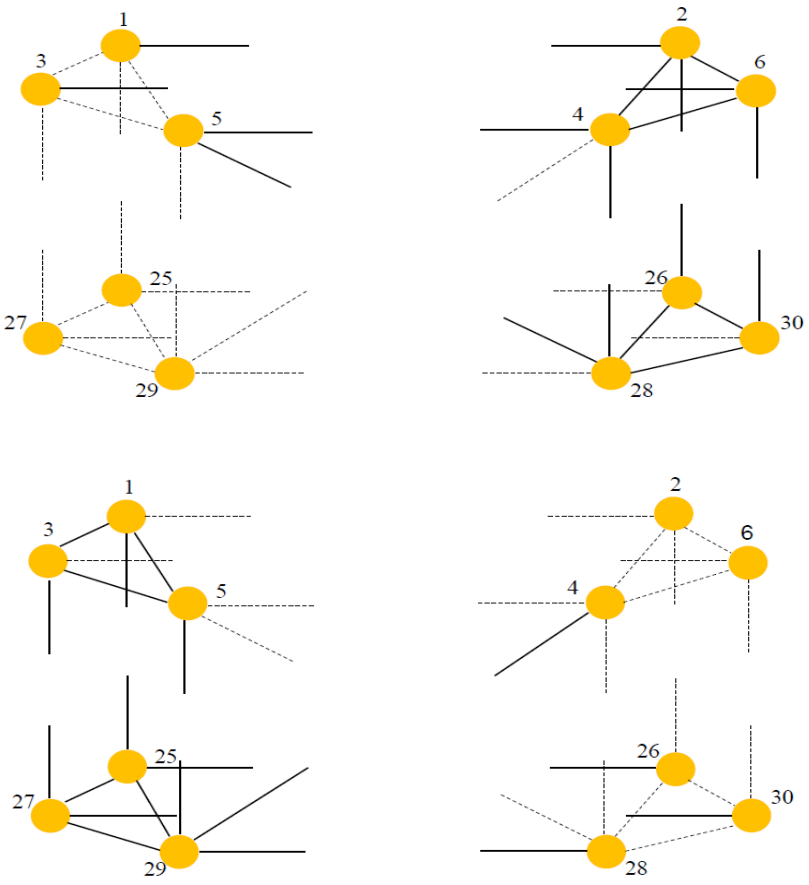


Figura 4.1: Scenario 1 e 2. In alto è rappresentata una rete appartenente al primo gruppo, in basso una appartenente al secondo gruppo. La linea continua indica un arco il cui peso è generato dalla variabile X, la linea tratteggiata un arco il cui peso è generato dalla variabile Y.

- peso di ogni arco che collega due nodi *dispari* generato in maniera indipendente dalla variabile casuale Y
- peso di ogni arco che collega i nodi *pari* ai primi 8 nodi *dispari* {1,3,...,13,15} generato in maniera indipendente dalla variabile casuale X
- peso di ogni arco che collega i nodi *pari* ai restanti nodi *dispari* {17,19,...,27,29} generato in maniera indipendente dalla variabile casuale Y

Le reti appartenenti al secondo gruppo presentano:

- peso di ogni arco che collega due nodi *pari* generato in maniera indipendente dalla variabile casuale Y
- peso di ogni arco che collega due nodi *dispari* generato in maniera indipendente dalla variabile casuale X
- peso di ogni arco che collega i nodi *pari* ai primi 8 nodi *dispari* {1,3,...,13,15} generato in maniera indipendente dalla variabile casuale Y
- peso di ogni arco che collega i nodi *pari* ai restanti nodi *dispari* {17,19,...,27,29} generato in maniera indipendente dalla variabile casuale X

Si fissano:

- $\{x_1, x_2, x_3\} = \{11, 12, 13\}$
- $\{y_1, y_2, y_3\} = \{4, 5, 6\}$
- $(\pi_1^x, \pi_2^x, \pi_3^x) = (\pi_1^y, \pi_2^y, \pi_3^y) = (0.25, 0.5, 0.25)$

- **Secondo scenario**

  Il processo generatore delle connessioni tra i nodi è identico a quello dello scenario 1, con l'unica variazione per il supporto della variabile X, ora fissato a $\{x_1, x_2, x_3\} = \{7, 8, 9\}$. L'intento è quello di aumentare la complessità dello scenario precedente: difatti, la variazione proposta implica strutture di comunità meno definite.

- **Terzo scenario**

  Sia X variabile aleatoria discreta che assume valori $\{x_1, x_2, x_3, x_4, x_5\}$ con probabilità $(\pi_1^x, \pi_2^x, \pi_3^x, \pi_4^x, \pi_5^x)$, e Y variabile aleatoria discreta che assume valori $\{y_1, y_2, y_3, y_4, y_5\}$ con probabilità $(\pi_1^y, \pi_2^y, \pi_3^y, \pi_4^y, \pi_5^y)$.

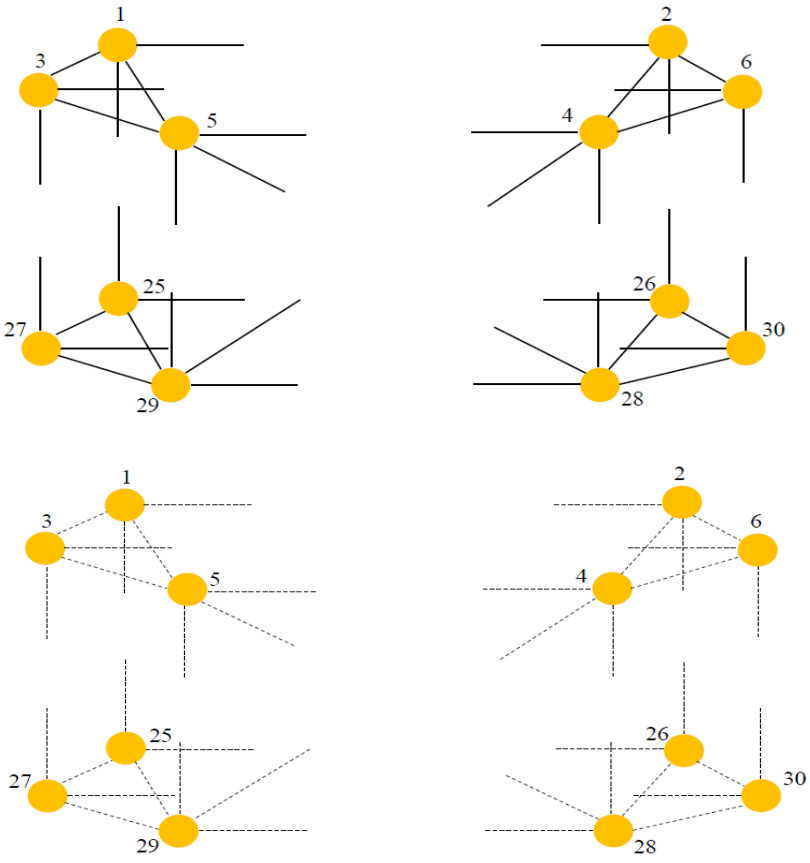


Figura 4.2: Scenari 3 e 4. Cfr Figura 4.1.

Le reti appartenenti al primo gruppo evidenziano connessioni tra i nodi generate in maniera indipendente dalla variabile casuale X, quelle appartenenti al secondo gruppo presentano invece collegamenti generati in maniera indipendente dalla variabile casuale Y.

Si fissano:

- $\{x_1, x_2, x_3, x_4, x_5\} = (8, 9, 10, 11, 12)$
- $\{y_1, y_2, y_3\} = \{3, 4, 5, 6, 7\}$
- $(\pi_1^x, \pi_2^x, \pi_3^x, \pi_4^x, \pi_5^x) = (\pi_1^y, \pi_2^y, \pi_3^y, \pi_4^y, \pi_5^y) = (0.2, 0.2, 0.2, 0.2, 0.2)$

- **Quarto scenario**

  Il processo generatore delle connessioni tra i nodi è identico a quello dello scenario 3, facendo variare:

  - $\{x_1, x_2, x_3, x_4, x_5\} = \{7, 8, 9, 10, 11\}$
  - $(\pi_1^x, \pi_2^x, \pi_3^x, \pi_4^x, \pi_5^x) = (\pi_1^y, \pi_2^y, \pi_3^y, \pi_4^y, \pi_5^y) = (0.05, 0.3, 0.3, 0.3, 0.05)$

  L'intento è quello di aumentare la complessità dello scenario precedente poiché si riduce la differenza, in termini di numero di connessioni medie, tra le reti dei due gruppi.

Ogni scenario è valutato per tre valori differenti della soglia, ossia $\alpha = 0.3, 0.5, 0.7$.

La metodologia viene confrontata con l'approccio *naive* introdotto nel paragrafo 3.2.1: esso consiste nel considerare ogni rete singolarmente, calcolarne la struttura di comunità ed unire le reti che presentano maggiore somiglianza nelle partizioni ottenute.

Gli scenari afferiscono a due situazioni sulle quali è necessario porre particolare attenzione. Negli scenari 1 e 2 le strutture di comunità che contraddistinguono le reti dei due gruppi sono ben distinte, e le connessioni tra le diverse coppie di nodi assumono lo stesso *range* di valori all'interno del campione. In questo contesto, la compressione nell'intervallo $[0, 1]$ dei pesi associati agli archi può risultare un accorgimento inutile in quanto la popolazione di riferimento non evidenzia alcuna delle problematiche messe in luce nel paragrafo 3.2.1.

Gli scenari 3 e 4 presentano un contesto ben diverso: le strutture di comunità sono completamente casuali per tutte le unità statistiche, ma le reti del primo gruppo si distinguono da quelle del secondo per la differente *media dei gradi* $\bar{d}$. Un metodo di raggruppamento efficace

dovrebbe individuare le due sottopopolazioni, cogliendo l'eterogeneità presente nei dati.

Le situazioni considerate rappresentano due aspetti che la metodologia cerca di bilanciare: da una parte la contestualizzazione delle singole reti all'interno di un campione statistico, dall'altra la perdita d'informazione che ne deriva.

La valutazione dei metodi sarà effettuata confrontando il raggruppamento ottenuto con la partizione teorica, imponendo come numero di gruppi $N_G = 2$; a tal proposito, un indicatore comunemente utilizzato per questo scopo è il già discusso ARI. Con riferimento alla parte computazionale, il lavoro è stato svolto con il linguaggio di programmazione R (R Core Team, 2016); il calcolo delle strutture di comunità con il metodo di Louvain è stato realizzato con il pacchetto *igraph* (Csardi & Nepusz, 2006), mentre il calcolo dell'ARI con il pacchetto *pdfCluster* (Azzalini & Menardi, 2014).

Per ogni scenario sono state svolte N=1000 simulazioni.

4.2 Risultati

La tabella 4.1 riporta i principali indici di posizione relativi alle N simulazioni condotte per ogni scenario.

Il primo evidenzia un miglioramento del metodo all'aumentare della soglia: la perdita d'informazione causata dalla compressione dei pesi nell'intervallo $[0, 1]$ deve essere bilanciata dall'introduzione di una soglia adeguata che permetta di considerare solo le relazioni particolarmente sorprendenti per la popolazione. L'elevata deviazione standard nei casi con $\alpha = 0.3, 0.5$ suggerisce particolare cautela in contesti come quello studiato. Viceversa, i risultati ottenuti con un valore alto per s sembrano comparabili con quelli conseguiti con il metodo *naive*, vero e proprio *target* in una situazione come quella analizzata: difatti, in tutte le N simulazioni quest'ultimo propone massima concordanza tra la partizione trovata e quella teorica. Un punto di forza è riscontrabile nel passaggio dal primo al secondo scenario: per valori appropriati della soglia ($\alpha = 0.5, 0.7$), la procedura non sembra peggiorare quando le strutture di comunità diventano meno nitide.

Il terzo scenario vede il completo fallimento dell'approccio *naive*: la natura marginale del metodo di Louvain non permette di ravvisare le differenze in termini di numero di connessioni totali tra i gruppi. Discorso simile vale per valori bassi della soglia, che propongono partizioni

	Scenario 1				**Scenario 2**			
	$s = 0.3$	$s = 0.5$	$s = 0.7$	*Naive*	$s = 0.3$	$s = 0.5$	$s = 0.7$	*Naive*
Minimo	0	0	0.029	1	0	0	0.008	1
1^o Quart.	1	1	1	1	0	1	1	1
Mediana	1	1	1	1	1	1	1	1
Media	0.834	0.915	0.993	1	0.689	0.946	0.993	1
Std. Dev	0.372	0.266	0.072	0	0.462	0.212	0.074	0
3^o Quart.	1	1	1	1	1	1	1	1
Max	1	1	1	1	1	1	1	1
	Scenario 3				**Scenario 4**			
	$s = 0.3$	$s = 0.5$	$s = 0.7$	*Naive*	$s = 0.3$	$s = 0.5$	$s = 0.7$	*Naive*
Minimo	-0.010	0	0	-0.010	-0.010	-0.010	0	-0.010
1^o Quart.	-0.005	0.005	0.012	-0.005	-0.002	-0.002	0.001	-0.004
Mediana	-0.001	0.085	0.085	-0.001	0	0	0.037	-0.001
Media	0	0.315	0.291	0	0.004	0.006	0.353	0
Std. Dev	0.01	0.397	0.376	0.010	0.016	0.018	0.444	0.009
3^o Quart.	0.001	0.605	0.458	0.001	0.003	0.007	1	0.001
Max	0.069	1	1	0.070	0.151	0.153	1	0.093

Tabella 4.1: Tabella che riporta i risultati delle simulazioni. Minimo, 1^o quartile, mediana, media, deviazione standard, 3^o quartile, massimo dell'ARI relativi alle 1000 simulazioni per ogni scenario.

completamente inesatte nella quasi totalità dei casi. Il miglioramento ottenuto con l'utilizzo di soglie più alte è incoraggiante: l'ARI medio non elevatissimo (rispettivamente 0.32 e 0.29) è dovuto all'assoluta complessità dello scenario studiato, in cui strutture di comunità casuali per tutto il campione implicano notevole similarità tra i due gruppi. Il quarto scenario conferma la necessità di una soglia elevata quando le differenze tra i gruppi tendono a diminuire.

Nel complesso, la metodologia offre un discreto compromesso nella gestione dei differenti aspetti che rendono particolarmente ostico il raggruppamento di reti: d'altro canto, l'influenza del valore della soglia sui risultati rappresenta un limite innegabile in un contesto non supervisionato.

Capitolo 5

Applicazione al gioco del calcio

5.1 I dati

Nel presente capitolo viene trattata l'applicazione della metodologia proposta al gioco del calcio. Sia per i semplici appassionati che per i *match analysts* un elemento di particolare interesse è il cosiddetto *schema di gioco*, riassumibile come modalità con cui una determinata squadra muove il pallone lungo il terreno di gioco nel corso di una partita. In termini statistici, tale aspetto può essere adeguatamente rappresentato da una rete, i cui nodi simboleggiano diverse zone del campo e gli archi descrivono le transizioni della palla tra queste.

Un primo delicato fattore da considerare è il numero di attori presenti in ogni unità statistica: difatti, porzioni di campo troppo estese comportano difficoltà nel rilevare potenziali differenze tra le strutture di comunità, viceversa zone di dimensioni particolarmente ridotte rischiano di rendere fortemente dissimili schemi di gioco del tutto somiglianti. Alla luce di ciò, un buon compromesso pare essere la scelta di $K = 9$ zone.

I dati di cui si dispone, forniti da InStat (http://instatfootball.com/), fanno riferimento alle 380 partite del campionato italiano di Serie A TIM, stagione 2015-2016; poiché ogni gara consta di due differenti reti - una per la squadra di casa e una per quella in trasferta -, il campione statistico risulta composto da 760 osservazioni.

Per ogni squadra ed ogni partita, i dati rilevano le coordinate spaziali (x, y) in cui sono avvenuti determinati *gesti tecnici* da parte dei singoli calciatori, con x posizione lungo l'asse orizzontale e y posizione lungo

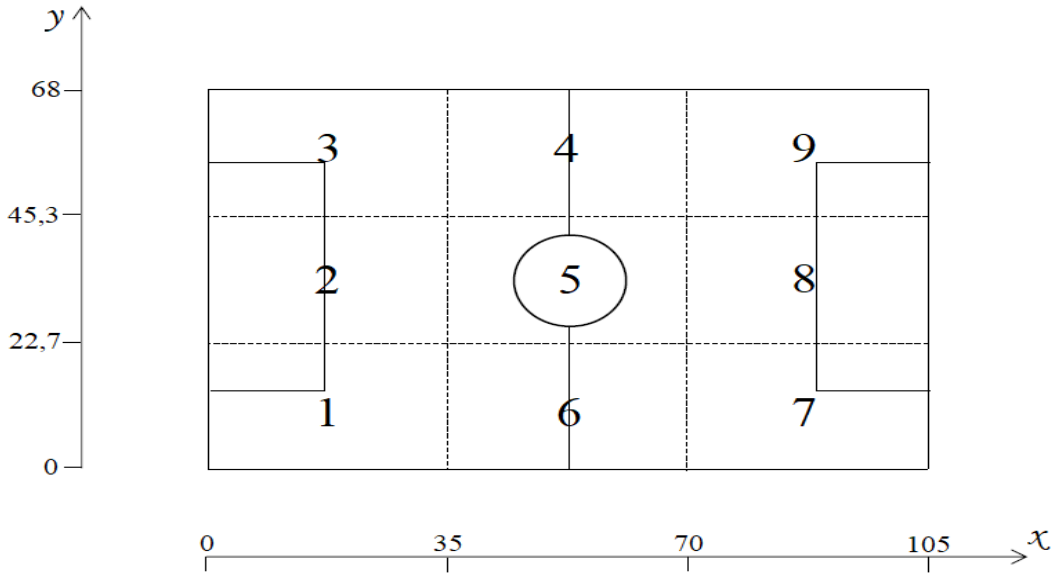


Figura 5.1: Rappresentazione della suddivisione del terreno di gioco. Le dimensioni del campo sono 105×68 m: le nove zone sono ottenute dividendo, sia in lunghezza che in larghezza, il terreno di gioco in tre parti uguali. Si suppone che le squadre attacchino da sinistra verso destra: pertanto, i rettangoli numerati da 1 a 3 rappresentano la *zona di difesa*, quelli da 4 a 6 la *zona di centrocampo* e quelli da 7 a 9 la *zona d'attacco*.

l'asse verticale: una rappresentazione grafica utile è proposta in Figura 5.1. L'evento ravvisato, in base alla sua tipologia, è classificabile in 12 distinte categorie, a loro volta riassumibili in 4 *macrocategorie*:

- **Passaggi** Categoria più numerosa, riporta la posizione di *partenza* di un passaggio. Comprende: passaggi riusciti, passaggi imprecisi, assist, cross riusciti, cross imprecisi
- **Dribbling** Riporta la posizione in cui è iniziato un dribbling. Comprende: dribbling riuscito, dribbling non riuscito
- **Contrasti di gioco** Riporta la posizione in cui è avvenuto un contrasto tra due giocatori avversari. Comprende: palloni rubati riusciti, palloni rubati non riusciti, duelli vinti, duelli persi
- **Tiri** Riporta la posizione in cui scocca il tiro

A partire dalle coordinate spaziali, è possibile assegnare ogni gesto tecnico ad una delle nove aree di Figura 5.1; quindi, il collegamento tra

il nodo i e il nodo j è dato dal conteggio delle coppie di eventi consecutivi avvenuti nelle zone i e j. Poiché la metodologia considera reti indirette, non risulta rilevante l'informazione riguardo quale dei due nodi abbia promosso la relazione e quale l'abbia ricevuta: la mancanza di direzionalità nelle connessioni rappresenta chiaramente un limite nel contesto calcistico, come chiarito nel paragrafo 5.5. Qualora un determinato evento non risulti né preceduto né seguito da un gesto tecnico di un giocatore della stessa squadra, questo non rientra nel conteggio in quanto fenomeno isolato; si noti, invece, come la struttura proposta permetta la creazione di *self-loops* quando $i = j$.

Le reti create rappresentano solamente un'approssimazione dello schema di gioco: i dati disponibili, difatti, ignorano del tutto i movimenti palla al piede da parte dei calciatori. Per tale motivo, si ipotizza che nel tempo intercorrente tra due eventi consecutivi, rilevati rispettivamente nelle zone i e j, il pallone sia transitato solamente tra quelle due porzioni di campo senza coinvolgerne altre. Basandosi su considerazioni unicamente calcistiche, è lecito attendersi che le squadre che propongono un gioco maggiormente incentrato sulle iniziative personali dei giocatori presentino caratteristiche differenti rispetto alle formazioni che limitano tale aspetto: pertanto, il grado di approssimazione del metodo di gioco può variare a seconda dell'unità statistica esaminata. D'altro canto è possibile considerare il numero complessivo di gesti tecnici riscontrati in ogni partita così da ottenere un'indicazione sulla perdita d'informazione dovuta alla mancata osservazione del reale schema di gioco: la media[1] degli eventi registrati nel corso delle 380 gare è di 23.96 al minuto, quantità decisamente soddisfacente.

5.2 Analisi esplorative

Le analisi esplorative che seguono hanno come scopo primario lo studio delle caratteristiche del campione statistico, con particolare riferimento alle peculiarità messe in luce nel paragrafo 3.2.1.

Il primo aspetto da esaminare è la distribuzione del peso medio associato agli archi nelle diverse unità statistiche: vista la differente qualità in termini di *abilità di palleggio* delle squadre italiane, è facilmente prevedibile che tale valore vari molto a seconda della rete, ossia a seconda

[1] La stima è ottenuta considerando la durata media *effettiva* di un incontro di calcio: la fonte che riporta tale quantità (Santini, 2014) si basa su una ricerca - di cui non si dispone - condotta dall'azienda Opta. Pertanto l'informazione vuole avere carattere puramente orientativo.

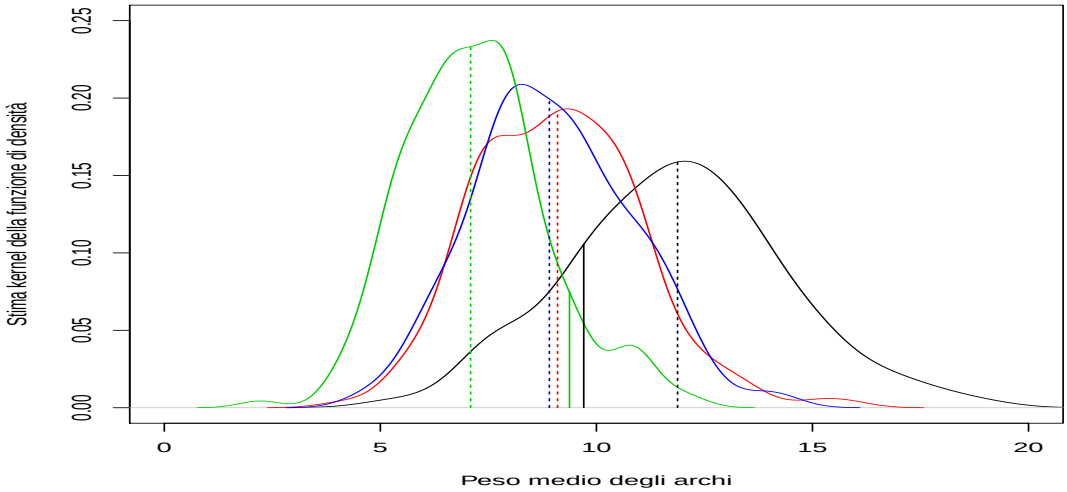


Figura 5.2: Stima della funzione di densità del peso medio degli archi per le squadre di alta (linea nera), medio-alta (linea rossa), medio-bassa (linea blu), bassa (linea verde) classifica. Le linee tratteggiate segnalano le mediane, mentre le linee continue indicano il primo quintile (linea nera) e il nono decile (linea verde) delle rispettive distribuzioni.

dello schema di gioco offerto dalla singola squadra in una specifica partita. La Figura 5.2 riporta la stima *kernel* della funzione di densità relativa all'intensità media delle connessioni tra i nodi a seconda della posizione in classifica del *team* a fine campionato. In particolare, le squadre sono state suddivise in quattro categorie, ossia *alta classifica* (posizione 1-5), *medio-alta classifica* (posizione 6-10), *medio-bassa classifica* (posizione 11-15), *bassa classifica* (posizione 16-20). I quattro grafici sembrano confermare quanto supposto in precedenza: le squadre intermedie (classifica medio-alta e medio-bassa) risultano abbastanza simili in termini distributivi e le squadre di fascia alta nettamente separate da quelle in zona retrocessione. Difatti, il primo quintile della distribuzione più a destra risulta superiore al nono decile di quella più a sinistra.

Un altro elemento di particolare interesse riguarda il numero di collegamenti tra le varie zone del campo: anche in questo caso, la situazione presenta un'eterogeneità elevata, con un valore medio che spazia da 0.11 - individuato nella relazione tra la zona sinistra di difesa 3 e la zona destra d'attacco 7, cfr. Figura 5.1 - e 39.31, osservabile nel *self-loop* relativo alla zona sinistra di centrocampo 4. Nello specifico, i due archi che collegano i nodi $(3, 7)$ e $(1, 9)$ - rispettivamente difesa sinistra/attacco

destro, difesa destra/attacco sinistro - presentano una distribuzione dei pesi alquanto asimmetrica: oltre ad un'alta inflazione degli zeri (rispettivamente 91% e 89%), queste ravvisano le varianze minime (rispettivamente 0.12 e 0.15) e il valore massimo inferiore, ossia 3. L'evidenza è tutt'altro che sorprendente in quanto i nodi coinvolti rappresentano zone di campo diametralmente opposte. La mancanza quasi assoluta di connessioni porta a pensare che le pochissime rilevate siano fortuite e del tutto assimilabili ai tentativi di passaggio non conteggiati in quanto terminati fuori campo. Conseguentemente, nell'applicazione del metodo di raggruppamento viene assegnato valore nullo a tutti i pesi relativi ai due archi sopraccitati in modo da non condizionare la struttura di comunità sulla base di eventi ritenuti accidentali.

Tra i vari impieghi, l'analisi grafico-descrittiva può essere utilizzata per considerazioni di più ampio respiro sebbene con la dovuta attenzione e cautela. Un quesito di capitale importanza riguarda l'utilità della metodologia proposta: l'approccio presentato nel capitolo 3.2 suggerisce - difatti - una serie di accorgimenti per gestire la dimensione campionaria del contesto studiato che, se scarsamente funzionali, rappresenterebbero un'inutile complicazione rispetto ad un semplice confronto delle strutture di comunità delle reti di partenza. A tal proposito, si considerino due squadre opposte per qualità di gioco: il Napoli e il Frosinone. La squadra partenopea rappresenta la formazione che ha espresso la trama di gioco indiscutibilmente più fitta, in quanto tutti i principali indici di posizione (minimo, massimo, mediana, media, primo e terzo quartile) - relativi al collegamento medio tra i nodi nel corso del campionato - risultano superiori a quelli dei *team* avversari. Viceversa la squadra laziale, malamente retrocessa al termine della stagione sportiva, presenta un *possesso palla* assai scadente come si evince dal valore minimo per la Serie A TIM 2015/2016 del primo, secondo e terzo quartile della distribuzione della stessa quantità. La differenza tra le due formazioni è così marcata che il valore massimo ravvisato per il Frosinone (9.16) è inferiore a quello minimo del Napoli (10.09). Quindi, per ciascuna delle 38 partite delle due squadre è possibile calcolare la percentuale del peso associato ad ogni arco rispetto al peso complessivo della rete in modo da scorgere eventuali differenze nello schema di gioco al netto del numero di connessioni. Confrontando i valori medi di tali quantità tra le due squadre, si può notare come i 15 archi con peso relativo minore - che costituiscono rispettivamente l'1.9% e il 3.3% del peso complessivo per Napoli e Frosinone - siano comuni ai due *team*; lo stesso vale per 13 delle 15 relazioni con peso relativo maggiore - rispettivamente 76.7% e 71.14%

del totale -. Alla luce di ciò, una considerazione appare lampante: le direzioni lungo cui si sviluppa maggiormente il gioco sono pressoché comuni a tutte le squadre a causa delle specifiche dinamiche che regolano il calcio ovvero, statisticamente parlando, a causa delle *caratteristiche proprie della popolazione* da cui il campione osservato è estratto . Pertanto, appare necessario un approccio al raggruppamento che consideri il contesto in cui la singola rete è collocata.

5.3 Applicazione del metodo di raggruppamento

5.3.1 La scelta della soglia

Con riferimento al metodo di raggruppamento proposto, è necessario innanzitutto scegliere la soglia: non avendo indicazioni di alcun genere a riguardo, la decisione si basa unicamente sul bilanciamento dei due aspetti contrapposti presentati nel paragrafo 3.2.1. Pertanto, il valore selezionato per α è 0.95. L'imposizione di una soglia così elevata è conseguenza della volontà di sacrificare alcune relazioni comunque evidenti - tutte quelle inferiori al novantacinquesimo percentile - pur di ottenere strutture di comunità caratterizzate da connessioni di straordinaria intensità tra i nodi, superiori al novantacinquesimo percentile: così facendo, la problematica relativa alla dimensione *marginale* del metodo di Louvain viene assolutamente ridimensionata. D'altro canto, la metodologia utilizzata ignora l'evidente eterogeneità presente nell'insieme molto ampio di valori inferiori alla soglia.

5.3.2 L'interpretazione dei gruppi

L'analisi del dendrogramma fornisce interessanti indicazioni sulla procedura di raggruppamento: in una prima fase può risultare utile esaminare il *numero di unioni con massima similarità*, ossia il numero di passi dell'algoritmo contraddistinti dall'unione di gruppi di reti che presentano ARI pari a 1. Nel caso analizzato, tale valore è 522, e il raggruppamento corrispondente presenta 52 *clusters* con più di un'unità statistica e 186 singoletti. Dal punto di vista interpretativo, tale partizione propone tutti i 238 differenti schemi di gioco ravvisabili nel campionato di Serie A TIM 2015-2016: le unioni successive, pertanto, raggruppano

questi *stili* di gioco proponendo insiemi di reti caratterizzati da crescente eterogeneità.

La selezione finale del numero di gruppi N_G, non disponendo di specifiche indicazioni a riguardo, deve bilanciare considerazioni prettamente statistiche e finalità pratiche. Come spiegato nel paragrafo 3.2.2, il criterio generalmente utilizzato in quest'ambito è quello relativo alla caduta di similarità, d'altro canto il numero di *clusters* non deve essere troppo elevato per agevolare le analisi in fase di interpretazione. Per esempio, l'ARI massimo in presenza di 20 gruppi è 0.30 e in presenza di 2 gruppi è 0: per questo motivo una scelta congrua a tal proposito sembra essere N_G=15, che presenta ARI massimo pari a 0.21. I valori assai contenuti dell'indice di similarità rappresentano certamente un gravoso limite a tutte le considerazioni future: le unità statistiche che compongono i gruppi sono caratterizzate da strutture di comunità fortemente dissimili tra loro, conseguenza prevedibile visto il valore assunto da N_G in rapporto all'eterogeneità del campione statistico.

Quest'ultima osservazione implica un'ulteriore problematica: la presenza di strutture di comunità diverse all'interno di un dato gruppo rende complicato ravvisare le peculiarità del *cluster* stesso, con ovvie conseguenze a livello interpretativo. Alla luce di ciò, il criterio utilizzato per identificare le principali caratteristiche dei raggruppamenti consiste nel valutare la percentuale di volte che le coppie e le terne di nodi delle reti di un dato gruppo sono state assegnate alla medesima comunità; valori alti di tale indicatore segnalano che il gruppo è caratterizzato da schemi di gioco con frequenti connessioni tra quelle coppie/terne di zone di campo.

Di seguito si riporta il commento dei 15 gruppi trovati. Per semplicità concettuale, questi sono stati collocati in 5 categorie (più una comprendente i 2 singoletti) che identificano le principali macro-tipologie di schema di gioco. La Tabella 5.1 presenta, per ogni squadra, il numero di reti assegnate a ciascun *cluster*.

Schema di gioco assente

- *Gruppo 1,* composto da 349 reti. Il gruppo comprende quasi metà delle osservazioni ed è formato dalle reti che non ravvisano alcuna particolare connessione tra le zone del campo; la numerosità del *cluster* non stupisce visto il valore assai elevato scelto per la soglia. E' composto principalmente da squadre di bassa classifica, che sono assegnate a questo gruppo il 67% delle volte; viceversa, Napoli (4

reti su 38 complessive), Roma (5/38) e Fiorentina (5/38) sono le squadre meno presenti

Schema di gioco con cambio di fascia

- *Gruppo 2,* composto da 132 reti. Lo schema di gioco si sviluppa soprattuto a centrocampo, con cambi di gioco (zone 4-6) e collegamenti con la difesa (3-5). È composto principalmente da squadre di alta classifica (46.2%)

- *Gruppo 3,* composto da 22 reti. Lo schema di gioco si sviluppa con cambi di fascia nella zona difensiva (zone 1-3). È composto principalmente da squadre di alta e medio-alta classifica (77%)

- *Gruppo 4,* composto da 45 reti. Le strutture di comunità delle reti che contraddistinguono questo gruppo sono alquanto differenti tra loro, rendendo difficile l'interpretazione. Principalmente lo schema di gioco si esprime con cambi di fascia (1-3-6) e azioni corali (2-4-8). Le squadre che prendono parte a questo gruppo sono distribuite in maniera omogenea all'interno della classifica

Schema di gioco offensivo

- *Gruppo 5,* composto da 21 reti. Lo schema di gioco connette fortemente centrocampo e attacco, sviluppandosi lungo le terne (5-6-7) e (4-8-9). È composto principalmente da squadre di alta classifica (67%)

- *Gruppo 6,* composto da 26 reti. Lo schema di gioco si sviluppa principalmente nella zona centrale di centrocampo e nella zona d'attacco centro-destra (5-7-8). Le squadre che prendono parte a questo gruppo sono distribuite in maniera omogenea all'interno della classifica

- *Gruppo 7,* composto da 21 reti. Lo schema di gioco interessa la zona centro-destra del centrocampo (5-6) per poi svilupparsi nella zona d'attacco centro-sinistra (5-8-9). Le squadre che prendono parte a questo gruppo sono distribuite in maniera omogenea all'interno della classifica

- *Gruppo 8,* composto da 7 reti. Lo schema di gioco si sviluppa per vie centrali (5-8). E' composto principalmente da squadre di alta e medio-alta classifica (86%)

- *Gruppo 9,* composto da 3 reti. Lo schema di gioco si sviluppa nella zona d'attacco (7-8-9), con alcuni lanci lunghi dalla difesa (1-7). Le reti che compongono il gruppo si riferiscono a singole partite di Udinese, Juventus, Chievo

Schema di gioco con lanci lunghi

- *Gruppo 10,* composto da 46 reti. Lo schema di gioco abbina scambi tra la zona destra della difesa e il centrocampo centrale (1-5) con lanci lunghi dalla difesa alla zona d'attacco (2-9). Le squadre che prendono parte a questo gruppo sono distribuite in maniera omogenea all'interno della classifica

- *Gruppo 11,* composto da 43 reti. Lo schema di gioco prevede la transizione della palla tra zone molto distanti tra loro, sia esternamente (4-7) che centralmente (2-8). E' composto principalmente da squadre di medio-alta o medio-bassa classifica (67%)

- *Gruppo 12,* composto da 4 reti. Lo schema di gioco combina tentativi di giro-palla (3-4,8-9) a lanci lunghi (1-8). Le reti che compongono il gruppo si riferiscono a singole partite di Roma, Napoli, Fiorentina e Milan

Schema di gioco sulle fasce

- *Gruppo 13,* composto da 39 reti. Lo schema di gioco prevede intensi scambi lungo le fasce laterali del campo (1-6,3-4). Questo gruppo rappresenta un *cluster* di assoluto interesse in quanto le strutture di comunità presenti nelle reti sono contraddistinte da un'omogeneità superiore a quella ravvisata in tutti gli altri gruppi: difatti, oltre l'87% delle reti segnala che le zone 3-4 o 1-6 appartengono alla stessa comunità. E' composto principalmente da squadre di alta classifica (49%), mentre quelle di bassa classifica sono praticamente assenti (8%). Una sintesi grafica è proposta in Figura 5.3

Singoletti

- *Gruppo 14,* composto dalla rete relativa allo schema di gioco dell'Atalanta in "Atalanta-Genoa". Lo schema di gioco connette la zona sinistra di difesa con la zona centro-offensiva destra (3-6-7)

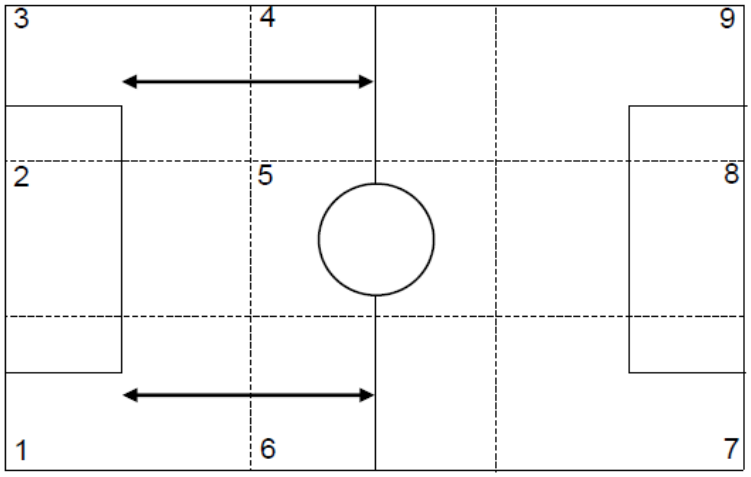


Figura 5.3: Sintesi grafica dello schema di gioco relativo al Gruppo 13. La costruzione del gioco riguarda principalmente le zone laterali del campo.

- *Gruppo 15,* composto dalla rete relativa allo schema di gioco della Fiorentina in "Lazio-Fiorentina". Lo schema di gioco, che ha portato la squadra toscana a segnare ben 4 reti in trasferta contro una squadra di medio-alta classifica, propone un possesso palla distribuito lungo tutto il campo (2-3-4, 5-8)

In appendice A sono riportati, a livello grafico, i 15 schemi di gioco tranne quelli relativi al Gruppo 13 e al Gruppo 1. Il primo, proposto in Figura 5.3, per la rilevanza di questo *cluster* nelle analisi successive; il secondo a causa della mancanza di relazioni considerevoli tra le zone del campo. A partire dalle considerazioni sopra esposte, è possibile trarre interessanti conclusioni sulle caratteristiche complessive del campionato o relative alle prestazioni delle singole squadre. Per esempio, la Juventus ha inanellato un'incredibile serie di vittorie (26 su 28) dopo un avvio di stagione disastroso contrassegnato da soli 12 punti in classifica nelle prime dieci giornate di campionato. Appare perciò sorprendente il fatto che solo due tra le dieci partite in cui la formazione campione d'Italia ha proposto uno schema di gioco "assente" (Gruppo 1) appartengano all'inizio travagliato di stagione. Tale riscontro, seppure nella sua semplicità, conferma un luogo comune relativo al calcio italiano: non sempre trionfa chi esprime il gioco maggiormente coeso.

	Gruppi														
Squadre	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Juventus	10	11	2	3	4	2	1	-	1	3	-	-	1	-	-
Napoli	4	14	-	1	2	-	2	2	-	3	1	1	8	-	-
Roma	5	9	4	1	3	5	1	-	-	3	2	1	4	-	-
Inter	13	8	2	2	3	1	-	1	-	1	3	-	4	-	-
Fiorentina	5	19	1	2	2	-	2	-	-	2	1	1	2	-	1
Sassuolo	17	9	-	3	-	1	-	1	-	1	4	-	2	-	-
Milan	22	4	1	1	-	1	-	-	-	2	3	1	3	-	-
Lazio	11	7	5	3	1	1	3	1	-	2	2	-	2	-	-
Chievo	21	7	2	1	-	1	-	1	1	1	2	-	1	-	-
Genoa	19	7	-	2	-	1	2	-	-	3	4	-	-	-	-
Empoli	20	5	-	-	-	-	4	-	-	6	2	-	1	-	-
Torino	9	12	1	2	2	2	-	-	-	2	5	-	3	-	-
Atalanta	23	1	1	2	-	-	-	-	-	4	3	-	3	1	-
Bologna	25	2	-	3	1	2	1	-	-	2	1	-	1	-	-
Sampdoria	17	1	2	5	2	4	1	1	-	1	3	-	1	-	-
Palermo	27	1	1	3	-	2	1	-	-	2	-	-	1	-	-
Udinese	26	2	-	4	-	1	1	-	1	-	2	-	1	-	-
Carpi	31	2	-	-	-	1	-	-	-	1	3	-	-	-	-
Frosinone	26	4	-	3	1	-	-	-	-	3	1	-	-	-	-
Verona	18	7	-	4	-	1	2	-	-	4	1	-	1	-	-
Totale	349	132	22	45	21	26	21	7	3	46	43	4	39	1	1

Tabella 5.1: Numero di reti assegnate a ciascun gruppo a seconda della squadra.

5.4 Modellazione del numero di gol segnati

La modellazione del risultato finale di una partita di calcio costituisce un ambito di forte interesse: è pertanto necessario valutare, sulla base di considerazioni calcistiche, quale siano le *variabili* che maggiormente influiscono sull'esito di un incontro per poi inserire tale informazione all'interno di un apposito modello statistico.

A tal fine, una prima formulazione è quella ideata da Maher (1982): nel suo articolo, il risultato finale di una gara (X, Y) è rappresentato da due variabili indipendenti con distribuzione di Poisson, i cui parametri dipendono dalla *forza offensiva* e da quella *difensiva* delle singole squadre a seconda che queste giochino in casa o in trasferta. Il modello sembra assolutamente ragionevole: infatti è lecito attendersi che il numero di gol segnati da una squadra dipenda dalla qualità della stessa e dal fatto di giocare sul proprio terreno di gioco o meno. A partire da queste osservazioni, è stata presentata una serie di sviluppi con l'obiettivo di cogliere in maniera sempre più dettagliata le reali dinamiche che determinano le prestazioni delle singole squadre: Dixon & Coles (1997), Baio & Blangiardo (2010) e Koopman & Lit (2015) sono solo alcuni dei numerosissimi esempi a riguardo.

Lo scopo delle successive analisi è quello di valutare se, ed eventualmente in che modo, lo schema di gioco influisca sul risultato di una partita. La trattazione si articola come segue: nel paragrafo 5.4.1 viene considerato un primo semplice modello, mentre nel paragrafo 5.4.2 viene proposta una modifica a quello ideato da Dixon e Coles.

5.4.1 Regressione di Poisson

Un modello capace di riproporre le peculiarità evidenziate da Maher (1982), con l'aggiunta dell'informazione riguardante lo schema di gioco, è quello della regressione di Poisson con legame canonico. Per una trattazione approfondita dei modelli lineari generalizzati si rimanda, ad esempio, a Azzalini (2004).

Sia $X_{ij} \sim Poisson(\lambda_{ij})$ il numero di gol segnati dalla squadra di casa i nella gara di campionato contro la squadra in trasferta j, e sia $Y_{ij} \sim Poisson(\mu_{ij})$ il numero di gol segnati dalla squadra fuori casa nella stessa sfida; poiché il campionato è costituito da 20 *team* complessivamente, risulta $1 \leq i \neq j \leq 20$. Per semplicità, si ipotizzi inoltre che il numero di gol segnati da una formazione sia indipendente dal numero di reti segnate in passato e dalle realizzazioni della

squadra avversaria. Considerando le 380 partite del campionato di Serie A TIM 2015-2016, si dispone quindi di un vettore di osservazioni $z = (x_{12}, x_{13}, \ldots, x_{20,18}, x_{20,19}, y_{12}, y_{13}, \ldots, y_{20,18}, y_{20,19})$. A causa della scarsa numerosità dei gruppi 8,9,12,14,15, vengono rimosse dal vettore z le 16 osservazioni che riportano il numero di marcature di una squadra che, in quella specifica partita, ha proposto uno schema di gioco classificato in uno dei suddetti *clusters*: il numero totale di osservazioni risulta pertanto $n = 744$.

Il modello di regressione si articola nel seguente modo:

$$log(\lambda_{ij}) = \beta_0 + \gamma + \sum_{f=2}^{4} \alpha_f r_{if} + \sum_{h \in A} \beta_h c_{ij,h}$$

$$log(\mu_{ij}) = \beta_0 + \sum_{f=2}^{4} \alpha_f r_{jf} + \sum_{h \in A} \beta_h t_{ij,h}$$

con $A = \{2, 3, 4, 5, 6, 7, 10, 11, 13\}$. In particolare:

- β_0 è l'intercetta del modello
- γ è il parametro che esprime il vantaggio dato dal giocare in casa, ipotizzato costante per tutte le squadre
- r_{if} è la variabile che identifica la fascia di classifica in cui la squadra i si è posizionata al termine del campionato. La *baseline* è costituita dalla fascia alta, mentre $f = 2$ indica la fascia medio-alta, $f = 3$ la fascia medio-bassa e $f = 4$ la fascia bassa
- $c_{ij,h}$ è la variabile che identifica il gruppo a cui è stato assegnato lo schema di gioco della squadra di casa i nella partita contro la squadra j; $t_{ij,h}$ quella che identifica il gruppo a cui è stato assegnato lo schema di gioco della squadra in trasferta j nella stessa gara. La *baseline* è costituita dal gruppo 1, ed i valori di h rappresentano il numero identificativo degli altri nove gruppi

In tabella 5.2 sono riportate le stime del modello. Innanzitutto, il valore dei primi cinque coefficienti stimati sembra essere in linea con le aspettative: la zona di classifica in cui una squadra si è posizionata, così come il fatto di giocare sul proprio terreno di gioco, influisce sul numero di gol segnati. Viceversa, le restanti stime non forniscono segnali confortanti, con livelli di significatività osservati estremamente elevati nella maggior parte dei casi. Ad una prima occhiata, l'evidenza sembra

	Stima	Std. Error	Stat. T	Valore p
β_0	0.3521	0.0860	4.09	0.0000
γ	0.2889	0.0658	4.39	0.0000
α_2	-0.3195	0.0873	-3.66	0.0003
α_3	-0.4353	0.0915	-4.76	0.0000
α_4	-0.5894	0.0991	-5.95	0.0000
β_2	0.0910	0.0919	0.99	0.3217
β_3	0.0408	0.1915	0.21	0.8311
β_4	0.0289	0.1453	0.20	0.8421
β_5	-0.0697	0.2017	-0.35	0.7296
β_6	0.0208	0.1797	0.12	0.9080
β_7	-0.2501	0.2299	-1.09	0.2767
β_{10}	0.1522	0.1341	1.14	0.2562
β_{11}	0.1859	0.1376	1.35	0.1767
β_{13}	0.2933	0.1340	2.19	0.0286

Tabella 5.2: Risultati della regressione di Poisson. Stime, Standard error, valori della statistica T e relativi *valori p*.

propendere per una considerazione in parte condivisibile: le squadre segnano perché sono "forti" - ossia perché hanno i calciatori migliori, gli allenatori più preparati *etc.* - e non perché propongono uno schema di gioco particolarmente efficace in fase realizzativa. L'unica eccezione appare essere quella relativa al gruppo 13, con un *valore p* inferiore alla soglia 0.05: già in fase di interpretazione dei gruppi si era evidenziato come questo *cluster* fosse di particolare interesse in quanto costituito da reti con strutture di comunità molto simili. Pertanto è possibile che l'eccessiva eterogeneità degli altri gruppi non permetta di ravvisare un effetto sulla variabile risposta; non sorprende che il coefficiente relativo al gruppo con minore omogeneità interna, ossia il numero 4, sia uno di quelli con *valore p* maggiore.

Questa formulazione presenta comunque una visione eccessivamente semplicistica del fenomeno, sulla quale non ci si soffermerà ulteriormente. Ciononostante, dalle analisi svolte si può trarre qualche interessante indicazione: nel seguente sottoparagrafo si valuterà l'effettiva incidenza dello schema di gioco relativo al gruppo 13 all'interno di un modello maggiormente elaborato.

5.4.2 Modifica del modello di Dixon-Coles

Si riporta di seguito la presentazione sintetica del succitato modello di Dixon-Coles.[2] Sia X_k il numero di gol segnati dalla squadra di casa al termine della gara k con $k = 1, \ldots, 380$ e sia Y_k il numero di reti messe a segno dalla squadra in trasferta durante la stessa partita; per praticità, le gare sono considerate in ordine temporale. Ad ogni partita k, quindi, è associata implicitamente una squadra i che ha giocato in casa e una squadra j che ha giocato in trasferta, con $1 \leq i \neq j \leq 20$.

Rivisitando parzialmente la formulazione proposta da Maher, il punto di partenza del modello prevede che il numero di gol segnati dalle due squadre sia realizzazione di due variabili di Poisson indipendenti:

$$\begin{aligned} X_k &\sim Poisson(\lambda_k) \\ Y_k &\sim Poisson(\mu_k) \end{aligned}$$

con:

$$\begin{aligned} log(\lambda_k) &= \gamma + \alpha_{i(k)} + \beta_{j(k)} \\ log(\mu_k) &= \alpha_{j(k)} + \beta_{i(k)} \end{aligned} \tag{5.1}$$

con γ parametro che esprime il vantaggio di giocare in casa, $\alpha_{i(k)}$,$\beta_{i(k)}$ parametri che esprimono la forza offensiva e difensiva di una squadra, $i(k)$, $j(k)$ indicatori rispettivamente della squadra di casa e di quella in trasferta.

La prima modifica sostanziale introdotta dai due autori è la specificazione di una struttura di dipendenza tra le variabili (X_k, Y_k): la motivazione della scelta risiede in uno studio preliminare, condotto sempre da Dixon e Coles, che evidenzia la violazione dell'ipotesi di indipendenza per i risultati di 0-0,0-1,1-0,1-1. Di conseguenza, la funzione di probabilità congiunta $P(X_k = x, Y_k = y) = P(X_k = x)P(Y_k = y)$ viene modificata nel seguente modo:

$$P(X_k = x, Y_k = y) = \tau_{\lambda_k,\mu_k}(x, y)P(X_k = x)P(Y_k = y)$$

[2] Il codice R relativo al modello di Dixon e Coles è stato tratto da Dandolo (2017).

$$\tau_{\lambda_k,\mu_k}(x,y) = \begin{cases} 1-\lambda_k\mu_k\rho & \text{se } x=y=0 \\ 1+\lambda_k\rho & \text{se } x=0, y=1 \\ 1+\mu_k\rho & \text{se } x=1, y=0 \\ 1-\rho & \text{se } x=y=1 \\ 1 & \text{altrimenti} \end{cases}$$

sotto i vincoli:

$$max\left(-\frac{1}{\lambda_k}, -\frac{1}{\mu_k}\right) \leq \rho \leq min\left(\frac{1}{\lambda_k\mu_k}, 1\right)$$

$$\frac{\sum_{i=1}^{20}\alpha_i}{20} = 1$$

di cui quest'ultimo introdotto per prevenire la sovraparametrizzazione del modello.

La seconda innovazione consta nel considerare lo stato di forma delle squadre nel corso del campionato: è difatti assolutamente lecito aspettarsi che le prestazioni delle varie formazioni varino nel corso della stagione, con periodi particolarmente esaltanti ed altri di maggiore difficoltà. Alla luce di ciò, i due autori dividono la stagione calcistica in una serie di sottoperiodi - o *istanti di tempo* - di lunghezza pari a metà settimana. La formulazione finale del modello prevede pertanto la costruzione di una 'pseudoverosimiglianza' per ciascun istante di tempo t:

$$L_t(\alpha_i, \beta_i, \rho, \gamma; i = 1, \ldots, 20) = \prod_{k \in A_t} \left[\tau_{\lambda_k,\mu_k}(x_k, y_k) e^{-\lambda_k} \lambda_k^{x_k} e^{-\mu_k} \mu_k^{y_k}\right]^{\phi(t-t_k)} \tag{5.2}$$

con t_k istante in cui si gioca la partita k, $A_t = \{k : t_k < t\}$ e $\phi(t)$ una funzione non decrescente. Si noti come la funzione (5.2) dipenda solamente dalle gare giocate prima dell'istante t, rendendo così utilizzabile il modello per prevedere il numero di gol segnati dalle varie squadre al tempo t. Per non appesantire eccessivamente la notazione, nella formula (5.2) non viene indicata la dipendenza temporale per i parametri $\alpha_i, \beta_i, \rho, \gamma$. Tra le tante funzioni possibili, $\phi(t)$ è definita come segue:

$$\phi(t) = e^{-\xi t}$$

con $\xi > 0$ opportuno parametro. La scelta di ξ risulta particolarmente difficoltosa: l'equazione (5.2) definisce al variare di t una sequenza ordinata di funzioni non indipendenti, che rende di fatto assai complicata la

selezione dello ξ in grado di massimizzare la capacità predittiva globale del modello. La metodologia proposta da Dixon e Coles per ovviare al problema, essendo di marginale interesse per i nostri scopi, è riportata in Appendice B.

Appare evidente la maggiore accuratezza del modello proposto se confrontato con la regressione di Poisson precedente: oltre al fondamentale ruolo svolto dalle funzioni $\tau_{\lambda,\mu}(x,y)$ e $\phi(t)$, in questo frangente la qualità di una determinata squadra non è più indicata sommariamente dalla fascia di classifica in cui si è collocata, ma risulta specifica per ogni formazione, separando la capacità offensiva da quella difensiva.

Al fine di verificare l'effetto dello schema di gioco "sulle fasce" sul numero di gol segnati, la modifica suggerita è concettualmente piuttosto semplice; a partire dalla formula (5.1), si ottiene:

$$\begin{aligned} log(\lambda_k) &= \gamma + \alpha_{i(k)} + \beta_{j(k)} + \delta c_{i(k)} \\ log(\mu_k) &= \alpha_{j(k)} + \beta_{i(k)} + \delta c_{j(k)} \end{aligned} \tag{5.3}$$

con $c_{i(k)} = 1$ se lo schema di gioco proposto dalla squadra i durante la partita k è assegnato al gruppo 13, 0 altrimenti; δ parametro che esprime l'effetto di tale schema sul numero di gol segnati.

A differenza del modello originario, d'ora in avanti gli istanti di tempo non saranno più di ampiezza fissa, ma t_k rappresenterà il preciso giorno dell'anno in cui si è svolta la partita k. Questo accorgimento è motivato dal fatto che a differenza di quanto avveniva negli anni '90, periodo in cui Dixon e Coles hanno effettuato lo studio, oggigiorno le singole giornate di campionato prevedono partite in giorni anche molto distanti tra loro, ed è quindi auspicabile una suddivisione più precisa della stagione sportiva.

Si noti innanzitutto come il modello con l'introduzione della modifica (5.3) non possa più essere utilizzato per prevedere in anticipo l'esito di una partita: l'informazione relativa alla tipologia di gioco, difatti, è ottenibile solamente al termine della gara stessa.

Poiché la quantità d'interesse è δ, l'attenzione sarà posta esclusivamente su tale parametro. La massimizzazione della pseudoverosimiglianza (5.2), opportunamente modificata tramite l'introduzione di δ, è ottenibile solamente per via numerica, così come l'originale: la stima puntuale del coefficiente in questione, come noto, non risulta sufficiente per trarre conclusioni di natura inferenziale. Pertanto, particolarmente utile è l'analisi della *verosimiglianza-profilo* (ad esempio, Cox & Barndorff-Nielsen, 1994), definita come:

$$L_t^P(\delta) = L_t(\delta, \hat{\alpha}_{i_\delta}, \hat{\beta}_{i_\delta}, \hat{\rho}_\delta, \hat{\gamma}_\delta) \tag{5.4}$$

con $\hat{\theta} = (\hat{\alpha}_{i_\delta}, \hat{\beta}_{i_\delta}, \hat{\rho}_\delta, \hat{\gamma}_\delta)$ insieme di valori che rende massima la funzione $L_t(\delta, \theta)$, con δ fissato. Come in precedenza, la dipendenza dei parametri dal tempo è data per sottintesa. Grazie ai noti risultati asintotici (ad esempio, Pace & Salvan, 2001), è possibile ottenere un intervallo di confidenza per δ basato sul log-rapporto di verosimiglianza profilo:

$$\left\{\delta : 2\big(l_t^P(\hat{\delta}) - l_t^P(\delta)\big) < \chi^2_{1;1-\alpha}\right\} \tag{5.5}$$

con $l_t^P(\delta) = log\big(L_t^P(\delta)\big)$. Tale intervallo può essere calcolato nei vari istanti di tempo t: per ottenere stime affidabili è però sensato attendere alcune giornate di campionato in modo da non trarre conclusioni fuorvianti sulla base di stime scarsamente attendibili. Inoltre, verranno ignorate le ultime cinque settimane del torneo: spesso il raggiungimento dell'obiettivo di squadra - sia esso la salvezza o la vittoria del campionato - porta ad assistere a sfide in cui una formazione offre una prestazione ben al di sotto delle proprie possibilità a causa di motivazione ed impegno scarsi. I dati di cui si dispone constano di $T = 97$ giorni distinti di gara, dal 22 Agosto 2015 al 15 Maggio 2016: quindi, pur disponendo delle stime dei parametri del modello relative alle giornate di gara fino al 13 Febbraio 2016 ($t = 60$), queste non saranno valutate per il suddetto motivo. Questa prima parte di campionato viene utilizzata per ottimizzare la scelta di ξ tramite la procedura descritta in Appendice B; il valore selezionato per tale parametro è 0.0008 e sarà considerato costante per tutto il resto della stagione sportiva. A partire dal 14 Febbraio 2016 ($t = 61$), gli intervalli di confidenza sono calcolati fino al 17 Aprile 2016 ($t = 84$), per un totale di 24 giorni distinti di gioco.

Il modello offre vantaggi non indifferenti: ponendosi di volta in volta all'istante t, è possibile ottenere stime estremamente accurate della quantità d'interesse sulla base di tutte le partite - opportunamente pesate - avvenute in precedenza. Inoltre, il progressivo aumento del numero di dati considerati comporta il restringimento degli intervalli di confidenza, con ovvi benefici a livello inferenziale.

Gli intervalli di confidenza di livello 0.95 sono riportati in Figura 5.4. Nell'interpretazione dei risultati è richiesta particolare cautela in quanto le verosimiglianze profilo, così come le pseudoverosimiglianze da cui sono calcolate, non sono indipendenti: ciononostante, l'evidenza di significatività (con α=0.05) nei vari istanti di tempo è tale che le conclusioni inferenziali risultano chiare. Difatti, solamente in un'occasione l'estremo inferiore dell'intervallo di confidenza è negativo, con valore estremamente prossimo allo zero ($IC_{17}^- = -0.0049$); alla luce di ciò, è possibile concludere che lo schema di gioco "sulle fasce" incide positi-

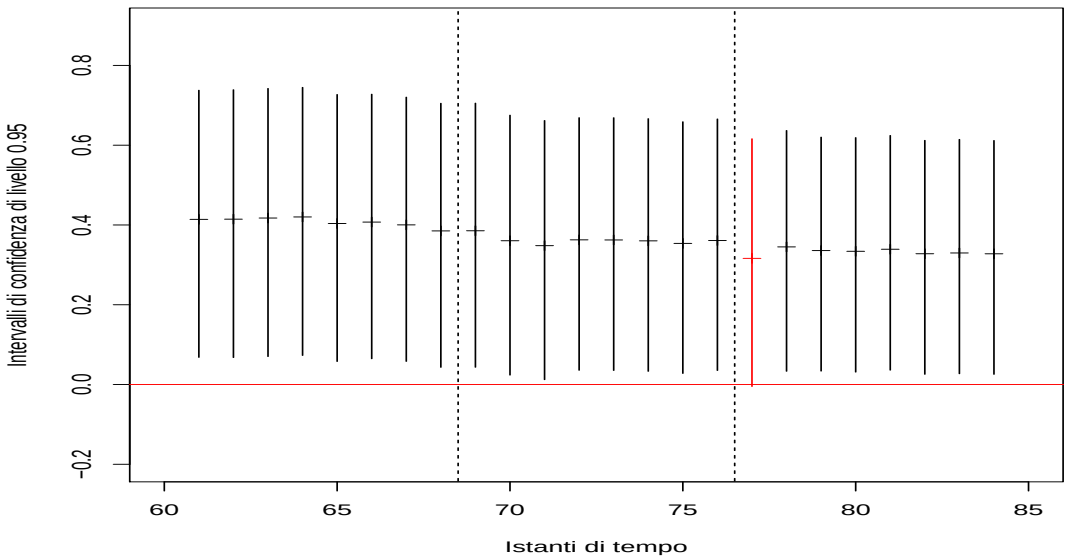


Figura 5.4: Intervalli di confidenza di livello 0.95 e stime puntuali (trattino orizzontale) del parametro δ. In rosso l'intervallo di confidenza che include lo 0. Le linee tratteggiate scandiscono i mesi dell'anno (Febbraio - Marzo - Aprile).

vamente sul numero di gol effettuati. Seppur mantenendo il carattere tipicamente oscillatorio, il trend delle stime puntuali appare decrescente: con il passare del tempo, quindi, l'influenza della tipologia di gioco '13' sulla capacità di segnare sembra diminuire. Tale fenomeno ha una possibile spiegazione pratica: dopo un inevitabile primo periodo di adattamento, le varie squadre mettono a punto una serie di accorgimenti tecnico-tattici finalizzati a cautelarsi contro questo specifico schema di gioco, rendendolo così meno efficace in fase realizzativa.

5.5 Riflessioni finali

I risultati ottenuti sono incoraggianti: la significatività del parametro δ sottolinea l'importanza dello studio della *rete di passaggi* di una squadra per quantificare il potenziale offensivo della stessa. Un raggruppamento costituito da un numero maggiore di *clusters* permetterebbe di ottenere gruppi internamente meno eterogenei; ciò consentirebbe di estendere l'analisi ad un più ampio insieme di tipologie di gioco, rendendo possibile confronti tra queste.

Un aspetto da esaminare, e che può risultare di particolare rilevanza per futuri sviluppi, è la non direzionalità delle connessioni che costituiscono le reti: è lecito attendersi che l'informazione riguardo quale zona del campo promuova la relazione e quale invece la riceva possa influire sui gruppi trovati, ed essere inoltre assai utile a fini interpretativi. Si pensi, ad esempio, alla differenza in termini di gioco di una squadra che inizia l'azione dalla difesa per poi sviluppare la trama offensiva a centrocampo, e una formazione che, per "temporeggiare" e ottenere un pareggio chiave in ottica salvezza, passa sistematicamente la palla dal centrocampo alla difesa.

Inoltre, è possibile considerare una divisione del terreno di gioco differente da quella proposta. Oltre che nel numero, le zone possono variare anche per forma o estensione: per esempio, può essere di particolare interesse valutare le transizioni riguardanti l'area di rigore, che nella Figura 5.1 interessa parte delle zone 1,2,3.

A differenza di quello originario di Dixon e Coles, il modello includente il parametro δ non può essere impiegato nell'ambito delle scommesse calcistiche poiché utilizza l'informazione relativa allo schema di gioco ottenuta al termine della partita di cui si vuole prevedere l'esito. In tal senso, un primo sviluppo può essere quello di determinare lo stile di gioco che un *team* proporrà in un dato incontro a partire dall'analisi del *giro-palla* nelle gare precedenti.

Appendice A

Rappresentazione grafica dei principali schemi di gioco

Di seguito è riportata una sintesi grafica dei principali schemi di gioco. Non è rappresentato il gruppo 1, in quanto privo di connessioni rilevanti tra le zone del campo, e il gruppo 13, il cui grafico è già stato presentato nel paragrafo 5.3.2.

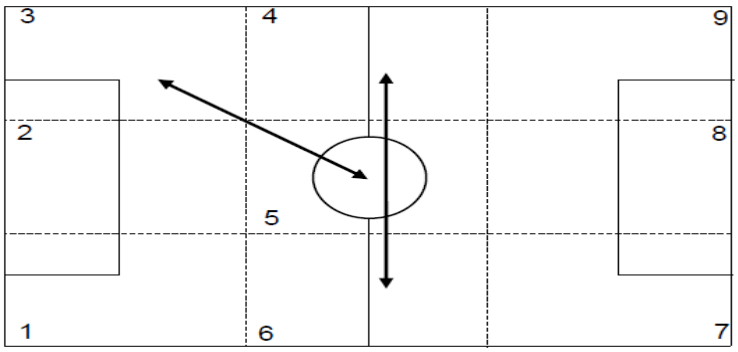


Figura A.1: Gruppo 2

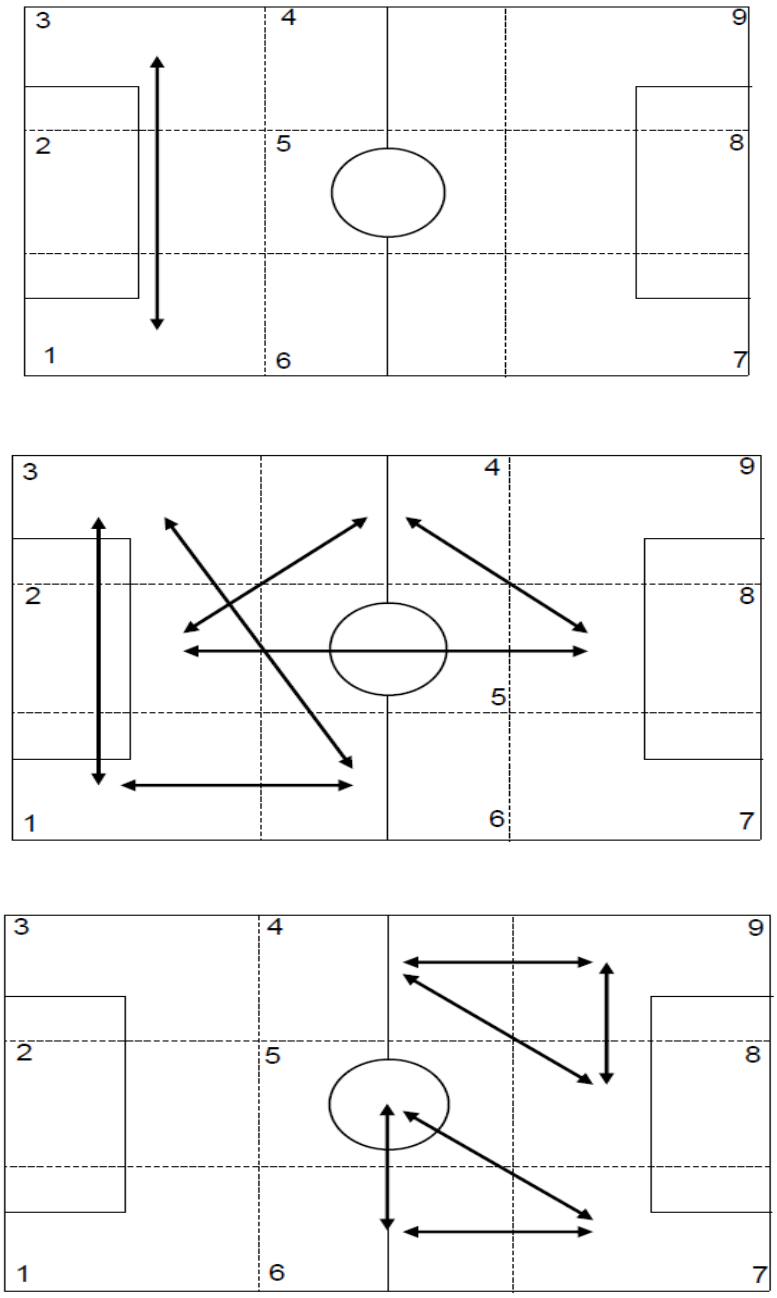


Figura A.2: Gruppo 3, in alto. Gruppo 4, al centro. Gruppo 5, in basso.

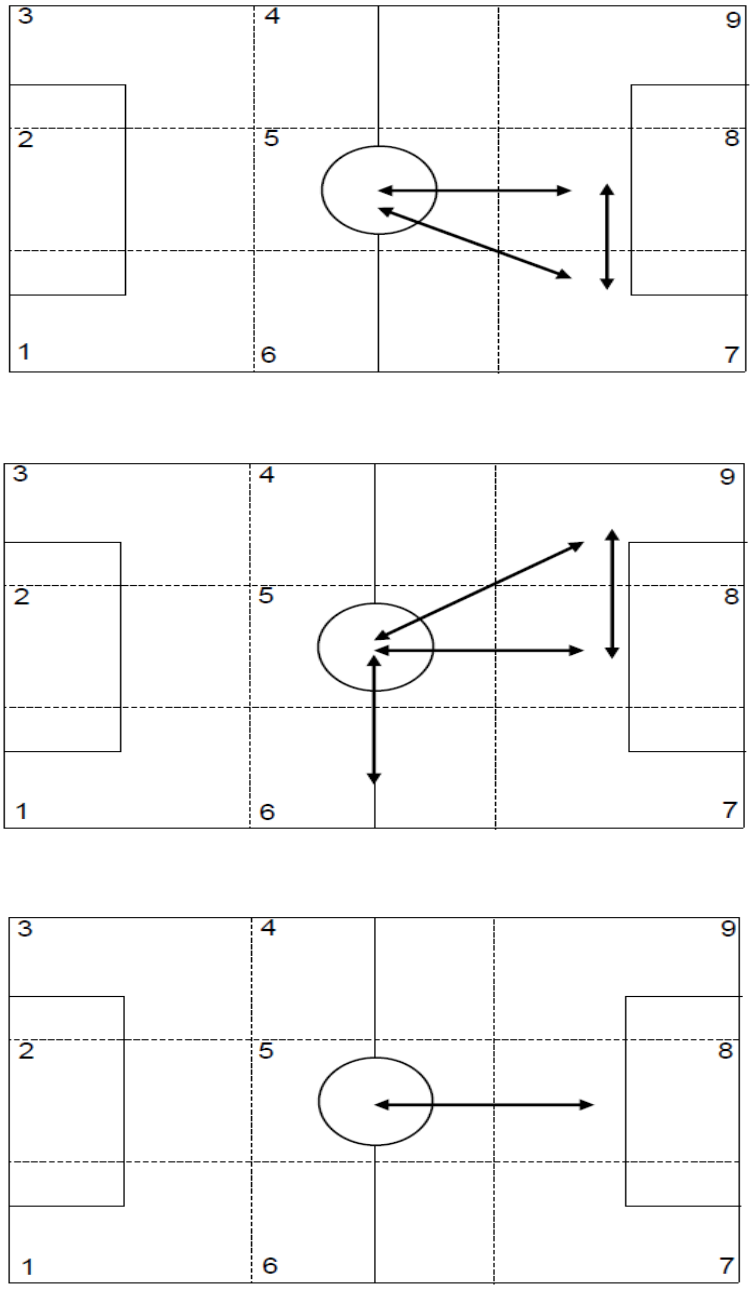


Figura A.3: Gruppo 6, in alto. Gruppo 7, al centro. Gruppo 8, in basso.

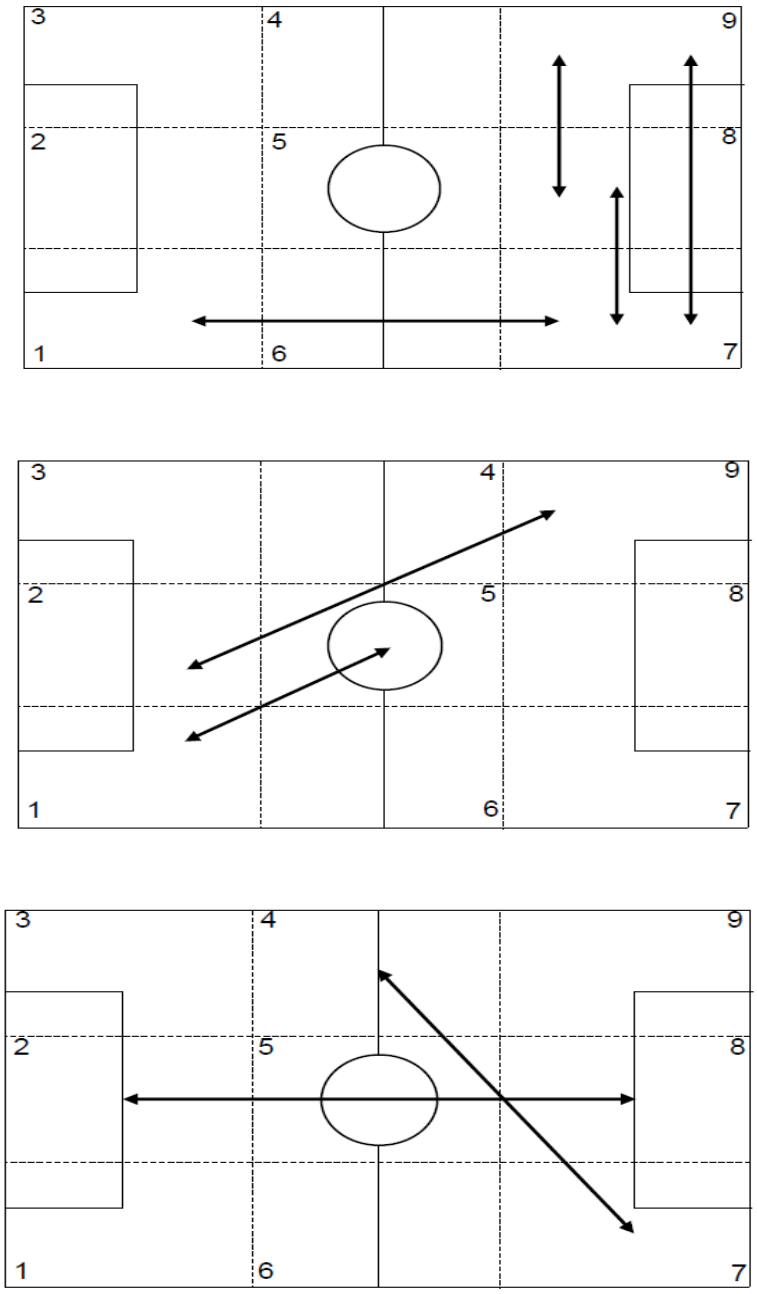


Figura A.4: Gruppo 9, in alto. Gruppo 10, al centro. Gruppo 11, in basso.

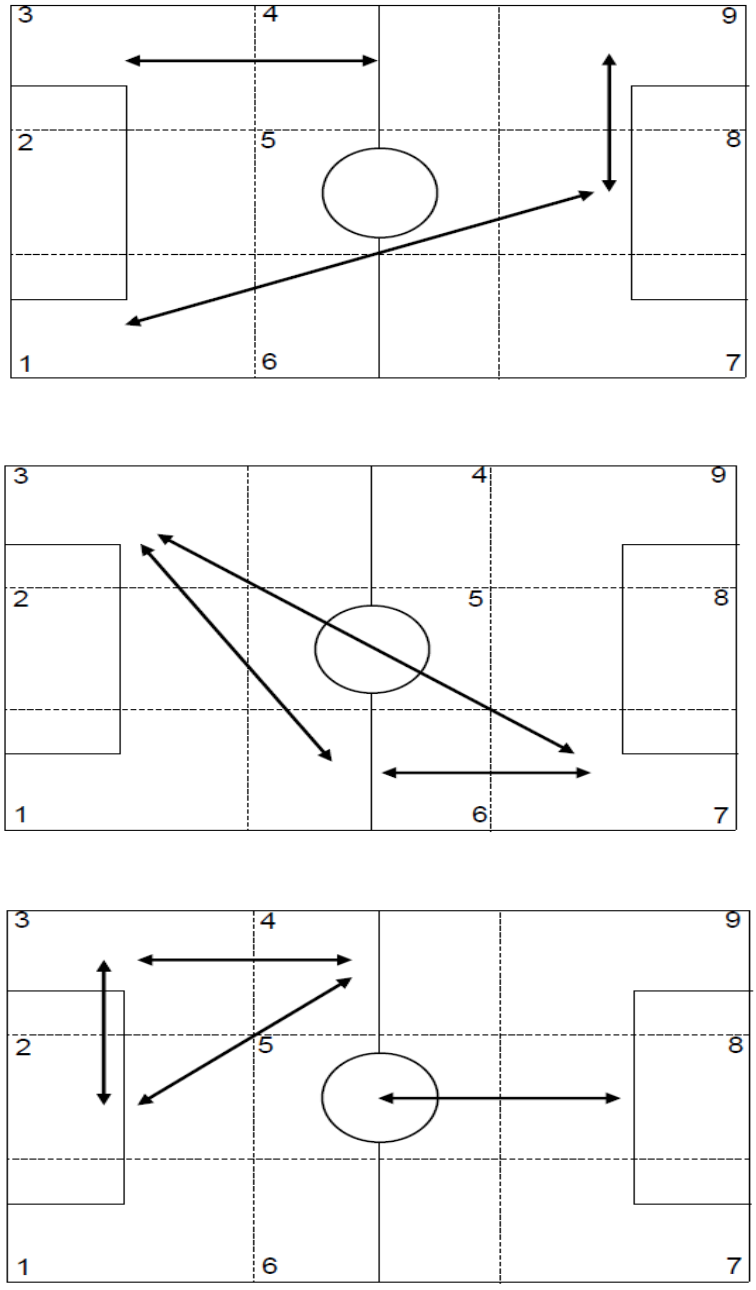


Figura A.5: Gruppo 12, in alto. Gruppo 14, al centro. Gruppo 15, in basso.

Appendice B

Ottimizzazione di ξ

Di seguito viene presentata la procedura proposta da Dixon & Coles (1997) per la selezione del valore ottimo di ξ. Poiché l'obiettivo degli autori è quello di sviluppare una strategia redditizia nello scommettere sull'esito di un incontro con riferimento alla vittoria-pareggio-sconfitta delle squadre, l'ottimizzazione di ξ può essere semplificata massimizzando una funzione diversa dalla (5.2). In particolare, il fine sarà quello di rendere massima la capacità complessiva del modello di prevedere la formazione vincente - o l'eventuale pareggio - di una gara e non i gol delle singole formazioni. Pertanto i due autori definiscono come stima della probabilità rispettivamente di vittoria, pareggio e sconfitta della squadra di casa nella gara k le seguenti quantità:

$$p_k^H(\xi) = \sum_{l,m \in B_H} Pr(X_k = l, Y_k = m)$$

$$p_k^D(\xi) = \sum_{l,m \in B_D} Pr(X_k = l, Y_k = m)$$

$$p_k^A(\xi) = \sum_{l,m \in B_A} Pr(X_k = l, Y_k = m)$$

con $B_H = \{(l,m) : l > m\}$, $B_D = \{(l,m) : l = m\}$ e $B_A = \{(l,m) : l < m\}$. Le probabilità dei gol segnati dalle due squadre $P(X_k = l)$ e $P(Y_k = m)$ sono determinate massimizzando la funzione (5.2) all'istante $t(k)$ in cui la partita k viene disputata, con ξ valore predefinito.

Quindi, è possibile calcolare per ogni valore di ξ:

$$S(\xi) = \sum_{k=1}^{N} (\delta_k^H log p_k^H + \delta_k^D log p_k^D + \delta_k^A log p_k^A)$$

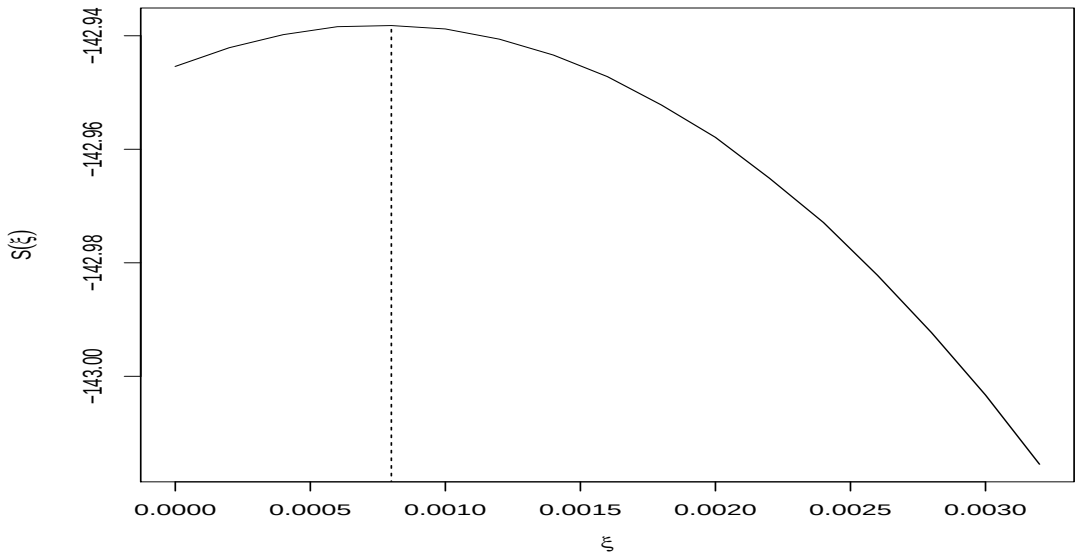


Figura B.1: Grafico della funzione $S(\xi)$. Il massimo si ottiene quando $\xi = 0.0008$.

con $\delta_k^H = 1 (\delta_k^D = 1, \delta_k^A = 1)$ se nella gara k la squadra di casa ha vinto (pareggiato, perso) e 0 altrimenti, N numero totale di gare. La scelta di ξ si riduce così al valore capace di massimizzare la funzione $S(\xi)$.

Come esplicitato nel corso della trattazione, la selezione di ξ è stata realizzata utilizzando i primi $t = 60$ istanti di tempo, con risultato $\xi = 0.0008$. Il grafico della funzione $S(\xi)$ è riportato in figura B.1.

Ringraziamenti

La stesura di questo lavoro si è resa possibile grazie ad alcune persone che mi hanno onorato con il loro contributo umano e professionale, e che ora mi sento in dovere di ringraziare.

Bruno Scarpa mi è stato di guida con consigli e correzioni, dandomi tutto il supporto possibile anche in attività non connesse a questa tesi.

Gli scambi di idee con Daniele Durante mi hanno aiutato a definire l'idea sottostante al presente lavoro.

Lorenzo Favaro, CEO di SportAnalisi, fin da subito mi ha dedicato il suo tempo e mi ha consentito l'accesso a dati con cui da tempo desideravo cimentarmi.

Ringrazio Lazar Petrov, rappresentante e responsabile italiano per InStat, fonte di tali dati.

Infine, un grazie a David Dandolo per l'ottimo codice R relativo al modello di Dixon e Coles.

Bibliografia

Azzalini, Adelchi. 2004. *Inferenza statistica: una presentazione basata sul concetto di verosimiglianza*. Springer Science & Business Media.

Azzalini, Adelchi, & Menardi, Giovanna. 2014. «Clustering via Nonparametric Density Estimation: The R Package pdfCluster». *Journal of Statistical Software*, **57**(11), 1–26.

Azzalini, Adelchi, & Scarpa, Bruno. 2012. *Data analysis and data mining: An introduction*. Oxfor University Press USA.

Baio, Gianluca, & Blangiardo, Marta. 2010. «Bayesian hierarchical model for the prediction of football results». *Journal of Applied Statistics*, **37**(2), 253–264.

Barnes, John Arundel. 1954. «Class and committees in a Norwegian island parish». *Human relations*, **7**(1), 39–58.

Bassett, Danielle Smith, & Bullmore, ED. 2006. «Small-world brain networks». *The neuroscientist*, **12**(6), 512–523.

Blondel, Vincent D, Guillaume, Jean-Loup, Lambiotte, Renaud, & Lefebvre, Etienne. 2008. «Fast unfolding of communities in large networks». *Journal of statistical mechanics: theory and experiment*, **2008**(10), P10008.

Boccaletti, Stefano, Latora, Vito, Moreno, Yamir, Chavez, Martin, & Hwang, D-U. 2006. «Complex networks: Structure and dynamics«. *Physics reports*, **424**(4), 175–308.

Boccaletti, Stefano, Bianconi, Ginestra, Criado, Regino, Del Genio, Charo I, Gómez-Gardenes, Jesús, Romance, Miguel, Sendina-Nadal, Irene, Wang, Zhen, & Zanin, Massimiliano. 2014. «The structure and dynamics of multilayer networks». *Physics Reports*, **544**(1), 1–122.

Brandes, Ulrik, Delling, Daniel, Gaertler, Marco, Görke, Robert, Hoefer, Martin, Nikoloski, Zoran, & Wagner, Dorothea. 2006. «Maximizing modularity is hard». *arXiv preprint physics/0608255*.

Clauset, Aaron, Newman, Mark EJ, & Moore, Cristopher. 2004. «Finding community structure in very large networks». *Physical review E*, **70**(6), 066111.

Cox, David Roxbee, & Barndorff-Nielsen, OE. 1994. *Inference and asymptotics*. Vol. 52. CRC Press.

Csardi, Gabor, & Nepusz, Tamas. 2006. «The igraph software package for complex network research». *InterJournal*, **Complex Systems**, 1695.

Dandolo, David. 2017. Modellazione statistica di risultati calcistici. *Relazione finale di laurea triennale.*

Dixon, Mark J, & Coles, Stuart G. 1997. «Modelling association football scores and inefficiencies in the football betting market». *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**(2), 265–280.

Dugué, Nicolas, & Perez, Anthony. 2015. *«Directed Louvain: maximizing modularity in directed networks»*. Ph.D. thesis, Université d'Orléans.

Durante, Daniele, Dunson, David B, & Vogelstein, Joshua T. 2016. «Nonparametric Bayes modeling of populations of networks». *Journal of the American Statistical Association.*

Fortunato, Santo, & Barthélemy, Marc. 2007. «Resolution limit in community detection». *Proceedings of the National Academy of Sciences*, **104**(1), 36–41.

Fowlkes, Edward B, & Mallows, Colin L. 1983. «A method for comparing two hierarchical clusterings». *Journal of the American statistical association*, **78**(383), 553–569.

Fraley, Chris, & Raftery, Adrian E. 1998. «How many clusters? Which clustering method? Answers via model-based cluster analysis». *The computer journal*, **41**(8), 578–588.

Friedman, Jerome H. 1998. «Data mining and statistics: What's the connection?». *Computing Science and Statistics*, **29**(1), 3–9.

Girvan, Michelle, & Newman, Mark EJ. 2002. «Community structure in social and biological networks». *Proceedings of the national academy of sciences*, **99**(12), 7821–7826.

Gómez, Sergio, Jensen, Pablo, & Arenas, Alex. 2009. «Analysis of community structure in networks of correlated data». *Physical Review E*, **80**(1), 016114.

Górski, AZ, Drożdż, S, & Kwapień, J. 2008. «Scale free effects in world currency exchange network». *The European Physical Journal B-Condensed Matter and Complex Systems*, **66**(1), 91–96.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition)*. Springer-Verlag, New York.

Hubert, Lawrence, & Arabie, Phipps. 1985. «Comparing partitions». *Journal of classification*, **2**(1), 193–218.

Jonsson, Pall F, Cavanna, Tamara, Zicha, Daniel, & Bates, Paul A. 2006. «Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis». *BMC bioinformatics*, **7**(1), 2.

Kaufman, Leonard, & Rousseeuw, Peter J. 1990. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Koopman, Siem Jan, & Lit, Rutger. 2015. «A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League». *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178**(1), 167–186.

Leicht, Elizabeth A, & Newman, Mark EJ. 2008. «Community structure in directed networks». *Physical review letters*, **100**(11), 118703.

Maher, Michael J. 1982. «Modelling association football scores». *Statistica Neerlandica*, **36**(3), 109–118.

McPherson, Miller, Smith-Lovin, Lynn, & Cook, James M. 2001. «Birds of a feather: Homophily in social networks». *Annual review of sociology*, **27**(1), 415–444.

Morey, Leslie C, & Agresti, Alan. 1984. «The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement». *Educational and Psychological Measurement*, **44**(1), 33–37.

Newman, Mark. 2010. *Networks: an introduction*. United States: Oxford University Press Inc., New York.

Newman, Mark EJ. 2004. «Fast algorithm for detecting community structure in networks». *Physical review E*, **69**(6), 066133.

Newman, Mark EJ, & Girvan, Michelle. 2004. «Finding and evaluating community structure in networks». *Physical review E*, **69**(2), 026113.

Pace, Luigi, & Salvan, Alessandra. 2001. *Introduzione alla statistica: Inferenza, verosimiglianza, modelli.xvi, 422 p.* Cedam.

Pons, Pascal, & Latapy, Matthieu. 2006. «Computing communities in large networks using random walks.». *J. Graph Algorithms Appl.*, **10**(2), 191–218.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, William M. 1971. «Objective criteria for the evaluation of clustering methods». *Journal of the American Statistical association*, **66**(336), 846–850.

Santini, Augusto. 2014. *Quando il calcio dura poco: in Serie A non si raggiunge l'ora di gioco, ma non siamo i peggiori.* http://www.goal.com/it/news/2/serie-a/2014/03/11/4676323/quando-il-calcio-dura-poco-in-serie-a-non-si-raggiunge-lora.

Traag, Vincent A, & Bruggeman, Jeroen. 2009. «Community detection in networks with positive and negative links». *Physical Review E*, **80**(3), 036115.

Viroli, Cinzia. 2011. «Finite mixtures of matrix normal distributions for classifying three-way data». *Statistics and Computing*, **21**(4), 511–522.

Wasserman, Stanley, & Faust, Katherine. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge University Press.

Wu, Fang, & Huberman, Bernardo A. 2004. «Finding communities in linear time: a physics approach». *The European Physical Journal B-Condensed Matter and Complex Systems*, **38**(2), 331–338.