Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione Finale

Analisi statistica nella pallamano: prestazioni dei giocatori e loro peso nell'economia di una partita

Relatore Prof. Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureando Alberto Lollo
Matricola 2008155

Anno Accademico 2022/2023

Indice

Introduzione

L'analisi statistica si sta diffondendo sempre di più negli ultimi anni nel mondo dello sport e costituisce un importante strumento per allenatori e giocatori volto a migliorare le prestazioni singolari e collettive e, quindi, incrementare la probabilità di vittoria di una partita. Tale aumento d'interesse si può riscontrare anche in sport meno popolari come la pallamano, come mostrato dalle ricerche svolte da Prieto et al. (2015a) e Saavedra (2018). La pallamano, in particolare, è uno sport molto complesso caratterizzato da un notevole numero di dettagli riguardanti il gioco individuale e corale che possono risultare decisivi al termine di una partita.

Obiettivo principale dell'analisi presentata in questo elaborato è, quindi, quello di comprendere quali siano gli aspetti del gioco che influenzano maggiormente l'esito di una partita e, successivamente, cercare di individuare quali ruoli dei giocatori risultano più determinanti sempre rispetto al risultato di una partita. Analisi statistiche di questo tipo, riguardanti il peso di determinate variabili sull'esito di una partita, sono già state svolte negli ultimi anni da autori che hanno preso in considerazione dati provenienti da manifestazioni internazionali come Olimpiadi (Saavedra et al., 2017) e campionati mondiali (Daza et al., 2017). Nel presente elaborato, invece, vengono presi in esame gli indici prestazionali dei giocatori nelle singole partite della stagione 2021/22 della *Liqui Moly HBL* (primo campionato tedesco), focalizzandosi sull'*Handball Performance Index* (HPI), introdotto dal campionato tedesco a partire dalla stagione 2020/21 e il quale costituisce una misura complessiva della prestazione di un giocatore in una determinata partita.

L'elaborato è strutturato come segue. Innanzitutto, si presenta lo sport della pallamano, non essendo uno sport tra i più diffusi e popolari. In particolare, dopo aver fornito alcuni cenni storici e una breve descrizione del gioco generale e del regolamento, si espone il progresso della ricerca statistica nella pallamano che ha caratterizzato gli ultimi anni.

Nel secondo capitolo si svolge, dopo una breve descrizione del dataset utilizzato, un'analisi esplorativa, focalizzata in primis sull'indice HPI, considerato sia dal punto di vista medio di squadra che dal punto di vista del singolo giocatore, e successivamente sui diversi aspetti del gioco, cercando quelli che presentano una maggiore differenza a seconda del risultato finale di una partita.

Infine, negli ultimi due capitoli si conduce un'analisi volta a cercare quali ruoli incidono maggiormente sull'esito di una partita, servendosi dei valori medi per posizione dell'indice HPI. Nello specifico, tale analisi si svolge considerando due diversi tipi di modelli: modello per variabili risposta categoriali ordinali e modello di Bradley-Terry. Nell'analisi svolta con quest'ultimo modello, inoltre, si stima un coefficiente di "abilità" per ogni squadra.

L'analisi è stata svolta utilizzando il software statistico R, nella sua versione 4.3.1.

Capitolo 1

Pallamano

1.1 Introduzione

La pallamano è uno sport olimpico di squadra caratterizzato da un elevato dinamismo dei singoli giocatori e da un ritmo di gioco incessante. Esso è uno sport di contatto in cui si affrontano due squadre composte da sedici giocatori ciascuna, di cui sette in campo, e che abbina peculiarità tipiche di altri sport come la pallacanestro e il calcio: obiettivo del gioco è, infatti, segnare il maggior numero di gol nella porta avversaria lanciando il pallone con le mani.

1.2 Storia

1.2.1 Origini

Diversi paesi rivendicano la paternità della pallamano avendo, tra la fine del diciannovesimo e l'inizio del ventesimo secolo, creato e introdotto attività simili ad essa in ambito scolastico. La versione moderna di questo sport, come riportato da Poto (2005) ha avuto uno sviluppo parallelo in paesi del centro Europa come la Cecoslovacchia (*"Hazena"*, ideata dagli insegnanti Josef Klenker e Václav Karas), la Danimarca (*"Handbold"* ideata da Holger Nielsen, medaglia di bronzo ai Giochi Olimpici di Atene nella scherma, ha avuto le sue prime partite sperimentali nel 1912) e la Germania (*"Raffballspiel"* ideata da Konrad Koch nel 1892), ma è a quest'ultima che viene attribuita la sua invenzione. In particolare, durante gli anni della prima guerra mondiale, Max Heiser diede origine al

gioco *"Torball"*, dal quale successivamente il professor Karl Schelenz creò l'*Handball*[1], il gioco che più si avvicina alla pallamano attuale, praticato su campi da calcio all'aperto e con undici giocatori per squadra.

1.2.2 Pallamano moderna

Dopo una comparsa alle Olimpiadi di Berlino del 1936, la versione del gioco della pallamano a undici giocatori venne utilizzata ufficialmente fino al campionato del mondo del 1966.

In contemporanea, a causa del clima rigido tipico dei paesi nordici si sviluppò quella che è la versione attuale della pallamano, giocata in campi indoor di dimensioni minori e con sette giocatori per squadra, per la quale si svolgevano manifestazioni internazionali separate.

Dopo essere stata inclusa ufficialmente tra gli sport olimpici ai Giochi del 1972 a Monaco di Baviera, ottenne fin da subito una discreta popolarità (Clanton & Dwight, 1997), la quale aumentò nel corso degli anni grazie all'evoluzione del gioco in termini di intensità e velocità. In particolare, è stato registrato un notevole incremento di praticanti negli ultimi quindici anni: dai circa 19 milioni registrati nel 2009 (Prieto et al., 2015a) agli oltre 30 milioni attuali secondo l'*International Handball Federation* (IHF, sito: https://www.ihf.info/marketing-homepage).

1.2.3 In Italia

In contrasto con quanto avvenuto nei Paesi circostanti, in Italia lo sviluppo della pallamano è stato decisamente più lento nel corso degli anni. Pioniere fu Aurelio Chiappero che promosse l'attività agonistica per la prima volta nel 1940 e successivamente nel 1945.

Dopo una paralisi del movimento avvenuta appena due anni dopo, la pallamano riprese ad essere introdotta nelle scuole soltanto negli anni Sessanta, fino alla costituzione della federazione nazionale (FIGH) avvenuta a fine 1969. Dopo essere entrata nel novero degli sport olimpici, venne riconosciuta dal CONI nel dicembre del 1974.

Lo sviluppo tardivo della disciplina si è tradotto negli anni in nessuna partecipazione alle Olimpiadi e in una sola partecipazione sia ai campionati mondiali (1997) che ai campionati europei (1998) da parte della sola nazionale maschile.

Ad oggi, dopo una fase di stallo registrata nei primi anni del nuovo millennio, la disciplina sta osservando un periodo di crescita, portando sempre più giocatori in campionati esteri ben più sviluppati e competitivi.

[1] Primo incontro ufficiale disputato il 13 settembre 1925

1.3 Regolamento

1.3.1 Campo da gioco

Il campo da gioco, rappresentato in Figura 1.1, ha forma rettangolare, misura 40 metri in lunghezza e 20 in larghezza e generalmente ha fondo in legno o in gomma. Al centro dei lati corti, detti anche linee di fondo, sono poste le porte, alte due metri e larghe tre, attorno alle quali sono delimitate le aree di porta tramite una linea continua posta a sei metri dalla porta stessa[2]. All'interno dell'area di porta è posta la linea di limite del portiere a quattro metri dal centro della linea di porta.

A tre metri dall'area di porta è disegnata una linea tratteggiata detta linea del tiro di punizione (o linea dei nove metri). Tra quest'ultima e l'area di porta, centrata, parallela alla porta e a sette metri da essa è posta una linea lunga un metro che costituisce la linea di tiro del rigore.

Infine, la linea mediana unisce i punti centrali dei lati lunghi del campo e al centro di essa vi è disegnato un cerchio (non riportato in Figura 1.1) di quattro metri di diametro che costituisce l'area del tiro d'inizio. Per ogni lato del campo rispetto alla linea mediana è tracciata, a quattro metri e mezzo da essa, la linea del cambio.

FIGURA 1.1: Campo da gioco regolare.

[2]La linea deve essere tracciata nel seguente modo:

- una linea della lunghezza di tre metri di fronte e parallela alla porta, dalla quale è distante sei metri
- due quarti di cerchio, ognuno del raggio di sei metri (misurato dall'angolo posteriore interno del palo), che congiungono la linea descritta al punto precedente con la linea di fondo;

1.3.2 Gioco e regole generali

Una partita, della durata di due tempi da 30 minuti effettivi ciascuno, vede affrontarsi due squadre composte ciascuna da sette giocatori di cui un portiere. Scopo del gioco, come già detto, è segnare nella porta avversaria, costruendo l'azione d'attacco per superare la squadra in difesa, generalmente schierata davanti all'area di porta.

I giocatori, al di fuori delle aree di porta, possono tenere la palla in mano compiendo non più di tre passi o sostando non più di tre secondi. In alternativa, possono passare la palla ad un compagno oppure palleggiare, senza però bloccare il palleggio per poi ricominciarlo (fallo di doppia). Ad essi non è concesso colpire la palla al di sotto del ginocchio e entrare nell'area di porta: in alternativa, al momento del tiro ad esempio, essi possono saltare dentro l'area liberandosi del pallone prima dell'atterraggio.

Se l'azione d'attacco, che non può andare oltre un determinato limite di tempo ("gioco passivo", deciso al momento dagli arbitri), termina con una rete, la squadra che la subisce deve rimettere in gioco il pallone dal cerchio di centrocampo, tenendo tutti i propri giocatori dietro la linea mediana. Se, invece, la palla termina sul fondo direttamente o dopo una deviazione del portiere il gioco riprende con una rimessa dall'area di porta. Nel caso in cui l'ultima deviazione fosse di un difensore, si riprende con un tiro d'angolo.

Per quanto riguarda i falli, molto frequenti vista la fisicità tipica dello sport, possono essere sanzionati in tre modi diversi:

- ammonizione;
- esclusione temporanea, della durata di due minuti in caso di fallo abbastanza grave;
- espulsione, in caso di fallo gravissimo oppure alla terza esclusione temporanea[3].

La penalità comminata può essere un tiro dai sette metri (o rigore), nel caso in cui il fallo blocchi una chiara occasione da gol, oppure una punizione, la quale viene battuta nel punto dove è stato commmesso il fallo oppure al di fuori della linea dei nove metri nel caso in cui il fallo fosse commesso tra essa e l'area di porta.

1.3.3 Ruoli di gioco

Tralasciando il ruolo del portiere, i restanti sei giocatori di campo si possono suddividere in tre gruppi distinti:

[3]Il giocatore non può più rientrare in campo, ma può essere sostituito da un'altro dopo due minuti di inferiorità numerica

- il primo formato dai due terzini (destro e sinistro), generalmente giocatori alti e potenti, e dal centrale. Durante l'azione d'attacco, organizzata dal centrale, si schierano al di fuori della linea dei nove metri dove iniziano la circolazione del pallone e da dove, grazie alla loro stazza, possono sorprendere la difesa anche con tiri dalla distanza;

- il secondo formato dalle due ali (sinistra e destra), giocatori di norma di stazza più piccola e dotati di una notevole agilità. Durante il gioco essi percorrono le fasce laterali e si posizionano agli angoli del campo in modo da allargare la difesa. Generalmente sono meno coinvolti nelle azioni offensive, nelle quali attendono che i terzini e il centrale liberino loro lo spazio necessario per andare a tirare in porta;

- l'ultimo formato dal solo pivot, giocatore molto robusto che si posiziona tra i giocatori in difesa cercando di creare spazi tra essi per portare al tiro il resto dei compagni di squadra.

1.4 Ricerca e analisi statistica nella pallamano

L'interesse scientifico nei confronti della pallamano è cresciuto, in particolar modo nell'ultimo decennio, parallelamente alla popolarità dello sport stesso. Nello specifico, dal 1900 al 2012 il *Web of Science* e dal 1950 al 2012 *MEDLINE* hanno raccolto un totale di 1054 articoli contenenti la parola "handball", di cui solo 373 aventi la pallamano come argomento principale (Prieto et al., 2015a). Invece, una ricerca analoga ha mostrato come il solo *Web of Science* dal 2013 ad Aprile 2018 abbia pubblicato ben 256 articoli incentrati sulla pallamano (Saavedra, 2018).

Oltre alla crescita in quanto a numeri, le stesse ricerche hanno registrato un cambio di tendenza rispetto ai temi trattati dagli articoli: la prima analisi ha mostrato come i temi principalmente trattati fossero gli infortuni (26.54% degli articoli) e le capacità fisiche (17.96%), con solo il 6.17% degli articoli concentrato sulla performance e le variabili di successo; l'analisi più recente, invece, ha mostrato una crescita del numero di pubblicazioni relative a quest'ultimo tema (12.69%), risultato il secondo più trattato dietro solo alle capacità fisiche.

Tale crescita di interesse riguardo all'analisi delle performance è stata resa possibile nell'ultimo ventennio dal parallelo sviluppo tecnologico, il quale ha portato i tecnici a ottenere preziosi indicatori volti a valutare le prestazioni dei singoli e della squadra (Prieto et al., 2015b).

1.4.1 Handball performance index

Uno degli esempi più significativi e recenti di risultato di un'analisi delle prestazioni è l'*Handball performance index* (HPI): a partire dalla stagione 2020/21 della *Liqui Moly HBL*, primo campionato tedesco, è stata composta una "task-force" formata da tecnici, giocatori in attività e non, statistici e performance analyst con lo scopo di definire i parametri per la creazione di un indice prestazionale oggettivo che rendesse confrontabili i singoli giocatori. Nacque così l'indice HPI, ancora oggi utilizzato, in una versione migliorata, per individuare i migliori giocatori del campionato.

1.4.1.1 Prima versione dell'indice HPI

La prima versione dell'indice HPI, utilizzata nella stagione 2020/21, si basava su un semplice accumulo di "punti HPI": nello specifico, partendo da un valore di 100 punti HPI assegnati ad ogni giocatore ad inizio partita, veniva sommato un determinato numero di punti per ogni azione positiva compiuta (come goal, assist, parata), mentre ne veniva sottratto un determinato numero per ogni azione negativa (come errori tecnici, tiri sbagliati, sanzioni). Tali aggiunte e detrazioni venivano ponderate principalmente rispetto alla probabilità a priori di compiere una determinata azione, tenendo però conto anche della frequenza con cui essa appare nel gioco.

Ad esempio, come mostrato nella Tabella 1.1, per il ruolo del portiere venivano aggiunti più punti in caso di parata effettuata su una tipologia di tiro a cui era stata assegnata una probabilità di parata minore. Di conseguenza, per un tipo di tiro con probabilità di parata minore, venivano sottratti meno punti in caso di goal subito.

Tabella 1.1: Punti HPI assegnati ai portieri per tipo di tiro affrontato

.

Tipo di tiro	Probabilità di parata	Punti per parata	Punti per goal subito
6 metri centrale	19%	8	-2
6 metri laterale	25%	8	-2
Rigore	24%	8	-2
9 metri centrale	30%	7	-3
9 metri laterale	34%	7	-3
Dietro metà campo	8%	9	-1
Punizione diretta	20%	5	-5
Contropiede	17%	8	-2
Dall'ala	36%	6	-4
Altro	18%	8	-2

Dalla Tabella 1.1 si può notare come i tiri del tipo "Punizione diretta" abbiano meno punti assegnati in caso di parata rispetto, ad esempio, ai tiri di "Rigore", nonostante la probabilità di parata sia minore: ciò accade data la bassissima frequenza con cui appare nel gioco la tipologia di tiro "Punizione diretta".

Seguendo lo stesso principio, ad un giocatore di campo venivano assegnati più punti in caso di goal e sottratti meno punti in caso di errore per tipologie di tiro con una probabilità di successo minore, come mostrato nella Tabella 1.2.

Tabella 1.2: Punti HPI assegnati ai giocatori di campo per tipo di tiro effettuato.

Tipo di tiro	Probabilità di goal	Punti per goal	Punti per assist	Punti per tiro sbagliato
6 metri centrale	71%	6	4	-7
6 metri laterale	65%	7	3	-7
Rigore	76%	6	4	-8
9 metri centrale	43%	10	0	-4
9 metri laterale	43%	10	0	-4
Dietro metà campo	66%	7	3	-7
Punizione diretta	14%	10	0	-1
Contropiede	83%	5	5	-8
Dall'ala	64%	7	3	-6
Altro	75%	6	4	-8

Ulteriori eventi che causano assegnazioni o detrazioni di punti HPI sono mostrati nella Tabella 1.3.

Tabella 1.3: Altri eventi che determinano l'assegnazione o detrazione di punti HPI.

Evento	Punti assegnati/detratti
Sospensione	-3
Espulsione	-10
Rigore causato	-7
Muro	7
Palla rubata	8
Infrazione tecnica	-8

1.4.1.2 Versione attuale dell'indice

A partire dalla stagione 2021/22, l'indice HPI venne modificato in modo da rendere possibile un confronto più accurato tra giocatori di ruoli diversi. Difatti, la prima versione dell'indice tendeva a premiare maggiormente i giocatori più coinvolti nelle trame di gioco (centrale e terzini).

Perciò, l'indice HPI, costituito in origine dai soli "punti HPI", venne riscalato in modo da avere valori compresi tra 100 e 50, utilizzando la formula

$$HPI = 50 + \big((HPIpoints - minHPI)50\big)/(maxHPI - minHPI) \tag{1.1}$$

in cui *HPIpoints* corrisponde ai punti HPI calcolati in una determinata partita della stagione attuale (analogo della prima versione dell'indice), mentre *maxHPI* e *minHPI* corrispondono al valore massimo e al valore minimo di punti HPI registrati nella stagione precedente per ogni gruppo di giocatori (gruppi specificati al paragrafo 1.3.3, con l'aggiunta del gruppo formato dai portieri).

Capitolo 2

Campionato tedesco di pallamano

2.1 I dati

Il dataset preso in esame contiene i dati relativi alle prestazioni di ogni singolo giocatore ad ogni partita effettuata nella stagione 2021/22 della *Liqui Moly HBL*. Il suddetto campionato è composto da diciotto squadre che giocano un totale di trentaquattro partite ciascuna (ogni squadra gioca contro le altre una volta in casa e una volta in trasferta) e che ad ogni partita possono schierare un numero massimo di sedici giocatori.

I dati sono stati ricavati dal sito ufficiale del campionato (https://www.liquimoly-hbl.de/de/) e sottoposti ad un'attenta pulizia. In particolare, nei dati originali venivano considerate anche le "prestazioni" di giocatori e allenatori che avevano solamente subito una sanzione dalla panchina (presumibilmente per proteste) durante il corso della partita, senza mai entrare effettivamente in campo. Le righe corrispondenti sono state eliminate, portando ad avere un dataset composto da 7679 osservazioni totali relative a 381 giocatori diversi che hanno preso parte ad almeno una partita del campionato.

In aggiunta agli indicatori più comuni è stato considerato nell'analisi anche l'indice HPI, avente una sezione dedicata nel suddetto sito. Nella trascrizione dei valori dell'indice, però, si sono notati dei dati mancanti rispetto alle prestazioni registrate: ad esempio, per la settima giornata di campionato non erano stati registrati i valori HPI di due intere squadre (*SC DHfK Leipzig* e *Rhein Neckar-Löwen*). Ciò ha portato ad avere ottantatre valori mancanti dell'indice.

Infine, per ogni giocatore sono stati aggiunti successivamente un indicatore relativo al risultato della partita giocata e un indicatore relativo al "fattore campo" (partita giocata in casa o in trasferta).

2.2 Le variabili

Tabella 2.1: Variabili del dataset con loro descrizione.

Nome	Descrizione
Matchday	numero della giornata di campionato
Player	nome del giocatore
Team	sigla della squadra a cui appartiene il giocatore
Pos	ruolo del giocatore
GK	indica se il giocatore è un portiere
Goal	il numero di goal segnati
Misses	numero di tiri sbagliati
Field Goals (FG)	numero di goal segnati con tiri "dal campo"
7mG	numero di goal segnati su rigore
Shoot_perc	percentuale di goal su tiri effettuati
Ast	numero di assist effettuati
Technical Fouls (TF)	numero di infrazioni tecniche commesse
Stl	numero di palloni rubati
Blk	numero di tiri murati
Yc	indica se il giocatore ha ricevuto un cartellino giallo
Susp	numero di sospensioni temporanee ricevute
Rc	indica se il giocatore ha ricevuto un cartellino rosso
Time	tempo di gioco effettivo in secondi
Saves	numero di parate messe a segno
Field Saves (FS)	numero di parate effettuate su tiri "dal campo"
7mS	numero di rigori parati
Goal_conc	numero di goal subiti
Save_perc	percentuale di parate su tiri in porta ricevuti
Shots per goal conceded (SxG)	tiri in porta subiti su goal concessi
HPI	"Handball Performance Index"
Res	fattore che indica il risultato della partita
Campo	indica se la squadra del giocatore ha giocato in casa

Le variabili presenti nel dataset assieme ad una loro breve descrizione sono riportate nella Tabella 2.1.

Le variabili *Misses*, *FG*, *7mG*, *Shoot_perc*, *TF* e *Stl* sono rilevate esclusivamente per i giocatori di campo mentre *Time*, *Saves*, *FS*, *7mS*, *Goal_conc*, *Save_perc*, *SxG* sono rilevate esclusivamente per i portieri. Inoltre, le variabili *GK*, *Yc*, *Rc* e *Campo* sono dicotomiche.

Con l'espressione "tiri dal campo" si fa riferimento a tutte le tipologie di conclusione verso la porta avversaria, escludendo i rigori, mentre, nella descrizione della variabile *Save_perc*, con "tiri in porta" escludiamo le conclusioni in cui non è stato centrato lo

specchio della porta, ovvero anche quelle in cui viene colpito un palo con successiva uscita della palla.

La variabile *SxG* può essere ricavata come combinazione delle variabili *Saves* e *Goal_conc* tramite la formula: $(Saves + Goal\_conc)/Goal\_conc$.

Infine, la variabile *Res* è un fattore a tre livelli: "V" in caso di vittoria, "P" in caso di pareggio e "S" in caso di sconfitta.

2.3 Analisi esplorativa

Obiettivo principale dell'analisi esplorativa svolta sul dataset era quello di individuare eventuali relazioni tra le statistiche prestazionali rilevate e il risultato finale della rispettiva partita.

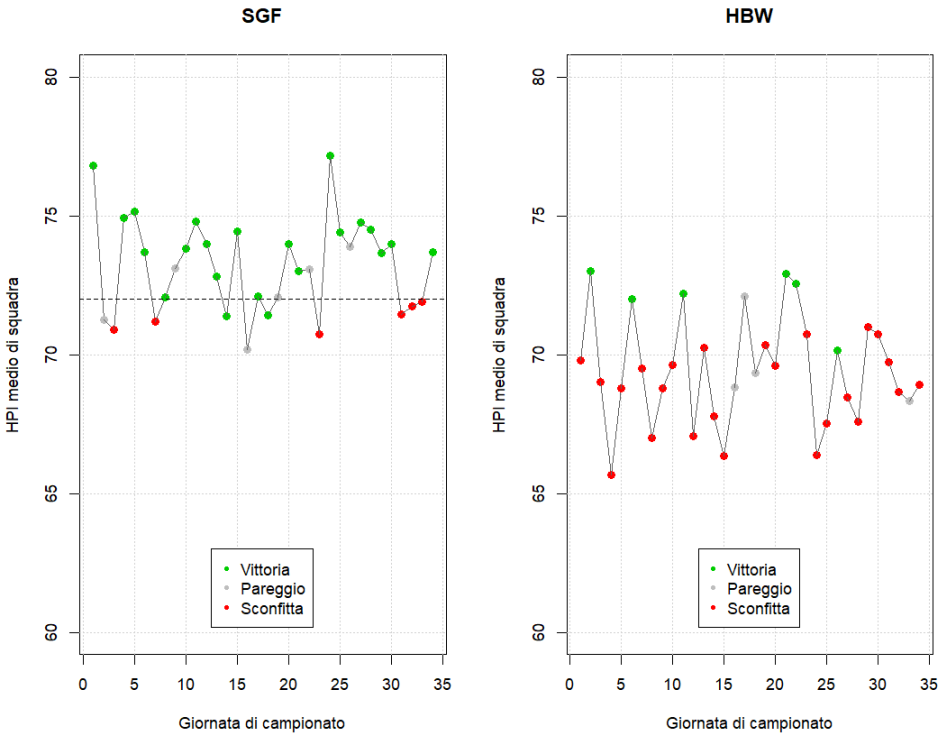

FIGURA 2.1: Andamenti del livello medio di squadra dell'indice HPI per *SG Flensburg-Handewitt* (a sinistra) e *HBW Balingen-Weilstetten* (a destra). Nel grafico di sinistra, linea tratteggiata in corrispondenza di un valore di HPI medio uguale a 72.

Inizialmente, l'analisi si è concentrata sul livello medio di squadra dell'indice HPI per ogni partita giocata e ha evidenziato come, nella maggior parte dei casi, registrare un HPI medio inferiore a 70 porti ad una sconfitta. Considerando innanzitutto le prime

quattro classificate a fine campionato (in ordine di classifica: *SC Magdeburg*, *THW Kiel*, *Füchse Berlin* e *SG Flensburg-Handewitt*), si è notato, ad esempio, che il Magdeburg ha perso le uniche partite della stagione in due casi su quattro in cui ha registrato un HPI medio di squadra inferiore a 70. Il Kiel su quattro partite con HPI medio inferiore a 70 ha collezionato due sconfitte (su quattro totali), un pareggio e una vittoria, mentre il Berlin, su tre partite sotto la suddetta soglia, ha registrato due sconfitte e un pareggio. Caso particolare, invece, è quello del Flensburg, unica squadra del campionato a non aver mai registrato un livello di HPI medio inferiore a 70, come si vede in Figura 2.1. Tuttavia, si può comunque notare come le sconfitte siano arrivate sotto una certa soglia di HPI medio: le sconfitte (sei) sono state registrate sotto un livello medio uguale a 72, assieme a due pareggi e 2 vittorie.

Considerando squadre di medio-alta classifica, come *TBV Lemgo Lippe* (sesta classificata) e *SC DHfK Leipzig* (nona classificata), si è constatato come entrambe abbiano perso tutte le partite in cui hanno registrato un HPI medio inferiore a 70, escluso un pareggio per la prima e un pareggio e una vittoria per la seconda.

Infine, analizzando squadre di bassa classifica come *Handball Sport Verein Hamburg* (quattordicesima classificata) e *HBW Balingen-Weilstetten* (penultima classificata), si è registrata una notevole diminuzione di partite con HPI medio maggiore di 70, nelle quali la prima ha collezionato otto vittorie, un pareggio e una sconfitta, mentre la seconda ha registrato le uniche vittorie del suo campionato (Figura 2.1).

Successivamente l'indagine si è focalizzata sulla serie dei valori dell'indice HPI per i singoli giocatori, esaminando coloro che, giocando almeno dieci partite, hanno registrato un HPI medio maggiore o uguale a 75. Nei diciassette giocatori che soddisfavano i suddetti requisiti, si è osservata una netta prevalenza di singoli che occupano ruoli meno coinvolti nelle trame di gioco: difatti, sono stati rilevati otto pivot, sei ali, due terzini e un portiere (centrale con miglior HPI medio: Jim Gottfridsson, 74.71).

Giocatore più costante nelle prestazioni di alto livello è stato Hans Lindberg, ala destra del *Füchse Berlin* che ha fatto registrare l' HPI medio più elevato con 80.35. Nonostante ciò, il premio di miglior giocatore della stagione è stato assegnato a Omar Ingi Magnusson, terzino destro del *SC Magdeburg*, il quale ha registrato il secondo HPI medio più elevato (77.82). A differenza di Lindberg, Magnusson ha avuto prestazioni meno costanti come evidenzia il trend crescente dei valori dell'indice HPI verso fine stagione, rappresentato in Figura 2.2.

Ulteriore serie interessante è quella dei valori HPI di Kevin Møller, unico portiere ad aver registrato un livello medio superiore a 75: egli infatti è il giocatore con più partite con HPI maggiore di 90 (cinque).

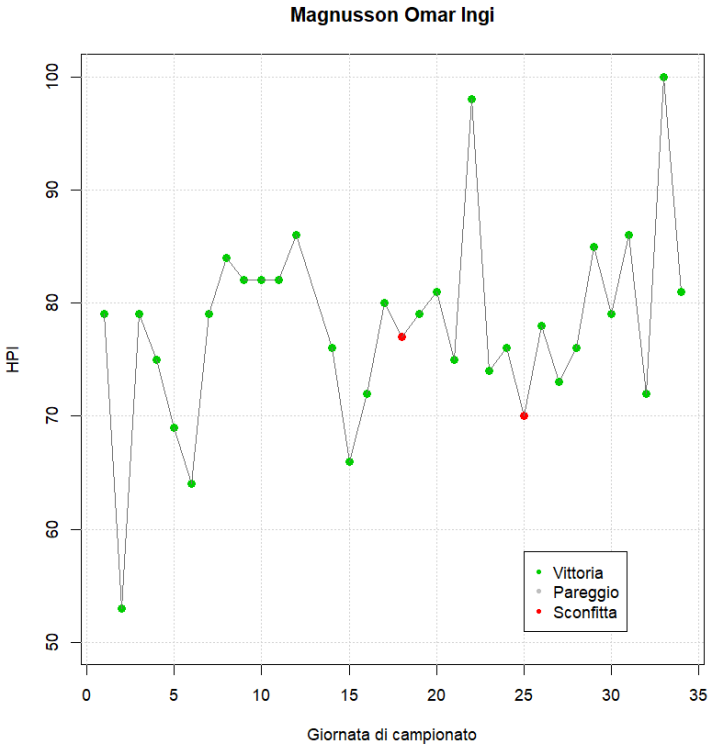

FIGURA 2.2: Serie dei valori dell'indice HPI di Omar Ingi Magnusson durante il corso della stagione

L'analisi è poi proseguita concentrandosi sulla distribuzione marginale dell'indice HPI e ha evidenziato come essa vari leggermente a seconda del risultato della partita: in particolare, il valore dell'indice tende ad essere più alto in caso di vittoria, comportamento in linea con la natura dell'indice stesso.

In parallelo è stata svolta la medesima analisi separando i giocatori di campo dai portieri e si è riscontrato che, indipendentemente dal risultato della partita, il valore HPI è mediamente più basso per i portieri (Figura 2.3). Questo risultato ha portato ad un controllo della distribuzione dell'indice in relazione al ruolo dei singoli, il quale ha evidenziato come i giocatori maggiormente coinvolti nel gioco (terzini, centrale e portiere) abbiano valori dell'indice mediamente inferiori rispetto agli altri (ali e pivot), come mostrato nella Tabella 2.2.

A seguire, l'analisi è stata sviluppata dividendo il dataset tra giocatori di campo e portieri, dal momento in cui le variabili misurate non sono le stesse per i due sottogruppi sopracitati.

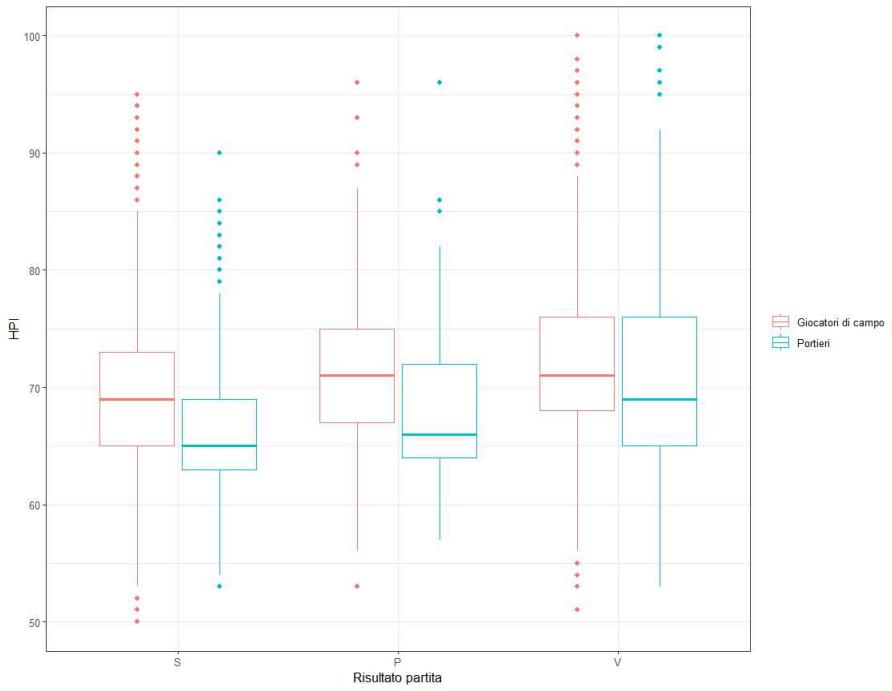

FIGURA 2.3: Boxplot del valore dell'indice HPI per portieri e giocatori di campo in relazione al risultato della partita.

TABELLA 2.2: Valori notevoli indice HPI per ogni posizione.

Ruolo	Min	1st Quart.	Mediana	Media	3rd Quart.	Max
Portiere (GK)	53.00	64.00	67.00	68.32	72.00	100.00
Terzino sinistro (LB)	51.00	66.00	69.00	69.90	73.00	94.00
Centrale (CB)	53.00	66.00	69.00	69.88	74.00	93.00
Terzino destro (RB)	53.00	65.00	69.00	69.93	74.00	100.00
Ala destra (RW)	50.00	67.00	70.00	71.10	74.00	95.00
Ala sinistra (LW)	51.00	67.00	70.50	71.47	75.00	97.00
Pivot (PI)	54.00	69.00	73.00	73.03	76.00	96.00

2.3.1 Giocatori di campo

In prima battuta, è stata svolta un'analisi esplorativa univariata per le variabili *Goal*, *Misses*, *FG*, *Shoot_perc*, *Ast*, *Blk*, *TF* e *Stl*. Essa ha mostrato come la distribuzione delle

frequenze di ognuna sia asimmetrica, con una elevata presenza di zeri.

L'analisi si è poi concentrata sulla ricerca di eventuali relazioni tra le singole variabili e il risultato delle partite. Delle variabili sopracitate, quella che denota la maggiore differenza a seconda del risultato è la percentuale al tiro (*Shoot_perc*): nello specifico, non considerando i giocatori con zero tiri tentati durante la partita (1309 giocatori, per i quali *Shoot_perc* vale zero), in caso di vittoria la media delle percentuali al tiro è circa del 72.56%, mentre, in caso di sconfitta, del 65.82%. Eseguendo la medesima analisi considerando tutti i giocatori di campo, si è registrata una discreta diminuzione dei valori del primo e del terzo quartile per le distribuzioni della variabile in caso di pareggio o sconfitta (Tabella 2.3).

Per le restanti variabili, invece, sono state riscontrate delle leggere differenze in media, quasi impercettibili per quelle che descrivono gli aspetti meno frequenti del gioco, come muri (*Blk*), palle rubate (*Stl*) e infrazioni tecniche (*TF*).

Tabella 2.3: Valori notevoli della percentuale al tiro condizionati al risultato della partita nei due sottogruppi.

Sottogruppo	Res	Min	1st Quart.	Mediana	Media	3rd Quart.	Max
Almeno un tiro	V	14.29	50.00	71.43	72.56	100.00	100.00
	P	16.67	50.00	66.67	69.51	100.00	100.00
	S	11.11	50.00	66.67	65.82	87.50	100.00
Tutti i giocatori	V	0.00	40.00	66.67	59.49	100.00	100.00
	P	0.00	33.33	60.00	56.74	85.71	100.00
	S	0.00	25.00	50.00	51.41	75.00	100.00

2.3.2 Portieri

A differenza dei giocatori di campo, il portiere ha a disposizione meno possibilità per influire nel match e ha un range di errore notevolmente ridotto: infatti, più un portiere para durante il corso della partita e più probabilità ha di avere un minutaggio elevato nella stessa. Questo aspetto è rispecchiato anche dai dati in nostro possesso: attraverso un'analisi esplorativa bivariata si è notata, infatti, l'elevata correlazione tra il tempo di gioco (*Time*) e il numero di parate totali (*Saves*) e dal campo (*FS*), in entrambi i casi di circa 0.87.

L'analisi è proseguita considerando le singole variabili in relazione col risultato della partita e si è constatato che in media un portiere gioca circa cinque minuti in più in caso di vittoria e pareggio rispetto al minutaggio in caso di sconfitta, come mostrato

dalla Tabella 2.4. Inoltre, come ci si poteva aspettare, le variabili inerenti alle parate del singolo portiere (*Saves*, *FS*, *Save_perc*, *SxG*), ad eccezione del numero di rigori parati (*7mS*), hanno in media valori superiori in caso di vittoria.

Risultato, invece, inaspettato riguarda la variabile *Goal_conc* relativa ai goal concessi: infatti, si è riscontrato che un portiere subisce più goal in media in caso di pareggio (16.56), piuttosto che in caso di sconfitta (15.69). Questo esito si potrebbe spiegare tenendo conto del fatto che un portiere gioca di meno in caso di sconfitta, riducendo quindi le possibilità di subire singolarmente un numero elevato di goal.

Tabella 2.4: Valori notevoli del minutaggio condizionati al risultato della partita.

Res	Min	1st Quart.	Mediana	Media	3rd Quart.	Max
V	2	776.5	2405.5	2043.36	3302	3600
P	20	1004.0	2381.0	2080.03	3226	3514
S	7	868.5	1726.0	1761.66	2668	3600

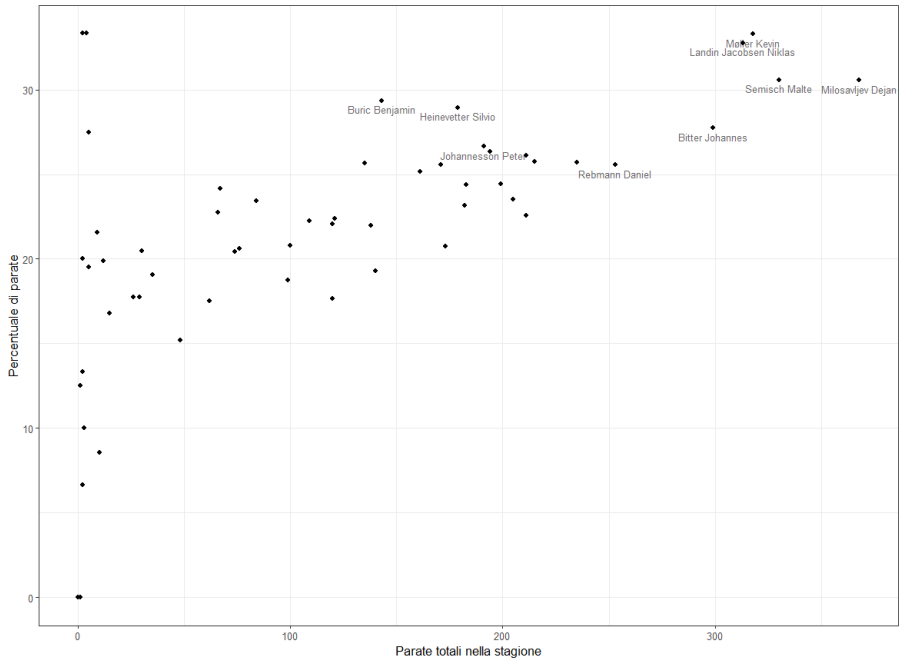

Figura 2.4: Grafico a dispersione per parate totali e percentuali di parata dei singoli portieri in tutta la stagione.

Infine, per mettere a confronto le prestazioni dei portieri durante l'arco della stagione sono stati calcolati il numero di parate totali effettuate assieme alla percentuale media

di parate di ciascuno di essi, rappresentati mediante un diagramma a dispersione in Figura 2.4.

Come si può notare dal grafico riportato, vi è un sottogruppo formato da quattro portieri (Kevin Møller, Niklas Landin Jacobsen, Malte Semisch e Dejan Milosavljev) le cui prestazioni sono state al di sopra della media, mettendo a segno oltre trecento parate con una percentuale superiore al 30% (ad essi si può aggiungere anche Johannes Bitter, il quale ha fatto registrare 299 interventi con una percentuale del 27.72%).

Capitolo 3

Modelli per variabili risposta categoriali ordinali

Nel presente capitolo l'obiettivo principale sarà quello di individuare i ruoli dei giocatori che influiscono maggiormente sul risultato finale di una partita e, in aggiunta, valutare l'effetto del "fattore campo" su quest'ultimo.

Per fare ciò, dopo una riorganizzazione dei dati in forma non raggruppata, si implementeranno modelli per risposte politomiche ordinali considerando il risultato di ogni partita come variabile risposta. In particolare, verrà implementato innanzitutto un modello per logit cumulati, per il quale l'interpretazione dei risultati è più semplice, e successivamente verranno proposti alcuni modelli alternativi. Infine, si valuterà quale dei modelli proposti si adatta meglio ai dati in nostro possesso.

3.1 Risposte politomiche su scala ordinale

Data una variabile risposta qualitativa con più di due modalità, detta anche politomica, il modello statistico più appropriato per la distribuzione di essa è il modello multinomiale. Le variabili categoriali possono disporre di un ordinamento naturale tra le loro modalità, caratteristica molto diffusa nell'ambito medico e in quello socio-economico, ad esempio nel caso in cui si vuole misurare il livello di soddisfazione delle persone rispetto ad un servizio o ad un prodotto oppure misurare l'intensità dei sintomi di una malattia in alcuni pazienti, come mostrato negli esempi al paragrafo 4.4 in Salvan et al. (2020).

Per applicare il modello nel caso generale in cui si ha una variabile risposta politomica, il valore di quest'ultima per l'i-esimo soggetto deve essere costituito dal vettore $y_i = (y_{i1}, \ldots, y_{ic})$, $i = 1, \ldots, n$, con n uguale al numero di osservazioni, $y_{ij} = 1$ se viene

osservata la j-esima modalità e $y_{ij} = 0$ altrimenti, $j = 1, \ldots, c$, con c uguale al numero di modalità della risposta. In questo modo si ha $\sum_{j=1}^{c} y_{ij} = 1$ e la variabile $Y_i = (Y_{i1}, \ldots, Y_{ic})$, di cui y_i è realizzazione, ha distribuzione multinomiale elementare con funzione di probabilità

$$p_{Y_i}(y_i; \pi_i) = Pr(Y_i = y_i) = \pi_{ij}^{y_{ij}} \ldots \pi_{ic}^{y_{ic}}, \tag{3.1}$$

definita sul supporto $\left\{ y_i \in \{0,1\}^c : \sum_{j=1}^{c} y_{ij} = 1 \right\}$ e con π_{ij} che indica la probabilità di osservare la j-esima modalità per l'i-esima osservazione, tale per cui $\pi_{ij} \in (0,1)$ e $\sum_{j=1}^{c} \pi_{ij} = 1$.

In presenza, invece, di variabili risposta qualitative su scala ordinale, risulta di notevole importanza valutare l'effetto delle variabili esplicative sull'ordinamento stocastico delle distribuzioni della risposta. L'ordinamento stocastico di una variabile qualitativa può essere definito tramite la sua funzione di ripartizione: date Y_i e Y_k, variabili che descrivono rispettivamente i valori della risposta per l'i-esimo e per il k-esimo soggetto (con $k \neq i$), se $Pr(Y_i \leq j) \geq Pr(Y_k \leq j)\ \forall j = 1, \ldots, c$, con la disuguaglianza stretta per almeno un j, allora la distribuzione di Y_i è stocasticamente più piccola della distribuzione di Y_k; se, invece, si inverte il verso della disuguaglianza, la distribuzione di Y_i è stocasticamente più grande della distribuzione di Y_k.

3.2 Modello per logit cumulati

La scelta più opportuna per modellare variabili politomiche ordinali risulta quella di adottare modelli che mettono in relazione le variabili esplicative con le probabilità cumulate. Il modello statistico più utilizzato in queste situazioni è il modello di regressione per logit cumulati, proposto da McCullagh (1980), il quale assume che

$$logit[Pr(Y_i \leq j)] = \alpha_j + x_i\beta, \tag{3.2}$$

per $j = 1, \ldots, c-1$, con $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{c-1}$ e $Pr(Y_i \leq j) = \pi_{i1} + \cdots + \pi_{ij}$. Nel suddetto modello, ogni logit cumulato ha una sua intecetta (termine α), mentre gli effetti β delle variabili esplicative sono gli stessi per ogni logit cumulato, costanti quindi al variare di j.

Se è soddisfatto il modello (3.2):

$$log \frac{Pr(Y_i \leq j | x_i = u)/Pr(Y_i > j | x_i = u)}{Pr(Y_i \leq j | x_i = v)/Pr(Y_i > j | x_i = v)} = (u - v)\beta, \tag{3.3}$$

ovvero, il logaritmo del rapporto delle quote, detto *rapporto delle quote cumulate*, è lineare rispetto alla differenza tra i valori assunti dalla variabile esplicativa. Da ciò deriva il fatto che il modello (3.2) è anche detto a *quote proporzionali*: difatti, per un incremento unitario di una variabile esplicativa, ad esempio x_{ir}, la quota cumulata risulta moltiplicata per e^{β_r}, costante in j, a parità delle altre variabili esplicative.

3.3 Modelli alternativi

Quando si trattano modelli che coinvolgono le probabilità cumulate la funzione logit risulta essere la scelta più frequente tra le diverse funzione di legame. Ciò è legato in particolar modo al fatto che l'utilizzo di essa rende più semplice l'interpretazione dei coefficienti di regressione stimati.

Sono possibili, però, implementazioni di diverse funzione di legame, tra le quali le più utilizzate sono la funzione *probit* (Agresti, 2015), che applica alle probabilità cumulate l'inversa della funzione di ripartizione di una normale, e la funzione *log-log complementare*, o *cloglog*, (Agresti, 2013) non simmetrica a differenza delle funzioni logit e probit, e per la quale il modello (3.2) diventa: $log\{-log[1-Pr(Y_i \leq j)]\} = \alpha_j + x_i\beta$. In entrambi i casi sopracitati l'interpretazione a quote proporzionali non può essere applicata.

Altro modello alternativo al (3.2) è quello per cui viene rilassata l'assunzione di proporzionalità delle quote: spesso, infatti, il modello a quote proporzionali risulta non avere un adattamento soddisfacente, in particolare quando la variabilità della risposta dipende dalle variabili esplicative. In questa situazione, i valori β diventano dipendenti da j, ovvero gli effetti delle variabili esplicative risultano differenti per ogni logit cumulato:

$$logit[Pr(Y_i \leq j)] = \alpha_j + x_i\beta_j. \tag{3.4}$$

Il suddetto modello, oltre a essere meno parsimonioso e più difficile da interpretare rispetto a quello a quote proporzionali, potrebbe portare alla non convergenza dell'algoritmo di massimizzazione della verosimiglianza usato per la stima dei parametri, dato che potrebbe violare la condizione $Pr(Y_i \leq j) \leq Pr(Y_i \leq j')$ per ogni coppia $j, j' = 1, \ldots, c-1$ con $j < j'$, portando alla perdita dell'ordinamento stocastico delle distribuzioni della risposta. Perciò, anche se il modello più ampio produce un miglioramento significativo, spesso si preferisce comunque il modello a quote proporzionali.

3.4 Applicazione al campionato tedesco

Obiettivo principale dell'analisi tramite modelli per risposte politomiche ordinali nel caso di studio è, come già detto in precedenza, quello di individuare i ruoli più incisivi nell'economia di una singola partita. Per fare ciò, si è studiata la relazione tra l'esito di una partita giocata da una specifica squadra in una specifica giornata e le prestazioni medie per ruolo registrate dai giocatori della suddetta squadra nella partita presa in considerazione. Inoltre, in parallelo a quest'analisi, si è valutato l'effetto del fattore campo sul risultato finale di una partita, tipicamente presente nella maggioranza degli sport.

Per implementare i modelli descritti si è resa necessaria una riorganizzazione dei dati in nostro possesso: in particolare, si è creato un dataset contenente dati non raggruppati dove ogni riga rappresenta la prestazione di una squadra in una precisa giornata. Per ogni osservazione, oltre al nome della squadra, al numero della giornata di campionato e se la squadra ha giocato in casa o meno, si sono riportati i valori medi per posizione dell'indice HPI registrato dai giocatori della squadra durante la partita. Infine, per ogni riga si è riportato l'esito della partita (variabile *Res* nel dataset di partenza), variabile risposta politomica su scala ordinale. A causa delle osservazioni mancanti dell'indice HPI nel dataset di partenza, una discreta parte delle 612 osservazioni ottenute (trentaquattro partite per ognuna delle diciotto squadre) conteneva almeno un valore "NA" tra le medie calcolate. Esse sono state eliminate per permettere la successiva implementazione dei modelli, producendo un dataset contenente 529 osservazioni.

Indicato con Y_i il risultato dell'i-esima osservazione, con modalità ordinate "S" ($j = 1$), "P" ($j = 2$) e "V" ($j = 3$) analoghe a quelle della variabile *Res* del dataset di partenza, inizialmente è stato implementato il modello per logit cumulati

$$logit[Pr(Y_i \leq j)] = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8}, \quad (3.5)$$

$i = 1, \ldots, 529$, $j = 1, 2$, dove $x_{i1}, \ldots, x_{i7}$ rappresentano i valori medi dell'indice HPI per ognuno dei sette ruoli (nell'ordine: ala sinistra, ala destra, centrale, terzino destro, terzino sinistro, pivot e portiere) e x_{i8} è la variabile dicotomica che rappresenta il fattore campo, la quale vale uno se una specifica squadra ha giocato in casa.

In primo luogo, nel "summary" del modello si sono notati dei valori decisamente elevati per le stime dei termini di intercetta α_1 e α_2, rispettivamente 55.860 e 56.492. Da ciò si deduce che, quando tutte le esplicative sono poste uguali a zero, la stima della probabilità di sconfitta $Pr(Y_i = 1)$ è approssimativamente uguale a 1 e quella della probabilità di pareggio $Pr(Y_i = 2)$ approssimativamente uguale a 0 (tale situazione non

può essere reale dato che il valore minimo che può essere registrato per l'indice HPI è 50. La stima della probabilità di sconfitta nel caso in cui le medie dell'indice per ogni posizione siano uguali a 50 e in caso di partita in trasferta è comunque approssimativamente uguale a 1).

TABELLA 3.1: Stime dei coefficienti di regressione del modello per logit cumulati assieme ai rispettivi standard error.

Parametro	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Stima	-0.067	-0.045	-0.158	-0.133	-0.112	-0.131	-0.152	-0.011
Std. error	0.017	0.017	0.025	0.021	0.022	0.027	0.018	0.205

Nella Tabella 3.1 sono riportate le stime dei coefficienti di regressione, i quali sono risultati tutti significativi ad eccezione di quello relativo alla variabile che descrive il fattore campo, la cui deviazione standard è notevolmente maggiore rispetto alle altre. Si può notare innanzitutto che le stime riportate sono tutte negative: quindi, per un incremento unitario della media dell'indice HPI per una posizione specifica, la probabilità cumulata $Pr(Y_i \leq j)$ per $j = 1, 2$ diminuisce a parità delle altre esplicative, comportamento in linea con la natura dell'indice HPI. Inoltre, le stesse probabilità cumulate diminuiscono, seppur di poco, quando una squadra gioca in casa rispetto a quando gioca in trasferta.

Si può osservare che le stime dei primi due coefficienti, relativi alle medie dell'indice per le ali sinistre e le ali destre, sono più vicine allo zero rispetto alle altre, mentre le stime dei coefficienti relativi alle posizioni di centrale e portiere sono quelle che si scostano maggiormente dallo zero. Di conseguenza, come mostrato dalla Tabella 3.2, un incremento unitario delle medie per le posizioni di ala produce un effetto moltiplicativo sulla quota cumulata $Pr(Y_i \leq j)/Pr(Y_i > j)$ abbastanza vicino all'uno: ciò significa che tali posizioni sono quelle che influiscono meno sul risultato finale delle partite. Al contrario, le posizioni che incidono maggiormente sono quelle del centrale e del portiere, i cui relativi effetti moltiplicativi sono i minori registrati (0.854 e 0.859 rispettivamente) e le quali, come mostrato nella Tabella 2.2 dell'analisi esplorativa, hanno registrato l'HPI medio minore durante l'intera stagione.

Sempre dalla Tabella 2.2 si può notare come i ruoli con HPI medio minore sono quelli con effetto moltiplicativo che si discosta maggiormente dall'uno, fatta eccezione per il ruolo del pivot: esso, infatti, ha regitrato il valore di HPI medio più elevato (73.03) ed uno degli effetti moltiplicativi che si discosta maggiormente dall'uno (0.878). Ciò è

probabilmente dovuto all'importanza tipica dei giocatori di questo ruolo anche in fase difensiva.

Dalla Tabella 3.2, inoltre, si può riscontrare un' inaspettata differenza tra gli effetti moltiplicativi relativi alle due posizioni di terzino. Nonostante i due ruoli abbiano funzioni analoghe nel gioco, il modello ha evidenziato una maggiore incisività da parte dei terzini destri rispetto a quelli sinistri (effetti moltiplicativi rispettivamente 0.875 e 0.894).

Per quanto riguarda, invece, la variabile relativa al fattore campo, essa presenta l'effetto moltiplicativo che più si avvicina all'uno (0.989) tra quelli delle variabili considerate, evidenziando come il fattore campo non influisca in maniera significativa sull'esito della singola partita, contrariamente a quanto ci si poteva aspettare.

TABELLA 3.2: Stime degli effetti moltiplicativi sulla quota cumulata dovuti ad un incremento unitario delle rispettive variabili esplicative.

Parametro	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
$exp(\beta_i)$	0.936	0.956	0.854	0.875	0.894	0.878	0.859	0.989

Per valutare l'accuratezza dei risultati ottenuti, è stata presa in considerazione come esempio la squadra del *SG Flensburg-Handewitt*. Essa è arrivata "solamente" quarta in classifica nonostante abbia potuto contare sull'apporto di ben quattro giocatori considerati i migliori nel loro ruolo al termine della stagione, due dei quali impiegati nelle posizioni più incisive secondo i risultati del modello: Jim Gottfridsson (centrale), Kevin Møller (portiere), Johannes Golla (pivot) e Hampus Wanne (ala sinistra). Tra questi, Gottfridsson è risultato il più determinante per il risultato di squadra: su otto delle undici partite giocate e non vinte ha registrato un HPI minore o uguale a 72. I restanti tre, invece, nei pareggi e nelle sconfitte hanno registrato valori HPI mediamente maggiori rispetto al loro centrale, con in particolare Golla che ha ottenuto un valore HPI inferiore a 72 in una sola occasione. In aggiunta a ciò, e probabilmente una delle cause dei risultati al di sotto delle aspettative, il Flensburg ha avuto i terzini destri e sinistri che nelle partite non vinte hanno registrato valori HPI mediamente bassi (rispettivamente 68 e 70.625).

In parallelo, per capire quanto il peso delle prestazioni delle ali possa essere basso rispetto a quello delle altre posizioni, è stato preso come esempio Hans Lindberg, eletto miglior ala destra del campionato al termine della stagione e, come già detto, giocatore con HPI medio più elevato durante la stagione. Egli, infatti, nelle partite non vinte ha

ottenuto un HPI medio di 80.90, più dieci punti superiore, ad esempio, rispetto a quello registrato da Gottfridsson nelle sconfitte e nei pareggi (70.09).

Nella fase successiva dell'analisi sono state proposte delle migliorie rispetto al modello con quote proporzionali sopracitato. In primo luogo si sono provati ad aggiungere i termini di interazione tra le variabili significative. Il modello risultante, oltre ad un notevole incremento dei parametri da stimare, ha registrato due soli coefficienti di regressione significativamente diversi da zero, riguardanti l'interazione tra le variabili relative ai ruoli di ala destra e terzino sinistro e l'interazione tra le variabili relative al ruolo di terzino sinistro e al fattore campo. Pertanto, si è deciso di implementare il modello per logit cumulati senza la variabile relativa al fattore campo, il quale ha prodotto stime abbastanza simili a quelle della Tabella 3.1 per i restanti coefficienti relativi alle medie dell'indice HPI.

A seguire, si è provata a rilassare l'ipotesi di proporzionalità delle quote per controllare se portasse a miglioramenti nell'adattamento ai dati. Considerando le stesse variabili specificate nel modello (3.5), l'algoritmo non convergeva, perciò si è deciso di considerare il modello senza la variabile relativa al fattore campo, la quale causava la non convergenza. Il modello risultante ha prodotto, come nel caso con quote proporzionali, tutte le stime dei coefficienti di regressione negative e significativamente diverse da zero. Però, esso ha mostrato delle anomalie nei risultati ottenuti: innanzitutto, le stime dei termini di intercetta α_1 e α_2 non soddisfavano la proprietà $\alpha_1 \leq \alpha_2$, presentando rispettivamente i valori 57.463 e 55.346; inoltre, il "summary" del modello mostrava dei valori NA per i valori della statistica Z[1] e del *p-value* del *t-test* dei coefficienti relativi al ruolo di ala destra. Pertanto, si è svolto il test del log rapporto di verosimiglianza per confrontare il modello a quote proporzionali con quello a quote non proporzionali (entrambi senza la variabile relativa al fattore campo), il quale ha segnalato un'evidenza abbastanza forte contro quest'ultimo.

Come ultima fase dell'analisi, prendendo come riferimento il modello (3.5), si sono valutati modelli con funzione di legame diverse dalla logistica, come la funzione probit e quella log-log complementare.

Nel primo modello, detto anche "per probit cumulati", si può notare come, analogamente ai modelli precedenti, le stime dei coefficienti di regressione siano tutte negative, come riportato nella Tabella 3.3. Come già riscontrato nel modello (3.5), la stima del coefficiente della variabile relativa al fattore campo non risulta significativa, mentre le restanti risultano tutte altamente significative, ad eccezione di quella del coefficiente relativo al ruolo dell'ala destra, per la quale si è registrato un p-value del t-test decisamente

[1] $z_i = \hat{\beta}_i / Std.Error(\hat{\beta}_i)$

Tabella 3.3: Stime dei coefficienti di regressione del modello per probit cumulati assieme ai rispettivi standard error.

Parametro	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Stima	-0.038	-0.025	-0.089	-0.075	-0.063	-0.076	-0.088	-0.033
Std. error	0.010	0.010	0.014	0.012	0.012	0.016	0.010	0.119

più elevato rispetto a quello delle altre stime significative (0.0136).

In aggiunta, confrontando la Tabella 3.3 con la Tabella 3.1 si può osservare come le prime sette stime dei coefficienti e deviazioni standard del modello per probit cumulati siano più piccole in valore assoluto rispetto a quelle del modello per logit cumulati. Ciò è dovuto al fatto che la distribuzione logistica standard ha deviazione standard uguale a 1.81, valore notevolmente maggiore rispetto a quello della normale standard, uguale a uno (Agresti, 2015).

Per il modello con funzione di legame log-log complementare, i commenti sono analoghi a quelli esposti per il modello per probit cumulati: anche in questo caso le stime dei coefficienti sono risultate tutte negative e significativamente diverse da zero, ad eccezione di quella relativa alla variabile del fattore campo, con quella relativa al ruolo dell'ala destra avente p-value ancor più elevato, anche rispetto al precedente modello (0.048, molto vicino alla soglia del 5%).

Confrontando tramite il criterio di Akaike (AIC) i modelli senza fattore campo e con le tre diverse funzione di legame si è scelto come modello finale il modello per logit cumulati, il quale ha fatto registrare l'AIC più basso, come riportato nella Tabella 3.4.

Tabella 3.4: Valori AIC dei modelli con diversa funzione di legame.

Legame	AIC
logit	759.47
probit	762.91
cloglog	767.90

Capitolo 4

Modello di Bradley-Terry

Nel presente capitolo si svilupperà ulteriormente l'analisi svolta nella sezione precedente, al fine di comprendere quali ruoli incidono maggiormente nell'esito di una partita mediante un modello che meglio si adatta al problema considerato. Nello specifico, verrà implementato il modello ipotizzato da Ralph A. Bradley e Milton E. Terry, il quale utilizza il metodo dei confronti a coppie (Bradley & Terry, 1952): tale tecnica si basa sul confronto di diversi "oggetti" organizzati a coppie e per il quale bisogna esprimere una preferenza. Nel caso di studio considerato gli oggetti sono le singole squadre mentre i confronti a coppie sono dati da ognuna delle 306 partite svolte durante la stagione. Per una corretta implementazione del modello, si è resa dunque necessaria un'ulteriore riorganizzazione dei dati, nonchè l'utilizzo di un nuovo dataset contenente i confronti a coppie considerati.

In aggiunta, come nel modello per logit cumulati verrà analizzata anche l'influenza del fattore campo sull'esito di una partita, problematica per la quale il modello di Bradley-Terry fornisce una disamina abbastanza agevole.

4.1 Introduzione al modello

Dati dei confronti di coppie di "oggetti" per le quali si vuole esprimere una preferenza il modello più opportuno è quello di Bradley-Terry, utilizzato in queste situazioni per classificare gli oggetti sopracitati secondo un determinato criterio.

Ambito in cui il modello risulta comunemente più utilizzato è quello sportivo, nel quale gli "oggetti" considerati sono le squadre o i giocatori (a seconda se si analizza uno sport collettivo o individuale) e "esprimere una preferenza" tra due oggetti messi a confronto significa determinare l'esito di una partita.

Tipicamente, il modello considera solamente due possibili esiti dei singoli confronti, che corrispondono alla preferenza di uno o dell'altro oggetto della coppia, non rendendo possibili pareggi (Agresti, 2013). Esistono, però, estensioni del modello che considerano una variabile risposta categoriale avente più di due modalità. In primis, sono stati sviluppati modelli che includono il pareggio tra gli esiti dei confronti a coppie, i più conosciuti dei quali sono quelli di Rao & Kupper (1967) e Davidson (1970). Più recentemente, invece, sono stati utilizzati modelli aventi variabili risposta che distinguono l'esito degli incontri in più categorie: ad esempio, certi autori in ambito calcistico hanno distinto i risultati delle partite tramite una variabile categoriale con cinque modalità, dividendo le diverse categorie a seconda della differenza reti registrata al termine della partita (Schauberg & Tutz, 2019).

Il modello ipotizzato da Bradley-Terry assume che, in un confronto tra due oggetti i e j selezionati da un insieme di K oggetti, la probabilità che i venga preferito a j è α_i/α_j, dove α_i e α_j sono parametri positivi che possono essere interpretati come le "abilità" dei singoli oggetti (Turner & Firth, 2012). In alternativa, il modello "non strutturato" può essere espresso come:

$$logit[Pr(\text{“}i\text{”}\ preferito\ a\ \text{“}j\text{”})] = \lambda_i - \lambda_j, \tag{4.1}$$

nel quale si suppone che $Pr(\text{“}i\text{”}\ preferito\ a\ \text{“}j\text{”}) + Pr(\text{“}j\text{”}\ preferito\ a\ \text{“}i\text{”}) = 1$ per ogni coppia di oggetti e dove $\lambda_i = log(\alpha_i)$ per ogni oggetto i.

In determinate circostanze, può risultare d'interesse valutare se alcune variabili esplicative hanno un effetto più o meno rilevante sui risultati dei confronti. Nel contesto del modello di Bradley-Terry, le variabili esplicative possono essere specifiche:

- degli oggetti confrontati;
- del soggetto che opera il confronto;
- del confronto.

Il modello che considera le variabili esplicative oggetto-specifiche, detto anche modello "strutturato" (Cattelan, 2012), date p variabili $x_{i1}, \ldots, x_{ip}$ relative all'i-esimo soggetto, descrive i parametri λ_i che rappresentano il valore di ogni singolo oggetto tramite la combinazione lineare $\lambda_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$. Il modello (4.1) perciò diventa

$$logit[Pr(\text{“}i\text{”}\ preferito\ a\ \text{“}j\text{”})] = (x_{i1} - x_{j1})\beta_1 + \cdots + (x_{ip} - x_{jp})\beta_p, \tag{4.2}$$

dove quindi i coefficienti β_r, con $r = 1, \ldots, p$, sono associati alle differenze tra i valori assunti dalle covariate per ogni coppia di oggetti messi a confronto.

Come mostrato da Turner & Firth (2012), le ultime librerie del software R contenenti funzioni per l'implementazione e l'analisi del modello di Bradley-Terry (come il pacchetto *BradleyTerry2*) permettono l'aggiunta di semplici effetti casuali al predittore lineare, i quali permettono la distinzione tra oggetti aventi uguali valori delle covariate e inducono correlazione tra confronti con un oggetto in comune. In caso di presenza di un effetto casuale, il precedente predittore lineare diventa:

$$\lambda_i = \sum_{r=1}^{p} \beta_r x_{ir} + U_i, \tag{4.3}$$

dove U_i sono errori indipendenti per cui $U_i \sim N(0, \sigma^2)$ per ogni oggetto i.

Ulteriore aspetto di cui il modello di Bradley-Terry permette di tenere conto è quello dell'ordine con cui vengono presentati gli oggetti nei confronti. In determinati ambiti applicativi, infatti, ad alcuni o a tutti i confronti è associata una distorsione dei risultati dovuta o all'ordine con cui vengono presentati gli oggetti o al luogo dove avviene un determinato confronto. Ad esempio, in ambito sportivo l'*order effect* (Turner & Firth, 2012) è associato al fattore campo e al tipico vantaggio dovuto allo giocare in casa una partita.

Il modello 4.1 nella suddetta situazione diventa

$$logit[Pr(\text{“}i\text{”}\ preferito\ a\ \text{“}j\text{”})] = \lambda_i - \lambda_j + \delta z \tag{4.4}$$

dove $z = 1$ se l'oggetto i gode del vantaggio supposto e $z = -1$ se invece ne gode l'oggetto j. Nel caso in cui il parametro δ risulti negativo significa che l'effetto dovuto all'ordine degli oggetti nel confronto in realtà è svantaggioso.

4.2 Applicazione al campionato tedesco

Nel nostro caso di studio il modello di Bradley-Terry, come già detto in precedenza, è stato utilizzato innanzitutto per esprimere una valutazione della capacità complessiva di ognuna delle diciotto squadre del campionato, le quali rappresentano gli “oggetti” del nostro modello, e successivamente per cercare una relazione tra le prestazioni medie stagionali per ruolo di ogni squadra e i risultati delle partite durante il corso della stagione, le quali, invece, costituiscono i “confronti” del modello. Analogamente alla sezione precedente, inoltre, è stata effettuata un'ulteriore valutazione dell'influenza sugli esiti delle partite del fattore campo.

Perciò, per implementare il modello, sono stati ulteriormente riorganizzati i dati: invece che calcolare per ogni squadra le medie per ruolo dell'indice HPI in ogni giornata

come svolto per l'analisi del capitolo precedente, sono state calcolate per ogni squadra le medie complessive stagionali per ruolo dell'indice. In più è stato creato un nuovo dataset contenente per ogni riga i dati relativi a ognuna delle 306 partite disputate durante la stagione. In particolare, per ogni partita vengono riportati in ordine:

- nome della squadra in casa;
- nome della squadra in trasferta;
- variabile associata alla squadra in casa con valore 1 in caso di vittoria, 0.5 in caso di pareggio e 0 in caso di sconfitta;
- variabile associata alla squadra in trasferta con valore 1 in caso di vittoria, 0.5 in caso di pareggio e 0 in caso di sconfitta.

Le ultime due variabili costituiscono la variabile risposta del nostro modello, per le quali i pareggi sono gestiti considerando metà valore di una vittoria per ciascuna delle squadre. Come riportato in Turner & Firth (2012), l'utilizzo di questo metodo ha portato a risultati molto simili rispetto all'applicazione di analisi più sofisticate.

In primo luogo è stata implementata la versione non strutturata del modello nella quale le probabilità associate agli esiti della singola partita sono ottenute esclusivamente sulla base delle abilità delle squadre. In Tabella 4.1 sono riportate le stime di massima verosimiglianza in ordine decrescente delle log-abilità λ_i assieme ai rispettivi standard error nel modello che considera il *Rhein-Neckar Löwen* come squadra di riferimento (la cui log-abilità è posta uguale a 0 per convenzione). Da essa si può notare come l'ordine delle log-abilità delle squadre corrisponda in gran parte alla classifica finale del campionato. Unica differenza tra le due "liste" ordinate è la posizione delle squadre *HSG Wetzlar* e *TBV Lemgo Lippe*: nello specifico, le due squadre nel campionato hanno terminato con due punti di differenza in favore del Lemgo (16 vittorie, 5 pareggi e 13 sconfitte contro le 16 vittorie, 3 pareggi e 15 sconfitte del Wetzlar), mentre il modello di Bradley-Terry ha stimato la stessa log-abilità per entrambe (0.481).

Le prime quattro squadre classificate hanno registrato un notevole distacco in termini di abilità rispetto alle altre contendenti, essendo le uniche per cui il modello ha stimato una log-abilità maggiore di uno. Questa differenza si può riscontrare anche nella classifica del campionato, dove la quarta classificata (*SG Flensburg-Handewitt*) ha terminato con ben dodici punti in più della quinta (*FRISCH AUF! Göppingen*). All'interno del gruppo di testa si possono riscontrare due ulteriori distacchi rilevanti tra le squadre: tra *THW Kiel* e le due squadre di *SG Flensburg-Handewitt* e *Füchse Berlin* e tra il Kiel stesso e il *SC Magdeburg*, rispettivamente seconda e prima classificata al termine della

stagione. Quest'ultima differenza pari a 1.003 tra le log-abilità delle prime due classificate, che corrispondenza ad una probabilità di prevalenza del Magdeburg del 73.16%, rimarca la bontà della stagione svolta dal Magdeburg, caratterizzata da 32 vittorie e due sole sconfitte, contro le 28 vittorie, due pareggi e 4 sconfitte messe a segno dal Kiel.

TABELLA 4.1: Stime dei coefficienti di abilità per ogni squadra coi relativi standard error (S.E.) nel modello non strutturato.

Squadra	Abilità	S.E.
SC Magdeburg	3.357	0.855
THW Kiel	2.354	0.668
Füchse Berlin	1.587	0.585
SG Flensburg-Handewitt	1.494	0.578
FRISCH AUF! Göppingen	0.551	0.529
HSG Wetzlar	0.481	0.527
TBV Lemgo Lippe	0.481	0.527
MT Melsungen	0.205	0.523
SC DHfK Leipzig	0.205	0.523
Rhein-Neckar Löwen	0.000	0.000
Bergischer HC	-0.068	0.523
HC Erlangen	-0.206	0.524
TSV Hannover-Burgdorf	-0.206	0.524
Handball Sport Verein Hamburg	-0.275	0.526
TVB Stuttgart	-0.416	0.529
GWD Minden	-0.864	0.548
HBW Balingen-Weilstetten	-1.027	0.558
TuS N-Lübbecke	-1.201	0.571

Analizzando invece la parte bassa della classifica, le uniche squadre con stima della log-abilità inferiore a -1 corrispondono alle due squadre retrocesse nella seconda lega al termine della stagione, ovvero *HBW Balingen-Weilstetten* e *TuS N-Lübbecke*.

In seguito, al modello standard si è aggiunta la variabile relativa al fattore campo, in modo da valutare la presenza di un eventuale vantaggio derivante dallo giocare in casa. La stima risultante del coefficiente relativo alla variabile sopracitata è significativamente diversa da zero (p-value circa 0.033) ed è pari a circa 0.3 (0.29869 per l'esattezza), evidenziando perciò l'esistenza di un leggero vantaggio rispetto all'esito delle partite per le squadre che giocano in casa. In particolare, a parità delle altre esplicative la probabilità di vittoria per la squadra di casa è del 57.41%. Analizzando i risultati delle partite giocate, il 50.33% delle vittorie registrate sono per la squadra di casa contro il 40.20% per la squadra in trasferta. Il restante 9.47% rispecchia la maggior sporadicità con cui avvengono pareggi nella pallamano rispetto che in altri sport come il calcio.

A questo punto dell'analisi si sono aggiunte al modello di Bradley-Terry le variabili esplicative oggetto-specifiche corrispondenti alle medie complessive per ruolo per ogni squadra dell'indice HPI. Le log-abilità λ_i per ogni squadra i quindi vengono modellate nel seguente modo:

$$\lambda_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}, \tag{4.5}$$

con $x_{i1}, \ldots, x_{i7}$ relative alle medie per ruolo dell'indice HPI per ogni squadra i (ruoli considerati nello stesso ordine del capitolo precedente). Il rispettivo modello quindi, nella forma del modello (4.2), considera come covariate le differenze tra le medie HPI per ruolo registrate dalla squadra in casa e quelle registrate da quella in trasferta, dato che l'ordine delle squadre in ogni partita è tale per cui la prima elencata è sempre quella che gioca in casa.

In Tabella 4.2 sono riportate le stime di massima verosimiglianza dei coefficienti di regressione del modello. I coefficienti relativi al ruolo di ala sinistra (0.012) e al ruolo di portiere (0.031) non risultano significativamente diversi da zero, nel caso di quest'ultimo andando contro i risultati del modello per logit cumulati: infatti, nel modello proposto nel capitolo precedente, il ruolo del portiere è risultato tra i due più incisivi rispetto al risultato di una partita. Nel modello di Bradley-Terry stimato, invece, la differenza tra le medie HPI dei portieri della squadra di casa e di quella in traserta non sembra avere un'effetto rilevante sulla probabilità di vittoria.

Tabella 4.2: Stime dei coefficienti di regressione nella forma strutturata del modello di Bradley-Terry assieme ai rispettivi standard error.

Parametro	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Stima	0.012	-0.164	0.248	-0.321	0.162	0.263	0.031
Std. error	0.059	0.063	0.111	0.083	0.079	0.132	0.100

Risultano correlate positivamente con la probabilità di vittoria di una partita le differenze tra le medie HPI relative alle posizioni di centrale, pivot e terzino sinistro, con quest'ultimo ruolo avente coefficiente lievemente inferiore agli altri due (0.162 contro 0.263 e 0.248 rispettivamente). Aspetto interessante dei risultati ottenuti riguarda i coefficienti relativi ai ruoli che occupano il lato destro del campo. Infatti, quelli riguardanti ala destra e terzino destro sono gli unici risultati negativi nel modello ed entrambi significativamente diversi da zero. Inoltre, il coefficiente relativo al terzino destro risulta essere il più elevato di tutti in termini di valore assoluto (-0.321).

Quest'ultimo risultato può essere dovuto al fatto che, fatte alcune eccezioni, il livello dei terzini destri del campionato è abbastanza normale in confronto al livello dei terzini sinistri e dei centrali, ruoli in cui si contano più giocatori di livello assoluto che si prendono più responsabilità durante una partita. Perciò, si potrebbe ipotizzare che i terzini destri registrano livelli HPI più elevati nelle partite in cui terzini sinistri e centrali giocano peggio, abbassando le possibilità di vittoria della propria squadra.

In Tabella 4.3 sono riportate le stime delle log-abilità delle singole squadre in ordine decrescente e si possono notare dei risultati interessanti. Innanzitutto, anche questa versione del modello di Bradley-Terry premia le prestazioni del Magdeburg, prima squadra della lista per distacco anche in questo caso. Subito dietro il Flensburg, squadra che, come già visto nell'analisi esplorativa e nell'implementazione del modello per logit cumulati, ha fatto registrare durante la stagione alcuni dei migliori valori medi HPI per ruolo.

TABELLA 4.3: Stime dei coefficienti di abilità per ogni squadra coi relativi standard error (S.E.) nel modello con variabili esplicative.

Squadra	Abilità	S.E.
SC Magdeburg	18.676	6.604
SG Flensburg-Handewitt	17.310	6.368
THW Kiel	17.211	6.379
HC Erlangen	17.185	6.232
Füchse Berlin	17.100	6.295
HBW Balingen-Weilstetten	16.995	6.336
FRISCH AUF! Göppingen	16.953	6.342
MT Melsungen	16.866	6.323
TBV Lemgo Lippe	16.864	6.363
TVB Stuttgart	16.711	6.244
SC DHfK Leipzig	16.658	6.406
HSG Wetzlar	16.566	6.176
TuS N-Lübbecke	16.540	6.347
Handball Sport Verein Hamburg	16.520	6.414
TSV Hannover-Burgdorf	16.445	6.299
Bergischer HC	16.072	6.241
GWD Minden	15.849	6.425
Rhein-Neckar Löwen	15.655	6.195

Al quarto posto, tra Kiel e Berlin, si inserisce inaspettatamente *HC Erlangen*, squadra classificata al dodicesimo posto al termine della stagione. Risultati ancor più imprevisti sono il sesto posto del *HBW Balingen-Weilstetten*, squadra retrocessa a fine campionato,

e l'ultimo posto del *Rhein-Neckar Löwen*, squadra che ha terminato il campionato a metà classifica.

Questi esiti in termini di log-abilità del modello strutturato fanno dubitare della bontà delle variabili prese in considerazione. Le medie complessive dell'indice HPI potrebbero essere una misura troppo sintetica delle prestazioni di una squadra durante il corso di un'intera stagione, non dando il giusto peso ad eventuali eventi determinanti rispetto al risultato di una partita.

Come ultima fase dell'analisi mediante il modello di Bradley-Terry, al modello strutturato si è provata ad aggiungere un'intercetta casuale per ogni squadra. La formula (4.5) per le log-abilità diventa: $\lambda_i = \sum_{r=1}^{7} \beta_r x_{ir} + U_i$, dove $U_i \sim N(0, \sigma^2)$ per ogni squadra i e le variabili esplicative sono le stesse considerate nella formula (4.5).

TABELLA 4.4: Stime dei coefficienti di regressione assieme ai rispettivi standard error nel modello di Bradley-Terry che considera un'intercetta casuale.

Parametro	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Stima	0.039	-0.206	0.311	-0.377	0.177	0.309	0.043
Std. error	0.157	0.168	0.302	0.219	0.208	0.363	0.276

Come si può notare in Tabella 4.4, le stime dei coefficienti di regressione sono più elevate in valore assoluto rispetto a quelle in Tabella 4.2, con un notevole incremento in parallelo anche dei rispettivi standard error. Ciò rende tutti i coefficienti non significativamente diversi da zero.

Per quanto riguarda invece l'intercetta casuale, essa risulta altamente significativa con stima e relativo standard error rispettivamente pari circa a 0.903 e 0.240. La stima discretamente elevata dell'intercetta casuale fa presumere che il modello abbia una bassa potenza esplicativa.

Conclusioni

In questo elaborato è stata svolta un'analisi statistica incentrata nelle singole prestazioni di giocatori di pallamano, studiando il loro peso nell'economia di una partita al fine di individuare quali ruoli dei giocatori incidono maggiormente sul risultato finale di essa.

Per fare ciò, è stato creato e preso in esame un dataset contenente le informazioni relative ad alcuni indici prestazionali registrati durante ogni singola partita dai giocatori del campionato tedesco durante la stagione 2021/22. In particolare, l'analisi si è focalizzata sull'indice HPI, nuova misura introdotta nel campionato tedesco a partire dalla stagione 2020/21 che sintetizza la prestazione di un giocatore e che racchiude in essa i diversi aspetti del gioco.

Dopo una breve presentazione riguardante lo sport della pallamano in generale, nel secondo capitolo è stata svolta un'analisi esplorativa in cui inizialmente si è andato a studiare il rapporto tra l'indice HPI medio di squadra e il risultato di una partita. Quanto emerso da questa prima analisi è che se una squadra registra un HPI medio inferiore a 70, aumenta notevolmente il rischio di subire una sconfitta. Nel prosieguo, dopo aver individuato i migliori giocatori in quanto a HPI medio stagionale, l'analisi si è concentrata sulla distribuzione marginale dell'indice rispetto all'esito di una partita e al ruolo dei singoli giocatori. Come ci si poteva aspettare, il valore dell'indice è mediamente più alto in caso di vittoria, comportamento in linea con la natura di esso. Inoltre, la sua distribuzione varia a seconda del ruolo del giocatore: infatti, le posizioni maggiormente coinvolte nel gioco tendono ad avere valori dell'indice mediamente inferiori.

Successivamente, tramite due diverse tipologie di modelli si è andata studiare la relazione tra le medie per ruolo dell'indice HPI e l'esito di una partita, cercando di individuare quali posizioni incidono maggiormente. In parallelo si è valutato anche l'effetto del "fattore campo" sul risultato di una partita.

Inizialmente, considerando l'esito di una partita come variabile categoriale su scala ordinale, è stato stimato un modello per logit cumulati che ha mostrato come i portieri e i centrali hanno un peso maggiore rispetto alla probabilità di vittoria. Ruoli, invece, con

peso minore sono quelli delle ali. Il suddetto modello, inoltre, ha evidenziato inaspettatamente come il giocare in casa non influisca in maniera significativa sull'esito di una partita, forse a causa della differenza di livello tra le squadre di alta classifica e quelle di medio-bassa classifica (riscontrata anche nel seguito dell'analisi). Si sono proposti poi dei modelli alternativi, considerando diverse funzione di legame o considerando diverse ipotesi (quote non proporzionali), dai quali si sono ottenuti risultati analoghi.

Come ultima fase dell'analisi è stato stimato il modello di Bradley-Terry in due diverse versioni: dapprima, si è implementata la forma non strutturata del modello per cogliere l'effetto del fattore campo e l'effetto delle abilità delle squadre sulla probabilità di vittoria e successivamente si sono aggiunte ad esso le covariate relative alle differenze, per ogni coppia di squadre che si sono affrontate, tra i valori medi complessivi per ruolo dell'indice HPI. Il primo ha evidenziato una notevole differenza in termini di capacità tra le prime quattro classificate e le restanti squadre del campionato oltre che all'esistenza di un vantaggio dovuto allo giocare in casa, risultato in contrasto con quanto ottenuto dal modello per logit cumulati. Il secondo ha portato a risultati interessanti: infatti, esso ha messo in evidenza una correlazione negativa tra le prestazioni dei giocatori dei ruoli del lato destro del campo (tipicamente mancini) e la probabilità di vittoria. Inoltre, contrariamente a quanto visto nel modello per logit cumulati, il ruolo del portiere è risultato avere un effetto non significativo rispetto alla probabilità di vittoria.

L'aggiunta successiva di un intercetta casuale al modello ha fatto presumere una bassa potenza esplicativa di esso, forse dovuta alla scelta delle covariate. I valori medi stagionali dell'indice HPI potrebbero essere troppo sintetici, portando ad una perdita di informazione. Ciò può essere preso come spunto per approfondire l'analisi svolta, ad esempio studiando l'andamento delle abilità delle squadre durante il corso della stagione, evitando di ricorrere quindi ad indicatori eccessivamente sintetici.

Inoltre, sarebbe interessante sviluppare il modello di Bradley-Terry non considerando esclusivamente l'indice HPI, ma utilizzando nel modello gli indici prestazionali riguardanti i diversi aspetti del gioco, cercando quali hanno una maggior correlazione con la probabilità di vittoria di una partita. Altrimenti, senza cambiare covariate, si potrebbe applicare il modello a campionati più equilibrati, senza evidenti disparità tra gruppi di squadre come visto nel campionato tedesco.

Bibliografia

Agresti, A. (2013). *Categorical Data Analysis*. 3rd edition. New York: Wiley.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. New York: Wiley.

Bradley, R. A. & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs, I: The Method of Pair Comparisons. *Biometrika* **39**, 324–345.

Cattelan, M. (2012). Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. *Statistical Science* **27**, 412–433.

Clanton, R. E. & Dwight, M. P. (1997). *Team handball: steps to success*. Champaign, Ilinois: Human Kinetics.

Davidson, R. R. (1970). On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association* **65**, 317–328.

Daza, G., Andrés, A. & Tarragó, R. (2017). Match Statistics as Predictors of Team's Performance in Elite competitive Handball. *Revista Internacional de Ciencias del Deporte* **13**.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B* **42**, 109–142.

Poto, D. (2005). Handball. In *Enciclopedia Treccani*, Enciclopedia dello sport.

Prieto, J., Gómez, M.- & Sampaio, J. (2015a). A bibliometric review of the scientific production in handball. *Cuadernos de Psicología del Deporte* **15**, 145–154.

Prieto, J., Gómez, M.- & Sampaio, J. (2015b). From a static to a dynamic perspective in handball match analysis: a systematic review. *The Open Sports Sciences Journal* **8**, 25–34.

Rao, P. V. & Kupper, L. (1967). Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *Journal of the American Statistical Association* **62**, 194–204.

Saavedra, J. M., , S., Kristjánsdóttir, H., Chang, M. & Halldórsson, K. (2017). Handball game-related statistics in men at Olympic Games (2004-2016): Differences and discriminatory power. *Retos* **32**, 260–263.

Saavedra, J. M. (2018). Handball research: State of the art. *Journal of Human Kinetics* **63**, 5–8.

Salvan, A., Sartori, N. & Pace, L. (2020). *Modelli lineari generalizzati*. Springer.

Schauberg, G. & Tutz, G. (2019). BTLLasso: A Common Framework and Software Package for the Inclusion and Selection of Covariates in Bradley-Terry. *Journal of Statistical Software* **88**, 1–29.

Turner, H. & Firth, D. (2012). Bradley-terry models in R: The BradleyTerry2 Package. *Journal of Statistical Software* **48**, 1–21.