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

Il modello Bradley-Terry dinamico penalizzato per l'analisi di risultati calcistici

Relatrice Prof.ssa Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureando: Matteo Casagrande
Matricola N 2004210

Anno Accademico 2021/2022

Indice

Introduzione

L'applicazione della statistica nel mondo del calcio sta prendendo sempre più piede, ma a differenza di altri sport risulta molto più difficile attribuire una correlazione evidente tra prestazione e risultato essendo uno sport "a punteggio basso" e quindi in un certo senso più imprevedibile dal punto di vista dell'esito dell'incontro. Possono bastare infatti anche solamente poche occasioni per decidere una partita (*Rivista undici* 2019).
L'analisi presentata in questo lavoro mira quindi a comprendere quali siano gli aspetti che più influenzano l'esito di una gara.
In particolare l'organizzazione dell'elaborato è strutturata come segue. Nel primo capitolo viene introdotto il problema affrontato fornendo alcune motivazioni che possono portare ad intraprendere un'analisi di questo tipo. A seguire vengono descritti i dati e la metodologia con cui essi sono stati raccolti, filtrati e rielaborati in modo da poter essere utilizzabili e trattabili ai fini richiesti.
Nel secondo capitolo dell'elaborato si conduce invece un'analisi esplorativa completa, andando a comprendere quali aspetti del gioco possano essere più interessanti da analizzare a fini statistici. Un punto sicuramente cruciale riguarda quali variabili spostino maggiormente l'equilibrio di una gara e ne determinino poi l'esito. In questo capitolo pertanto si procede stimando un coefficiente per ogni variabile calcolato sull'intera stagione sportiva utilizzando il modello Bradley-Terry (Bradley e Terry 1952).
Analisi riguardanti l'importanza di alcune covariate nelle partite di calcio sono state proposte già da diversi autori (Goddard e Asimakopoulos 2004).
Sempre in questa prima parte dell'elaborato si stima inoltre un coefficiente di "abilità" diverso per ogni squadra, al netto dell'effetto delle covariate e centrato sullo zero in modo da rendere il confronto tra le squadre più agevole. Questo aspetto era già stato introdotto nella letteratura statistica (Kuk 1995), utilizzando però due parametri di abilità per ogni squadra e separando quindi per le partite in casa ed in trasferta.

A questo punto dell'analisi si avrà quindi un'idea dell'impatto che hanno

l'abilità di ogni squadra e le diverse covariate sulla probabilità di vittoria.
Nel Capitolo 3 ci si concentrerà quindi nel cogliere la componente temporale per quanto riguarda l'abilità delle squadre nel corso della stagione.
Per non tralasciare informazioni importanti inoltre è stato considerato inizialmente un trend temporale distinto per le partite giocate in casa e in trasferta.
In particolare si è scelto di utilizzare un processo EWMA per modellare l'andamento delle squadre nell'arco della stagione, permettendo così di cogliere anche un aspetto legato allo stato di forma in base ai risultati ottenuti negli incontri precedenti.
In questo modo c'è un netto calo anche del numero di parametri che descrivono le abilità delle squadre, in particolare passando dall'avere un coefficiente diverso per ognuna all'avere solamente una coppia di coefficienti che descrivono il processo.
Un altro dei vantaggi dell'utilizzare un trend temporale è dato dal riuscire a tenere conto della dipendenza presente tra le partite di una stessa squadra, che ovviamente va diminuendo con l'allontanarsi delle partite nel tempo.
Un approccio analogo era già stato adottato nella letteratura statistica (Barry e Hartigan 1993) e anche piuttosto recentemente (Cattelan, Varin e Firth 2013).
Per concludere il capitolo viene presentata anche un'altra formulazione del modello utilizzando un unico processo temporale che combini sia le partite giocate in casa che quelle giocate in trasferta.

L'obiettivo del Capitolo 4 invece è quello di ridurre il numero di covariate che influenzano la probabilità di vincere una gara e cogliere gli effetti comuni ad alcune squadre per determinate variabili esplicative.
Questi obiettivi possono essere raggiunti tramite l'utilizzo di una tecnica di penalizzazione per cogliere le variabili esplicative che maggiormente influenzano l'esito delle partite.
L'applicazione di questa metodologia inoltre permette anche di creare clusters di squadre che si comportano in modo simile rispetto ad una o più variabili, cogliendo variazioni importanti rispetto al modello iniziale che invece tiene i coefficienti delle esplicative fissati per tutte le squadre.

Un'analisi simile, anche se con poche variabili e quindi più a scopo puramente didattico, è stata proposta anche da altri autori nella letteratura statistica (Schauberg e Tutz 2019).
Nel Capitolo 5 è stato poi stimato il modello Bradley-Terry che utilizza un unico processo temporale e la penalizzazione di tipo lasso, unendo quindi le analisi svolte nei capitoli precedenti.
Infine sono state riepilogate le conclusioni dello studio e sono state proposte alcune possibili estensioni del progetto.
L'intera analisi è stata svolta con l'ausilio del software statistico R versione 4.0.4 (*R software* 2022).

1 Motivazioni e dataset

1.1 Analisi di dati sportivi

L'analisi di eventi sportivi è sempre stata oggetto di studio e di grande interesse nel mondo della statistica.
Negli ultimi anni, vista anche la crescita delle reti di scouting per le squadre professioniste e l'utilizzo della tecnologia nei campi da gioco (sensori e sistemi di tracking soprattutto), ha acquisito ancora maggior importanza la figura del Data Analyst sportivo (*Corriere comunicazioni* 2020).
Questa professione che sta prendendo sempre più rilievo svolge le sue funzioni principali nell'analisi delle partite, la valutazione delle performance e il recruitment dei giocatori.
I primi due compiti in particolare hanno come obiettivo quello di migliorare le prestazioni atletiche, tecniche e tattiche individuali e di squadra. L'ultimo aspetto invece è legato all'elaborazione dei dati al fine di individuare e valutare i giocatori di prospettiva da seguire ed eventualmente acquistare e integrare nel club.

In questo lavoro ci si concentrerà sul primo di questi due aspetti d'interesse, ossia la comprensione di cosa possa influenzare principalmente l'esito di una partita, o ancora meglio di un campionato.
Questo potrebbe essere utile anche alle società sportive per comprendere come modificare il proprio stile di gioco, su cosa investire la maggior parte del tempo e degli allenamenti e quali tattiche adoperare poi in partita.
Spesso per esempio ci si chiede nelle partite di calcio quale sia l'importanza dei calci d'angolo, il numero di tiri o il peso delle sanzioni disciplinari.
In particolare, la possibile relazione tra numero di tiri in porta e classifica del campionato è un aspetto che spesso emerge nelle analisi sportive, tuttavia ci sono sempre le dovute eccezioni del caso (*Oubliette Magazine* 2022).
Un altro fattore d'interesse riguarda sicuramente anche il monitoraggio dello stato di forma di una squadra durante il campionato, che può essere causato da fattori più evidenti come cambi allenatore o infortuni di giocatori chiave,

oppure da aspetti più difficili da analizzare come cambiamenti nei piani di gioco.

Posti gli obiettivi dell'analisi, bisogna ora definire le metodologie utilizzate.
La scelta del modello infatti non è mai una scelta banale, soprattutto quando si vuole cogliere l'andamento di un fenomeno di cui non si ha una ben chiara idea a priori.
Uno dei modelli che meglio si adatta al problema in questione è sicuramente quello proposto da Bradley e Terry in cui si utilizza il metodo dei confronti a coppie (Bradley e Terry 1952).
Questa particolare tecnica si basa sul confronto di diversi "oggetti" (nel nostro caso squadre) presentati a coppie e per il quale bisogna indicare una preferenza.
Ovviamente la scala di preferenza può avere un numero arbitrario di modalità. Per gli eventi sportivi in genere si utilizza una modalità per ogni possibile esito dell'incontro.
Nella pallacanestro per esempio si hanno solo due modalità: vittoria o sconfitta. Nel calcio invece si ha anche una terza possibilità, ovvero il pareggio. A pallavolo si hanno addirittura quattro possibili esiti poichè in base a quanti set vincono le due squadre, esse possono totalizzare da zero a tre punti.
Questa tipologia di modelli pertanto risulta adatta per analizzare eventi sportivi poichè confronta le due squadre che giocano la partita e per ognuna viene riportata la squadra vincitrice, o nel caso in cui siano ammissibili, anche i pareggi.
La variabile risposta può essere codificata in modo che distingua l'esito dell'incontro in più categorie, tenendo conto per esempio della differenza reti qualora si trattasse di una partita di calcio.
Una proposta che hanno adoperato alcuni autori riguarda infatti la distinzione di 5 possibili modalità per l'esito dell'incontro, distinguendo ulteriormente nel caso in cui la vittoria, o equivalentemente la sconfitta, fosse avvenuta con uno scarto superiore ai due gol (Schauberg e Tutz 2019).
Nel presente elaborato considereremo solamente tre modalità, tuttavia come

appena suggerito, ulteriori estensioni sono facilmente implementabili.

In questo lavoro viene analizzato un dataset relativo alle partite disputate in Premier League nella stagione sportiva 2014-2015.
La Premier League è la massima serie calcistica del campionato maschile inglese, nonchè a detta di molti esperti il miglior campionato al mondo nella stagione sportiva da noi considerata (*Colgados por el futbol* 2014). E non è un caso che tutt'ora tre delle prime quattro squadre nel ranking UEFA per club giochino proprio in Premier League (*Ranking UEFA* 2022).
Come accennato in precedenza, si possono fare diversi tipi di analisi in ambito sportivo. Lo studio svolto in questo elaborato mira a comprendere quali variabili influenzino maggiormente gli esiti delle partite lungo tutta la stagione, ed un aspetto di particolare interesse è sicuramente quello legato al giocare in casa o in trasferta.
Nella letteratura statistica infatti sono presenti diversi articoli dove si affronta quest'ultima tematica (Clarke e Norman 1995).
Questo aspetto merita quindi un'analisi approfondita ed ancora una volta il modello Bradley-Terry garantisce una trattazione piuttosto agevole del problema permettendo di tenere conto dell'ordine delle squadre nella definizione dell'incontro.

1.2 Raccolta e descrizione del dataset

I dati utilizzati, come già accennato, riguardano le partite disputate nel massimo campionato inglese di calcio noto come Premier League nella stagione sportiva 2014-2015.
I dati in questione sono stati estratti da un dataset molto più corposo che comprende tutte le partite disputate nei cinque principali campionati europei (Inghilterra, Spagna, Germania, Francia e Italia) dalla stagione 2011-2012 fino a circa metà della stagione 2016-2017.
È stata scelta la stagione 2014-2015 poichè era l'ultimo campionato per cui erano disponibili i dati interamente (infatti per quanto riguardava la stagione

successiva non erano disponibili i dati relativi ad alcune partite).
La scelta del Paese invece è ricaduta sull'Inghilterra poichè la Premier League è uno dei campionati più combattuti, giocato ad alti ritmi e in cui sono presenti molte squadre di ottimo livello.

I dati utilizzati sono stati reperiti online su Kaggle. Essa è una famosa piattaforma gratuita di data science, nata nel 2010 ed in seguito acquisita da Google nel 2017 (*Football Events* 2017).
Nel dataset disponibile per il lavoro sono contenute informazioni relative ad oltre 9000 partite, dalle quali sono stati estratti unicamente i dati relativi al campionato d'interesse per la nostra analisi creando un nuovo dataset più maneggiabile.
Il dataset ricavato da Kaggle e successivamente filtrato è composto pertanto da 34 906 righe e 22 colonne.
Esso raccoglie le informazioni relative a tutte le 380 partite giocate durante la stagione ed è strutturato in modo tale che in ogni riga venga rappresentato un diverso "evento" accaduto durante un'azione di gioco ed identificato attraverso una serie di variabili che lo descrivono.
Per evento si intende tutto ciò che può accadere durante una partita ad un livello di dettaglio sufficientemente ammissibile (effettuare un tiro, fare un gol, un assist o un fallo, ricevere un'ammonizione, guadagnare un calcio d'angolo, etc).
Il dataset è costruito in modo tale che per ognuno dei 34 906 eventi le variabili assumano valori codificati come riportato in Tabella 1.
Per alcune di esse che lo richiedono, inoltre è fornita in seguito una descrizione supplementare e più esaustiva.

Bisogna notare che quello che interessa ai fini di quest'analisi sono solamente gli eventi che riguardano le partite (le unità statistiche), non i giocatori le cui variabili saranno pertanto inutilizzate. In questo lavoro non si è interessati a chi svolge l'evento, ma solamente a quale squadra.
Le variabili appena descritte possono essere combinate per creare eventi più specifici e più informativi. Allo scopo di ottenere una migliore indagine e ri-

Variabile	**Descrizione**
id odsp	Codice identificativo della partita
id event	Codice identificativo dell'evento
sort order	Ordine cronologico degli eventi nella partita
time	Minuto della partita
text	Descrizione dell'evento
event type	Evento primario
event type2	Evento secondario
side	1 se la squadra che svolge l'evento gioca in casa, 2 altrimenti
event team	La squadra che svolge l'evento
opponent	La squadra avversaria
player	Giocatore coinvolto nell'evento principale
player2	Giocatore coinvolto nell'evento secondario
player in	Giocatore che effettua la sostituzione
player out	Giocatore che viene sostituito
shot place	Destinazione del tiro specifica
shot outcome	Destinazione del tiro generica
is goal	1 se è gol, 0 altrimenti
location	Zona del campo dove è avvenuto l'evento
bodypart	1 se colpisce col piede destro, 2 col sinistro o 3 di testa
assist method	Tipologia di assist (se il tiro è assistito)
situation	Situazione di gioco
fastbreak	1 se l'azione avviene in contropiede, 0 altrimenti

Tabella 1: Variabili nel dataset originale.

sultati più affidabili infatti è sicuramente utile analizzare anche i sottoeventi. Come sembra intuitivo, al fine di comprendere l'effetto che ha compiere un maggiore numero di tiri sull'esito della partita, non si può dare lo stesso peso ad un tiro da vicino rispetto ad uno da fuori area, o da una posizione centrale rispetto ad una molto defilata.
Per questo motivo nella scelta delle esplicative e la costruzione del dataset finale, sono state prese in considerazione variabili ottenute come combinazione di alcune delle variabili precedentemente descritte.
Per esempio si è deciso di considerare sia i tiri che i tiri in porta, utilizzando oltre alla variabile "event type" anche la variabile "shot outcome". Ulteriori distinzioni sono state fatte per esempio sulla posizione da cui è stato effettuato il tiro e con che parte del corpo.
La motivazione e la descrizione delle variabili è comunque trattata più approfonditamente nel paragrafo seguente.
Si procede quindi con la descrizione delle 22 colonne del dataset originale riportando il significato di ogni valore assunto da queste variabili. Si prosegue successivamente con una descrizione più precisa di alcune di esse.
La prima e più importante è sicuramente "event type". Essa presenta le 11 possibili modalità riportate in Tabella 2.

La modalità "Annuncio" tuttavia non si presenta mai nell'intero dataset. Alcune di queste modalità, come i "tocchi di mano", non sono comunque state utilizzate nell'analisi perchè si presentavano troppe poche volte. Infatti sono state registrate nel dataset solamente 16 infrazioni commesse per un tocco di mano, probabilmente solo le più rilevanti.
Proseguendo con le variabili successive, si ha "event type2" chiaramente collegata con la precedente. Essa descrive degli eventi più specifici che si possono verificare solo in concomitanza con alcune delle modalità descritte dalla variabile precedente come in occasione di un tiro o di un fuorigioco.
La variabile "event type2" pertanto descrive la modalità con cui è avvenuto un determinato evento così come riportato in Tabella 3.

Procedendo in ordine c'è la variabile "shot outcome". Essa definisce in

Modalità	Descrizione
0	Annuncio
1	Tiro
2	Calcio d'angolo
3	Fallo
4	Cartellino giallo
5	Secondo cartellino giallo
6	Cartellino rosso
7	Sostituzione
8	Calcio di punizione ottenuto
9	Fuorigioco
10	Tocco di mano
11	Rigore concesso

Tabella 2: Modalità della variabile "event type".

Modalità	Descrizione
12	Passaggio chiave
13	Passaggio filtrante sbagliato
14	Mandato fuori
15	Autogol

Tabella 3: Modalità della variabile "event type2".

modo generico la destinazione del tiro ed è stata di grande utilità nell'analisi. Essa è stata codificata come riportato in Tabella 4.

Modalità	**Descrizione**
1	In porta
2	Fuori
3	Bloccato
4	Sul palo

Tabella 4: Modalità della variabile "shot outcome".

La variabile "shot place" invece identifica in modo più preciso la locazione finale del tiro. Con queste informazioni infatti è stato possibile costruire variabili separate in base a quando il tiro è terminato nell'angolo alto della porta, in quello basso o centrale per esempio, permettendo una comprensione anche a livello tattico di dove possa essere più conveniente mirare per l'attaccante.
Utilizzando questa variabile inoltre è possibile per esempio comprendere se un tiro è terminato ampiamente fuori dallo specchio della porta o se ci è andato vicino. Ovviamente anche questo tipo di informazioni permette di pesare in modo diverso i tiri effettuati.
Andando più nel dettaglio, la variabile in questione è descritta in Tabella 5.

La colonna successiva del dataset identifica la variabile relativa alla zona del campo di gioco in cui si è verificato l'evento.
Quest'informazione è stata utilizzata in particolare nella creazione di variabili specifiche relative alla posizione da cui veniva effettuato il tiro.
La variabile "location" presenta 19 modalità che sono descritte in Tabella 6.

Tra le ultime variabili rimaste, ci sono quindi la variabile "assist method" utilizzata durante l'analisi per distinguere quando un tiro fosse assistito o creato magari da una giocata individuale del calciatore, oltre a comprendere l'importanza dei cross rispetto magari a giocare "palla a terra".
Tutti questi aspetti vanno poi confrontati chiaramente con la rosa a dispo-

Modalità	Descrizione
1	Di poco alta
2	Bloccata
3	Angolo basso sinistro
4	Angolo basso destro
5	Centro della porta
6	Alta e larga
7	Sul palo
8	Fuori a sinistra
9	Fuori a destra
10	Troppo alta
11	Parte alta e centrale della porta
12	Angolo alto sinistro
13	Angolo alto destro

Tabella 5: Modalità della variabile "shot place".

Modalità	Descrizione
1	Metà campo offensiva
2	Metà campo difensiva
3	Centro area
4	Ala sinistra
5	Ala destra
6	Angolo difficile e lontano
7	Angolo difficile a sinistra
8	Angolo difficile a destra
9	Lato sinistro dell'area
10	Lato sinistro dell'area piccola
11	Lato destro dell'area
12	Lato destro dell'area piccola
13	Molto vicino
14	Calcio di rigore
15	Fuori area
16	Da lontano
17	Oltre 35 yards
18	Oltre 40 yards
19	Non registrato

Tabella 6: Modalità della variabile "location".

sizione dell'allenatore cercando di sfruttare i punti di forza della propria squadra.
Nulla toglie poi che nel caso in cui una società debba licenziare il proprio allenatore durante la stagione sportiva, allora cambi anche lo stile di gioco suggerito dal nuovo arrivato.
Di seguito quindi le modalità relative alla variabile "assist method" riportate in Tabella 7.

Modalità	**Descrizione**
0	Nessuno
1	Passaggio
2	Cross
3	Passaggio di testa
4	Passaggio filtrante

Tabella 7: Modalità della variabile "assist method".

Infine la variabile "situation" permette di distinguere da che tipo di situazione è scaturito l'evento. Nell'analisi svolta è stata utilizzata in particolare per cogliere l'effetto dei tiri diretti su calcio di punizione.
La variabile può assumere quindi quattro valori codificati come riportato in Tabella 8.

Modalità	**Descrizione**
1	Azione di gioco
2	Schema di gioco
3	Corner
4	Calcio di punizione

Tabella 8: Modalità della variabile "situation".

1.3 Dataset finale

Come spiegato in precedenza, dalle variabili presenti nel dataset scaricato da Kaggle ne sono state ricavate altre più dettagliate.
In questo modo è stato elaborato un dataset in cui in ogni riga sono presenti le unità statistiche, ossia ogni riga rappresenta una diversa partita, mentre nelle colonne abbiamo alcune informazioni relative alla partita oltre a tutte le variabili create precedentemente e separate per la squadra che gioca in casa e quella che gioca in trasferta.
Il dataset così creato pertanto ha 380 righe e 47 colonne.
Per convenzione le partite sono state ordinate in base alla data partendo dalla prima giornata disputata il 16/08/2014 fino all'ultima giocata il 24/05/2015.
Le prime 5 colonne del nuovo dataset creato riguardano rispettivamente:

- Data in cui è stata giocata la partita;
- Squadra che ha giocato in casa;
- Squadra che ha giocato in trasferta;
- Gol fatti dalla squadra in casa;
- Gol fatti dalla squadra in trasferta.

Il campionato di Premier League infatti è strutturato in modo tale che ogni squadra giochi contro tutte le squadre rispettivamente una volta in casa e una volta in trasferta.
Non vi sono mai pertanto partite giocate in campo neutro, come invece può capitare nelle finali delle coppe o nei tornei organizzati per nazioni come mondiali o europei in cui c'è sempre uno (o più) Stato ospitante.
Le variabili esplicative che sono state ritenute d'interesse per l'analisi, e che corrispondono quindi alle successive 42 colonne del dataset, sono state separate distinguendo in base a quando l'evento è stato svolto dalla squadra in casa e quando invece è stato svolto dalla squadra in trasferta.
Esse sono in ordine:

- Tiri;
- Tiri in porta;
- Pali colpiti;
- Tiri di testa;
- Tiri da fuori area;
- Tiri assistiti;
- Passaggi filtranti chiave;
- Calci d'angolo;
- Falli commessi;
- Ammonizioni ricevute;
- Fuorigioco commessi (intesa come abilità della difesa di mandare i giocatori avversari in fuorigioco);
- Tiri da centro area;
- Tiri dalla parte sinistra dell'area;
- Tiri dalla parte destra dell'area;
- Tiri su punizione;
- Tiri da posizioni difficili;
- Cross effettuati;
- Ammonizioni ricevute prima del 45° minuto di gioco;
- Tiri bloccati (intesa come abilità del difensore di fermare i tiri prima che arrivino a destinazione);
- Tiri negli angoli alti della porta;

- Tiri negli angoli bassi della porta.

La maggior parte delle variabili si riferisce a varie sfaccettature dei tiri effettuati in base alle modalità con cui sono avvenuti, la parte del corpo, la posizione o la destinazione. Questo perchè chiaramente per vincere una partita, l'unico modo è tirare e pertanto acquisisce particolare importanza il modo in cui sia più efficace farlo. Ne segue quindi che esso sia l'aspetto di maggiore interesse.
Tuttavia, come già accennato, è opportuno fare distinzioni e considerare le varie sottocategorie precedentemente definite.

L'importanza di analizzare in una partita il numero tiri e di tiri in porta o che colpiscono i pali della porta, è quindi piuttosto evidente. Altri aspetti invece possono sembrare più criptici e pertanto vanno spiegati più accuratamente.
La possibilità di analizzare la posizione da dove viene effettuato il tiro permette di dare un peso potenzialmente differente ai tiri che provengono da posizioni centrali rispetto ai tiri da posizioni laterali e quindi più facilmente parabili dal portiere. Discorso simile vale anche per i tiri effettuati in prossimità dell'area rispetto ai tiri effettuati da posizioni più lontane dallo specchio della porta.
I tiri su punizione sono un altro aspetto interessante poichè mi aspetto che possa presentare una discreta variabilità da squadra a squadra vista la presenza di alcuni "specialisti" in determinate formazioni.
La destinazione dei tiri in porta può essere un altro aspetto d'interesse soprattutto nella preparazione e nell'allenamento dei giocatori poichè potrebbe essere che mirare agli angoli alti per esempio sia più determinante rispetto a mirare agli angoli bassi della porta nell'esito di un tiro.
Il numero di tiri di testa insieme al numero di cross, o al contrario il numero di passaggi filtranti, possono identificare due diversi stili di gioco e pertanto potrebbe essere utile comprendere quale dei due può essere più efficace ai fini pratici del risultato.
Fortemente collegato al numero di colpi di testa c'è ovviamente anche il nu-

mero di calci d'angolo ottenuti solitamente da un maggiore impegno della difesa e del portiere avversari. Anche questo aspetto può essere d'interesse perchè in genere va a premiare anche le squadre più fisiche e organizzate nello sfruttare gli schemi preparati in allenamento.
Un ultimo aspetto legato all'ambito dei tiri è quello del numero delle conclusioni assistite. Potrebbe essere per esempio che alcune squadre si affidino maggiormente alle capacità individuali dei propri giocatori, mentre altre creino più occasioni attraverso i passaggi.
Un altro aspetto d'interesse nell'analisi è sicuramente quello legato all'aggressività delle squadre e quindi il numero di falli commessi e di ammonizioni ricevute. Non tutte le sanzioni però pesano allo stesso modo in una gara e pertanto è stata costruita un'ulteriore variabile relativa alle ammonizioni ricevute prima del 45° minuto e che quindi limiteranno i giocatori che le hanno ricevute almeno per un tempo di gioco.
L'esito di una gara tuttavia non è determinato unicamente dalle abilità offensive, ma anche da quelle difensive che tuttavia sono più complicate da misurare.
Per questo motivo sono state considerate anche le capacità di mandare i giocatori avversari in fuorigioco e la capacità di bloccare i tiri avversari prima che arrivino in porta. La prima in particolare potrebbe essere interessante da analizzare per comprendere se questo stile di gioco e questa impostazione difensiva possano essere efficaci nell'arco di una partita nonostante i rischi che comporta.

Per concludere la descrizione del dataset utilizzato viene proposta una breve analisi esplorativa per le variabili appena descritte e utilizzate nel prosieguo dell'analisi.
In Tabella 9 sono quindi riportati i valori di alcune statistiche descrittive per quanto riguarda il numero di gol a partita e le variabili esplicative descritte precedentemente, separate in base al fatto che la squadra abbia giocato in casa o in trasferta.

Variabile	Media	Dev std
Gol casa	1.474	1.263
Gol trasferta	1.092	1.070
Tiri casa	14.592	5.533
Tiri trasferta	11.282	4.576
Tiri in porta casa	4.437	2.576
Tiri in porta trasferta	3.497	2.076
Pali colpiti casa	0.216	0.460
Pali colpiti trasferta	0.182	0.412
Tiri di testa casa	2.292	1.639
Tiri di testa trasferta	1.608	1.367
Tiri dal limite casa	5.861	3.173
Tiri dal limite trasferta	4.871	2.677
Tiri assistiti casa	10.800	4.406
Tiri assistiti trasferta	8.463	3.893
Passaggi filtranti casa	0.279	0.525
Passaggi filtranti trasferta	0.268	0.515
Corner casa	5.947	3.111
Corner trasferta	4.608	2.515
Falli casa	10.437	3.171
Falli trasferta	10.600	3.400
Ammonizioni casa	1.658	1.228
Ammonizioni trasferta	2.084	1.389
Fuorigioco casa	1.887	1.573
Fuorigioco trasferta	1.979	1.618
Tiri dal centro casa	4.855	2.572
Tiri dal centro trasferta	3.429	2.169
Tiri da sinistra casa	1.179	1.218
Tiri da sinistra trasferta	0.976	1.023
Tiri da destra casa	1.095	1.192
Tiri da destra trasferta	0.813	0.985

Continua nella pagina successiva

Continua dalla pagina precedente

Variabile	**Media**	**Dev std**
Tiri su punizione casa	0.276	0.514
Tiri su punizione trasferta	0.239	0.463
Tiri da posizioni difficili casa	0.332	0.591
Tiri da posizioni difficili trasferta	0.213	0.476
Cross casa	2.829	1.931
Cross trasferta	1.895	1.483
Ammonizioni primo tempo casa	1.037	0.998
Ammonizioni primo tempo trasferta	1.326	1.111
Tiri bloccati casa	3.032	2.216
Tiri bloccati trasferta	4.068	2.629
Tiri negli angoli alti casa	0.692	0.909
Tiri negli angoli alti trasferta	0.547	0.693
Tiri negli angoli bassi casa	2.174	1.733
Tiri negli angoli bassi trasferta	1.639	1.445

Tabella 9: Media e deviazione standard per il numero di gol e per ogni variabile esplicativa, separate per la squadra che ha giocato in casa o in trasferta.

Ogni partita in media vengono segnati 2.566 gol, ma chiaramente tale ammontare può variare notevolmente da una partita ad un'altra. L'incontro con il maggior numero di gol nella stagione sportiva considerata è stato disputato il 30/08/2014 tra le squadre Everton e Chelsea con i secondi che hanno vinto fuori casa per 6 a 3. Essa è stata anche la partita con il maggior numero di gol per una squadra che ha giocato in trasferta nel campionato 2014-2015. La partita con il maggior numero di gol per la squadra in casa è stata invece Southampton - Sunderland disputata il 18/10/2014 e terminata 8 a 0.

La distribuzione del numero di gol a partita separato per la squadra che ha

giocato in casa e la squadra che ha giocato in trasferta, è riportata in Figura 1.

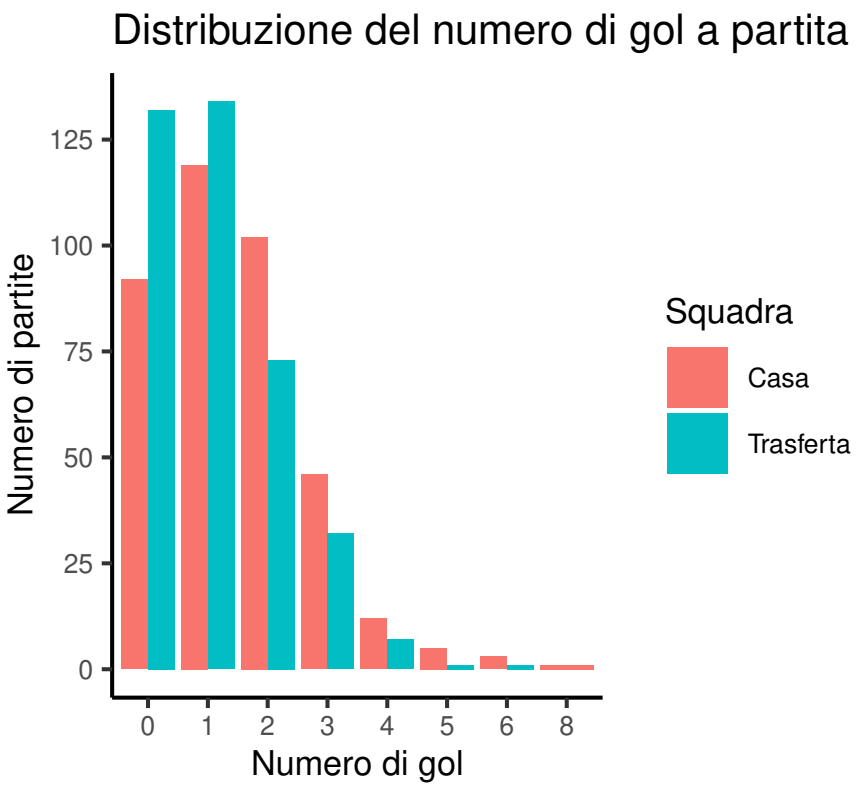

Figura 1: Distribuzione del numero di gol a partita separato per squadra in casa e squadra in trasferta.

Così come per il numero di gol a partita, anche la maggior parte delle variabili esplicative utilizzate presentano in media valori più alti per la squadra in casa rispetto alla squadra in trasferta, come si evince in Tabella 9. Fanno eccezione infatti solamente le covariate riferite al numero di falli e le conseguenti sanzioni, oltre al numero di tiri bloccati e il numero di volte in cui si è mandato un giocatore avversario in fuorigioco.
Tranne per le variabili appena elencate inoltre, i valori assunti dalle covariate per la squadra in casa sembrano variare maggiormente rispetto a quanto

accade per la squadra fuori casa.

Bisogna anche notare che molte variabili esplicative sono correlate tra loro. Per esempio sembra logico aspettarsi che il numero di falli ed il numero di ammonizioni ricevute abbiano una relazione positiva. Discorso simile vale anche per esempio per la relazione che esiste tra il numero di conclusioni di testa ed il numero di calci d'angolo battuti.
Da una prima analisi esplorativa si possono anche già trarre informazioni sulla correlazione presente non solo tra le variabili esplicative, ma anche tra la variabile risposta e le variabili indipendenti.
In Figura 2 per esempio viene messa in evidenza l'associazione positiva tra la differenza tra il numero di tiri in porta effettuati e il numero di tiri in porta subiti rispetto all'esito dell'incontro.
In particolare si nota come differenze maggiori portino in genere ad esiti migliori dell'incontro. Al contrario i valori maggiormente negativi sono tutti associati a sconfitte subite.

1.4 L'importanza di giocare in casa

Prima di introdurre il modello utilizzato è bene comprendere la struttura dei dati che sono stati raccolti.
La Premier League è un campionato composto da 20 squadre che si affrontano ognuna due volte per campionato (una in casa e una in trasferta) per un totale di 380 partite disputate nell'arco dell'intera stagione sportiva, da cui deriva il numero di righe di cui è composto il dataset.
Ad ogni partita la squadra vincente guadagna 3 punti, la squadra sconfitta ne guadagna 0, mentre in caso di pareggio la posta in palio viene divisa equamente garantendo un punto ad entrambe le compagini in gara.
Come facilmente intuibile inoltre la maggior parte delle vittorie sono per la squadra che gioca in casa: 45,3 % mentre solamente il 30,2 % per la squadra che gioca fuori casa. Il restante 24,5% di partite ovviamente si sono concluse con un pareggio.

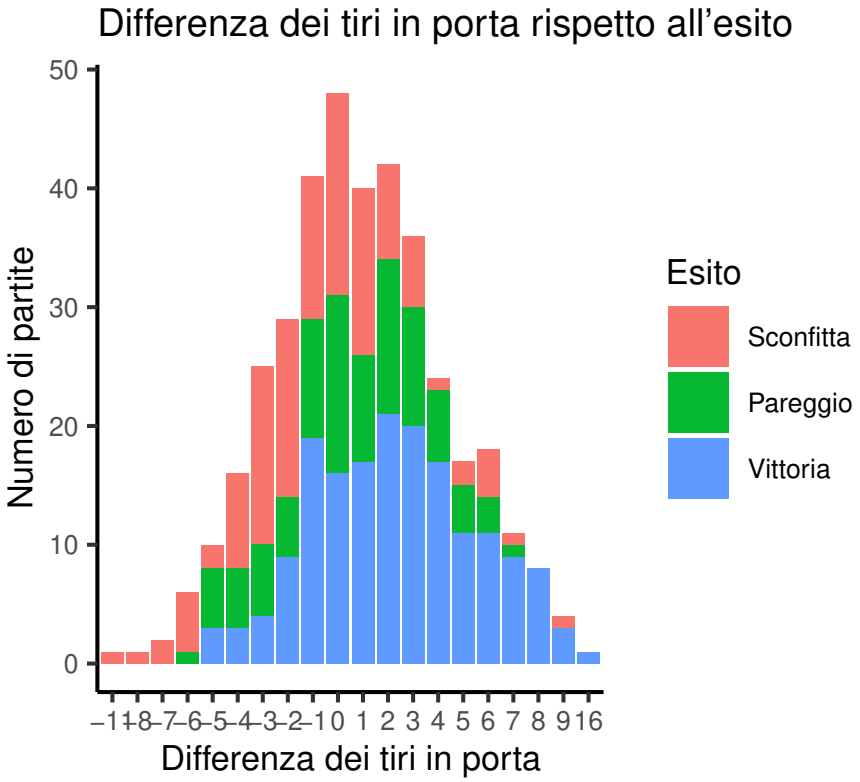

Figura 2: Differenza tra tiri in porta effettuati e subiti rispetto al risultato.

Le vittorie per la squadra in casa sono circa una volta e mezzo quelle delle squadre fuori casa. Questo è sicuramente un risultato tutt'altro che banale, di cui va tenuto conto nella definizione del modello e che va analizzato per capire se ci sia un fattore di "favoritismo" per chi gioca in casa costante per ogni squadra o se ci possano essere differenze anche significative tra le varie squadre.
Da una prima analisi esplorativa si può già notare che alcune squadre sembrerebbero risentire più di altre di questo aspetto, come emerge in Tabella 10. Una squadra addirittura (Crystal Palace) ha vinto più partite in trasferta che non in casa.
La percentuale di vittorie ottenute in casa sul totale delle vittorie varia infatti da un minimo di 46.2% ad un massimo del 75% per QPR e West Ham. Da notare anche la presenza del Manchester United tra le prime posizioni, nonostante al termine del campionato risulti nella parte alta della classifica e per cui ci si aspetterebbe che mantenesse buoni risultati anche quando gioca lontano da casa.

Di queste differenze se ne terrà conto soprattutto nei Capitoli 3 e 4.
Nel modello iniziale proposto nel capitolo seguente invece si stimerà un parametro relativo al fattore campo calcolato sui dati dell'intero campionato e costante per ogni squadra.

Squadre	**% vittorie in casa**
QPR	75.0%
West Ham	75.0%
Manchester United	70.0%
Newcastle	70.0%
Stoke City	66.7%
Leicester City	63.6%
West Brom	63.6%
Hull City	62.5%
Southampton	61.1%
Everton	58.3%
Manchester City	58.3%
Chelsea	57.7%
Burnley	57.1%
Sunderland	57.1%
Swansea	56.2%
Liverpool	55.6%
Arsenal	54.5%
Tottenham	52.6%
Aston Villa	50.0%
Crystal Palace	46.2%

Tabella 10: Percentuale di vittorie ottenute in casa per ogni squadra.

2 Modello Bradley-Terry

2.1 Introduzione al modello Bradley-Terry

Il modello Bradley-Terry è stato sviluppato per analizzare i risultati di confronti a coppie per le quali bisogna esprimere una preferenza.
Questa metodologia si basa sul fatto che colui che svolge l'analisi abbia a disposizione una serie di oggetti e che voglia ordinarli sulla base di un qualche criterio.
La problematica che affronta tale modello riguarda l'impossibilità, o comunque l'incertezza, nel disporre gli oggetti se confrontati tutti contemporaneamente. Pertanto la procedura è quella di paragonare gli oggetti a coppie e determinare di volta in volta se un oggetto sia preferibile all'altro o se siano sullo stesso livello.
Nel nostro caso gli "oggetti" sono le 20 squadre che disputano il campionato. Il numero di confronti fatti è quindi il numero di partite disputate, ossia 380. Stabilire una preferenza per una delle due squadre che giocano l'incontro significa pertanto determinare l'esito della partita.
In formule questo implica di stabilire la probabilità che la variabile risposta, che indica il confronto (risultato dell'incontro), assuma un determinato valore.
Nel caso specifico delle partite di calcio possiamo avere quindi tre esiti che possiamo assumere come modalità ordinate.

Sia Y_i la variabile risposta relativa all'i-esima partita per $i = 1, \ldots, 380$, con le sue tre modalità: 0 se la squadra che ha vinto è la squadra che ha giocato in trasferta, 1 se la partita è terminata in pareggio e 2 se la squadra in casa ha vinto.
A questo punto la scelta sulla funzione di legame utilizzata deve tenere conto sia della presenza di tre possibili modalità sia del fatto che sono ordinate.
Rifacendosi sempre alla letteratura statistica disponibile (Koning 2000), la scelta è ricaduta su un modello logit cumulato per il quale è possibile scrivere la probabilità come:

$$pr(Y_i \leq y_i) = \frac{\exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i})}{1 + \exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i})}, \tag{2.1}$$

con $y_i \in \{0, 1, 2\}$.

I parametri δ_{y_i} sono tre coefficienti (δ_0, δ_1 e δ_2) chiamati anche "punti di soglia", il parametro η rappresenta l'effetto del giocare in casa, mentre a_k rappresenta l'abilità della k-esima squadra. Più nello specifico a_{h_i} indica l'abilità della squadra che gioca in casa, mentre a_{v_i} indica l'abilità della squadra che gioca in trasferta.
I "punti di soglia" sono i valori ottimali in cui dividere le unità statistiche e che vengono calcolati massimizzando la verosimiglianza partizionata rispetto a tutti i possibili valori.
La verosimiglianza partizionata in particolare è una quantità data dalla somma delle verosimiglianze calcolate per le osservazioni prima e dopo i punti di soglia (Hugall e Lee 2003).
Un aspetto fondamentale di questi parametri è che devono avere valori perfettamente simmetrici al fine di garantire che se due squadre ugualmente forti giocassero in un campo neutro allora avrebbero la stessa probabilità di vincere la partita.
Pertanto nel modello viene stimato un unico valore del parametro δ_1 e si fissa $\delta_2 = -\delta_1$, mentre δ_0 è sempre pari a $-\infty$.
Il parametro η in questo modello viene stimato costante per tutte le squadre e calcolato sui dati dell'intero campionato. Nei successivi capitoli verranno considerate diverse opzioni per cogliere meglio tale effetto che sembra variare tra le squadre e nel corso della stagione.
Inoltre ai parametri a_{h_i} e a_{v_i}, relativi rispettivamente alle stime per l'abilità della squadra in casa e la squadra in trasferta, viene imposto un vincolo per l'identificabilità, garantendo che $\sum_{k=1}^{n} a_k = 0$.

Nel dataset considerato, la stima dei coefficienti relativi ai "punti di soglia" risulta pertanto $\hat{\delta_1}$ = -0.491 e di conseguenza $\hat{\delta_2}$ = 0.491.

Il parametro η invece viene stimato pari a $\hat{\eta} = 0.390$ con standard error pari a 0.103, costante per tutte e 20 le squadre.
In questa prima formulazione del modello bisogna notare che non entrano in gioco le variabili esplicative. Le probabilità associate ai tre possibili esiti dell'incontro sono pertanto ottenute unicamente sulla base del "fattore campo" e delle capacità delle squadre.
In Tabella 11 sono infine riportate le stime delle abilità a_k per ogni squadra. Oltre agli standard error delle stime, sono stati calcolati anche i rispettivi Quasi-standard error (QSE), ossia un metodo parsimonioso e intuitivo che permette di fare inferenza in modo approssimato per i confronti in termini di abilità tra diverse squadre (Firth e Menezes 2004).
Più nello specifico, i QSE sono un metodo che approssima gli standard error, particolarmente utile nel caso in cui si abbiano variabili dummy, come in questo caso, e si voglia fare inferenza su differenze tra diversi livelli.
Il problema in questi casi è che bisognerebbe riportare tutta la matrice di varianze e covarianze dello stimatore. I Quasi-standard errorssono quindi un'approssimazione che permette di fare inferenza come se fossero due variabili indipendenti.
Le stime dei coefficienti per le abilità delle singole squadre sono quindi riportate in ordine in Tabella 11, affiancate dai relativi standard error (SE), Quasi-standard error (QSE) e le Quasi-varianze (QV) con i quali si possono fare confronti più accurati.

Le squadre che presentano le stime maggiori in termini di abilità risultano essere in ordine: Chelsea, Manchester City, Arsenal, Manchester United e Tottenham.
A supporto del modello e del buon lavoro svolto finora, basta guardare la classifica al termine della stagione e si noterà come le prime cinque posizioni in classifica saranno esattamente queste (*Guerin sportivo* 2015).
Utilizzando l'approssimazione fornita dai QSE si può facilmente determinare se le differenze in termini di abilità tra due squadre sono statisticamente significative.
Se si confrontano per esempio le stime dei coefficienti di abilità delle prime

Squadre	**Abilità**	**SE**	**QSE**	**QV**
Chelsea	1.529	0.464	0.350	0.123
Manchester City	1.140	0.460	0.341	0.116
Arsenal	0.940	0.000	0.321	0.103
Manchester United	0.748	0.443	0.313	0.098
Tottenham	0.462	0.446	0.317	0.101
Liverpool	0.358	0.443	0.315	0.099
Southampton	0.262	0.451	0.317	0.101
Swansea	0.025	0.449	0.308	0.095
Stoke City	0.022	0.446	0.307	0.094
West Ham	-0.180	0.440	0.298	0.089
Everton	-0.182	0.437	0.299	0.090
Crystal Palace	-0.221	0.449	0.313	0.098
West Brom	-0.307	0.444	0.304	0.093
Sunderland	-0.353	0.432	0.287	0.082
Leicester City	-0.527	0.448	0.314	0.098
Newcastle	-0.623	0.450	0.310	0.096
Hull City	-0.624	0.447	0.308	0.095
Aston Villa	-0.697	0.455	0.318	0.101
Burnley	-0.722	0.449	0.308	0.095
QPR	-1.051	0.466	0.332	0.110

Tabella 11: Coefficienti di abilità stimati per ogni squadra coi relativi standard error (SE), Quasi-standard error (QSE) e Quasi-varianze (QV).

due squadre in classifica si ha che i relativi QSE sono rispettivamente pari a 0.350 per il Chelsea e 0.341 per il Manchester City, mentre la differenza tra le due stime di abilità (in valore assoluto) è pari a $1.529 - 1.140 = 0.389$.
Utilizzando quindi quest'approssimazione si può calcolare lo standard error relativo alla differenza tra i due coefficienti che risulta pari a $(0.350^2 + 0.341^2)^{1/2} = 0.489 > 0.389$ e pertanto il divario in termini di abilità tra le due squadre è non significativo da un punto di vista statistico, nonostante le due squadre abbiano finito il campionato con 8 punti di distacco (rispettivamente 87 e 79).

2.2 Modello Bradley-Terry con variabili esplicative

Nel modello descritto in precedenza chiaramente manca un apporto importante dovuto alle variabili esplicative raccolte durante le partite.
Aggiungendo pertanto le covariate al modello, la Formula (2.1) utilizzata in precedenza viene opportunamente modificata nel seguente modo:

$$pr(Y_i \leq y_i) = \frac{\exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i} + \sum_{j=1}^{p} \gamma_j x_{i,j})}{1 + \exp(\delta_{y_i} + \eta + a_{h_i} - a_{v_i} + \sum_{j=1}^{p} \gamma_j x_{i,j})}, \tag{2.2}$$

con $y_i \in \{0, 1, 2\}$.

Mantenendo valide le definizioni riguardanti i coefficienti visti finora, gli ultimi parametri di cui approfondire il significato sono pertanto i γ_j associati agli $x_{i,j}$.
Questi ultimi sono le differenze tra i valori che assumono le $p = 21$ covariate per le squadre che giocano in casa e in trasferta in ognuna delle 380 partite di campionato.
Per esempio nella prima partita della stagione la squadra di casa (West Bromwich) ha effettuato 10 tiri, mentre la squadra fuori casa (Sunderland) ne ha

effettuati 7. Essendo la variabile relativa ai tiri la prima nel dataset, si ha che il valore di $x_{1,1} = 3$.
Di conseguenza i γ_j rappresentano i valori dei coefficienti associati mantenuti costanti per tutte le squadre e stimati sui dati dell'intero campionato.
L'ordine delle squadre per ogni partita infatti è fissato in modo tale che la prima squadra elencata è sempre la squadra che gioca in casa.

Con l'aggiunta delle covariate al modello chiaramente sono cambiate anche le stime di tutti gli altri parametri.
In particolare, mantenendo valido il discorso sulla simmetria dei parametri di soglia, si ha che la stima ora è pari a $\hat{\delta_1}$ = -0.154.
Una modifica interessante l'ha subita anche il parametro relativo al fattore campo. In particolare esso è aumentato ed è ora pari a $\hat{\eta} = 0.442$ con standard error pari a 0.139, dando ancora maggiore importanza al fatto di giocare in casa, al netto chiaramente di tutti gli altri fattori.
In Tabella 12 sono infine riportate le stime delle abilità delle squadre al netto dell'effetto delle covariate x_j raccolte durante le partite e le stime dei coefficienti γ_j associati alle x_j con i relativi standard error (SE), Quasi-standard error (QSE) e Quasi-varianze (QV).
Le stime relative alle variabili esplicative sono state riportate in ordine decrescente al fine di evidenziare in modo più immediato quali siano le variabili che influenzano maggiormente l'esito di una partita da un lato come dall'altro.
Si notano subito infatti alcune variabili con un effetto decisamente più forte sulla variabile risposta, mentre altre hanno un effetto quasi nullo e non significativo. In questo punto del lavoro non è comunque d'interesse fare selezione delle variabili, pertanto verranno mantenuti tutti i coefficienti stimati nel modello.

Variabile/Squadra	Stima	SE	QSE	QV
Chelsea	1.344	0.553	0.497	0.247
Manchester City	1.263	0.576	0.527	0.278
Arsenal	0.983	0.000	0.009	0.000
Manchester United	0.897	0.515	0.459	0.210
Tottenham	0.848	0.555	0.498	0.248
Stoke City	0.172	0.558	0.491	0.241
Liverpool	0.088	0.548	0.491	0.241
West Ham	0.087	0.567	0.497	0.247
Southampton	-0.073	0.558	0.494	0.244
Swansea	-0.077	0.563	0.493	0.243
Newcastle	-0.292	0.538	0.478	0.229
Sunderland	-0.319	0.570	0.496	0.246
West Brom	-0.486	0.557	0.488	0.238
Leicester City	-0.490	0.563	0.497	0.247
Burnley	-0.521	0.564	0.495	0.245
Crystal Palace	-0.563	0.562	0.489	0.239
QPR	-0.674	0.597	0.526	0.276
Hull City	-0.682	0.575	0.504	0.254
Everton	-0.811	0.544	0.479	0.229
Aston Villa	-1.183	0.576	0.506	0.256
Tiri angoli alti	0.585	0.170	-	-
Tiri	0.489	0.097	-	-
Tiri angoli bassi	0.394	0.096	-	-
Tiri Bloccati	0.137	0.050	-	-
Ammonizioni	0.115	0.120	-	-
Assist	0.068	0.055	-	-
Tiri in porta	0.022	0.070	-	-
Tiri punizione	-0.001	0.198	-	-
Falli	-0.036	0.032	-	-
Tiri testa	-0.073	0.100	-	-

Continua nella pagina successiva

Continua dalla pagina precedente

Variabile/Squadra	**Stima**	**SE**	**QSE**	**QV**
Offside	-0.120	0.062	-	-
Ammonizioni primo tempo	-0.134	0.135	-	-
Corner	-0.136	0.037	-	-
Pali colpiti	-0.144	0.202	-	-
Cross	-0.198	0.094	-	-
Tiri da destra	-0.388	0.134	-	-
Tiri dal centro	-0.425	0.097	-	-
Tiri da sinistra	-0.508	0.125	-	-
Tiri dal limite	-0.551	0.107	-	-
Tiri posizioni difficili	-0.575	0.190	-	-
Passaggi filtranti	-0.583	0.239	-	-

Tabella 12: Coefficienti stimati con il modello Bradley-Terry con i relativi standard error (SE), quasi-standard error (QSE) e quasi-varianze (QV).

Il numero di tiri ovviamente è fortemente correlato con la probabilità di vittoria della partita come ci si aspettava.
Mirare negli angoli alti della porta inoltre sembra essere più producente rispetto a tirare negli angoli bassi, anche se chiaramente entrambe hanno un effetto positivo perchè probabilmente la maggior parte dei gol è arrivato tirando in quelle zone, essendo i tiri centrali più facilmente parabili.
Altro aspetto interessante è legato al gioco di squadra ed in particolare al coefficiente associato al numero di tiri assisti. Esso presenta un debole effetto positivo sulla variabile risposta.
Può stupire inoltre il fatto che i tiri in porta e il numero di pali colpiti abbiano un coefficiente così modesto e chiaramente non significativo al netto delle altre esplicative.
Gli standard error infatti sono maggiori delle stime dei coefficienti in valore

assoluto e pertanto non si può dire che il coefficiente sia significativamente diverso da zero.
Sembrerebbe quindi che l'effetto dovuto al numero di tiri in porta sulla variabile risposta venga colto principalmente dai coefficienti relativi ai tiri mirati negli angoli.
Un aspetto difensivo importante ai fini della gara pare essere il numero di tiri bloccati, con coefficiente chiaramente positivo.
Le variabili relative all'aggressività di una squadra mostrano come il numero di falli e di ammonizioni nel primo tempo abbiano una stima leggermente negativa al netto delle altre variabili. Tuttavia le ammonizioni ricevute in tutta la partita hanno una stima positiva, anche se inferiore in valore assoluto a quella attribuita alle ammonizioni nel primo tempo.
Anche il coefficiente associato al numero di tiri di testa ha un valore leggermente negativo e non significativo da un punto di vista statistico. I tiri effettuati di testa infatti probabilmente arrivano da zone molto vicine alla porta, tuttavia in genere sono più deboli e imprecisi.
Associato al coefficiente relativo ai colpi di testa sicuramente ci sono anche quelli relativi ai calci d'angolo e ai cross, entrambi nettamente negativi.
Il valore del coefficiente associato ai tiri effettuati direttamente da calcio di punizione è praticamente nullo al netto delle altre covariate, mentre il coefficiente associato alla capacità delle squadre di mandare gli avversari in fuorigioco è chiaramente negativo.
Le ultime variabili con effetto negativo e con valori decisamente importanti sono invece quelle legate alla posizione da cui si effettua il tiro.
Esse presentano valori comunque maggiori se confrontati con i successivi coefficienti relativi alla distanza dalla porta e l'angolazione da cui viene effettuato il tiro.
Qualsiasi posizione di tiro che non sia molto vicina alla porta ha un effetto negativo sulla risposta. In particolare i tiri da fuori area, di cui spesso si abusa, non portano quasi mai ai risultati sperati così come i tiri da posizioni e angolazioni complicate.
I passaggi filtranti inoltre presentano il coefficiente maggiormente negativo al netto dell'effetto delle altre esplicative, segno di come queste giocate la

maggior parte delle volte non vengano sfruttate o comunque non siano la strategia migliore per vincere le partite.
Il coefficiente associato tuttavia ha una variabilità considerevole. Infatti è la stima con lo standard error più elevato tra i coefficienti relativi alle variabili esplicative. Potrà essere d'interesse nel prosieguo dell'analisi, ed in particolare nel Capitolo 4, comprendere se tale variabilità possa essere associata alle diverse squadre che giocano l'incontro.

Osservando i coefficienti riportati in Tabella 12 si può quindi determinare l'effetto marginale che avrebbe la variazione di un regressore sulla variabile risposta. Si consideri ora per esempio l'effetto marginale dovuto all'effettuare un tiro in più.
In particolare se ci si pone nella condizione in cui le due squadre abbiano la stessa "abilità", si ottiene che la probabilità di vittoria per la squadra in casa è pari al $57,1\%$. Se tale formazione effettuasse un tiro da centro area più degli avversari, la sua probabilità di vittoria aumenterebbe dell'$1,6\%$. Se poi tale tiro per esempio fosse diretto nell'angolino alto della porta, la probabilità di vittoria arriverebbe addirittura al $72,3\%$. Ovviamente tale percentuale diminuirebbe se per esempio il tiro fosse effettuato da fuori area o di testa.
Tuttavia il numero di tiri negli angoli alti per partita è molto basso come si è visto in Tabella 9. Esso infatti è pari a 0.692 per la squadra che gioca in casa e 0.547 per la squadra ospite.
Prendendo ora un esempio pratico dal dataset, si osserva la prima partita della stagione tra West Bromwich (in casa) e Sunderland (in trasferta). Si ottiene che nonostante i padroni di casa abbiano un coefficiente di abilità inferiore, grazie al fattore campo hanno una probabilità di vittoria del 53% secondo le stime del modello qualora considerassi i valori delle esplicative tutti nulli.
Se si volesse anche in questo caso calcolare l'effetto che avrebbe effettuare un tiro in più da centro area diretto nell'angolino alto per i padroni di casa, al netto anche dei valori delle abilità delle squadre e del fattore campo, l'effetto marginale sarebbe pari ad un aumento del $15,8\%$ della probabilità di

vittoria, quindi leggermente superiore a quanto visto per due squadre con la stessa abilità.

Per comprendere se il modello Bradley-Terry con l'aggiunta delle variabili esplicative possa avere un adattamento migliore per gli esiti delle partite dell'intero campionato rispetto al modello stimato senza covariate, si può utilizzare un indicatore noto come Ranked Probability Score (RPS).
Esso è molto utilizzato quando si hanno variabili risposta ordinali, come in questo caso (Czado, Gneiting e Held 2009).
Il RPS misura quanto le previsioni, espresse come una distribuzione di probabilità, si avvicinano ai valori osservati.
Definito K come il numero di modalità della variabile risposta y, la funzione RPS può essere descritta come:

$$RPS(y, \hat{\pi}(k)) = \sum_{k=1}^{K} (\hat{\pi}(k) - I(y \leq k))^2, \tag{2.3}$$

dove $I()$ indica la funzione indicatrice pari a 1 se l'argomento all'interno delle parentesi è vero e 0 altrimenti, mentre $\hat{\pi}(k)$ rappresenta la probabilità cumulata tale che $\hat{\pi}(k) = pr(y \leq k)$ stimata.
Pertanto il valore calcolato per il RPS è sempre positivo e più esso si avvicina allo zero, migliore sarà la bontà di adattamento per il modello utilizzato.
Il modello Bradley-Terry senza l'utilizzo delle covariate ha una stima del RPS pari a 0.388, mentre il modello che utilizza le informazioni ottenute dalle variabili esplicative durante le partite ha un valore inferiore e pari a 0.339.
Le covariate inserite nel modello pertanto sembrano migliorare la capacità previsiva per il dataset in questione.

Anche in questo caso inoltre si possono confrontare diverse squadre in termini di abilità utilizzando i relativi QSE come fatto anche nel paragrafo precedente.
Se si confrontano anche in questo caso le prime due squadre per compren-

dere se vi sia o meno una differenza in termini di abilità si possono ripetere i calcoli fatti in precedenza e si ottiene che la differenza tra i coefficienti di Chelsea e Manchester City è pari a $1.344 - 1.263 = 0.081$ che va confrontato con $(0.497^2 + 0.527^2)^{1/2} = 0.525$.
Pertanto la differenza di abilità tra le due squadre sembra essere ampiamente non significativa una volta considerato l'effetto delle covariate.

Il modello appena trattato viene definito "statico" nel senso che considera un parametro fisso di abilità per ogni squadra, costante per l'intera stagione. Nel capitolo seguente si vuole ipotizzare invece un approccio dinamico, permettendo a tali parametri di "aggiustarsi" nel corso del campionato seguendo un particolare processo.

3 Modello Bradley-Terry dinamico

3.1 Variazione temporale dell'abilità delle squadre

L'utilizzo di un modello dinamico punta a risolvere tre problematiche emerse finora e che verranno presentate man mano.
Il primo aspetto riguarda i valori delle abilità delle squadre che chiaramente non sono costanti lungo l'intera stagione.
La visione proposta pocanzi coi parametri a_k misurati sull'intero campionato rappresenta una versione semplicistica di quanto accade durante l'intera stagione sportiva che nella fattispecie è durata dal 16/08/2014 al 24/05/2015 quindi oltre 9 mesi, nei quali possono essere accaduti vari episodi.
Non è difficile infatti ipotizzare acquisti o cessioni di giocatori importanti, infortuni più o meno lunghi, cambi di allenatori o semplicemente periodi di forma differenti sia per motivi psicologici che per motivi fisici a cui per esempio possono contribuire un maggior numero di gare ravvicinate a causa anche di altre competizioni.
Manchester City, Arsenal e Chelsea per esempio hanno continuato a giocare in UEFA Champions League fino al 18/03/2015 (*UEFA Champions League* 2015). Tottenham e Chelsea inoltre sono arrivate fino in finale nel trofeo nazionale Football League Cup e pertanto hanno terminato la competizione il giorno 01/03/2015 (*Flashscore* 2015).
La presenza di squadre anche in altre competizioni non riguarda esclusivamente le compagini considerate più forti. L'Aston Villa infatti è arrivata in finale di FA Cup, altra competizione calcistica che si tiene in Gran Bretagna, dove ha perso contro l'Arsenal il 30/05/2015 (*Diretta* 2015).
In questo senso sicuramente avere una rosa equilibrata e ben assortita permette alle squadre migliori di turnare i propri giocatori in modo da concedere loro momenti di riposo, preservarne la freschezza e la lucidità, evitando così anche possibili infortuni o eventuali ricadute.
I cambiamenti all'interno della squadra inoltre possono riguardare sia i giocatori che i membri dello staff tecnico. L'acquisto più costoso nel corso della stagione sportiva analizzata è stato effettuato dal Manchester City acqui-

stando Wilfried Bony dalla squadra Swansea e subito a seguire dal Chelsea che ha comprato Juan Cuadrado dalla squadra Fiorentina (*Transfermarkt* 2015).
I cambi relativi agli allenatori invece hanno riguardato principalmente le squadre di media o bassa classifica (*Transfermarkt* 2015). In particolare è accaduto per Crystal Palace e Aston Villa (3 volte), West Bromwich, QPR e Newcastle (2 volte) Leicester City, Sunderland e West Ham (una volta).

I grafici in Figura 3 mostrano quindi questo aspetto relativo allo "stato di forma" per alcune squadre, andandone ad analizzare l'andamento in termini di "abilità" ricalcolando le stime fornite dal modello Bradley-Terry senza covariate ogni 5 giornate fino al termine del campionato, considerando sia le partite in casa che le partite in trasferta congiuntamente.
Per ogni squadra viene quindi stimato un modello Bradley-Terry per le prime 5 partite, le prime 10, le prime 15 e procedendo in tal modo fino al termine della competizione e modellando poi la curva utilizzando le splines (Perperoglou et al. 2019).
Le abilità misurate in questo modo sono calcolate senza l'utilizzo delle covariate e, come si può intuire, sono chiaramente legate al numero di punti ottenuti.

Come emerge dal grafico alcune squadre come l'Everton non hanno iniziato il campionato nel migliore dei modi, ma sono riuscite poi a recuperare e procedere in crescendo.
Altre squadre come il West Ham al contrario sono partite bene e poi invece sono calate.
Alcune compagini come il Sunderland hanno avuto un andamento piuttosto irregolare, faticando parecchio soprattutto all'inizio e verso la metà del girone di ritorno in cui hanno riportato parecchi risultati negativi.
Altre squadre invece come Leicester City, Crystal Palace e Stoke City hanno avuto un andamento in crescendo dopo un inizio sotto tono.
Il Crystal Palace in particolare ha risalito la classifica grazie soprattutto all'ottimo rendimeno ottenuto fuori casa dove nel girone di ritorno ha con-

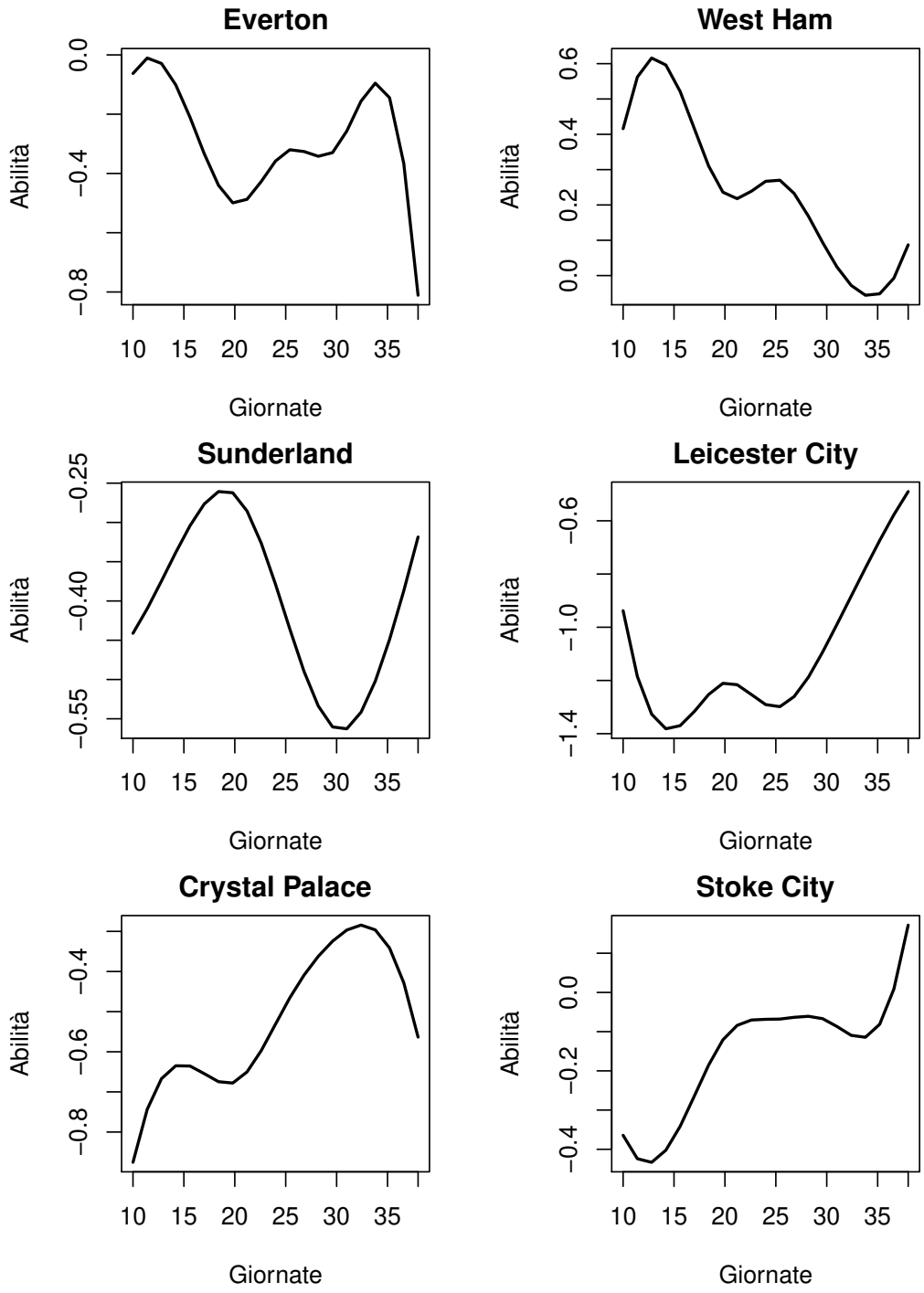

Figura 3: Andamento delle abilità di alcune squadre durante il campionato.

quistato addirittura 19 punti sui 27 disponibili.
Nel cambio di rotta della squadra sicuramente ha influito anche la decisione di cambiare l'allenatore esattamente al termine del girone d'andata (*Transfermarkt* 2015).

Ecco che allora sembra delinearsi uno scenario in cui le abilità delle squadre subiscono un processo temporale durante l'arco del campionato e che soprattutto possiamo stimare.
Alcune squadre infatti potrebbero aver cambiato molto l'andamento nel corso del torneo per diverse motivazioni.
Oltre ai fattori già proposti come la condizione fisica, gli infortuni e i trasferimenti, a volte la ragione invece risiede semplicemente nel "calendario" delle partite dove magari alcune squadre devono affrontare formazioni più forti per alcune partite di fila, per poi magari ritrovarsi alcune partite più facili nelle settimane successive.
L'andamento in crescendo del West Ham tra la 10° e la 15° giornata in cui ha ottenuto buoni risultati per esempio è sicuramente collegato anche al fatto che in quella striscia di partite ha affrontato solamente squadre di medio-bassa classifica.
Può essere interessante osservare anche l'andamento delle squadre che hanno terminato il campionato in cima alla classifica. Esse vengono illustrate in Figura 4.

Chelsea e Manchester City hanno avuto un netto calo nei loro risultati dopo un ottimo inizio. Nonostante questo peggioramento il Chelsea è comunque riuscito a vincere il campionato, anche grazie al declino nel rendimento della concorrenza.
Il Manchester City infatti ha ottenuto sicuramente meno punti nel girone di ritorno, così come il Chelsea che però è riuscito a mantenere la vetta della classifica grazie soprattutto alla partenza lanciata in cui è rimasto imbattuto fino al 06/12/2014.
Più altalenante l'andamento invece di Tottenham e Liverpool.
Una delle poche squadre del campionato che invece ha avuto un andamento

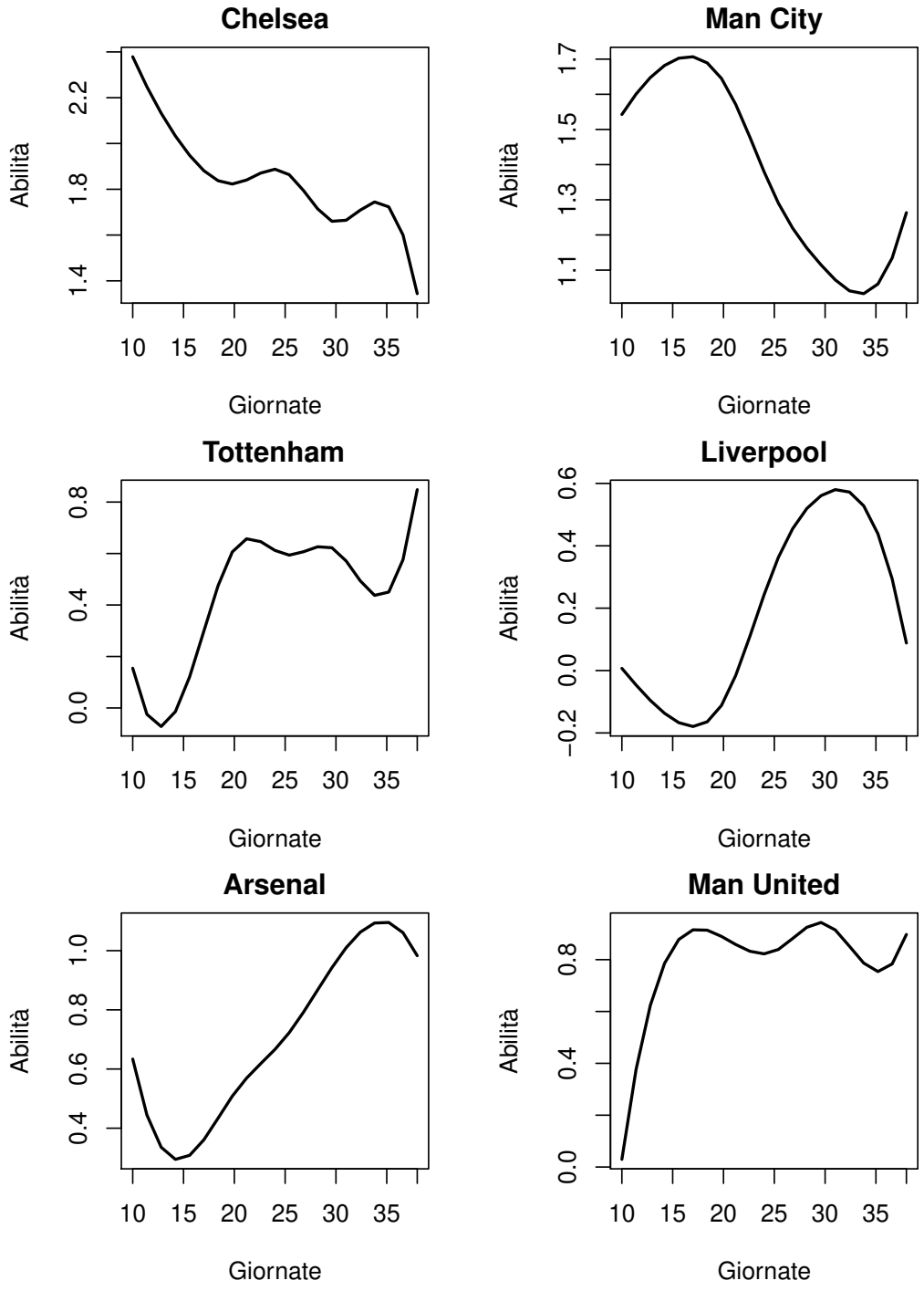

Figura 4: Andamento delle abilità delle prime squadre classificate durante il campionato.

in crescendo per oltre metà campionato è stata l'Arsenal. Tutto ciò le ha permesso di raggiungere il terzo posto al termine del campionato nonostante un avvio sotto tono (solo 2 vittorie nelle prime 8 partite).
Il Manchester United invece ha faticato un po' nelle prime partite, soprattutto fuori casa dove non ha mai trovato la vittoria nei primi 5 incontri, per poi rimanere su buoni livelli per il resto del campionato.

L'utilizzo di tali modellazioni tuttavia garantisce anche altri vantaggi. Infatti il secondo motivo per cui utilizzare un processo temporale per modellare le abilità delle squadre può avere un senso, è legato al numero di parametri stimati.
Nel modello considerato finora ci sono 42 coefficienti (20 per le abilità delle squadre, 21 per le variabili esplicative e 1 per il fattore campo) come riportato nella precedente Tabella 12.
Utilizzare un processo temporale, come quello presentato in seguito, che modelli l'andamento delle capacità delle squadre ridurrebbe drasticamente il numero di coefficienti associati ad esse.
Nell'analisi proposta inizialmente verranno stimati due processi separati per le partite in casa e fuori casa. In questo modo si passerà dall'avere un coefficiente per ogni squadra più un parametro relativo al fattore campo all'avere solamente una coppia di coefficienti, oltre chiaramente alla stima dei due parametri di tuning associati.
In questo modo, al fronte di una leggera quanto inevitabile perdita di informazioni sulle singole abilità di ogni squadra, se ne guadagnerebbe sicuramente in termini di interpretabilità e si potrebbe cogliere in maniera più dettagliata anche l'effetto del giocare in casa osservandone la sua evoluzione nel tempo. Questa in particolare era la terza ed ultima problematica che l'utilizzo di un modello Bradley-Terry dinamico ci permette di risolvere in quest'analisi.

3.2 Processo EWMA per le abilità

Come già trattato nella letteratura statistica (Cattelan, Varin e Firth 2013), un processo adeguato a questo tipo di eventi è il processo EWMA (Exponentially Weighted Moving Average).
Come suggerisce l'acronimo inglese, si tratta di un processo a media mobile pesato esponenzialmente. Il che significa che ad ogni istante temporale il processo prende tutte le osservazioni dall'inizio del campionato fino alla giornata in cui ci si trova, pesandole con un fattore che decresce esponenzialmente retrocedendo nel tempo, dando quindi un peso maggiore ai risultati ottenuti nelle ultime partite che caratterizzano lo stato di forma delle squadre.
Nel modello proposto inizialmente, per determinare lo stato di forma della squadra, è stato scelto di guardare esclusivamente le partite in casa o in trasferta a seconda di dove giochi la squadra coinvolta, anzichè prenderle tutte.
La Formula (2.1) vista nel capitolo precedente viene modificata per tenere conto di questo aspetto e diventa la seguente, in cui la variabile risposta all'istante i è condizionata a tutti i valori precedenti di essa:

$$pr(Y_i \leq y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1) = \frac{\exp(\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i))}{1 + \exp(\delta_{y_i} + a_{h_i}(t_i) - a_{v_i}(t_i))}, \quad (3.1)$$

dove $a_{h_i}(t_i)$ descrive il coefficiente di abilità della squadra h_i che gioca in casa contro la squadra v_i al tempo t_i.
In questo caso la scelta è stata quella di modellare solamente i parametri relativi all'abilità delle squadre.
Per i parametri relativi alle covariate è stato scelto di utilizzare un altro approccio nel capitolo seguente in cui essi vengono stimati mediante una regolarizzazione di tipo lasso.

I parametri di maggiore interesse in questa sezione sono quindi $a_{h_i}(t_i)$ e $a_{v_i}(t_i)$ che vengono modellati attraverso le successive equazioni.
A fini illustrativi le formule vengono riportate solamente per il fattore di abilità casalingo, tuttavia i ragionamenti sono identici anche per le partite

disputate fuori casa.
Il parametro $a_{h_i}(t_i)$ viene quindi modellato come segue:

$$a_{h_i}(t_i) = \lambda_1 \mu_{h_i}(t_i) + (1 - \lambda_1) a_{h_i}(t_i^{(-1)}), \tag{3.2}$$

dove λ_1 è un parametro di lisciamento da stimare e tale che $\lambda_1 \in [0,1]$, mentre la notazione $a_{h_i}(t_i^{(-1)})$ indica l'abilità della squadra in casa al tempo della partita precedente giocata in casa.
Il coefficiente $\mu_{h_i}(t_i)$ invece rappresenta la media delle abilità della squadra che gioca in casa calcolata esclusivamente in base ai risultati delle precedenti partite giocate in casa dalla squadra h_i.
Essa viene calcolata come:

$$\mu_{h_i}(t_i) = \beta_1 r_{h_i}(t_i^{(-1)}), \tag{3.3}$$

tale che β_1 è un parametro calcolato unicamente per le partite in casa, mentre $r_{h_i}(t_i)$ è il numero di punti che ha conquistato la squadra h_i nell'ultima partita in casa.
Quest'ultima variabile deve essere inizializzata e, senza aggiungere complicazioni, è stato assunto questo valore di partenza uguale per tutte le squadre e pari alla media punti totalizzati in casa da tutte le squadre durante il precedente campionato.
Non è possibile, tra l'altro, calcolare la media punti separatamente per ogni squadra per l'anno precedente poichè ogni stagione alcune squadre retrocedono nella serie inferiore, mentre altre vengono promosse in Premier League.
Questi valori di partenza sono pertanto pari a $r_h = 1.641$ per le squadre in casa e $r_v = 1.153$ per le squadre fuori casa. Essi vengono assegnati uguali per tutti in partenza.
Gli stessi calcoli e le stesse deduzioni vanno infatti riproposte anche per a_{v_i} determinata sulla base del parametro β_2 associato alle partite giocate lontano da casa.

Anche in questo caso chiaramente bisogna prima determinare il valore di $\lambda_2 \in [0,1]$ e per farlo, così come per λ_1, è necessario costruirsi le funzioni di verosimiglianza profilo.
Il parametro λ_2 corrisponde al parametro di regolazione che viene utilizzato per pesare i risultati ottenuti durante il campionato per le partite giocate in trasferta, allo stesso modo di come opera λ_1 in Formula (3.2) per le partite giocate in casa.

Per procedere nel determinare i valori dei due parametri di tuning è utile definire $a_{h_i}(t_i)$ e $a_{v_i}(t_i)$ in un modo analogo a quanto proposto in Formula (3.2).
Ancora una volta le equazioni vengono riportate unicamente per a_{h_i} per rendere il discorso più fluido.
In particolare si indica con K il numero di partite giocate in casa prima dell'incontro disputato al tempo t_i. Si può quindi riformulare il modello descritto dalle equazioni precedenti come segue.
Sia quindi

$$a_{h_i}(t_i) = \beta_1 z_{h_i}(t_i; \lambda_1), \tag{3.4}$$

dove $z_{h_i}(t_i; \lambda_1)$ è una media pesata dei risultati precedenti ottenuti nelle partite giocate in casa $r_{h_i}(t_i^{(-k)})$ con pesi $\lambda_1(1\text{-}\lambda_1)^{k-1}$ che decrescono retrocedendo nel tempo.
Più nello specifico si ottiene

$$a_{h_i}(t_i) = \beta_1 \left\{ \lambda_1 \sum_{k=0}^{K-1} (1-\lambda_1)^k r_{h_i}(t_i^{(-k-1)}) + (1-\lambda_1)^K \bar{r}_h \right\}, \tag{3.5}$$

con $\bar{r}_h$ che definisce la media delle variabili $r_{h_i}(t)$ calcolate per la stagione sportiva precedente.
Il valore stimato per il parametro di lisciamento λ_1 in particolare definisce quanto sia persistente la dipendenza rispetto alle partite giocate precedente-

mente in casa.
Analizzando i due casi limite si ha che se poniamo $\lambda_1 = 0$, la Formula (3.5) si riduce a $a_{h_i}(t_i) = \beta_1 \bar{r}_h = \eta$ costante per ogni squadra e per ogni t_i.
Nel caso in cui invece $\lambda_1 = 1$ la Formula (3.5) si semplifica ancora una volta e risulta che $a_{h_i}(t_i) = \beta_1 r_{h_i}(t_i^{(-1)})$ per cui l'abilità stimata per la squadra in casa dipende unicamente dalla partita precedente giocata in casa.
Procedimenti analoghi chiaramente vanno svolti anche per $a_{v_i}(t_i) = \beta_2 z_{v_i}(t_i; \lambda_2)$.

La scelta del metodo da utilizzare per stimare i parametri di lisciamento sicuramente non è unica. In questo caso sono stati determinati utilizzando il criterio della massima verosimiglianza profilo (Murphy e Van Der Vaart 2000).
Una volta definiti quindi $\omega = (\beta_1, \beta_2, \delta)^T$, $\lambda = (\lambda_1, \lambda_2)^T$ e $\theta = (\omega^T, \lambda^T)^T$, si può definire nel modo più tradizionale la funzione di verosimiglianza per θ come:

$$\mathcal{L}(\theta; y) = pr(Y_1 = y_1; \theta) \prod_{i=2}^{n} pr(Y_i = y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1; \theta), \qquad (3.6)$$

dove n è il numero di partite disputate durante la stagione sportiva e pari quindi a 380.

Definiti questi aspetti e dando per noti i valori dei due parametri di lisciamento, si può quindi calcolare la probabilità condizionata del risultato dell'i-esima partita con la formula del modello di regressione logistica cumulata con i punti di soglia fissati δ_{y_i}, attraverso la seguente formula:

$$\begin{aligned} pr(Y_i \leq y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1; \theta) = \\ = \frac{\exp\{\delta_{y_i} + \beta_1 z_{h_i}(t_i; \lambda_1) - \beta_2 z_{v_i}(t_i; \lambda_2)\}}{1 + \exp\{\delta_{y_i} + \beta_1 z_{h_i}(t_i; \lambda_1) - \beta_2 z_{v_i}(t_i; \lambda_2)\}}. \end{aligned} \qquad (3.7)$$

A questo punto la stima dei parametri $\hat{\omega}_\lambda$ si basa su un massimizzazione della verosimiglianza che avviene in due passi.

Per prima cosa vengono stimati i parametri di lisciamento attraverso la massimizzazione della verosimiglianza profilo $\mathcal{L}(\hat{\omega}_\lambda, \lambda; y)$ e risultano essere pari a $\hat{\lambda}_1 = 0.0418$ e $\hat{\lambda}_2 = 0.0335$.
La Figura 5 in particolare mostra le curve della log-verosimiglianza profilo utilizzata per determinare i valori ottimali dei parametri di lisciamento λ_1 e λ_2, ossia quelli che massimizzano la misura della log-verosimiglianza i cui valori sono riportati lungo le curve.

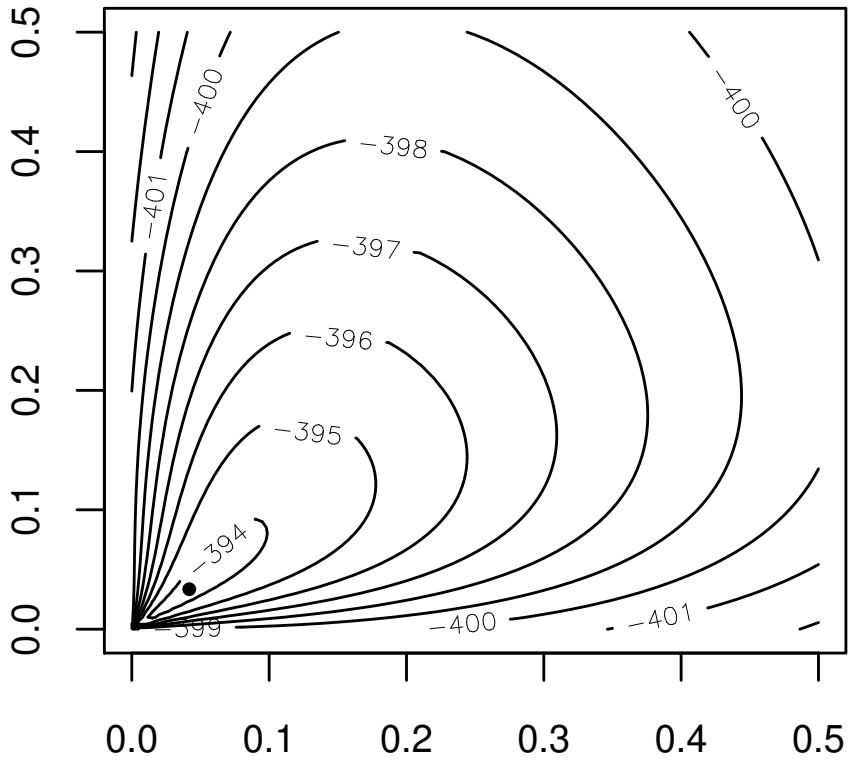

Figura 5: Curve della funzione di log-verosimiglianza profilo per λ_1 e λ_2.

Una volta stimati i valori dei due parametri di lisciamento si può quindi procedere col secondo passaggio determinando il processo a media mobile e stimando i valori di $\hat{\omega}_\lambda$.
Le stime che si ottengono sono rispettivamente $\hat{\beta}_1 = 1.964$ e $\hat{\beta}_2 = 2.464$,

con standard error stimati rispettivamente pari a 0.412 e 0.574, e pertanto i coefficienti risultano significativamente diversi da zero.
β_1 e β_2 in particolare danno una misura di quanto aumenta la probabilità di vittoria in relazione ai risultati delle partite precedenti. Quindi tendenzialmente migliori risultati ottiene una squadra negli ultimi incontri e più aumenta la probabilità di vincere anche la partita successiva, essendo le stime dei β positive.
La stima di β_2 maggiore di β_1 suggerisce anche che lo stato di forma della squadra in trasferta condiziona maggiormente l'esito della gara rispetto a quanto accade invece per la squadra ospitante a parità di risultato.
Quest'ultimo aspetto probabilmente è dovuto però anche ad una maggiore variabilità nei risultati ottenuti fuori casa rispetto a quelli ottenuti giocando in casa.
In questo modo pertanto il numero di coefficienti per l'abilità delle squadre è stato ridotto da 21, ossia uno specifico per ognuna ed un parametro relativo al fattore campo, a solamente β_1 e β_2 con i quali è stato colto l'andamento temporale durante l'intera stagione. Esso è stato poi rappresentato per alcune squadre in Figura 6.
In particolare la curva tratteggiata rappresenta l'andamento fuori casa, mentre la curva disegnata col tratto continuo riguarda le partite giocate in casa.

Dal grafico emergono alcuni aspetti interessanti sulle squadre proposte. Per esempio si può notare come il Crystal Palace e il Tottenham in certi frangenti del campionato abbiano ottenuto risultati migliori in trasferta anzichè nel proprio stadio. Come emerso inizialmente in Tabella 10 infatti esse sono state tra le squadre che hanno ottenuto meno vittorie quando hanno giocato in casa.
Al contrario per esempio il Manchester United, come avevamo ravvisato nel Capitolo 1, ha faticato parecchio nelle partite giocate lontano dal proprio stadio, considerato anche il fatto che ha terminato il campionato quarto in classifica.
Altre squadre come il Newcastle invece presentano differenze tra il rendimento in casa e in trasferta, più o meno costanti durante l'anno.

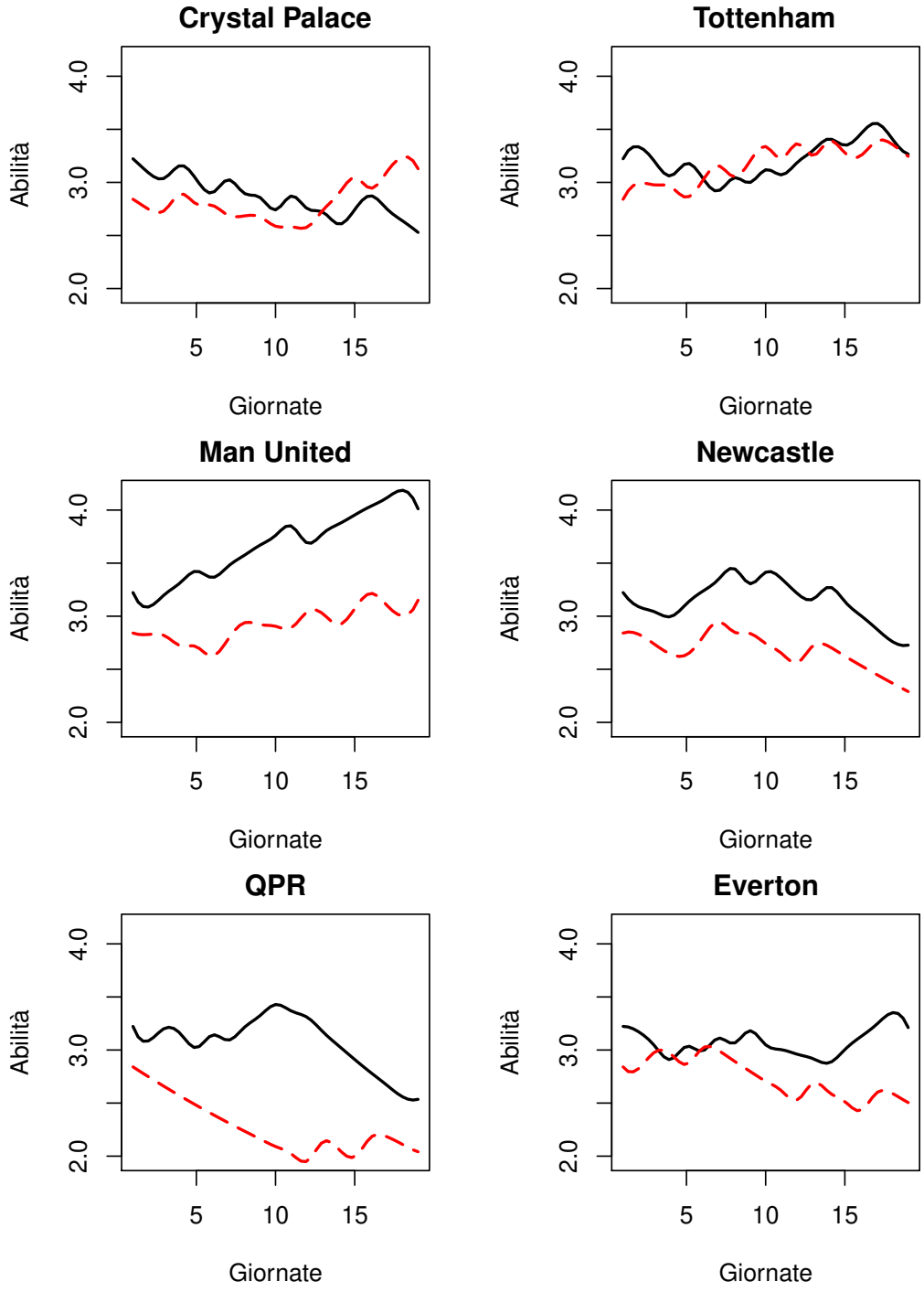

Figura 6: Abilità lisciate per alcune squadre durante il campionato (tratto continuo per le partite in casa, tratteggiato per le partite in trasferta).

La squadra che però forse presenta la differenza più interessante e più irregolare è il Queens Park Ranger (QPR). La prima vittoria fuori casa per loro è arrivata dopo quasi 6 mesi dall'inizio del campionato, per esattezza il 10/02/2015. Da quel momento ha poi totalizzato 6 vittorie consecutive in casa, a fronte di una sola vittoria in trasferta.
In questa inversione di tendenza sicuramente ha giocato la sua parte il cambio allenatore avvenuto esattamente la settimana prima (*Transfermarkt* 2015).

Una volta determinati i valori di β_1 e β_2 è stato possibile stimare i valori assunti dalle abilità dinamiche per ogni squadra al termine del campionato e riportarli in Tabella 13.

Squadre	**Abilità dinamiche**
Chelsea	0.615
Manchester City	0.429
Arsenal	0.296
Manchester United	0.278
Southampton	0.225
Tottenham	0.177
Liverpool	0.166
Swansea	0.060
West Ham	0.042
Stoke City	-0.008
Newcastle	-0.106
Everton	-0.108
Crystal Palace	-0.169
West Bromwich	-0.182
Aston Villa	-0.204
Sunderland	-0.234
Hull City	-0.256
Leicester City	-0.329
Burnley	-0.345
QPR	-0.346

Tabella 13: Coefficienti di abilità dinamica per ogni squadra a fine campionato sia per le partite in casa che in trasferta.

La classifica proposta può essere confrontata con quella ottenuta nel Capitolo 2 per le abilità statiche senza l'utilizzo di covariate.
Come si vede le posizioni in classifica variano poco rispetto a quelle ottenute in precedenza, soprattutto le prime posizioni rimangono pressochè invariate. Ciò che cambia in parte è la differenza tra le stime dei vari coefficienti che per alcune squadre risulta più marcata rispetto a prima com'è il caso tra Chelsea e Arsenal, decisamente più labile invece per esempio la differenza tra Burnley e QPR.
Anche per il modello Bradley-Terry con abilità dinamiche è stato calcolato il valore della statistica RPS per confrontarlo con quello dei modelli stimati nel Capitolo 2.
In questo caso la stima del Ranked Probability Score risulta pari a 0.398 quindi molto simile a quella ottenuta per il modello Bradley-Terry senza covariate, con il vantaggio che in questo caso sono stati utilizzati solamente 2 coefficienti anzichè 21.
Essendoci differenze tra i risultati ottenuti in casa e in trasferta, e soprattutto non necessariamente simili per tutte le squadre, può essere interessante osservare anche come cambiano i coefficienti di abilità stimati per i due casi. In particolare in Tabella 14 vengono riportate in ordine decrescente le stime di abilità per ogni squadra in base a quando ha giocato nel proprio stadio e quando invece fuori casa, calcolate a fine campionato.

Dal confronto che emerge in Tabella 14 si vedono sostanzialmente le differenze osservate in Figura 6 espresse in modo quantitativo.
Il QPR infatti è nettamente ultimo come rendimento fuori casa, mentre è circa a metà classifica a livello di performance nel proprio stadio.
Il Chelsea ha nettamente il miglior coefficiente per gli incontri in casa dal momento che non ha mai perso nel proprio stadio in tutte le 19 partite disputate. Subito alle sue spalle il Manchester United che invece passa dal secondo miglior coefficiente in casa a solamente il settimo nelle partite in trasferta.
QPR e Manchester United infatti sono le due squadre con la maggiore differenza tra le abilità in casa e fuori casa. Le due squadre invece che hanno

Squadra	**Abilità casa**	**Squadra**	**Abilità trasferta**
Chelsea	0.687	Chelsea	0.543
Man United	0.480	Man City	0.493
Man City	0.364	Tottenham	0.319
Arsenal	0.344	Arsenal	0.249
Southampton	0.229	Liverpool	0.224
Swansea	0.147	Southampton	0.220
Liverpool	0.108	Man United	0.075
West Ham	0.097	Stoke City	0.022
Tottenham	0.035	West Ham	-0.013
Stoke City	-0.038	Crystal Palace	-0.021
Newcastle	-0.039	Swansea	-0.028
Everton	-0.105	Aston Villa	-0.083
QPR	-0.117	Everton	-0.111
West Brom	-0.243	Sunderland	-0.118
Leicester City	-0.304	West Brom	-0.120
Crystal Palace	-0.316	Newcastle	-0.174
Hull City	-0.325	Hull City	-0.188
Aston Villa	-0.326	Leicester City	-0.355
Burnley	-0.329	Burnley	-0.362
Sunderland	-0.349	QPR	-0.574

Tabella 14: Coefficienti di abilità dinamica in casa e in trasferta per ogni squadra a fine campionato.

mantenuto un rendimento più simile tra casa e trasferta sono Everton e Southampton.
Discorso inverso per le formazioni del Tottenham, Crystal Palace, Aston Villa e Sunderland che si ritrovano tutte 6 posizioni più alte nella classifica delle abilità in trasferta.

3.3 Modello dinamico con un unico processo

Nell'ultima parte del capitolo viene proposta un'alternativa al modello Bradley-Terry dinamico visto finora.
Nel paragrafo precedente sono stati costruiti due processi distinti per stimare le abilità delle squadre a seconda che giochino in casa o in trasferta.
In questo modo sicuramente si ha un quadro dettagliato e completo. In alcuni casi però si può pensare ad un modello più semplice e soprattutto più facile da interpretare in cui le abilità delle squadre seguano un unico processo EWMA durante l'intero campionato, senza guardare pertanto dove giocano.
In questo modo l'andamento viene descritto da un unico coefficiente β che dà una misura di quanto aumenti la probabilità di vittoria in base ai risultati precedentemente ottenuti, prendendo appunto tutte le partite precedenti, non solamente quelle in casa o in trasferta.
L'alternativa proposta in questo paragrafo può apparire meno appropriata della precedente visto che le differenze in termini di abilità dinamica tra le partite giocate in casa e le partite giocate in trasferta non sono risultate costanti per tutte le squadre lungo l'intera stagione, come evidenziato nella precedente Figura 6.
Alcune squadre come il Newcastle infatti hanno mantenuto una differenza di abilità tra casa e trasferta pressochè identica per tutta la stagione.
Il che significa che se in un momento del campionato la squadra ha faticato ad ottenere punti, o ha conquistato una buona striscia di risultati utili, lo ha fatto sia in casa che fuori casa.
Altre squadre invece hanno chiaramente avuto una differenza tra i risultati ottenuti in casa e quelli ottenuti in trasferta che è risultata variabile durante

l'anno.
Il Manchester United per esempio ha alzato notevolmente il proprio rendimento casalingo nel corso della stagione sportiva, mantenendo un andamento in trasferta praticamente costante. Il QPR invece ha mantenuto un pessimo rendimento in trasferta da inizio campionato, mantenendo invece un buon rendimento nel proprio stadio solamente per una parte della stagione.
Tuttavia facendo un test per verificare se i parametri β_1 e β_2 siano statisticamente diversi tra loro, risulterebbe un rifiuto dell'ipotesi a livello 5% e pertanto le due stime non sembrano essere significativamente diverse tra loro. Pertanto si potrebbe, a maggior ragione, pensare di utilizzare un unico processo EWMA che controlli l'andamento delle squadre durante il campionato sia per le partite giocate in casa che per le partite disputate lontano dal proprio stadio.
Il modello così stimato sarà inoltre di grande utilità nel definire il modello dinamico penalizzato nel Capitolo 5 poichè permette di ridurre notevolmente i tempi di calcolo dovendo stimare un parametro di tuning in meno, garantendo comunque risultati apprezzabili.

Il modello proposto nel presente paragrafo viene descritto in modo simile a quanto visto per il modello dinamico con due processi proposto in Formula (3.7). In particolare in questo caso si ha che la probabilità condizionata del risultato dell'i-esima partita dati i risultati delle partite precedenti è descritta come:

$$\begin{aligned} pr(Y_i \leq y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1; \theta) = \\ = \frac{\exp\{\delta_{y_i} + \eta + \beta z_{h_i}(t_i;\lambda) - \beta z_{v_i}(t_i;\lambda)\}}{1 + \exp\{\delta_{y_i} + \eta + \beta z_{h_i}(t_i;\lambda) - \beta z_{v_i}(t_i;\lambda)\}}, \end{aligned} \tag{3.8}$$

in cui ovviamente sono presenti un unico parametro di regolarizzazione λ e di conseguenza un unico parametro β.
Anche le covariate $z_{h_i}(t_i;\lambda)$ e $z_{v_i}(t_i;\lambda)$ vengono calcolate allo stesso modo di

quanto visto in Formula (3.5), sulla base però di uno stesso valore λ.

Nel processo proposto nel presente paragrafo, l'unico parametro di tuning utilizzato è stato stimato massimizzando la funzione di verosimiglianza ed è pari a $\hat{\lambda} = 0.06$. Di conseguenza la stima per il coefficiente β è uguale a $\hat{\beta} = 1.507$ con errore standard associato pari a 0.253, quindi anche in questo caso la stima risulta ampiamente significativa.
Ancora una volta inoltre essa è positiva, quindi chiaramente ottenere buoni risultati nelle ultime partite aumenta la probabilità di fare bene anche nella partita successiva al netto dell'avversario.
Tuttavia la stima è più ridotta rispetto a quanto era emerso nell'analisi con β_1 e β_2, pertanto la dipendenza con i risultati precedenti sembra meno netta quando non si considera l'aspetto legato al fattore campo.
Il parametro di soglia δ_1 viene invece stimato pari a -0.610, mentre la stima del parametro η relativo all'effetto del giocare in casa è pari a 0.335.
Fatte quindi le giuste premesse, vengono ora riportate in Tabella 15 le abilità dinamiche calcolate utilizzando un unico processo EWMA per ogni squadra al termine del campionato.

L'ordine identificato in questo modo è leggermente diverso da quello proposto in Tabella 13 dove vengono utilizzati due processi temporali separati per le partite in casa e in trasferta.
La classifica appena proposta per le abilità dinamiche sembrerebbe più simile alla reale classifica al termine del campionato.
Si può osservare questo aspetto anche in modo quantitativo calcolando la correlazione tra le classifiche stimate tramite le abilità e la classifica reale al termine del campionato.
In particolare si possono usare alcune misure note come "indice di Kendall" (Kendall 1938) e "indice di Spearman" (Spearman 1987).
Tali indicatori sono basati su metodi non parametrici e pertanto consentono una particolare flessibilità nelle condizioni necessarie per le variabili analizzate.
In particolare essi non richiedono che entrambe le variabili da confrontare

Squadre	**Abilità dinamiche**
Chelsea	0.483
Manchester City	0.430
Tottenham	0.411
Southampton	0.258
Manchester United	0.241
Arsenal	0.230
Liverpool	0.129
West Ham	0.119
Stoke City	0.079
Crystal Palace	0.020
Swansea	0.020
West Bromwich	-0.103
Newcastle	-0.184
Aston Villa	-0.193
Hull City	-0.208
Sunderland	-0.293
Everton	-0.298
Leicester City	-0.362
Burnley	-0.369
QPR	-0.412

Tabella 15: Coefficienti di abilità dinamica stimati a fine campionato per ogni squadra utilizzando un unico processo EWMA.

siano quantitative. Esse possono essere pertanto due classifiche, ossia variabili qualitative ordinali dal momento che la differenza tra una posizione e l'altra è sempre la stessa.
Questi due indici hanno peculiarità diverse e talvolta uno può venire preferito all'altro in base al contesto. Ad ogni modo in quest'analisi essi danno le stesse informazioni, rafforzando maggiormente le conclusioni che se ne traggono.
Entrambi gli indicatori variano da -1 a 1 come accade per il più tradizionale "indice di Pearson". Più la stima tende all'estremo negativo e più le due variabili sono inversamente correlate, più la stima tende all'estremo positivo e più le due variabili sono correlate positivamente. Se la stima è prossima allo 0 allora invece non vi è alcuna concordanza.
Il confronto può essere fatto oltre che tra i modelli dinamici, anche con il modello Bradley-Terry statico proposto nel Capitolo 2 senza variabili esplicative.
In particolare gli indici di Kendall e Spearman per misurare la correlazione tra la classifica delle abilità statiche riportata in Tabella 11 e la classifica reale, sono pari rispettivamente a 0.474 e 0.552.
Per quanto riguarda la correlazione tra la classifica delle abilità dinamiche calcolate tramite la Formula (3.5) che utilizza due processi EWMA e la classifica finale del campionato, le stime dei due indici sono rispettivamente uguali a 0.227 e 0.311.
In entrambi i casi vi è quindi una moderata concordanza positiva tra le abilità stimate delle squadre e la classifica effettiva al termine della stagione. Nel primo caso in cui si utilizza un coefficiente per squadra anzichè solo una coppia, le stime risultano chiaramente più alte.
Infine calcolando gli indici di correlazione tra la classifica delle abilità dinamiche stimate tramite un unico processo temporale e la classifica del campionato, essi risultano pari rispettivamente a 0.358 e 0.478.
Le stime così calcolate sono superiori a quelle calcolate per il modello dinamico che utilizza due processi e sono addirittura più vicine ai valori ottenuti con il modello statico.
I modelli Bradley-Terry dinamici stimati in questo capitolo chiaramente possono venire confrontati con i modelli stimati nel Capitolo 2 anche tramite

l'indicatore Ranked Probability Score. In particolare in questo modo il confronto può essere fatto anche con il modello che utilizza le informazioni derivate dalle variabili esplicative.
I valori del RPS per i modelli statici erano stati stimati rispettivamente pari a 0.388 per il modello senza covariate e 0.339 per il modello completo.
Il valore del Ranked Probability Score risulta inevitabilmente maggiore per i modelli dinamici stimati nel presente capitolo. Esso è infatti pari a 0.398 per il modello che utilizza due processi EWMA distinti, mentre risulta pari a 0.414 per il modello con un unico processo.
La prima formulazione del modello pertanto sembra adattarsi meglio ai dati analizzati partita per partita, come ci si poteva attendere.

Infine viene riportato in Figura 7 l'andamento delle abilità lisciate tramite splines per alcune squadre con il modello che utilizza un unico trend temporale.
Come si vede l'andamento è molto irregolare per qualsiasi squadra rispetto ai grafici riportati in Figura 6 in cui si erano utilizzati due processi.

Procedendo quindi con l'analisi, dopo aver ridotto il numero di coefficienti per le abilità delle squadre, il prossimo passo è quello di cercare di ridurre anche l'ammontare di variabili esplicative utili all'analisi, oltre a cogliere possibili effetti in comune tra diverse squadre per alcune covariate.
Per farlo ci si servirà di una delle tecniche di regolarizzazione più famose: il Lasso. Questa tecnica è stata adattata a moltissime situazioni a partire dagli ultimi anni dello scorso secolo (Tibshirani 1996).
Nel prossimo capitolo verrà quindi trattata questa tematica, servendosi del pacchetto BTLLasso (Schauberg e Tutz 2019) disponibile per il software R.

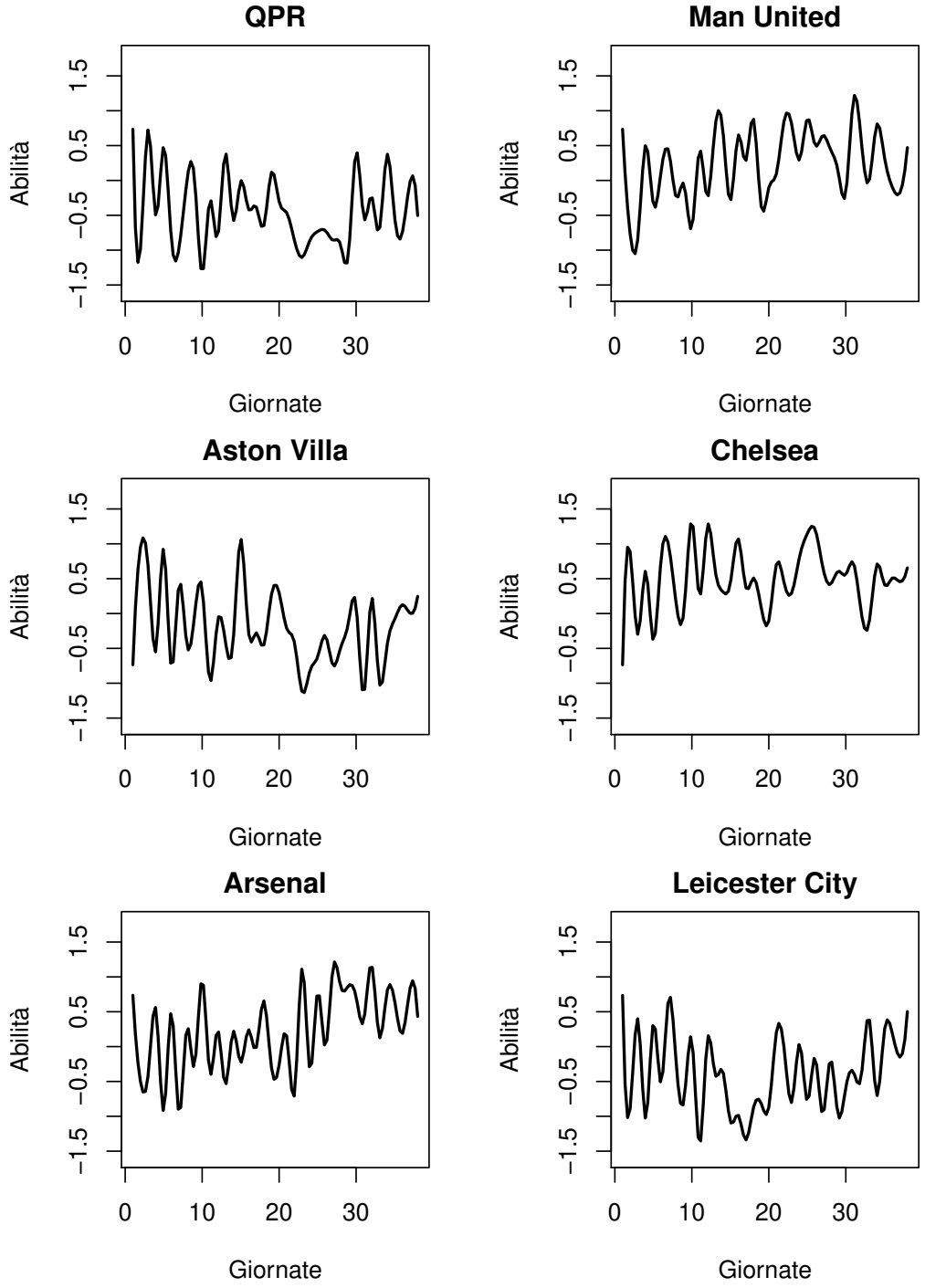

Figura 7: Abilità lisciate per alcune squadre durante il campionato per il modello con un unico processo EWMA.

4 Selezione delle variabili esplicative mediante lasso

4.1 Tipologia delle covariate

L'aggiunta di covariate al modello permette di comprendere meglio quali siano le caratteristiche che favoriscono una squadra rispetto ad un'altra in una partita.
La maggiore efficacia del modello che utilizza le covariate si può apprezzare anche tramite il confronto in termini di Ranked Probability Score fatto nel Capitolo 2.
Sembra abbastanza intuitivo per esempio che una squadra che tira di più in porta rispetto all'avversario, sia più propensa a segnare e a vincere la partita. Pertanto aggiungendo la covariata in questione ci si aspetterebbe che il coefficiente associato sia nettamente positivo come si è visto anche nelle precedenti analisi.
Sicuramente sarà di interesse anche valutare la significatività di tale coefficiente e se presenta variazioni importanti tra le diverse squadre.
Nell'analisi svolta nel Capitolo 2 infatti si sono osservati alcuni coefficienti relativi alle covariate non significativi. Bisogna quindi comprendere se tale variabilità delle stime possa essere dovuta a differenze sostanziali tra le squadre riguardo tale aspetto del gioco o se semplicemente la covariata sia ininfluente ai fini dell'analisi.
Per esempio due squadre potrebbero vincere lo stesso numero di partite, ma con una grande differenza di tiri.
Il QPR infatti ne è la conferma. Di fatto esso risulta quinto nella speciale classifica dei tiri a fine campionato, nonostante abbia concluso il torneo come ultimo (*Guerin sportivo* 2015), come mostrato in Tabella 16.

Un altro aspetto che può essere d'interesse avendo considerato una variabile risposta con sole tre modalità, riguarda il fatto che non conta di quanto una squadra vinca o perda.

Classifica tiri	**Numero tiri**	**Classifica campionato**
Manchester City	669	Chelsea
Arsenal	611	Manchester City
Liverpool	593	Arsenal
Chelsea	564	Manchester United
QPR	531	Tottenham
Tottenham	524	Liverpool
Manchester United	510	Southampton
Southampton	507	Swansea
Stoke City	498	Stoke City
West Ham	485	Crystal Palace
Everton	481	Everton
Newcastle	467	West Ham
Leicester City	456	West Bromwich
Crystal Palace	436	Leicester City
Burnley	426	Newcastle
Hull City	426	Sunderland
Swansea	424	Aston Villa
Aston Villa	412	Hull City
West Bromwich	408	Burnley
Sunderland	404	QPR

Tabella 16: Confronto classifica dei tiri per squadra e classifica del campionato.

Per fare un esempio, effettuare 10 tiri più degli avversari in una partita e vincere 3-0 o fare lo stesso numero di tiri e vincere 1-0, nell'analisi proposta danno lo stesso tipo di informazioni.
Nel campionato analizzato, nel $12,4\%$ dei casi la partita è terminata con almeno 3 gol di scarto. Queste informazioni potrebbero infatti venire utilizzate in un nuovo modello che tenga conto di più modalità per la variabile dipendente.
L'analisi svolta in questo lavoro invece punta a comprendere cosa possa maggiormente influire sulla probabilità di vittoria di un incontro. A tale fine sono state raccolte le 21 variabili esplicative descritte nel Capitolo 1.

Nei modelli Bradley-Terry vi sono due entità in particolare che vengono chiamate "soggetti" e "oggetti".
Le prime nell'analisi proposta sono le 38 giornate della competizione. Le seconde invece sono gli elementi tra cui avviene il confronto, ossia le 20 squadre di Premier League nel nostro caso.
In questo modo è facile quindi distinguere tre tipologie di covariate. Esse possono essere relative unicamente al soggetto, all'oggetto o ad una relazione tra soggetto e oggetto.
Ci sono quindi situazioni in cui l'esito di un confronto a coppie può venire deciso da un soggetto, tuttavia questo non è il caso in questione.
Infatti è difficile immaginare una variabile relativa alla partita, come potrebbe essere l'orario, il mese o la condizione atmosferica, che però abbia un effetto diverso sulle due squadre.
La seconda categoria è quella delle variabili relative unicamente all'oggetto. Esse sono variabili che dipendono solamente dalla squadra e sono costanti per tutta la durata della stagione.
Alcuni autori hanno svolto analisi di eventi sportivi con modelli Bradley-Terry utilizzando variabili relative unicamente all'oggetto ed in particolare assumendo che i valori di mercato delle varie squadre fossero costanti per l'intera stagione, considerando per esempio i valori ad inizio o a fine campionato, senza tenere conto di acquisti o cessioni avvenute nella finestra di mercato invernale (Schauberg e Tutz 2019).

A chiudere il quadro l'ultima tipologia, ossia le variabili che considerano una relazione tra soggetto e oggetto. Esse sono tutte quelle considerate in quest'analisi. Infatti dal dataset a disposizione sono disponibili tutti i valori di ogni variabile esplicativa per ognuna delle 380 partite.

4.2 Determinazione del parametro di tuning

L'obiettivo che ci si prefigge in questo capitolo non è solo quello di stimare i coefficienti associati alle covariate misurate durante le partite, ma è anche quello di fare selezione delle variabili tenendo nel modello solo quelle che maggiormente influiscono sulle probabilità di vittoria.
Per operare questa procedura si è deciso di utilizzare la tecnica del Lasso (Least Absolute Shrinkage and Selection Operator), introdotta verso la fine dello scorso secolo (Tibshirani 1996).
Questa tecnica permette di ridurre la variabilità delle stime a fronte di un leggero, e solitamente tollerabile, aumento della loro distorsione.
Tuttavia questo non è l'unico aspetto rilevante di tale approccio. Nel corso di quest'analisi infatti apparirà evidente anche come per alcune variabili si possano formare clusters di squadre con caratteristiche omogenee a gruppi.
Un'altra curiosità infatti potrebbe riguardare quali squadre presentano coefficienti estremamente diversi rispetto ad alcune variabili.
Per esempio alcune squadre quando vincono commettono più falli, mentre altre ottengono gli stessi risultati essendo meno aggressive, o ancora più banalmente riprendendo l'esempio precedente del numero di tiri effettuati si può notare una direzione di base comune a molte squadre, a cui fa eccezione per esempio il QPR.
L'applicazione di tecniche di regolarizzazione come quella del lasso permette di fare fronte anche a queste eventualità e la sua trattazione viene affrontata nel paragrafo seguente.

L'idea del lasso è quella di penalizzare la funzione di verosimiglianza con un termine ζ che dipende dalla complesssità del modello ed un termine $J(\epsilon)$

che penalizza specifiche strutture del vettore dei parametri, come spiegato in seguito.
Una delle problematiche di queste metodologie riguarda infatti la stima del parametro di tuning.
La scelta proposta nella fattispecie è quella di determinare il valore di ζ andando a massimizzare la verosimiglianza penalizzata $l_p(\epsilon)$, dove $\epsilon = (a_1, \ldots, a_m, \gamma_{1,1}, \ldots, \gamma_{p,m}, \eta_1, \ldots, \eta_m)$ indica l'intero vettore di parametri del modello.
La verosimiglianza penalizzata viene espressa nella seguente equazione, in modo piuttosto generico, poichè il termine di penalizzazione può essere più o meno complesso a seconda delle esigenze.
Essa è pertanto definita come:

$$l_p(\epsilon) = l(\epsilon) - \zeta J(\epsilon), \tag{4.1}$$

dove il primo termine indica la tradizionale funzione di log-verosimiglianza, mentre il secondo termine indica il prodotto tra il parametro di penalizzazione e una penalità costruita appositamente.
Come al solito ζ è un valore positivo che ha il ruolo di specificare quanto "pesa" la penalizzazione sulle stime.

Il parametro in questione può essere stimato in diversi modi. In questo lavoro il parametro viene determinato attraverso una procedura di convalida incrociata con 10 gruppi (Browne 2000) che risulta essere piuttosto onerosa, anche senza utilizzare un numero elevato di covariate.
Questa tecnica trova ampio utilizzo in ambito statistico poichè permette di ottenere misure di bontà di previsione per un modello evitando sia problemi di sovradattamento che di campionamento asimmetrico (e quindi affetto da distorsione).
Utilizzando la convalida incrociata si suddivide il campione osservato in gruppi della stessa numerosità, se ne esclude iterativamente uno alla volta e si forniscono delle stime per tali osservazioni con il modello stimato usando i gruppi non esclusi.

Il fine di queste operazioni è quello di verificare la bontà del modello di predizione utilizzato e determinare quindi il parametro di tuning associato.
Il modello ritenuto migliore per la presente analisi è quindi quello per cui la funzione di perdita presenta un valore inferiore.
La scelta sulla funzione da ottimizzare è ricaduta sul Ranked Probability Score introdotto nel Capitolo 2.

Il termine di penalizzazione $J(\epsilon)$ riportato in Formula (4.1) è composto da più termini combinati, ognuno dei quali penalizza una diversa componente del modello.
In particolare si ha:

$$J(\epsilon) = \sum_{l=1}^{L} \psi_l P_l, \tag{4.2}$$

dove P_l sono le differenti penalizzazioni, mentre ψ_l sono i pesi da applicare e L indica il numero di diverse penalizzazioni applicate.
Ovviamente prima di procedere con queste operazioni le variabili devono essere standardizzate in modo che il termine di penalità abbia lo stesso peso per tutte le covariate.
Le penalità possono essere applicate ai parametri di abilità e a tutte le tipologie di variabili esplicative descritte nel paragrafo precedente, oltre che ai termini che identificano l'ordine con cui si presentano gli oggetti e che in questo caso corrisponde al fattore campo.
Le penalizzazioni utilizzate possono essere sia sul valore assoluto dei singoli coefficienti che sul valore assoluto della differenza tra coefficienti, in base alle opzioni che si vogliono utilizzare.
In particolare in quest'analisi sono state utilizzate per la maggior parte le opzioni di default, con qualche modifica in modo da sfruttare l'ordine degli oggetti per tutte le partite (effetto del giocare in casa) e penalizzare la differenza tra le abilità delle squadre (per ridurre il numero di coefficienti).
Le penalità, utilizzate con lo scopo di fare clustering o selezione delle varia-

bili, e poi combinate per costruire $J(\epsilon)$ sono pertanto le seguenti tre.
La prima penalizzazione, P_1, viene applicata alle $m = 20$ stime delle abilità delle squadre a_k. Essa è definita come:

$$P_1(a_1, ..., a_m) = \sum_{r \leq s} |a_r - a_s|. \tag{4.3}$$

Si utilizza una penalizzazione applicata ai valori assoluti delle differenze tra i coefficienti per cercare di formare clusters di squadre con abilità simile.
La seconda penalizzazione, P_2, viene invece applicata ai coefficienti $\gamma_{k,j}$ relativi alle $p = 21$ variabili esplicative descritte nel Capitolo 1 che ora possono essere assunti diversi tra le varie squadre. In particolare il coefficiente $\gamma_{k,j}$ si riferisce alla k-esima covariata e alla j-esima squadra.
Tale penalizzazione viene definita come:

$$P_2(\gamma_{1,1}, ..., \gamma_{p,m}) = \sum_{j=1}^{p} \sum_{r \leq s} |\gamma_{r,j} - \gamma_{s,j}| + \sum_{j=1}^{p} \sum_{r=1}^{m} |\gamma_{r,j}|. \tag{4.4}$$

In questo caso la penalizzazione è applicata non solo ai valori assoluti delle differenze per cercare gruppi di squadre con effetti simili per alcune variabili, ma anche al valore assoluto di tali coefficienti per cercare di fare selezione delle variabili ed eliminare quindi dal modello quelle che presentano una stima inferiore ad una certa soglia chiaramente determinata dal valore di ζ.
Infine l'ultima penalizzazione viene applicata ai coefficienti che misurano l'effetto del giocare in casa. La penalizzazione P_3 viene quindi definita come:

$$P_3(\eta_1, ..., \eta_m) = \sum_{r \leq s} |\eta_r - \eta_s| + \tau \sum_{r=1}^{m} |\eta_r|, \tag{4.5}$$

dove η_k sono i coefficienti che stimano l'effetto dovuto all'ordine degli oggetti, ossia il fattore campo come definito nel Capitolo 2, mentre τ è una

variabile che può assumere valore 0 o 1 a seconda che si voglia utilizzare anche la penalizzazione in valore assoluto o meno.
In quest'analisi la penalizzazione è stata applicata solamente al valore assoluto delle differenze tra i coefficienti, per cercare squadre che ottengono un vantaggio simile dal giocare in casa in termini di probabilità di vittoria, pertanto è stato posto $\tau = 0$.

4.3 Risultati selezione variabili

Il primo aspetto che verrà descritto riguardo il modello Bradley-Terry in questione, a cui è stata applicata un procedura lasso, è la capacità di fare selezione delle variabili.
La procedura di penalizzazione applicata al modello ha permesso quindi di eliminare alcune variabili, tenendo solo quelle significative nell'analisi.
In particolare le covariate eliminate dal modello in questo modo sono i tiri di testa e i tiri assistiti.
Queste due variabili sono state ritenute ininfluenti ai fini dell'analisi per ogni squadra della Premier League.
Nelle analisi effettuate nel Capitolo 2 utilizzando il modello statico con tutte le covariate a disposizione, queste variabili avevano coefficienti rispettivamente pari a -0.073 (con standard error pari a 0.100) per i tiri di testa e 0.068 (con standard error 0.055) per i tiri assistiti.
Prendendo i classici valori di riferimento per testare la significatività delle stime, si può osservare come entrambe non fossero significative ad un livello di confidenza pari al 5%.
Coefficienti associati ad altre variabili invece sono nulli per molte squadre, ma non tutte, mentre altri sono molto contenuti ma comunque diversi da zero.

Quest'analisi più approfondita può dare indicazioni migliori su cosa sia più importante per l'esito di una gara. Per esempio se è più determinante tirare in generale o tirare in porta, se è più importante mirare agli angoli bassi o alti della porta, l'importanza di tirare da vicino piuttosto che da lontano,

o da un lato rispetto che dall'altro.
Sicuramente sarà interessante osservare anche se le stime ottenute in questo modello sono in linea con quelle ottenute in Tabella 12 per il modello Bradley-Terry senza penalizzazione.
Purtroppo a differenza di quanto visto nel Capitolo 2, in questo caso non è possibile calcolare anche gli standard error delle stime in modo agevole.
Il modo più conveniente infatti sarebbe attraverso una procedura di tipo bootstrap (Henderson 2005), la quale però risulterebbe particolarmente onerosa da un punto di vista computazionale dato il numero relativamente alto di variabili.
Il bootstrap in particolare è una tecnica utile ad approssimare la distribuzione campionaria di una statistica.
In genere si estraggono dalle osservazioni del dataset un numero elevato (M) di campioni casuali con reinserimento e si calcolano le stime di interesse. Alla fine pertanto si avranno M stime per ognuna delle nostre statistiche d'interesse da cui è possibile calcolare media, varianza, standard error e percentili per esempio.
Le stime dei parametri sia per le abilità delle squadre che per le covariate sono riportate in Tabella 17.
Come accennato in precedenza, non tutte le variabili hanno un'unica stima per tutte le squadre, pertanto verrà indicato l'insieme di squadre che hanno lo stesso valore per quel parametro.

Variabile/Squadra	**Stima**
Casa	0.402
Arsenal	0.666
Aston Villa, Burnley, Hull City, Leicester City, QPR	-0.350
Chelsea	1.089
Crystal Palace, Sunderland	-0.222
Everton	-0.349

Continua nella pagina successiva

Continua dalla pagina precedente

Variabile/Squadra	**Stima**
Liverpool	0.158
Man City	0.809
Man United	0.503
Newcastle	-0.337
Southampton	-0.169
Swansea, Stoke City	-0.093
Tottenham	0.293
West Bromwich	-0.169
West Ham	-0.114
Tiri	0.335
Tiri in porta	0.192
Tiri sul palo (Arsenal)	-0.022
Tiri sul palo (Aston Villa, Everton, Leicester City, Liverpool, Swansea, Tottenham, Newcastle)	0.004
Tiri sul palo (Burnley, Chelsea, Hull City, Man City, Man United, QPR, West Bromwich)	-0.077
Tiri sul palo (Crystal Palace)	0.115
Tiri sul palo (Southampton)	0.124
Tiri sul palo (Stoke City, Sunderland, West Ham)	-0.019
Tiri di testa	0.000
Tiri dal limite (Newcastle, QPR)	-0.096
Tiri dal limite (Everton)	-0.147
Tiri dal limite (Tutte le altre)	-0.152
Tiri assistiti	0.000
Passaggi filtranti (Aston Villa, Crystal Palace)	-0.200
Passaggi filtranti (Burnley)	-0.125
Passaggi filtranti (Chelsea, Hull City, Man United, Southampton)	0.000
Passaggi filtranti (Newcastle)	-0.447
Passaggi filtranti (Sunderland)	-0.389

Continua nella pagina successiva

Continua dalla pagina precedente

Variabile/Squadra	**Stima**
Passaggi filtranti (Swansea)	0.053
Passaggi filtranti (Tutte le altre)	-0.026
Corner (Everton)	-0.378
Corner (Tutte le altre)	-0.321
Falli (Everton, Liverpool, Sunderland, Swansea)	-0.032
Falli (Tutte le altre)	-0.076
Ammonizioni (Burnley)	0.014
Ammonizioni (Leicester City, West Ham)	-0.006
Ammonizioni (Tutte le altre)	0.000
Fuorigioco (Chelsea)	-0.098
Fuorigioco (Tutte le altre)	-0.205
Tiri dal centro	-0.006
Tiri da sinistra (Leicester City)	-0.028
Tiri da sinistra (Tutte le altre)	-0.020
Tiri da destra	0.109
Tiri su punizione (Everton)	0.000
Tiri su punizione (Leicester City)	0.244
Tiri su punizione (Tottenham)	0.319
Tiri su punizione (Tutte le altre)	-0.039
Tiri da posizioni difficili (QPR)	-0.240
Tiri da posizioni difficili (Southampton)	0.196
Tiri da posizioni difficili (Stoke City)	-0.051
Tiri da posizioni difficili (Burnley, Leicester City)	-0.099
Tiri da posizioni difficili (Man United)	-0.100
Tiri da posizioni difficili (Tutte le altre)	0.000
Cross (QPR)	0.000
Cross (Southampton)	-0.129
Cross (Tutte le altre)	-0.237
Ammonizioni primo tempo (Burnley)	0.234

Continua nella pagina successiva

Continua dalla pagina precedente

Variabile/Squadra	**Stima**
Ammonizioni primo tempo (Stoke City)	-0.152
Ammonizioni primo tempo (Tutte le altre)	-0.092
Bloccare tiri (QPR)	0.434
Bloccare tiri (Tutte le altre)	0.272
Tiri negli angoli alti	0.192
Tiri negli angoli bassi (Tottenham)	0.382
Tiri negli angoli bassi (Swansea)	0.321
Tiri negli angoli bassi (Burnley)	0.160
Tiri negli angoli bassi (Tutte le altre)	0.381

Tabella 17: Stime dei coefficienti di squadre e covariate mediante lasso.

Le stime maggiori per il coefficiente di abilità delle squadre sono anche in questo caso in ordine: Chelsea, Manchester City, Arsenal e Manchester United.
La maggior parte delle squadre ha un coefficiente differente, fatta eccezione per Swansea e Stoke City, Crystal Palace e Sunderland, e un gruppo di squadre di bassa classifica che presentano il valore peggiore. Esse sono: Leicester City, Aston Villa, Hull City, Burnley e Queens Park Rangers, che hanno terminato il campionato rispettivamente 14°, 17°, 18°, 19° e 20°.
Il coefficiente relativo all'abilità delle squadre ha ancora un peso piuttosto alto. La parte più interessante da analizzare attraverso le variabili esplicative potrebbe quindi riguardare maggiormente le squadre che hanno terminato circa a metà classifica.
Per esempio ci si potrebbe chiedere come mai Crystal Palace e Sunderland che hanno lo stesso coefficiente di abilità hanno però terminato il campionato con 10 punti di differenza e rispettivamente 10° e 16° rischiando quest'ultima addirittura la retrocessione.
Altro aspetto curioso potrebbe riguardare l'Everton arrivato 11° nonostante

abbia un coefficiente di abilità praticamente identico al gruppo di squadre che chiude la classifica.

La variabile che influisce maggiormente sulla probabilità di vittoria in una partita sembra essere il fattore campo che viene stimato costante per tutte le squadre e pari a 0.402.
Altre variabili con coefficienti relativamente alti sono il numero di tiri, tiri in porta, tiri da destra, il numero di tiri bloccati dal difensore, i tiri negli angoli alti e i tiri negli angoli bassi.
In particolare per queste ultime due variabili vengono capovolti i risultati emersi nell'analisi svolta nel Capitolo 2. Ora in particolare si ha che tirare negli angoli bassi della porta sembra essere più efficace che non mirare a quelli alti.
A conferma di quanto emerso in Tabella 12 nel modello Bradley-Terry classico, la probabilità di vittoria aumenta non solo col numero di tiri, ma anche con la loro precisione ed in particolare se essi sono diretti in porta.
Tirare dal lato sinistro dell'area si conferma meno efficace, così come i tiri da fuori area e i tiri da posizioni difficili (fatta eccezione per il Southampton).
L'unica squadra, inoltre, che ha un effetto decisamente negativo per i tiri da posizioni difficili è il QPR. Quest'osservazione si può collegare a quanto visto precedentemente quando si diceva che era una squadra che tirava tanto ma non vinceva. Probabilmente questo è dovuto ad un abuso dei tiri da posizioni complicate e che risultano ovviamente meno efficaci.

La stima dell'effetto del numero di tiri su punizione è nulla o leggermente negativa per tutte le squadre e pari a -0.039 tranne che per il Tottenham per la quale risulta pari a 0.319 e per il Leicester City per cui risulta pari a 0.244.
Il valore di questi parametri dipende probabilmente dalla presenza di "specialisti" in quelle due squadre che hanno saputo far valere le proprie qualità, più che da capacità collettive della squadra.
Anche in questo caso è interessante fare un confronto con i coefficienti in Tabella 12 dove la stima (costante per tutte le squadre) era stata assegnata praticamente nulla (-0.001) ma con uno standard error decisamente grande

(0.198). Già da questa prima analisi infatti si poteva pensare che tale coefficiente potesse variare molto da squadra a squadra, motivo per cui non è stato escluso dal modello.
Dall'analisi emerge anche il fatto che un tiro qualora venga assistito o costruito tramite una giocata personale non è particolarmente rilevante al fine dell'esito, probabilmente dipende più dalle caratteristiche del singolo giocatore.
Costruire il proprio gioco tramite passaggi filtranti o cross sembra anch'essa essere una strategia non ideale per la maggior parte delle squadre. Fa appunto eccezione la formazione del Swansea.
Probabilmente associato al numero di cross c'è anche il numero di corner che presenta anch'esso un coefficiente negativo per tutte le squadre anche se in modo differente.
Il numero di falli e di sanzioni disciplinari ha un peso modesto e negativo per la maggior parte delle squadre. Questo anche probabilmente per il fatto che una squadra meno tempo passa ad attaccare, più tempo passa nella metà campo difensiva e quindi più è incline a commettere falli e ricevere provvedimenti disciplinari.
Le ammonizioni ricevute nel primo tempo hanno chiaramente un peso superiore come si poteva immaginare. Infatti sia il coefficiente relativo alle sanzioni ricevute in tutta la partita, sia quello relativo solamente al primo tempo, sono entrambi negativi per la maggior parte delle squadre e il secondo ha un coefficiente maggiore in valore assoluto.
Il fatto più curioso riguarda ancora il Burnley che presenta un coefficiente positivo per questa variabile.
Continuando l'analisi della fase difensiva è interessante osservare anche il coefficiente relativo al fuorigioco che è negativo, mentre la capacità di bloccare i tiri è un aspetto decisamente utile al fine di incrementare la probabilità di vittoria, soprattutto per la squadra QPR.
Il numero di tiri sul palo si conferma infine una variabile di difficile interpretazione e infatti presenta stime molto basse in valore assoluto e differenti tra le squadre. Per questa variabile vengono addirittura formati 6 clusters infatti.

L'indicazione che si può trarre da questo coefficiente potrebbe comunque essere legata più ad un aspetto motivazionale della squadra. Alcuni giocatori infatti potrebbero reagire negativamente per l'opportunità sprecata, altri potrebbero invece motivarsi.
La maggior parte delle indicazioni date da queste stime comunque sembra in linea con quanto osservato nella prima analisi svolta nel Capitolo 2, tuttavia come evidenziato dai numerosi clusters che si formano per alcune variabili c'è una grande variabilità nelle stime in base alla squadra che gioca.
Infine chiaramente anche i parametri di soglia hanno assunto stime differenti. In particolare nel modello Bradley-Terry con il lasso si ha che $\hat{\delta}_1 = -0.308$.

Dall'analisi effettuata pertanto emerge che non tutte le squadre ottengono gli stessi risultati da determinate situazioni. Per alcune variabili addirittura il coefficiente associato è negativo per alcune squadre e positivo per altre.
Questo aspetto viene evidenziato ancora più chiaramente nelle Figure 8-13.
I grafici mostrano, per alcune variabili esplicative, come cambiano le stime dei coefficienti associati al variare del parametro di regolazione espresso in scala logaritmica.
Per valori di ζ più vicini allo zero, i coefficienti associati ad una stessa variabile risultano più diversi tra loro, fino al caso più radicale in cui quando $\zeta = 0$ si ha che per ogni squadra viene riportato un valore differente.
Al contrario invece quando la penalità è grande le stime si stabilizzano costanti per ogni squadra. Vi è quindi una relazione inversa tra numero di coefficienti del modello ed entità della penalizzazione.
Il parametro di tuning è stato stimato cercando di trovare il compromesso ideale tenendo conto di questi aspetti.
La retta verticale tratteggiata nei grafici indica il valore ottimale del parametro di regolazione del modello, stabilito attraverso una convalida incrociata con 10 gruppi come spiegato nella sezione precedente.
Tenendo conto delle penalità utilizzate, il valore ottimale per il parametro di tuning descritto in Formula (4.1) è pari a $\hat{\zeta} = 2.356$.

In Figura 8 viene mostrato il percorso relativo alle stime delle abilità delle

squadre per le quali si formano come visto alcuni cluster per le formazioni di medio e bassa classifica.
Nelle Figure 9 e 10 sono rappresentati i percorsi relativi alle ammonizioni date nei primi 45 minuti di gioco e il numero di tiri bloccati. In particolare si può notare come il coefficiente associato alle squadre Burnley e Stoke City in Figura 9 e la squadra QPR in Figura 10 siano diversi da quelli di tutte le altre.
In Figura 11 viene proposto il percorso relativo al coefficiente della variabile associata al numero di tiri direttamente da calcio di punizione, in cui si evidenziano due squadre (Tottenham e Leicester City) che si discostano nettamente dal trend comune come già accennato.
In Figura 12 invece viene rappresentato l'andamento dei coefficienti per la variabile "tiri da posizioni difficili" per la quale si formano 6 clusters nel modello ottimo. In particolare si nota la curva relativa alla squadra Southampton in alto e quella relativa alla squadra QPR in basso.
Infine, in Figura 13 è stato rappresentato il percorso relativo al coefficiente associato al numero di cross, stimato negativo per tutte le squadre eccetto il Queens Park Rangers per cui è nullo.

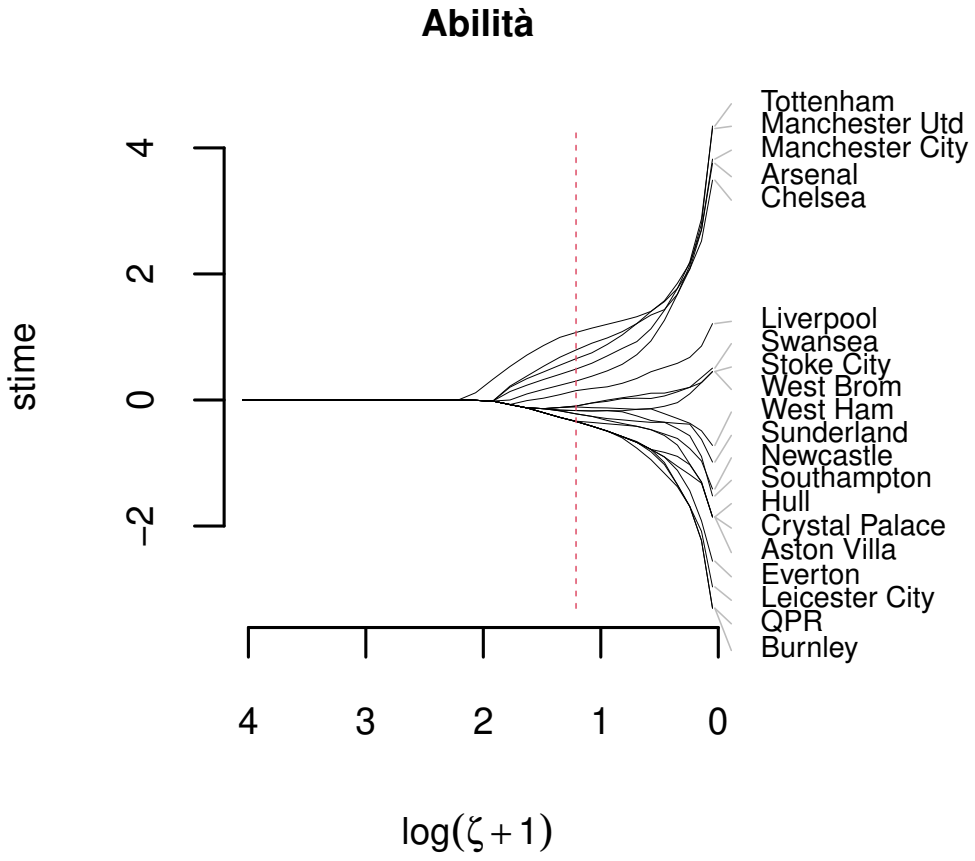

Figura 8: Percorso dei coefficienti relativi alle abilità delle squadre.

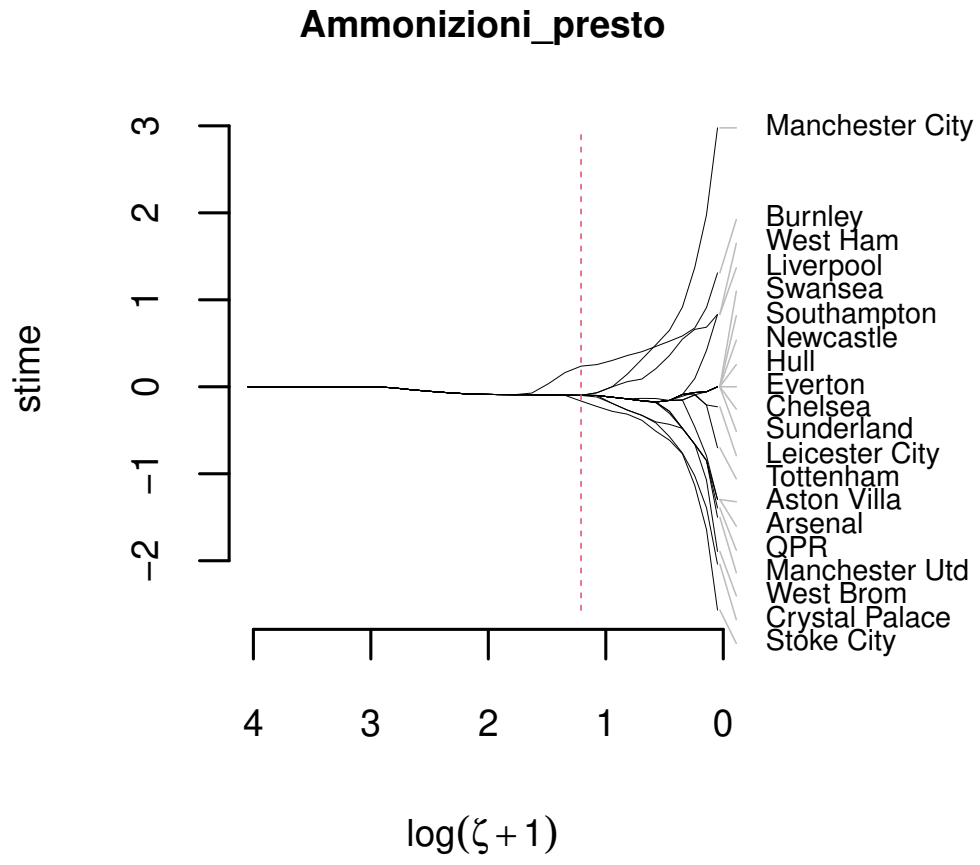

Figura 9: Percorso dei coefficienti relativi alle ammonizioni nei primi 45 minuti.

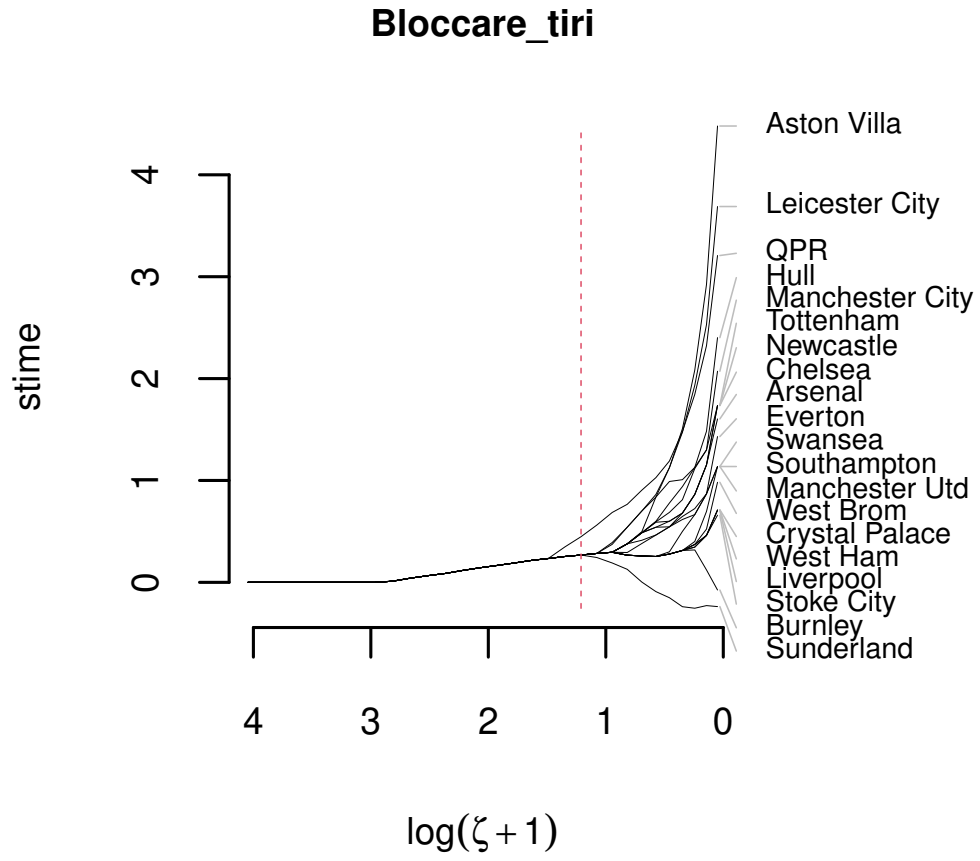

Figura 10: Percorso dei coefficienti relativi ai tiri bloccati dalla difesa.

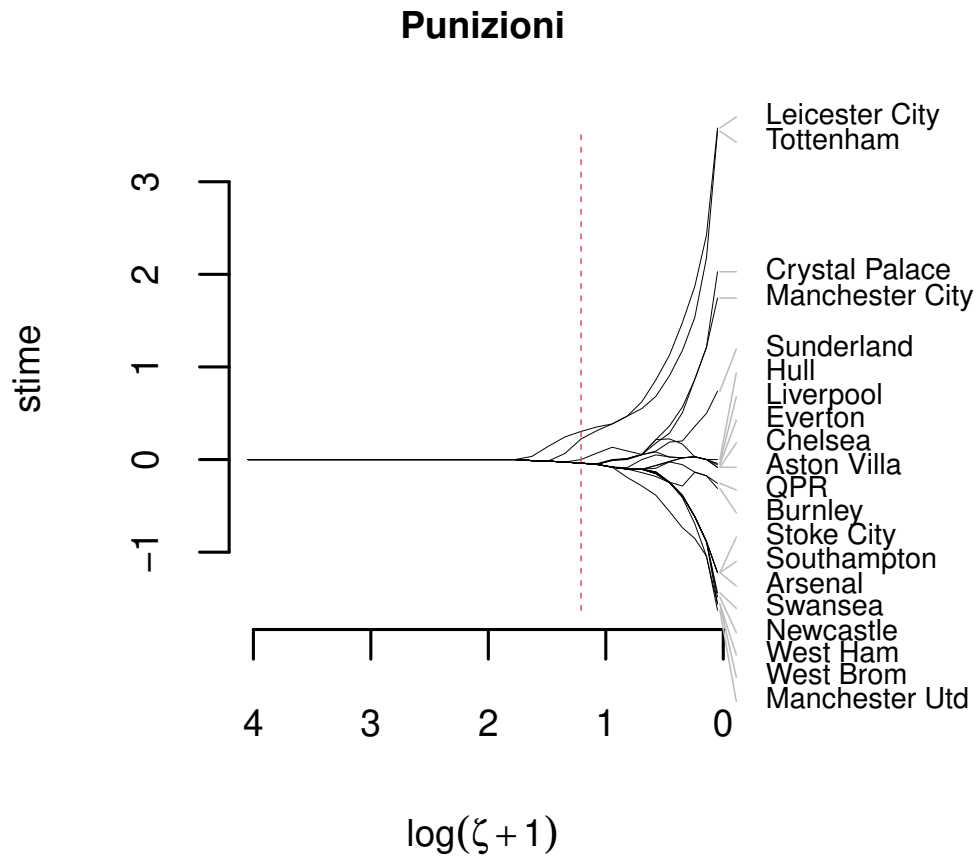

Figura 11: Percorso dei coefficienti relativi ai tiri su punizione.

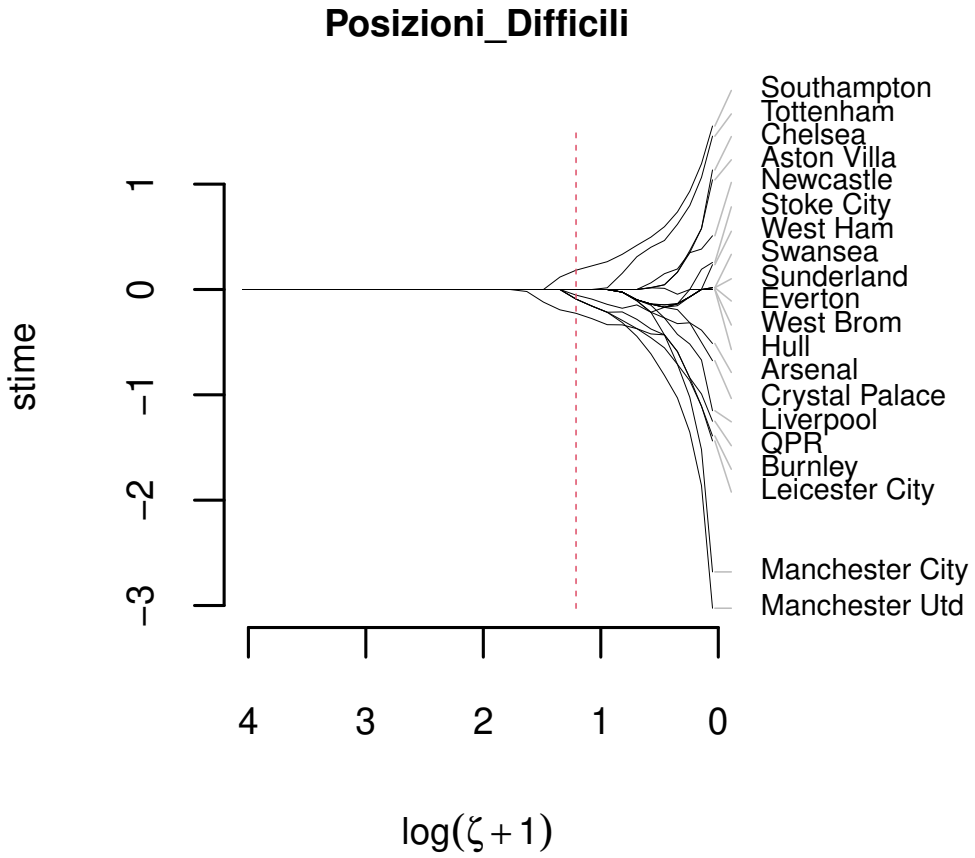

Figura 12: Percorso dei coefficienti relativi ai tiri da posizioni difficili.

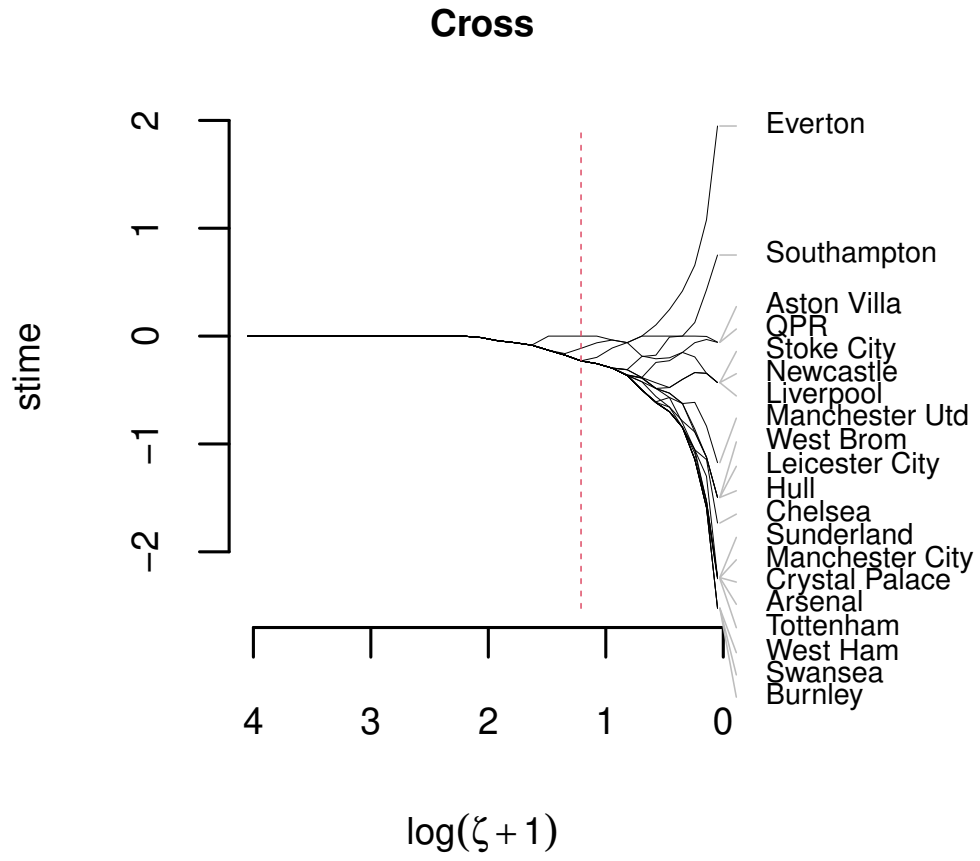

Figura 13: Percorso dei coefficienti relativi ai cross.

5 Modello dinamico e penalizzato

In quest'ultimo capitolo viene infine stimato il modello Bradley-Terry in cui le abilità delle squadre vengono modellate attraverso un processo EWMA, mentre i coefficienti relativi all'effetto del giocare in casa e alle variabili esplicative vengono stimati utilizzando una penalizzazione di tipo lasso.
La scelta è stata quella di utilizzare un unico processo temporale che tenesse conto dei punti conquistati da ogni squadra per tutte le partite giocate sia in casa che in trasferta.
La stima del modello infatti è piuttosto onerosa a livello computazionale, pertanto si è deciso di rendere le cose più agevoli stimando un unico processo temporale e quindi un unico parametro di tuning associato λ, oltre ovviamente a ristimare il parametro ζ relativo alla componente lasso.
I parametri di penalizzazione sono stati stimati tramite una griglia di valori selezionando quelli con cui il modello sembrava adattarsi meglio ai dati tramite il criterio del Ranked Probability Score.
I valori così stimati sono $\hat{\lambda} = 0.50$ e $\hat{\zeta} = 4.889$, mentre il parametro di soglia è pari a $\hat{\delta}_1$ = -0.600.
I valori dei parametri di tuning risultano essere decisamente più elevati rispetto a quelli stimati nei capitoli precedenti. Ciò comporta pertanto una maggiore penalizzazione dei coefficienti ponendo molti di essi pari a zero.
A differenza del modello Bradley-Terry con lasso stimato nel Capitolo 4, in questo caso viene fatta una selezione delle variabili decisamente più drastica ed in particolare non vi è nessuna variabile esplicativa per cui si formano clusters di squadre.
Anche il coefficiente di abilità stimato tramite il processo EWMA risulta dare un'unica stima di abilità per tutte le squadre al netto delle covariate.
Di conseguenza le probabilità di vittoria, sconfitta o pareggio per ogni partita vengono calcolate tramite il modello appena proposto unicamente sulla base dei valori delle covariate e dell'effetto dovuto al fattore campo che viene stimato positivo e costante per tutte le squadre come nel modello descritto nel Capitolo 4.
Esse sono quindi calcolate tramite la seguente equazione che offre una leg-

gera modifica alla Formula (3.8) per tenere conto dell'apporto dato dalle p_1 covariate stimate diverse da zero.
La probabilità condizionata dell'esito dell'i-esima partita dati i risultati degli incontri precedenti è descritta come

$$pr(Y_i \leq y_i | Y_{i-1} = y_{i-1}, ..., Y_1 = y_1; \theta) = \\ = \frac{\exp\left\{\delta_{y_i} + \eta + \beta z_{h_i}(t_i;\lambda) - \beta z_{v_i}(t_i;\lambda) + \sum_{j=1}^{p_1} \gamma_j x_{i,j}\right\}}{1 + \exp\left\{\delta_{y_i} + \eta + \beta z_{h_i}(t_i;\lambda) - \beta z_{v_i}(t_i;\lambda) + \sum_{j=1}^{p_1} \gamma_j x_{i,j}\right\}}, \tag{5.1}$$

dove i coefficienti γ_j sono costanti per ogni squadra poichè non si formano clusters per nessuna covariata.

Dopo aver stimato 11 coefficienti relativi ad alcune variabili esplicative pari a zero, il modello ottenuto utilizza solamente 12 coefficienti (un parametro η, un β e 10 γ_j).
Il calo relativo al numero di coefficienti perciò è veramente notevole se confrontato per esempio col modello Bradley-Terry formulato nel Capitolo 2 in cui si avevano 22 parametri da stimare, o rispetto addirittura ai 71 coefficienti del modello proposto nel Capitolo 4.
Nel modello stimato unicamente tramite lasso infatti si erano formati molti clusters di squadre ed erano state eliminate dal modello solamente 2 covariate delle 21 totali, ossia le variabili relative ai tiri effettuati di testa e ai tiri assistiti, che anche in questo modello vengono confermati ininfluenti nell'esito dell'incontro e pertanto i coefficienti associati vengono posti pari a zero.
Le stime relative a tutti i parametri del modello Bradley-Terry dinamico penalizzato sono riportate in Tabella 18.

Dall'analisi finale sul modello segue pertanto che la stima positiva di β

Variabili	**Stime**
β	0.337
Casa	0.185
Tiri	0.143
Tiri in porta	0.172
Pali colpiti	0.000
Tiri di testa	0.000
Tiri dal limite	0.000
Tiri assistiti	0.000
Passaggi filtranti	0.000
Corner	-0.237
Falli	0.000
Ammonizioni	0.000
Offside	-0.161
Tiri dal centro	0.000
Tiri da sinistra	0.000
Tiri da destra	0.104
Tiri su punizione	0.000
Tiri posizioni difficili	0.000
Cross	-0.071
Ammonizioni primo tempo	-0.105
Tiri bloccati	0.182
Tiri angoli alti	0.106
Tiri angoli bassi	0.287

Tabella 18: Coefficienti stimati per il modello dinamico e penalizzato.

suggerisce la presenza di un trend temporale che determina le abilità delle squadre ed in particolare rispecchia l'importanza dello stato di forma di una squadra prima di una partita, sia che essa giochi nel proprio stadio, sia che giochi in trasferta.
Infatti avere ottenuto buoni risultati negli incontri precedenti, aumenta la probabilità di vincere la partita successiva.
La stima di β risulta però chiaramente inferiore rispetto a quella stimata nei capitoli precedenti per i modelli dinamici in cui non si teneva conto dell'apporto delle covariate.
Lo standard error della stima $\hat{\beta}$ pari a 0.090 suggerisce peraltro una significatività importante di tale parametro da un punto di vista statistico.
Il coefficiente η relativo all'effetto del giocare in casa risulta anche in questo caso diverso da zero e chiaramente positivo. Giocare nel proprio stadio pertanto aiuta ad ottenere risultati migliori anche al netto dell'effetto stimato da β e dalle esplicative.
Infine vengono confermati anche la maggior parte dei risultati ottenuti sull'effetto delle covariate. Il numero di tiri, di tiri in porta e di tiri negli angoli alti e bassi presentano tutti stime decisamente positive, così come la capacità dei difensori di bloccare i tiri avversari.
Anche in questo caso inoltre emerge come sia più conveniente tirare da destra rispetto che dal centro o da sinistra, oltre alla conferma di risultati migliori quando si mira agli angoli bassi della porta rispetto a quelli alti.
I coefficienti negativi al netto delle altre covariate sono invece associati al numero di cross, di calci d'angolo, il numero di fuorigioco concessi e le ammonizioni ricevute nei primi 45 minuti di gioco.
Tutte le stime osservate per questo modello sono quindi in linea con quanto stimato nei modelli precedenti.

Una volta confrontati i valori dei coefficienti può essere sicuramente d'interesse paragonare i valori del Ranked Probability Score per i modelli utilizzati ed osservare quanto si discostano tra loro.
La stima del RPS per il modello analizzato in questo capitolo è pari a 0.374 quindi inferiore ad entrambi i valori ottenuti con i modelli dinamici proposti

ed inferiore anche al valore calcolato per il modello Bradley-Terry statico che non utilizza le covariate.
L'unico che presenta una stima migliore in termini di Ranked Probability Score è il modello statico con l'aggiunta delle informazioni date dalle 21 variabili esplicative presentato nel Capitolo 2.
Tuttavia il vantaggio rispetto a tale modello in termini di interpretazione è sicuramente rilevante poichè in questo caso vengono utilizzati solamente 12 coefficienti a fronte dei 42 stimati nel modello iniziale.

Conclusioni

In questo lavoro si è parlato di molti aspetti riguardanti il modello Bradley-Terry e le sue applicazioni nel mondo del calcio.
Per svolgere le analisi è stato utilizzato un dataset disponibile su Kaggle in cui sono disponibili informazioni riguardo gli eventi accaduti durante ogni partita del campionato di Premier League 2014-2015 (*Football Events* 2017).
Inizialmente è stato stimato un modello per cogliere l'effetto del giocare in casa e l'effetto dell'abilità di ogni squadra in termini di probabilità di vittoria.
In seguito sono state aggiunte 21 covariate relative ad eventi tipici di una partita di calcio per comprendere quali aspetti del gioco influenzino maggiormente l'esito di una gara.
Nel prosieguo dell'analisi sono state proposte anche alcune estensioni sia per cogliere l'andamento delle abilità delle squadre all'interno della stagione, sia per cogliere quali variabili effettivamente influiscono nei risultati e come variano tra le squadre.
Quanto emerso dalle analisi svolte è che l'andamento delle abilità delle squadre in casa e in trasferta nel corso del campionato non può essere ritenuto costante.
Esso infatti è stato modellato utilizzando un processo EWMA (Exponentially Weighted Moving Average) definito nel Capitolo 3.
Nel corso del lavoro sono stati proposti due tipi di modelli dinamici differenti. In particolare un'alternativa prevede l'utilizzo di due processi distinti a seconda che la squadra avesse giocato la partita in casa o in trasferta. La seconda proposta invece utilizza un unico processo per tutte le partite del campionato.
In questo modo si è riusciti non solo a cogliere l'aspetto dinamico, ma anche a ridurre notevolmente il numero di parametri relativi alle abilità delle squadre.
Utilizzando un processo temporale pertanto si è potuto cogliere la dipendenza che intercorre tra due partite ravvicinate e si è rivelato un aspetto non trascurabile in un'analisi riguardante un'intera stagione sportiva.

Nel Capitolo 4 invece è stata utilizzata una procedura di tipo lasso per selezionare le variabili importanti e sono state analizzate quelle che influivano maggiormente sull'esito dell'incontro.
Una volta stimato il parametro di tuning tramite convalida incrociata, sono state calcolate le stime dei coefficienti associati alle covariate descritte nel Capitolo 1. Alcune di esse sono state eliminate dalla procedura lasso applicata al modello perchè ritenute poco informative a fini statistici.
Le variabili esplicative quindi maggiormente correlate in senso positivo con la variabile risposta sono principalmente l'effetto del giocare in casa, il numero di tiri e in particolare quelli terminati negli angoli alti e bassi della porta, oltre alla capacità di intercettare i tiri degli avversari.
Interessante può essere anche il fatto che l'effetto attribuito al fattore campo è costante per tutte le squadre, nonostante le differenze sostanziali emerse in Tabella 10.
Gli effetti che invece maggiormente penalizzano l'esito dell'incontro, al netto delle altre esplicative, sembrano essere il numero di calci d'angolo, di cross e la tendenza a voler mandare in fuorigioco i giocatori avversari. Per alcune squadre inoltre ha un effetto particolarmente negativo anche il numero di passaggi filtranti.
Attraverso questa procedura infatti si sono anche formati clusters di squadre per cui gli effetti di alcune covariate sono diversi.
Un'analisi interessante per esempio riguarda i coefficienti associati al numero di tiri su calcio di punizione e il numero di ammonizioni nel primo tempo, negativi per tutte le squadre tranne alcune eccezioni.

Infine nel Capitolo 5 è stato stimato il modello completo in cui c'è sia la parte di tipo lasso, sia l'aspetto dinamico per modellare le abilità delle squadre.
In particolare per quest'ultimo aspetto si è scelto di utilizzare un unico processo EWMA per ridurre l'onere computazionale.
Infatti non è stato utilizzato il modello dinamico proposto inizialmente con due trend separati seppure esso desse risultati migliori in termini di Ranked

Probability Score. La decisione è principalmente dovuta a fini pratici.
La costruzione del nuovo modello infatti in quel modo avrebbe previsto la stima di tre diversi parametri di regolarizzazione: due per il processo EWMA e uno per il lasso.
Come descritto nel Capitolo 4, per stimare il modello ottimale con la penalizzazione di tipo lasso, si è usata una convalida incrociata con 10 gruppi. Sono state quindi calcolate le stime per 30 diversi valori di ζ ed in seguito sono stati confrontati i risultati ottenuti in termini di Ranked Probability Score.
Il tempo necessario per la stima del modello è stato calcolato tramite la funzione *system.time()* disponibile in R ed è pari a circa 3 ore. Il tempo necessario alla CPU per eseguire le istruzioni dell'utente è pari a circa 45 minuti, mentre il tempo necessario alla CPU per eseguire i comandi per conto del sistema è pari a circa 5 minuti.
Per la stima del modello proposto nel Capitolo 5, questa procedura deve essere ripetuta per ogni valore di λ. Il carico computazionale pertanto crescerebbe notevolmente se si stimasse il modello utilizzando le abilità dinamiche calcolate tramite due processi EWMA perchè bisognerebbe calcolare le stime per diverse combinazioni di λ_1 e λ_2.
Tuttavia questo potrebbe essere uno sviluppo futuro per quest'analisi.
La selezione delle covariate nel modello descritto nel Capitolo 5 è stata molto più drastica, andando ad eliminare 11 variabili delle 21 analizzate.
I risultati ottenuti con tale modello inoltre forniscono una stima costante delle abilità, del fattore campo e di tutti i coefficienti relativi alle esplicative per ogni squadra.
Le covariate che hanno un effetto significativo nel modello proposto sono pertanto il numero di tiri, tiri in porta, calci d'angolo, fuorigioco, tiri da destra, cross, ammonizioni nel primo tempo, tiri bloccati dal difensore, tiri negli angoli alti e tiri negli angoli bassi, oltre chiaramente al fattore campo.

L'analisi proposta in questo lavoro può essere integrata ed approfondita sicuramente con molti spunti interessanti.
Sarebbe utile per esempio ripetere le analisi invece che con una variabile risposta con tre modalità, utilizzandone una con cinque considerando per

esempio se la partita è finita con più o meno di due gol di scarto.
Altre estensioni ovviamente potrebbero essere fatte anche per gli altri campionati principali (Serie A, La Liga, Ligue 1, Bundesliga) o per alcuni campionati magari più equilibrati in cui per esempio si risenta in maniera minore magari dei dislivelli di "abilità" tra le squadre.
Ovviamente il modello si può adattare facilmente anche ad altri sport facendo le opportune modifiche.
Sarebbe sicuramente interessante osservare anche quale sarebbe il modello in altre stagioni sportive più recenti, visto anche il ruolo sempre maggiore che sembra avere l'aspetto finanziario per alcuni club creando divari sempre maggiori tra alcune squadre.
Di particolare interesse potrebbero essere le ultime stagioni in cui a causa dell'epidemia di Covid-19, si sono giocate diverse partite a "stadi chiusi" e quindi senza pubblico, per osservare come cambia l'effetto dovuto al fattore campo.
Curioso sarebbe anche analizzare l'andamento temporale non solo all'interno della singola stagione, ma anche tra i vari anni per le squadre che in quell'arco temporale non vengono mai retrocesse e rimangono sempre in Premier League per esempio.
Un'altra proposta da cui si potrebbero trarre conclusioni leggermente differenti per esempio potrebbe riguardare una diversa inizializzazione dei parametri r_{h_i} e r_{v_i} da cui poi modellare il processo sui dati del'intero campionato. Tuttavia da un'analisi effettuata per il girone di ritorno, basandosi esclusivamente sui valori ottenuti nel girone d'andata, sembrerebbe che le differenze non siano poi così marcate.
Oltre all'aspetto temporale, potrebbe essere decisivo anche reperire informazioni su altre variabili chiave ma forse di più difficile rilevazione come la percentuale di possesso palla o il numero di passaggi completati per ogni partita, così come il numero di contrasti vinti o di dribbling riusciti.
Un'ultima possibilità, applicabile però solamente utilizzando un calcolatore più potente, potrebbe essere anche quella di costruire degli intervalli di confidenza per i parametri stimati dal modello Bradley-Terry con lasso. Per farlo bisogna utilizzare un approccio di tipo bootstrap sfruttando l'opzione

disponibile nel pacchetto BTLLasso presente per il software R.
Come visto pertanto ci sono diverse possibili applicazioni e sviluppi per questo lavoro, garantendo la versatilità anche del modello utilizzato e le molteplici sfaccettature che un'analisi di questo tipo può suggerire.
L'obiettivo di questo elaborato, come chiarito già nel primo capitolo, riguarda unicamente l'aspetto esplorativo basato sull'analisi di quanto accade durante le partite. Nulla toglie però che questo tipo di approccio, con le dovute estensioni del caso, possa venire utilizzato anche a scopo previsivo.

Bibliografia

Barry, D. e J.A. Hartigan (1993). *Choice models for predicting divisional winners in major league baseball*. In: *Journal of the American Statistical Association* 88.423, pp. 766–774.

Bradley, R.A. e M.E. Terry (1952). *Rank Analysis of Incomplete Block Designs, I: The Method of Pair Comparisons*. In: *Biometrika* 39.3, pp. 324–345.

Browne, M.W. (2000). *Cross-Validation Methods*. In: *Journal of Mathematical Psychology* 44.1, pp. 108–132.

Cattelan, M., C. Varin e D. Firth (2013). *Dynamic Bradley-Terry modelling of sports tournaments*. In: *Journal of the Royal Statistical Society* 62.1, pp. 135–150.

Clarke, S.R. e J.M. Norman (1995). *Home Ground Advantage of Individual Clubs in English Soccer*. In: *Wiley for the Royal Statistical Society* 44.4, pp. 509–521.

Czado, C., T. Gneiting e L. Held (2009). *Predictive model assessment for count data*. In: *Biometrics* 65.4, pp. 1254–1261.

Firth, D. e R.X. de Menezes (2004). *Quasi-variances*. In: *Biometrika* 91.1, pp. 65–80.

Goddard, J. e I. Asimakopoulos (2004). *Forecasting football results and the efficiency of fixed-odds-betting*. In: *Journal of Forecasting* 23.1, pp. 51–66.

Henderson, A.R. (2005). *The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data*. In: *Clinica Chimica Acta* 359.1-2, pp. 1–26.

Hugall, A.F. e M.S.Y. Lee (2003). *Partitioned Likelihood Support and the Evaluation of Data Set Conflict*. In: *Systematic Biology* 52.1, pp. 15–22.

Kendall, M.G. (1938). *A new measure of rank correlation*. In: *Biometrika* 30.1-2, pp. 81–93.

Koning, R.H. (2000). *Balance in competitions in Dutch soccer*. In: *Statistician* 49.3, pp. 419–431.

Kuk, A.Y.C (1995). *Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players*. In: *Statistician* 44.4, pp. 523–528.

Murphy, S.A. e A.W. Van Der Vaart (2000). *On Profile Likelihood*. In: *Journal of the American Statistical Association* 95.450, pp. 449–465.

Perperoglou, A. et al. (2019). *A review of spline function procedures in R*. In: *BMC Medical Research Methodology* 19.46, pp. 19–46.

Schauberg, G. e G. Tutz (2019). *BTLLasso: A Common Framework and Software Package for the Inclusion and Selection of Covariates in Bradley-Terry Models*. In: *Journal of Statistical Software* 88.9, pp. 1–29.

Spearman, C. (1987). *The proof and measurement of association between two things*. In: *American Journal of Psychology* 100.3-4, pp. 441–471.

Tibshirani, R. (1996). *Regression Shrinkage and Selection Via the Lasso*. In: *Journal of the Royal Statistical Society* 58.1, pp. 267–288.

Sitografia

Colgados por el futbol (2014). URL: https://colgadosporelfutbol.com/it/que-liga-es-la-mejor-del-mundo/.

Corriere comunicazioni (2020). URL: https://www.corrierecomunicazioni.it/lavoro-carriere/competenze/il-calcio-e-sempre-piu-digitale-arriva-il-football-data-analyst/.

Diretta (2015). URL: https://www.diretta.it/calcio/inghilterra/fa-cup-2014-2015/.

Flashscore (2015). URL: https://www.flashscore.it/calcio/inghilterra/fa-cup-2014-2015/.

Football Events (2017). URL: https://www.kaggle.com/datasets/secareanualin/football-events?select=events.csv.

Guerin sportivo (2015). URL: https://www.guerinsportivo.it/pagine-gialle/classifica/premier-league-8-2014.

Oubliette Magazine (2022). URL: https://oubliettemagazine.com/2022/04/13/chi-tira-di-piu-in-porta-vince-sempre-analisi-della-statistica-nella-serie-a-in-corso/.

R software (2022). URL: https://www.r-project.org/.

Ranking UEFA (2022). URL: https://it.uefa.com/nationalassociations/uefarankings/club/#/yr/2022.

Rivista undici (2019). URL: https://www.rivistaundici.com/2019/04/03/calcio-statistiche/.

Transfermarkt (2015). URL: https://www.transfermarkt.it/premier-league/transfers/wettbewerb/GB1/plus/?saison_id=2014&s_w=w&leihe=1&intern=0&intern=1.

Transfermarkt (2015). URL: https://www.transfermarkt.it/premier-league/trainerwechsel/wettbewerb/GB1/plus/?saison_id=2014.

UEFA Champions League (2015). URL: https://it.uefa.com/uefachampionsleague/history/seasons/2015/matches/.