UNIVERSITA' DEGLI STUDI DI TRIESTE

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" (DEAMS)

Statistica e Informatica per l'azienda, la finanza e l'assicurazione

---

Modelli statistici per la previsione di risultati calcistici: implementazione in R del modello Dixon–Coles ed applicazioni al campionato di Serie A 2021-2022

Relatore:
Egidi Leonardo

Laureando:
Fantuzzi Giulio

Anno accademico 2022/2023

Indice

Introduzione

Il calcio è senza dubbio uno degli sport più amati a livello mondiale, e l'analisi statistica per comprenderne le dinamiche di gioco e prevederne i risultati sta assumendo un'importanza sempre più evidente. I modelli statistici per le previsioni dei risultati calcistici sono in costante aggiornamento e possono essere applicati in svariati ambiti, tra cui scommesse, analisi di dati sportivi e persino valutazione dei giocatori. L'obiettivo principale di questo progetto di tesi sarà quello di valutare l'efficacia dei modelli statistici di tipo *goal-based*, ovvero quei modelli che cercano di prevedere il numero di goal segnati dalle due squadre che si affrontano in una partita di calcio. Per prima cosa cercheremo di trovare una distribuzione di probabilità che risulti adatta a descrivere il numero di goal segnati, mentre nel seguito prenderemo in esame le principali proposte di modello presenti in letteratura scientifica. Partiremo dal modello introdotto da [1], analizzandone i punti di forza ed i limiti, in modo da comprendere quali siano gli aspetti più rilevanti da includere nella modellazione statistica dei risultati calcistici. Tra questi vedremo come sarà opportuno considerare alcuni fattori di forza d'attacco e di difesa di ciascuna squadra, la presenza di un beneficio associato al proprio stadio di casa (*home effect*) e l'influenza dello stato di forma delle squadre sulle loro prestazioni. Tutto ciò condurrà al modello [2], uno dei modelli goal-based più rinomati ed utilizzati nel mondo della statistica sportiva. Nel Capitolo 1 presenteremo i modelli da un punto di vista teorico, mentre nel Capitolo 2 analizzeremo la loro implementazione in R, che è stata realizzata adottando un approccio di tipo *from scratch*. A partire dai dati storici sul campionato di Serie A 2021-2022, nel Capitolo 3 approfondiremo alcune applicazioni concrete del modello, esplorando anche ambiti differenti dalla sola previsione dei risultati. Questo progetto di tesi, in sintesi, ambisce a fornire una panoramica completa sui modelli *goal-based* e di contribuire alla loro comprensione tramite alcune applicazioni concrete al campionato di Serie A 2021-2022, dando una dimostrazione empirica della loro efficacia.

Capitolo 1

I modelli goal-based

Nella costruzione di modelli statistici per i risultati calcistici esistono due approcci principali: quello *goal-based*, in cui si modella il numero di goal segnati dalle due squadre, e quello *result-based*, in cui si modella l'esito finale della partita (vittoria-pareggio-sconfitta). Quest'ultimo è chiaramente annidato rispetto al primo: il risultato di una partita di calcio è difatti implicato dai goal segnati e subiti, mentre la conoscenza del semplice esito non fornisce informazioni sui goal segnati dalle due squadre. Nonostante il dibattito sull'approccio da preferire sia ancora in corso [3], in questo capitolo ci concentreremo solamente sui modelli *goal-based*.

1.1 La variabile casuale di Poisson

La variabile casuale di Poisson è comunemente usata per modellare fenomeni che evolvono nel tempo (o nello spazio) e che implicano conteggi delle realizzazioni di un evento aleatorio. Assumiamo di osservare le manifestazioni di un dato fenomeno in un generico intervallo di tempo $[0, t]$. Il numero di realizzazioni (dette *arrivi*) nell'intervallo considerato è una variabile aleatoria discreta, che indicheremo con X. La funzione di probabilità di X dipende chiaramente dal modo in cui gli eventi si verificano, per cui formuleremo alcune assunzioni:

i) In un intervallo di tempo sufficientemente breve di lunghezza Δt, possono presentarsi le seguenti due situazioni: o non si verificano eventi oppure se ne verifica solo 1;

ii) La probabilità che esattamente un evento si verifichi nell'intervallo Δt è proporzionale alla lunghezza dell'intervallo (Δt) mediante una costante di proporzionalità ν (detta *intensità del processo*). Tale probabilità risulta dunque pari a $\nu \Delta t$;

iii) Eventi in intervalli di lunghezza Δt non sovrapposti sono tra loro indipendenti.

Se valgono tali condizioni, comunemente note come *assunzioni per un processo di Poisson di parametro* ν, è possibile definire la seguente variabile aleatoria discreta:

Definizione 1 (**Variabile aleatoria di Poisson**)**.** *Una variabile aleatoria che descrive il numero* X *di eventi in un intervallo di tempo* t *è detta variabile aleatoria di Poisson di parametro* $\lambda = \nu t$*, dove* ν *è il numero medio di arrivi per unità di tempo. Essa si indica con* $X \sim Pois(\lambda)$.

1.2 La distribuzione di Poisson

Basandoci ora sulle assunzioni introdotte nella Sezione 1.1, possiamo pensare di suddividere l'intervallo di tempo t in n sotto-intervalli di uguale ampiezza $\left(\Delta t = \frac{t}{n}\right)$ e di considerare la realizzazione dell'evento aleatorio in ciascuno di essi. Indicata ora con p la probabilità (costante) che l'evento si verifichi in un qualsiasi sotto-intervallo, possiamo descrivere il numero di realizzazioni nell'intervallo di ampiezza t come il risultato di n prove Bernoulliane indipendenti, ciascuna con probabilità di successo $p = \nu \cdot \Delta t = \nu \cdot \frac{t}{n}$. Segue allora:

$$P(X = x) = \binom{n}{x}\left(\frac{\nu t}{n}\right)^x \left(1 - \frac{\nu t}{n}\right)^{n-x}, \text{ con } X \sim Bin\left(n, p = \frac{\nu t}{n}\right) \quad . \tag{1.1}$$

A questo punto, calcolando il limite della (1.1) per $\Delta t \to 0$ (ossia per $n \to +\infty$), otteniamo:

$$\begin{aligned} P(X = x) &= \lim_{n\to+\infty} \binom{n}{x} \cdot \left(\frac{\nu t}{n}\right)^x \cdot \left(1 - \frac{\nu t}{n}\right)^{n-x} \\ &= \lim_{n\to+\infty} \frac{n!}{x!(n-x)!} \cdot \frac{(\nu t)^x}{n^x} \cdot \left(1 - \frac{\nu t}{n}\right)^{n} \cdot \left(1 - \frac{\nu t}{n}\right)^{-x} \\ &= \lim_{n\to+\infty} \frac{n \cdot (n-1) \cdot \ldots \cdot (n-x+1)}{n^x} \cdot \frac{(\nu t)^x}{x!} \cdot \left(1 - \frac{\nu t}{n}\right)^{n} \cdot \left(1 - \frac{\nu t}{n}\right)^{-x} \\ &= \frac{(\nu t)^x}{x!} e^{-\nu t} = \frac{\lambda^x}{x!} e^{-\lambda} \quad . \end{aligned} \tag{1.2}$$

La (1.2) è la funzione di probabilità per una Poisson di parametro $\lambda = \nu t$. Nello specifico, essa fornisce la probabilità che si verifichino esattamente x arrivi nell'intervallo continuo $(0, t]$.

1.3 La distribuzione di probabilità per il numero di goal

In una partita di calcio, ogni volta che una squadra è in possesso di palla ha l'opportunità di costruire un'azione in attacco e di segnare un goal. La storia di questo sport insegna che soltanto un ristretto numero di azioni offensive si traduce effettivamente in goal, per cui la probabilità di segnare è, in generale, una misura piuttosto bassa. Se assumiamo che tale probabilità sia costante (indichiamola con p) e che le azioni offensive siano indipendenti tra loro, ecco che il numero di goal segnati potrebbe seguire una distribuzione di probabilità di tipo binomiale. Inoltre, l'approssimazione della distribuzione binomiale con la distribuzione di Poisson, in questo contesto, è particolarmente adatta. In linea generale infatti, essa risulta valida quando il numero di prove (nel nostro caso le azioni offensive) è molto grande e quando la probabilità di successo (nel nostro caso segnare un goal) è molto piccola. Nel corso degli anni, l'idea di approssimare il numero di goal tramite una distribuzione di Poisson è stata più volte messa in discussione. Già nel 1951 [4] si rese conto di alcuni limiti della Poisson (es: l'uguaglianza tra media e varianza) e propose di optare verso una distribuzione di tipo binomiale negativa. Nonostante i vari dubbi iniziali, i suoi studi dimostrarono come l'adattamento della Poisson ai dati empirici fosse, in realtà, più che soddisfacente.

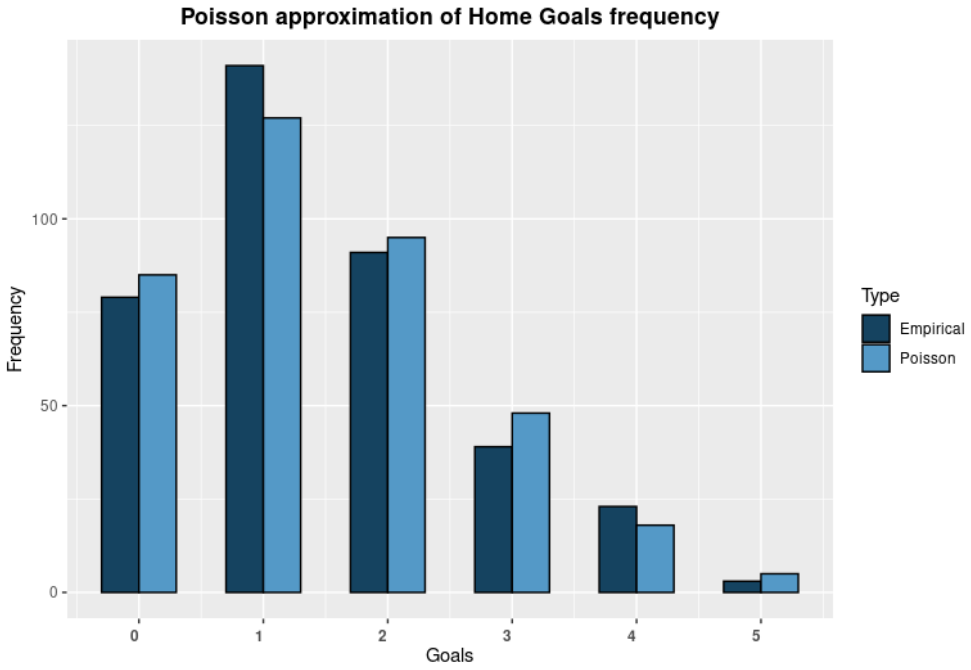

Figura 1.1: Confronto tra numero di goal osservati e teorici

In riferimento al campionato di Serie A 2021-2022, la Figura 1.1 confronta la frequenza empirica dei goal segnati dalle squadre di casa con quella stimata da una Poisson con parametro λ pari alla media aritmetica dei goal segnati dalle squadre in casa (un risultato analogo si è ottenuto dalle frequenze empiriche e teoriche dei goal segnati dalle squadre in trasferta).

La variabile di Poisson rappresenta dunque un compromesso eccellente tra facilità di modellazione ed adattamento ai dati empirici. Non a caso, infatti, essa risulta una delle scelte più comuni nell'ambito della modellazione statistica sportiva (non strettamente calcistica).

1.4 Il modello di Maher

Nel panorama dei modelli calcistici di tipo goal-based, un'interessante proposta fu introdotta nel 1982 da Maher [1], il quale sviluppò un modello in cui i punteggi delle due squadre vengono descritti mediante due distribuzioni di Poisson tra loro indipendenti.

1.4.1 Formulazione del modello

Consideriamo una partita di calcio tra due squadre, in cui per convenzione indicheremo con i la squadra di casa e con j la squadra in trasferta. Definiamo ora le seguenti variabili:

$$X_{ij} \sim Pois(\lambda_{ij}) \quad , \quad Y_{ij} \sim Pois(\mu_{ij}) \quad ,$$

dove X_{ij} rappresenta il numero di goal segnati dalla squadra i contro la squadra j ed Y_{ij} rappresenta il numero di goal segnati dalla squadra j contro la squadra i. Pertanto, il risultato osservato per la partita in questione è esprimibile mediante la coppia $(x; y)$, in cui x e y sono determinazioni rispettivamente delle variabili aleatorie X_{ij} e Y_{ij}.

Per essere più precisi, $(x; y)$ è una determinazione della coppia aleatoria $(X_{ij}; Y_{ij})$, dunque si potrebbe pensare di associare a ciascun risultato $(x; y)$ un valore di probabilità congiunta $Pr(X_{ij} = x\,; Y_{ij} = y)$. Ricordando l'ipotesi di indipendenza tra le variabili, è possibile fattorizzare la funzione di probabilità congiunta nel prodotto delle marginali ed ottenere:

$$Pr(X_{ij} = x\,; Y_{ij} = y) = Pr(X_{ij} = x) \cdot Pr(Y_{ij} = y) = \frac{(\lambda_{ij})^x}{x!} e^{-\lambda_{ij}} \cdot \frac{(\mu_{ij})^y}{y!} e^{-\mu_{ij}} \quad . \tag{1.3}$$

La notazione adottata nella (1.3) evidenzia come la media dei goal segnati da una squadra non sia una misura caratteristica della singola squadra, ma come sia influenzata dalle caratteristiche di entrambe le squadre coinvolte nella partita. Ciò risulta particolarmente intuitivo, in quanto una squadra di un certo livello avrà maggiori probabilità di segnare contro una squadra più debole, mentre incontrerà maggiori difficoltà nel farlo contro una squadra più forte. Per tenere conto di questo aspetto dovremo specificare meglio i parametri delle variabili di Poisson (λ_{ij} e μ_{ij}), esprimendoli in funzione di alcuni coefficienti che rappresentano le abilità in attacco e in difesa delle squadre coinvolte. Più precisamente, introdurremo:

- α_i: la "forza" in attacco della squadra i;
- α_j: la "forza" in attacco della squadra j;
- β_i: la "debolezza" difensiva della squadra i;
- β_j: la "debolezza" difensiva della squadra j.

È abbastanza naturale assumere che la probabilità di segnare un goal per una squadra sia proporzionale alla sua forza in attacco e alla debolezza difensiva della squadra che affronta. Considerando dunque la partita tra le squadre i e j, avremo:

$$\lambda_{ij} = \alpha_i \cdot \beta_j \quad ; \quad \mu_{ij} = \alpha_j \cdot \beta_i \quad . \tag{1.4}$$

Mettendo insieme (1.3) e (1.4), otteniamo:

$$Pr(X_{ij} = x; Y_{ij} = y) = Pr(X_{ij} = x) \cdot Pr(Y_{ij} = y) = \frac{(\alpha_i \cdot \beta_j)^x}{x!} e^{-\alpha_i \cdot \beta_j} \cdot \frac{(\alpha_j \cdot \beta_i)^y}{y!} e^{-\alpha_j \cdot \beta_i} \quad . \tag{1.5}$$

Dopo aver osservato un sufficiente numero di partite (N) è possibile ricavare i coefficienti di attacco e di difesa di tutte le squadre (n) attraverso metodi di stima di massima verosimiglianza. Si tratterà dunque di massimizzare la seguente funzione di verosimiglianza:

$$L(\alpha_i, \beta_i, i = 1, ...n) = \prod_{k=1}^{N} \frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \cdot \frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \quad , \tag{1.6}$$

con $\lambda_k = \alpha_{i(k)}\beta_{j(k)}$ e $\mu_k = \alpha_{j(k)}\beta_{i(k)}$, in cui $i(k)$ e $j(k)$ denotano rispettivamente gli indici della squadra di casa e trasferta che si affrontano nella k-esima partita.

1.4.2 Limiti del modello di Maher

Il modello di Maher è stato criticato fin da subito per la sua tendenza a sottostimare i risultati di pareggio tra le squadre ed in generale i risultati con pochi goal (0-0, 1-0, 0-1, 1-1). Inoltre, l'ipotesi di indipendenza tra i goal segnati dalle due squadre non appare così ragionevole. Spesso infatti accade che una squadra aumenti la propria intensità in attacco quando si trova in svantaggio, e che si richiuda in fase difensiva per proteggere il proprio vantaggio. Sulla base di queste considerazioni, lo stesso Maher proponeva di migliorare il modello sostituendo le due Poisson indipendenti con una Poisson bivariata, in modo da introdurre una componente di correlazione tra i goal segnati dalle due squadre che si affrontano in una partita.

1.4.3 Home effect

Il modello di Maher considera indifferentemente la squadra che gioca in casa con quella che gioca in trasferta. Tuttavia, analizzando i dati sportivi emerge una forte evidenza empirica verso l'esistenza di un "effetto casa" (*home effect*). Si tratta di un fenomeno riconosciuto per cui le squadre, quando giocano nel proprio stadio di casa, vincono più partite e segnano più goal rispetto a quando sono in trasferta [5]. Sebbene le cause dell'*home effect* ed il modo in cui esso influisca sulle prestazioni delle squadre non siano ancora del tutto chiari, possiamo ipotizzare numerosi fattori per spiegare questo beneficio, come il maggior supporto dei propri tifosi, l'assenza di fatica dovuta al viaggio, la familiarità con il proprio campo ed altri fattori sia fisici che psicologici (per maggiori dettagli si consulti [6]).

A partire dai dati relativi al campionato di Serie A 2021-2022, sono stati realizzati alcuni grafici per mostrare l'effettiva esistenza di un *home effect* e per cercare di quantificare, anche se solo graficamente, il suo impatto verso i goal fatti, i goal subiti ed i risultati utili conquistati da ciascuna squadra. Tali grafici sono proposti nelle pagine seguenti, nelle Figure 1.2- 1.7.

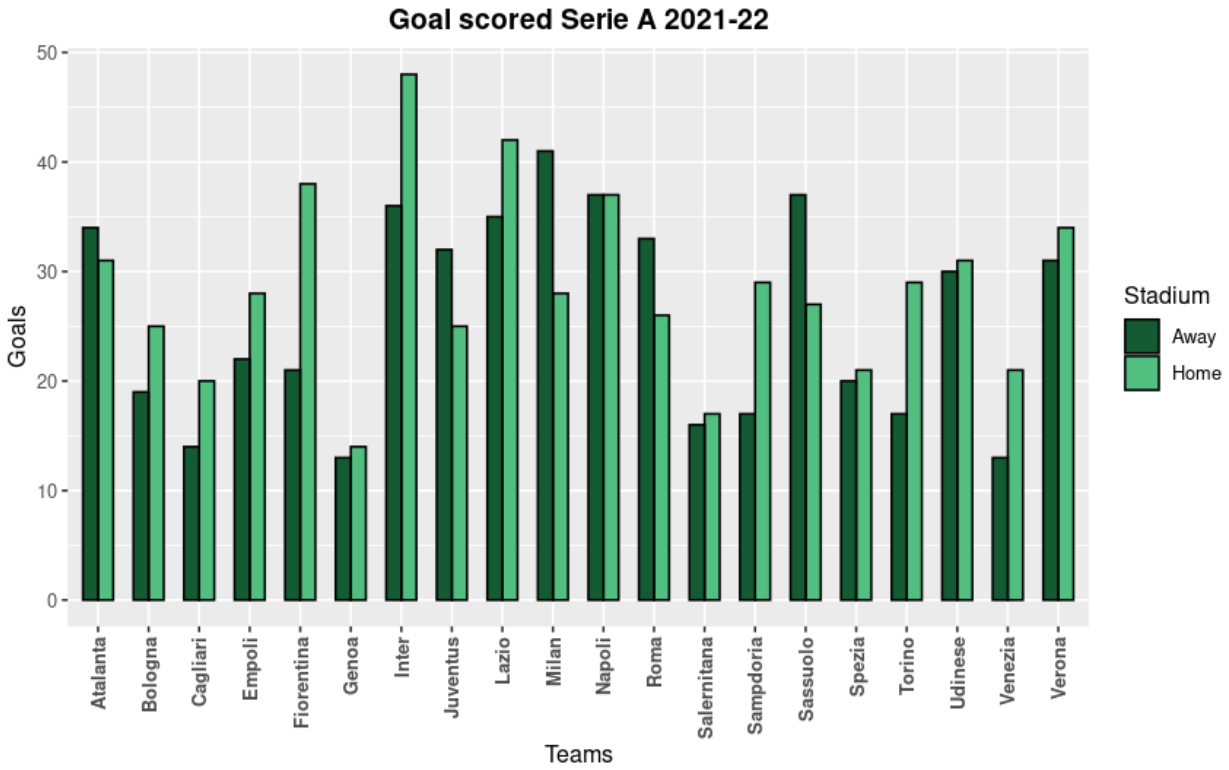

Figura 1.2: Goal segnati

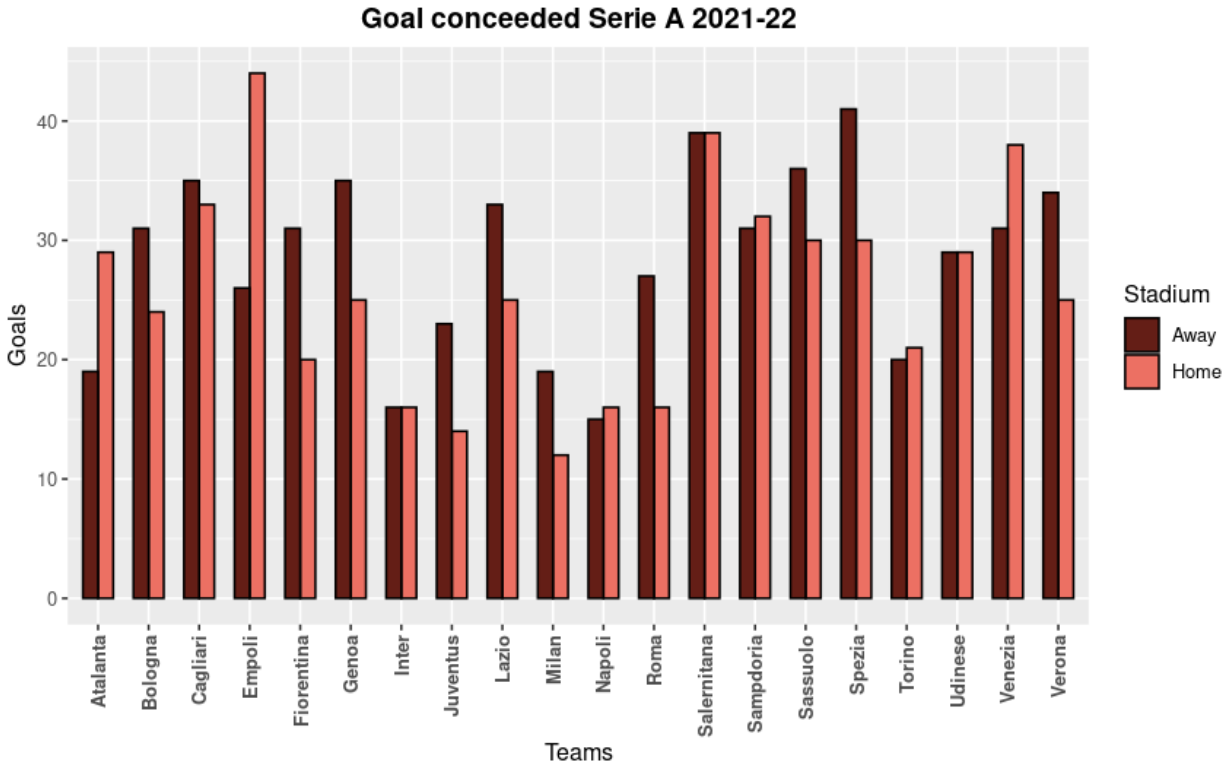

Figura 1.3: Goal subiti

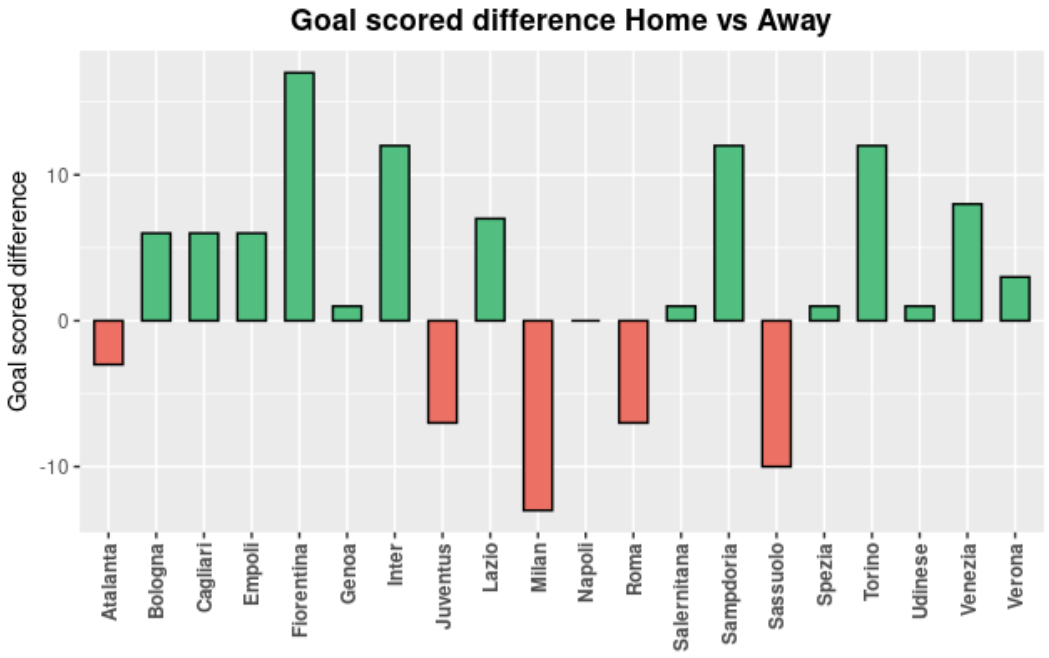

Figura 1.4: Goal segnati: differenza casa e trasferta

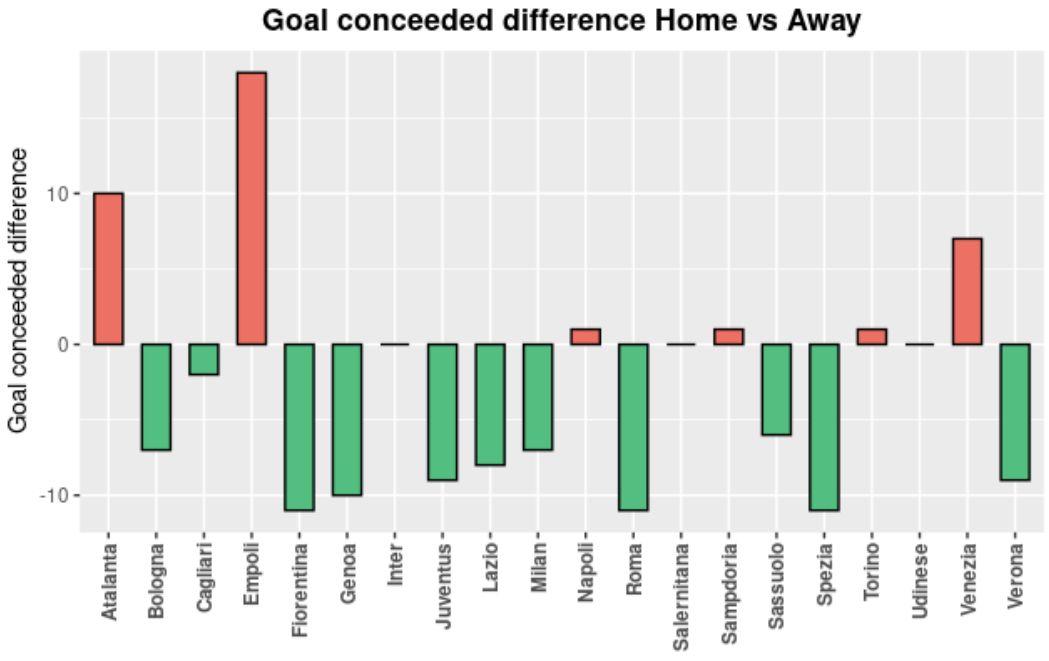

Figura 1.5: Goal subiti: differenza casa e trasferta

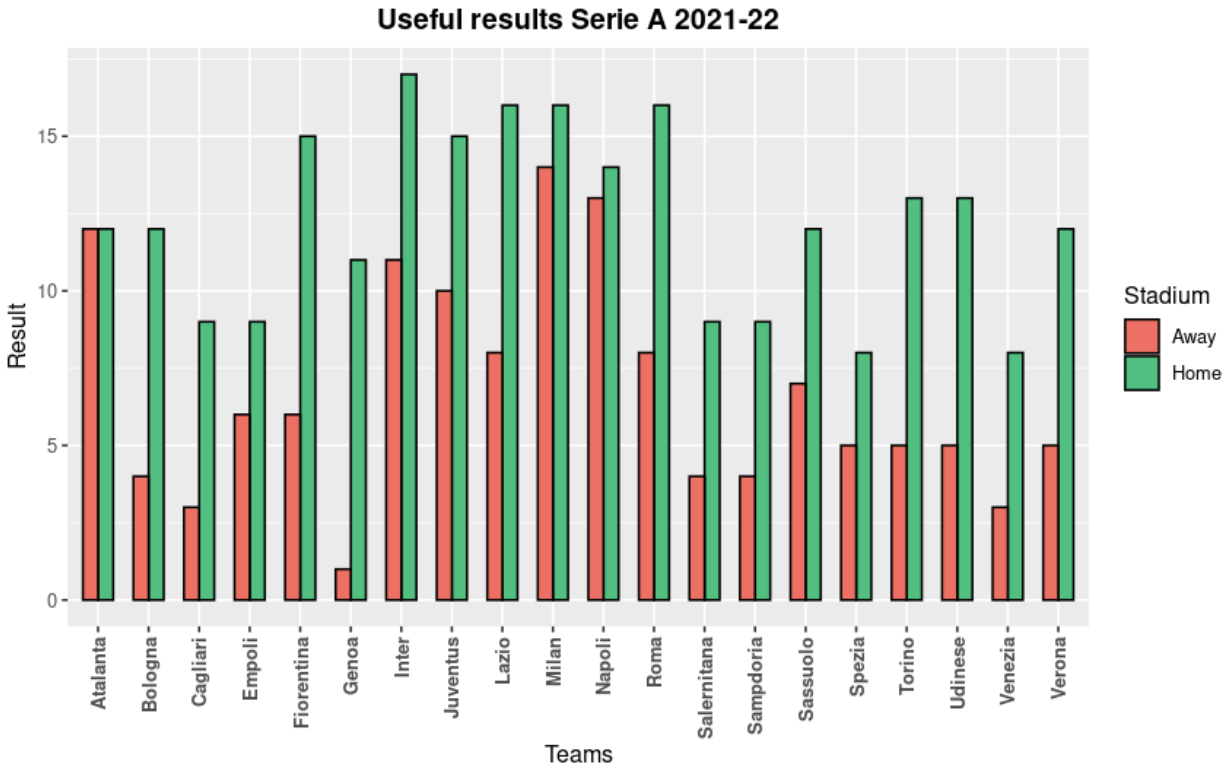

Figura 1.6: Risultati utili

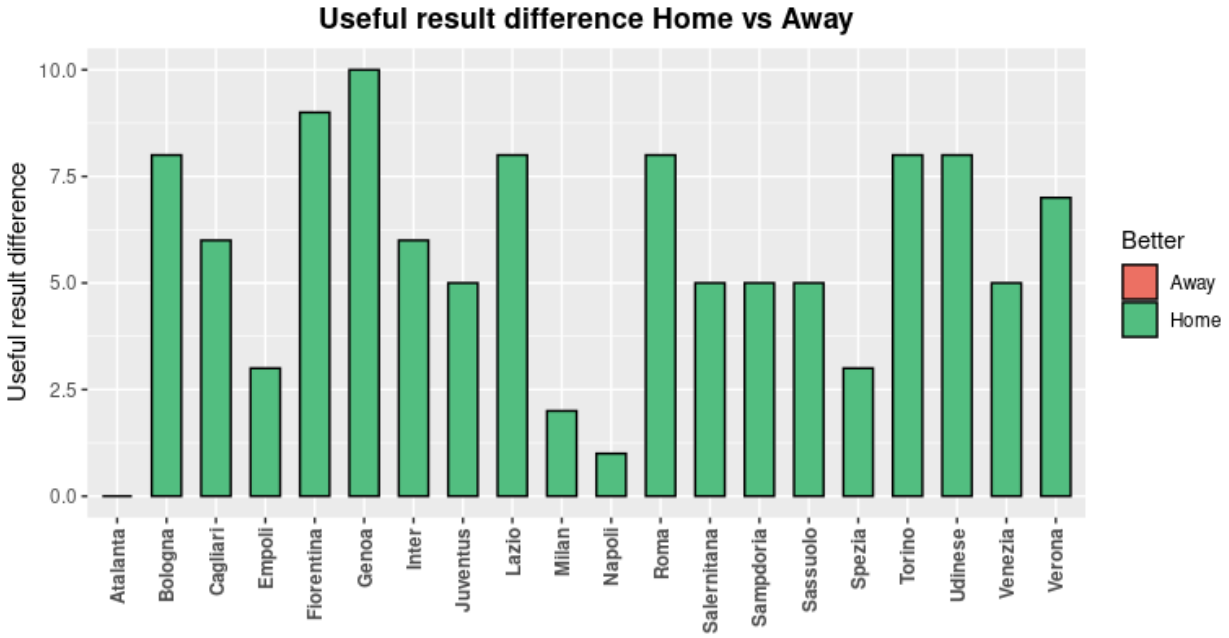

Figura 1.7: Risultati utili: differenza casa e trasferta

La Figura 1.2 mostra il numero di goal realizzati da ciascuna squadra, distinguendo quelli segnati in casa da quelli in trasferta. Analogamente, la Figura 1.3 mostra il numero di goal subiti da ciascuna squadra, distinguendo quelli concessi in casa da quelli in trasferta.

Al fine di visualizzare l'*home effect* in modo ancora più immediato, si è valutato di rappresentare per ciascuna squadra la differenza tra i goal segnati (subiti) in casa da quelli in trasferta. Nello specifico, dalla Figura 1.4 emerge come soltanto cinque squadre (sulle venti totali) abbiano segnato più goal in trasferta che in casa. Eccetto il Napoli, che in casa e trasferta ha segnato lo stesso numero di goal, tutte le altre squadre hanno avuto prestazioni offensive migliori giocando nel proprio stadio. Simmetricamente, dalla Figura 1.5 emerge come soltanto sei squadre abbiano subito più goal in casa che in trasferta. Tutte le altre squadre hanno manifestato prestazioni difensive migliori giocando nel proprio stadio (Inter, Salernitana e Udinese, in realtà, hanno concesso lo stesso numero di goal in casa e in trasferta).

Infine, le Figure 1.6 e 1.7 rappresentano l'*home effect* in relazione ai risultati utili (vittorie e pareggi) conseguiti da ciascuna squadra. In particolare, il risultato mostrato dalla Figura 1.7 è sorprendente: ad eccezione dell'Atalanta (in cui la differenza è pari a 0), tutte le squadre presentano differenze positive. In altri termini, nessuna squadra ha guadagnato dalle partite in trasferta più punti rispetto a quelli ottenuti dalle partite disputate nel proprio stadio.

In sintesi, le squadre sembrano avere un miglior rendimento quando giocano davanti al proprio pubblico. Tale aspetto si può apprezzare sia in termini di prestazioni offensive (più goal segnati) sia in termini di prestazioni difensive (meno goal subiti). Ciò si riflette anche sui risultati utili conseguiti, a dimostrazione di quanto l'*home effect* sia un aspetto fondamentale da dover introdurre in un modello statistico per la previsione di risultati calcistici.

1.5 Il modello di Dixon e Coles

Nell'ambito dei modelli statistici *goal-based*, un enorme contributo venne apportato da Dixon e Coles [2] nel 1997. Come abbiamo avuto modo di anticipare nella Sezione 1.4, il modello di Maher presentava forti limiti nella sua formulazione generale ed in alcune sue ipotesi, ma allo stesso tempo costituiva un ottimo punto di partenza per lo sviluppo di modelli più articolati. Dixon e Coles, partendo proprio dai risultati di Maher, svilupparono un modello che ne superasse i limiti e che tenesse conto di ulteriori aspetti, come ad esempio lo stato di forma delle squadre. Nelle sezioni seguenti approfondiremo meglio la formulazione e le caratteristiche del modello Dixon-Coles che, ad oggi, rappresenta una delle proposte più rinomate ed utilizzate nel panorama della statistica sportiva.

1.5.1 Formulazione del modello base

In analogia a quanto proposto nella Sezione 1.4.1, consideriamo una partita di calcio tra due squadre, indicando sempre con i la squadra di casa e con j la squadra in trasferta. Definiamo:

$$X_{ij} \sim Pois(\lambda_{ij}) \quad , \quad Y_{ij} \sim Pois(\mu_{ij}) \quad ,$$

dove X_{ij} rappresenta il numero di goal segnati dalla squadra i contro la squadra j ed Y_{ij} rappresenta il numero di goal segnati dalla squadra j contro la squadra i. Come nel modello di Maher, esprimiamo λ_{ij} e μ_{ij} in funzione di alcuni coefficienti che sintetizzano le abilità offensive e difensive delle squadre che si affrontano. Alla luce di quanto osservato nella Sezione 1.4.3, vogliamo tenere conto anche del vantaggio che generalmente hanno le squadre che giocano in casa. Possiamo dunque specificare i parametri delle Poisson come segue:

$$\lambda_{ij} = \alpha_i \cdot \beta_j \cdot \gamma \quad , \quad \mu_{ij} = \alpha_j \cdot \beta_i \quad , \tag{1.7}$$

in cui gli α_i sono i coefficienti di forza d'attacco ed i β_i sono i coefficienti di debolezza difensiva, mentre γ è un fattore che rappresenta l'*home effect*. Riflettendo sul significato teorico dell'*home effect*, si deduce facilmente come debba valere $\gamma > 0$. Notiamo inoltre che γ, a differenza di λ_{ij} e μ_{ij}, non è indicizzato, dunque l'effetto stadio viene assunto in egual misura per tutte le (n) squadre. Inizialmente Dixon e Coles provarono a considerare

un fattore γ_i specifico per ogni squadra, ma si accorsero che i risultati, in termini previsivi, non miglioravano significativamente. Non valeva dunque la pena inserire nel modello $n-1$ parametri aggiuntivi per descrivere l'*home effect*.

1.5.2 Correzione per la non indipendenza: la funzione Tau

Come anticipato nella Sezione 1.4.2 il modello di Maher tende a sottostimare i risultati di pareggio tra le squadre ed i risultati con pochi goal (0-0, 1-0, 0-1, 1-1), oltre al fatto che l'ipotesi di indipendenza tra i goal non risulta particolarmente ragionevole. Con l'obiettivo di inserire una correzione che tenesse conto dell'interdipendenza per i risultati sopracitati, Dixon e Coles apportarono una modifica sostanziale al modello. Nello specifico, introdussero un parametro ρ ed una funzione τ. La funzione di probabilità diventava così:

$$Pr(X_{ij}=x;Y_{ij}=y)=\tau_{\lambda_{ij},\mu_{ij}}(x,y)\cdot\frac{(\alpha_i\cdot\beta_j\cdot\gamma)^x}{x!}e^{-\alpha_i\cdot\beta_j\cdot\gamma}\cdot\frac{(\alpha_j\cdot\beta_i)^y}{y!}e^{-\alpha_j\cdot\beta_i} \quad , \tag{1.8}$$

dove:

$$\tau_{\lambda_{ij},\mu_{ij}}(x,y)=\begin{cases}1-\lambda_{ij}\cdot\mu_{ij}\cdot\rho & x=y=0\\ 1+\lambda_{ij}\cdot\rho & x=0,y=1\\ 1+\mu_{ij}\cdot\rho & x=1,y=0\\ 1-\rho & x=y=1\\ 1 & \text{altrimenti}\end{cases} \tag{1.9}$$

e con ρ tale che:

$$\max\left\{-\frac{1}{\lambda_{ij}},-\frac{1}{\mu_{ij}}\right\}\leq\rho\leq\min\left\{\frac{1}{\lambda_{ij}\mu_{ij}},1\right\} \quad . \tag{1.10}$$

ρ rappresenta quindi un parametro di dipendenza: se $\rho=0$ la funzione τ vale 1 (a prescindere dai valori di x ed y), dunque la (1.8) si riduce alla formulazione nel caso di indipendenza (1.3). Inoltre, Dixon e Coles dimostrarono che nonostante l'introduzione della funzione τ, le distribuzioni marginali di X_{ij} ed Y_{ij} rimanevano ancora $X_{ij}\sim Pois(\lambda_{ij})$ ed $Y_{ij}\sim Pois(\mu_{ij})$.

Se consideriamo un campionato composto da n squadre, il modello introdotto prevede n parametri d'attacco $(\alpha_1,...,\alpha_n)$, n parametri difensivi $(\beta_1,...,\beta_n)$, un parametro per rappresentare l'*home effect* (γ) ed un parametro di dipendenza(ρ). In totale ci saranno dunque $2n+2$ parametri, da stimare sulla base dei dati mediante tecniche di massima verosimiglianza. Dopo aver osservato un sufficiente numero di partite (N), si dovranno ricercare i parametri che massimizzano la seguente funzione di verosimiglianza (o la relativa log-verosimiglianza):

$$L(\alpha_i,\beta_i,\gamma,\rho,i=1,...n)=\prod_{k=1}^{N}\tau_{\lambda_k,\mu_k}(x_k,y_k)\frac{(\lambda_k)^{x_k}}{x_k!}e^{-\lambda_k}\cdot\frac{(\mu_k)^{y_k}}{y_k!}e^{-\mu_k} \quad , \tag{1.11}$$

con $\lambda_k=\alpha_{i(k)}\beta_{j(k)}\gamma$ e $\mu_k=\alpha_{j(k)}\beta_{i(k)}$, in cui $i(k)$ e $j(k)$ denotano rispettivamente gli indici della squadra di casa e trasferta nella k-esima partita. Inoltre, per garantire che il modello statistico risulti identificabile, viene imposto un ulteriore vincolo sui parametri:

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i=1 \quad ; \quad \frac{1}{n}\sum_{i=1}^{n}\beta_i=1 \quad . \tag{1.12}$$

Nelle sezioni seguenti la (1.12) verrà considerata in una forma del tutto equivalente, ottenibile mediante una semplice traslazione. Nello specifico, considereremo il vincolo:

$$\sum_{i=1}^{n-1}\alpha_i=-\alpha_n \quad ; \quad \sum_{i=1}^{n-1}\beta_i=-\beta_n \quad . \tag{1.13}$$

Si noti che in questo modo i parametri da stimare non saranno più $2n+2$, ma $2n$.

1.5.3 Lo stato di forma delle squadre

Un limite strutturale del modello (1.11) può essere identificato nella staticità dei suoi parametri. Essa, infatti, implica che le squadre abbiano un rendimento costante nel corso del tempo, sia in fase offensiva che in fase difensiva. Chiaramente un approccio di questo tipo non è molto plausibile: nel corso di una stagione sportiva ogni squadra è caratterizzata inevitabilmente da momenti più o meno prolifici. Non bisogna dunque trascurare l'importante impatto che il condizionamento fisico e mentale possono avere nei confronti delle prestazioni delle squadre che si affrontano in una partita. Lo stato di forma delle squadre diventa pertanto un elemento determinante da includere nel modello, in quanto può influenzare significativamente l'esito di una partita di calcio e, più in generale, di una qualunque competizione sportiva.

Alla luce delle precedenti considerazioni, le abilità delle squadre ora verranno assunte come dei parametri dinamici che evolvono nel corso della stagione. Nell'ambito particolare della previsione dei risultati, l'idea generale su cui ci si basa è che le prestazioni di una squadra siano correlate maggiormente ai risultati delle partite più recenti rispetto a quelli di partite più distanti nel tempo. Esaminiamo ora un esempio(tratto da [7]) per chiarire tale assunzione. Consideriamo una squadra che ha giocato 6 partite e valutiamo due diverse serie di risultati: V-V-P-P-S-S e P-P-S-S-V-V, in cui V indica una vittoria, P un pareggio ed S una sconfitta. In entrambi i casi la squadra ha ottenuto 2 vittorie, 2 pareggi e 2 sconfitte. Tuttavia, se nel primo caso la sequenza di risultati mostra un chiaro calo di forma della squadra, nel secondo caso sembra emergere un trend positivo delle prestazioni della squadra. Per prevedere con maggiore accuratezza l'esito della settima partita, diventa dunque essenziale tenere conto anche di questa informazione sullo stato di forma della squadra considerata.

Dixon e Coles, nel loro articolo, spiegarono come tale approccio dinamico potesse essere modellato formalizzando uno sviluppo stocastico dei parametri del modello. Tuttavia, data la dimensionalità del modello e dato il loro interesse ad usare il modello solo in ottica di scommesse, essi si limitarono ad un approccio più semplicistico. Tale approccio prevede che i parametri siano localmente costanti nel tempo e che le informazioni passate abbiano meno valore di quelle più recenti. Inoltre, i parametri ad ogni istante di tempo t sono stimati sulla base di tutte partite giocate fino a quel momento. Dixon e Coles apportarono così una sostanziale modifica all'equazione (1.11), in modo da includere nella funzione di verosimiglianza anche le informazioni relative alla data in cui vengono disputate le partite.

$$L(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \prod_{k \in A_t} \left[\tau_{\lambda_k, \mu_k}(x_k, y_k) \frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \cdot \frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \right]^{\phi(t-t_k)} , \qquad (1.14)$$

in cui t_k è la data in cui è stata disputata la k-esima partita, $A_t \doteq \{k : t_k \leq t\}$ e $\phi(\cdot)$ è una funzione non-crescente che serve a pesare i dati in base alla loro collocazione temporale. I parametri λ_k e μ_k, invece, sono definiti esattamente come nella (1.11). Nonostante il leggero abuso di notazione, è importante notare che i parametri α, β, γ e ρ sono ora dipendenti dal tempo di riferimento t. Se si modifica t, infatti, si modificherà anche l'insieme A_t dei dati che vengono considerati nella (1.11), dunque cambierà anche la stima dei parametri.

In definitiva, massimizzare la funzione (1.11) alla data t ci permetterà ancora di trovare le stime di massima verosimiglianza dei parametri, ma questa volta le stime saranno riferite solo a quella precisa data. Tale procedura può essere ripetuta nel corso del tempo, e valutando la serie delle stime per diverse date di riferimento è possibile ottenere una sorta di serie storica dell'andamento delle squadre (lo approfondiremo meglio nella Sezione 3.2).

Come anticipato, Dixon e Coles modificarono il modello base inserendo una funzione non-crescente $\phi(\cdot)$, il cui ruolo è quello di pesare in modo diverso (nella funzione di verosimiglianza) le partite più lontane nel tempo (rispetto all'istante di riferimento t). La prima soluzione proposta dai due studiosi consiste in una funzione così fatta:

$$\phi(t) = \begin{cases} 1 & t \leq t_0 \\ 0 & t > t_0 \end{cases} \tag{1.15}$$

In questo modo vengono inclusi nella funzione di verosimiglianza solo gli incontri per cui la differenza $t - t_k$ è minore di una soglia massima (fissata a priori) t_0, mentre vengono esclusi tutti quegli incontri "lontani" in cui tale differenza supera la soglia. Tuttavia, dalla formulazione della (1.15) emerge un grosso limite: i dati inclusi nella verosimiglianza vengono pesati allo stesso modo, contraddicendo molte delle considerazioni fatte precedentemente.

La seconda soluzione proposta da Dixon e Coles, invece, include tutte le partite precedenti alla data di riferimento, ponderandole sulla base della distanza $(t - t_k)$. In particolare:

$$\phi(t) = e^{-\xi t} \quad , \tag{1.16}$$

in cui il parametro ξ rappresenta un fattore di lisciamento, al variare del quale è possibile dare più o meno peso alle partite più lontane. In generale, aumentando ξ si assegna maggior peso alle partite più recenti e minor peso alle partite più distanti (e viceversa diminuendo ξ). Notiamo inoltre che quando $\xi = 0$, la (1.14) si riduce al modello statico (1.11).

La stima del parametro ξ, infine, non risulta affatto un procedimento semplice, ma questo problema verrà affrontato più nel dettaglio nella Sezione 2.2.3.1.

Capitolo 2

Implementazione in R dei modelli

Nel presente capitolo verrà approfondita l'implementazione, tramite il software R, dei modelli goal based presentati nel Capitolo 1 . L'approccio adottato è stato di tipo *from scratch*, con l'obiettivo di costruire le principali funzioni da zero, garantendo una maggiore comprensione dei meccanismi sottostanti ed una maggior flessibilità in termini di modellazione. Nello specifico, sono state implementate le funzioni di log-verosimiglianza dei diversi modelli e sono stati realizzati alcuni script di codice per stimare i parametri ottimali secondo la teoria della massima verosimiglianza. I risultati ottenuti dall'implementazione dei modelli sono stati poi convalidati e confrontati tra loro facendo riferimento ai dati della Serie A 2021-2022.

La parte computazionale di questo progetto è disponibile alla seguente repository GitHub: https://github.com/giuliofantuzzi/BSc-thesis.

2.1 Cenni alla teoria della verosimiglianza

La teoria della verosimiglianza è un concetto fondamentale dell'inferenza statistica e fornisce un approccio formale per stimare i parametri di un modello statistico in base ai dati osservati. Nello specifico, essa si basa su una funzione di verosimiglianza, che esprime la probabilità di osservare i dati campionari, dato un insieme specifico di parametri del modello.

Definizione 2 (**Funzione di verosimiglianza**)**.** *Osservato un campione casuale* $(y_1, y_2, ..., y_n)$ *di dimensione n da Y che è distribuita secondo il modello* $f(y;\theta)$*, ove* $\theta \in \Theta$ *è uno scalare e*

$\Theta \subset \mathbb{R}^m$, *si definisce funzione di verosimiglianza per* θ *la funzione:*

$$L(\theta) = f(y_1, y_2, ..., y_n; \theta) \quad .$$

Essa è una funzione definita su Θ *e a valori in* $\mathbb{R}^+$*. Nel caso di campionamento casuale, per la condizione di indipendenza delle variabili aleatorie da cui sono tratte le realizzazioni campionarie (nonché per l'identica distribuzione delle stesse) si ha:*

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta) \quad .$$

Per ogni $\theta \in \Theta$, la funzione $L(\theta)$ fornisce una misura di quanto è plausibile (verosimile) che il campione osservato sia generato da un modello con quello specifico parametro θ: i valori di θ per i quali $L(\theta)$ è più alto risultano allora quelli più plausibili. Sulla base di questa idea si può pensare di scegliere come stima del parametro θ quel valore $\hat{\theta}_{MV} : L(\hat{\theta}_{MV}) \geq L(\theta) \forall \theta \in \Theta$. Esso è detto *stima di massima verosimiglianza (SMV)*, ed è definito formalmente come segue:

$$\hat{\theta}_{MV} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(y_i; \theta) \quad . \tag{2.1}$$

Alcune note pratiche:

- se la funzione di verosimiglianza $L(\theta)$ viene moltiplicata per una costante c non dipendente da θ, il valore della SMV non cambia. Analogamente, se una funzione di verosimiglianza è esprimibile a meno di una costante moltiplicativa c (non dipendente da θ), tale costante può essere tranquillamente ignorata per determinare $\hat{\theta}_{MV}$;
- per determinare $\hat{\theta}_{MV}$ potrebbe essere più conveniente lavorare sulla funzione di log-verosimiglianza, definita come $l(\theta) = \log L(\theta)$. Molto spesso ciò semplifica i calcoli, ed essendo il logaritmo una trasformazione monotona crescente in senso stretto, non cambia assolutamente nulla in termini di massimizzazione della funzione.

2.2 La stima dei parametri

2.2.1 Modello di Maher

Nella Sezione 1.4.1 avevamo ottenuto la seguente funzione di verosimiglianza (1.6)

$$L(\alpha_i, \beta_i, i = 1, \dots n) = \prod_{k=1}^{N} \frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \cdot \frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \quad .$$

Da essa si può facilmente ricavare la funzione di log-verosimiglianza:

$$l(\alpha_i, \beta_i, i = 1, \dots n) = \sum_{k=1}^{N} \log\left(\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k}\right) + \log\left(\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k}\right) \quad . \tag{2.2}$$

In riferimento al campionato di Serie A 2021-2022, le seguenti tabelle mostrano i coefficienti del modello di Maher che sono stati stimati massimizzando la (2.2), considerando tutte le partite della stagione. Ricordando il significato dei coefficienti α e β, è stato poi possibile ordinare le squadre sulla base delle loro abilità offensive (Tabella 2.1) e difensive (Tabella 2.2).

Squadra	α	**s.e.**	**IC (95%)**
Udinese	0.641	0.082	[0.633 ; 0.650]
Napoli	0.628	0.073	[0.621 ; 0.635]
Inter	0.594	0.065	[0.588 ; 0.601]
Lazio	0.582	0.066	[0.576 ; 0.589]
Sampdoria	0.576	0.085	[0.567 ; 0.584]
Milan	0.571	0.069	[0.564 ; 0.578]
Verona	0.540	0.067	[0.533 ; 0.547]
Atalanta	0.532	0.066	[0.525 ; 0.539]
Sassuolo	0.525	0.066	[0.518 ; 0.531]
Juventus	0.523	0.069	[0.516 ; 0.530]
Empoli	0.516	0.073	[0.509 ; 0.524]
Bologna	0.494	0.075	[0.487 ; 0.502]
Torino	0.487	0.072	[0.480 ; 0.495]
Roma	0.485	0.063	[0.478 ; 0.491]
Spezia	0.481	0.075	[0.473 ; 0.488]
Fiorentina	0.469	0.061	[0.463 ; 0.475]
Cagliari	0.457	0.078	[0.449 ; 0.465]
Salernitana	0.451	0.079	[0.443 ; 0.459]
Genoa	0.448	0.086	[0.439 ; 0.457]
Venezia	0.421	0.072	[0.413 ; 0.428]

Tabella 2.1: Stime dei coefficienti d'attacco (ordinati dal migliore al peggiore)

Squadra	β	**s.e.**	**IC (95%)**
Milan	0.363	0.065	[0.357 ; 0.370]
Atalanta	0.395	0.057	[0.390 ; 0.401]
Spezia	0.405	0.048	[0.400 ; 0.410]
Inter	0.446	0.079	[0.438 ; 0.454]
Juventus	0.469	0.077	[0.461 ; 0.477]
Verona	0.481	0.063	[0.475 ; 0.488]
Torino	0.490	0.077	[0.482 ; 0.498]
Roma	0.491	0.075	[0.484 ; 0.499]
Cagliari	0.493	0.060	[0.487 ; 0.499]
Fiorentina	0.506	0.071	[0.499 ; 0.513]
Sampdoria	0.516	0.065	[0.510 ; 0.523]
Genoa	0.517	0.067	[0.511 ; 0.524]
Bologna	0.523	0.071	[0.516 ; 0.530]
Napoli	0.528	0.095	[0.518 ; 0.537]
Sassuolo	0.547	0.067	[0.540 ; 0.554]
Lazio	0.567	0.075	[0.559 ; 0.574]
Venezia	0.573	0.069	[0.566 ; 0.580]
Empoli	0.576	0.069	[0.569 ; 0.583]
Udinese	0.580	0.076	[0.572 ; 0.587]
Salernitana	0.585	0.066	[0.579 ; 0.592]

Tabella 2.2: Stime dei coefficienti di difesa (ordinati dal migliore al peggiore)

I valori di queste tabelle, confrontati con il reale rendimento delle squadre nel campionato considerato, non forniscono risultati sempre coerenti e verosimili. Ad esempio, il Milan e il Napoli sono le due squadre che in totale hanno subito il minor numero di goal (31). Il coefficiente di debolezza difensiva del Milan effettivamente risulta il più basso, ma quello del Napoli non è sicuramente tra i migliori. Riguardo la fase offensiva, invece, l'Udinese appare la squadra con il più alto coefficiente di forza in attacco, ma nel campionato reale altre squadre hanno performato decisamente meglio, realizzando in totale molti più goal. Altrettanto improbabile risulta il coefficiente difensivo dello Spezia: sebbene nella Tabella 2.2 esso appaia in terza posizione, la squadra ligure si è rivelata essere la seconda peggior difesa del campionato (dopo la Salernitana), con un totale di 71 goal subiti.

Grazie a questi semplici esempi possiamo dunque intuire come il modello di Maher sia un modello ancora troppo semplice per ottenere risultati soddisfacenti. Come già anticipato nella Sezione 1.4.2, esso presenta infatti forti limiti nella sua formulazione generale, ma risulta comunque un ottimo punto di partenza per lo sviluppo di modelli più sofisticati.

2.2.2 Modello Dixon-Coles statico

Consideriamo intanto la versione più semplice del modello Dixon-Coles, ossia quella che non include il fattore temporale. In quel caso, avevamo ottenuto la verosimiglianza (1.11):

$$L(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \prod_{k=1}^{N} \tau_{\lambda_k, \mu_k}(x_k, y_k) \frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \cdot \frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \quad .$$

I vari x_k ed y_k sono dati di cui si dispone, dunque risultano delle costanti note che non dipendono dai parametri del modello. Per questo motivo, nell'articolo di Dixon e Coles veniva proposta una versione più semplice della (1.11), ossia la funzione di (pseudo)verosimiglianza:

$$L(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \prod_{k=1}^{N} \tau_{\lambda_k, \mu_k}(x_k, y_k) (\lambda_k)^{x_k} e^{-\lambda_k} \cdot (\mu_k)^{y_k} e^{-\mu_k} \quad . \tag{2.3}$$

La rispettiva log-verosimiglianza risultava così:

$$l(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \sum_{k=1}^{N} \log[\tau_{\lambda_k, \mu_k}(x_k, y_k)] + x_k \log(\lambda_k) - \lambda_k + y_k \log(\mu_k) - \mu_k \quad . \tag{2.4}$$

In questo progetto di tesi si è valutato di stimare i parametri massimizzando la funzione di log-verosimiglianza, ma senza ricorrere alla versione semplificata proposta da Dixon e Coles. Infatti, è risultato molto più semplice partire dalla (1.11) e scrivere la log-verosimiglianza:

$$l(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \sum_{k=1}^{N} \log[\tau_{\lambda_k, \mu_k}(x_k, y_k)] + \log\left[\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k}\right] + \log\left[\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k}\right] , \tag{2.5}$$

così da sfruttare, a livello computazionale, la funzione *dpois()* già presente nel software R. Inoltre, anziché esprimere λ_{ij} e μ_{ij} come nella (1.7), si è fatto riferimento alla formulazione:

$$\lambda_{ij} = e^{\alpha_i + \beta_j + \gamma} \quad , \quad \mu_{ij} = e^{\alpha_j + \beta_i} \quad . \tag{2.6}$$

Essendo l'esponenziale una trasformazione strettamente crescente, non cambia nulla in termini di massimizzazione. Inoltre, ricordiamo che il parametro di una Poisson deve assumere valori in $\mathbb{R}^+$. La (2.6) garantisce che tale vincolo venga rispettato senza dover imporre condizioni aggiuntive sui parametri α e β (che invece potranno assumere valori negativi). In questo modo non c'è il rischio che R restituisca avvisi e/o errori durante la fase di ottimizzazione.

In analogia a quanto fatto per il modello di Maher, si sono ottenute le Tabelle 2.3 e 2.4:

Squadra	α	**s.e.**	**IC (95%)**
Inter	0.504	0.112	[0.493 ; 0.515]
Lazio	0.440	0.117	[0.428 ; 0.451]
Napoli	0.358	0.119	[0.346 ; 0.370]
Milan	0.291	0.123	[0.279 ; 0.304]
Verona	0.253	0.127	[0.240 ; 0.266]
Atalanta	0.246	/	/
Sassuolo	0.245	0.128	[0.232 ; 0.258]
Udinese	0.181	0.131	[0.167 ; 0.194]
Fiorentina	0.135	0.133	[0.122 ; 0.149]
Roma	0.133	0.133	[0.120 ; 0.147]
Juventus	0.083	0.136	[0.069 ; 0.097]
Empoli	-0.017	0.145	[-0.031 ; -0.002]
Sampdoria	-0.124	0.152	[-0.139 ; -0.109]
Torino	-0.153	0.152	[-0.168 ; -0.138]
Bologna	-0.177	0.155	[-0.192 ; -0.161]
Spezia	-0.247	0.161	[-0.263 ; -0.231]
Cagliari	-0.460	0.179	[-0.478 ; -0.442]
Venezia	-0.463	0.179	[-0.481 ; -0.445]
Salernitana	-0.494	0.182	[-0.512 ; -0.475]
Genoa	-0.736	0.204	[-0.756 ; -0.715]

Tabella 2.3: Stime dei coefficienti d'attacco (ordinati dal migliore al peggiore)

Squadra	β	**s.e.**	**IC (95%)**
Milan	-0.584	0.188	[-0.603 ; -0.565]
Napoli	-0.579	0.188	[-0.598 ; -0.560]
Inter	-0.522	0.185	[-0.540 ; -0.503]
Juventus	-0.388	0.171	[-0.406 ; -0.371]
Torino	-0.283	0.161	[-0.300 ; -0.267]
Roma	-0.214	0.157	[-0.230 ; -0.198]
Atalanta	-0.074	/	/
Fiorentina	-0.028	0.144	[-0.042 ; -0.013]
Bologna	0.050	0.138	[0.036 ; 0.064]
Udinese	0.126	0.135	[0.113 ; 0.140]
Genoa	0.133	0.132	[0.120 ; 0.146]
Lazio	0.144	0.135	[0.130 ; 0.157]
Verona	0.150	0.133	[0.136 ; 0.163]
Sampdoria	0.201	0.129	[0.188 ; 0.214]
Cagliari	0.270	0.124	[0.258 ; 0.283]
Sassuolo	0.271	0.126	[0.259 ; 0.284]
Venezia	0.283	0.123	[0.271 ; 0.296]
Empoli	0.313	0.122	[0.301 ; 0.326]
Spezia	0.317	0.121	[0.305 ; 0.329]
Salernitana	0.412	0.116	[0.401 ; 0.424]

Tabella 2.4: Stime dei coefficienti di difesa (ordinati dal migliore al peggiore)

Nota: nelle Tabelle 2.3 e 2.4 mancano alcuni valori associati all'Atalanta. Ciò è dovuto all'implementazione del vincolo (1.13): anziché inserirlo esplicitamente nell'algoritmo di ottimizzazione (ciò avrebbe aumentato di molto i tempi di stima), esso è stato considerato implicitamente ponendo i coefficienti dell'Atalanta (sia di attacco che di difesa) uguali all'opposto della somma degli altri. In realtà, i valori mancanti si potevano ricavare attraverso metodi di stima empirici (*es: bootstrap*), ma ciò andava oltre gli obiettivi di questo progetto.

Ricordiamo che nel modello di Dixon e Coles sono presenti anche un parametro associato all'*home effect* (γ) ed un parametro di dipendenza (ρ). Le stime ottenute per tali parametri (oltre che i rispettivi s.e. ed intervalli di confidenza) sono riassunte nella Tabella 2.5:

Parametro	**Stima**	**s.e.**	**IC(95 %)**
γ	0.305	0.044	[0.301 ; 0.309]
ρ	-0.071	0.077	[-0.079 ; -0.063]

Tabella 2.5: Stime dei coefficienti γ e ρ

I valori delle Tabelle, confrontati con il reale rendimento delle squadre nel campionato considerato, appaiono molto più plausibili rispetto a quelli ottenuti nel modello di Maher (Tabelle 2.1 e 2.2). Dal punto di vista offensivo, l'Inter è risultata la squadra che nell'intera stagione ha segnato più goal in assoluto (84), seguito da Lazio (77), Napoli (74), Milan (69) e Verona (65). Al contrario, le squadre meno prolifiche in fase realizzativa sono state Venezia (34), Salernitana (33) e Genoa (27). Tutto questo si rispecchia nell'ordine dei coefficienti d'attacco presenti nella Tabella 2.3. Dal punto di vista difensivo, invece, il Milan e il Napoli sono le squadre che nell'intera stagione hanno subito meno goal in assoluto (31), seguite da Inter (32), Juventus (37) e Torino (41). Al contrario, le squadre più deboli in fase difensiva sono risultate Empoli (70), Spezia (71) e Salernitana (78). Anche in questo caso emerge una forte corrispondenza con l'ordine dei coefficienti difensivi presenti nella Tabella 2.4. Inoltre, la stima del parametro associato all'*home effect* presenta un valore strettamente positivo (0.305). Ciò appare in perfetta linea con il significato teorico di tale coefficiente, in quanto esso esprime un vantaggio/beneficio che le squadre hanno quando giocano nel proprio stadio di casa. Si tratta dunque di un effetto "positivo" del parametro γ. Infine, la stima del parametro ρ presenta un valore negativo (-0.071), in analogia al risultato ottenuto da Dixon e Coles (l'effetto "negativo" di tale parametro era già stato introdotto nella Sezione 1.4.2).

2.2.3 Modello Dixon-Coles dinamico

Concentriamoci ora sulla versione dinamica del modello di Dixon e Coles, ossia quella che tiene conto di un fattore temporale. Tale modello è identificato dalla verosimiglianza (1.14) :

$$L(\alpha_i, \beta_i, \gamma, \rho, i = 1, \dots n) = \prod_{k \in A_t} \left[\tau_{\lambda_k, \mu_k}(x_k, y_k) \frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \cdot \frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \right]^{\phi(t-t_k)} \quad .$$

Assumendo $\phi(t - t_k) = e^{-\xi(t-t_k)}$, si può facilmente ricavare la funzione di log-verosimiglianza:

$$l(\alpha_i, \beta_i, \gamma, \rho) = \sum_{k \in A_t} e^{-\xi(t-t_k)} \left[\log[\tau_{\lambda_k, \mu_k}(x_k, y_k)] + \log\left(\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \right) + \log\left(\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \right) \right] \quad (2.7)$$

Come discusso nella Sezione 1.5.3, massimizzare la funzione (1.14) (o, equivalentemente, la (2.7)) alla data t consente di trovare le stime di massima verosimiglianza dei parametri, con l'accortezza che le stime sono riferite a quella precisa data. Rispetto ai modelli precedenti, tuttavia, non ci si può più limitare ad inserire la funzione nell'algoritmo di ottimizzazione (funzione *optim()* di R). Infatti, il rischio è che la (2.7) venga massimizzata banalmente aumentando ξ all'infinito e assegnando valori qualsiasi agli altri parametri. Osserviamo che:

$$\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \in [0,1] \implies \log\left[\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \right] \leq 0 \quad , \quad (2.8)$$

$$\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \in [0,1] \implies \log\left[\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \right] \leq 0 \quad . \quad (2.9)$$

Se ρ è molto piccolo, inoltre, si ha $\tau_{\lambda_k, \mu_k}(x_k, y_k) \approx 1$. Ciò risulta valido a livello empirico, ma è naturale pensare che sia così anche ragionando sul significato teorico del parametro ρ. Passando al logaritmo riesce pertanto $\log[\tau_{\lambda_k, \mu_k}(x_k, y_k)] \approx 0$. Definiamo ora la quantità:

$$P_k = \underbrace{\log\left[\tau_{\lambda_k, \mu_k}(x_k, y_k)\right]}_{\approx 0} + \underbrace{\log\left(\frac{(\lambda_k)^{x_k}}{x_k!} e^{-\lambda_k} \right)}_{\leq 0} + \underbrace{\log\left(\frac{(\mu_k)^{y_k}}{y_k!} e^{-\mu_k} \right)}_{\leq 0} \leq 0 \quad \forall k \quad . \quad (2.10)$$

Possiamo così riscrivere la (2.7) nella forma:

$$l(\alpha_i, \beta_i, \gamma, \rho) = \sum_{k \in A_t} e^{-\xi(t-t_k)} \cdot P_k \quad . \quad (2.11)$$

A questo punto, grazie alla (2.11), è immediato dimostrare che la log-verosimiglianza (2.7) è una funzione crescente di ξ. Infatti, studiando il segno della derivata parziale rispetto a ξ:

$$\begin{aligned}\frac{\partial l}{\partial \xi} &= \frac{\partial}{\partial \xi}\left(\sum_{k\in A_t} e^{-\xi(t-t_k)}\cdot P_k\right) \\ &= \sum_{k\in A_t}\frac{\partial}{\partial \xi}\left(e^{-\xi(t-t_k)}\cdot P_k\right) \\ &= \sum_{k\in A_t}\underbrace{e^{-\xi(t-t_k)}}_{\geq 0}\cdot\underbrace{P_k}_{\leq 0}\cdot\underbrace{(t_k-t)}_{\leq 0} \geq 0 \quad \forall \xi \quad .\end{aligned} \tag{2.12}$$

Ricapitolando, la versione "dinamica" del modello di Dixon e Coles presenta alcune problematiche relative alla stima dei parametri ottimali. A causa della monotonia rispetto a ξ, passare la log-verosimiglianza (2.7) come parametro alla funzione *optim()* di R non garantirebbe un risultato affidabile. Dixon e Coles, per risolvere questo problema, adottarono una prospettiva differente, la quale ricerca il valore di ξ in modo da massimizzare la capacità predittiva del modello: il valore ottimale di ξ sarà quello che permette al modello di fare le previsioni più accurate. Tale approccio verrà presentato nel dettaglio nella sezione seguente.

2.2.3.1 Verosimiglianza profilo per la stima di ξ

Come anticipato, il valore ottimale di ξ è tale da massimizzare la capacità predittiva complessiva del modello. Nella pratica si adotterà un approccio molto simile alla *cross-validation*, per cui si dividono i dati in un *training-set* per stimare il modello ed un *test-set* per valutare l'accuratezza delle previsioni. Supponiamo ad esempio di voler prevedere le partite della giornata i. Assegnato un valore specifico a ξ, è possibile usare tutti i dati della stagione fino alla giornata $i-1$ (training-set) per stimare a massima verosimiglianza i parametri del modello. Si potranno poi confrontare le previsioni del modello per la giornata i con gli effettivi risultati di quella giornata (test-set). A questo punto, si può pensare di passare alla giornata successiva e valutare le previsioni del modello (sempre con lo stesso ξ) per la giornata $i+1$. Chiaramente i parametri stimati in precedenza non saranno più validi, poiché nel modello dinamico la validità dei parametri si ha solamente a livello "locale". Bisognerà dunque usare tutti i dati della stagione fino alla giornata i (inclusa) così da ri-stimare i parametri del modello, per poi valutarne la sua capacità predittiva in riferimento ai risultati della giornata $i+1$. Ripetendo questo processo fino al termine della stagione considerata, si ottiene la cosiddetta

verosimiglianza profilo per ξ (indicata con $S(\xi)$), ossia una misura della validità complessiva del modello associato a quello specifico valore del parametro ξ. L'idea sarà ora quella di ripetere tutto questo procedimento facendo variare ξ, così da associare ai diversi ξ la loro rispettiva $S(\xi)$. Il valore del parametro ξ che massimizza $S(\xi)$ sarà quello che garantisce la migliore capacità predittiva al modello, per cui sceglieremo:

$$\xi^* = \underset{\xi \in \mathbb{R}^+}{\operatorname{argmax}}\ S(\xi) \quad .$$

Ora che abbiamo definito l'idea generale per la stima del modello ottimale, approfondiamo la formulazione matematica di $S(\xi)$. Indicato con N il numero totale di partite utilizzate per la "validazione" del modello, Dixon e Coles definirono $S(\xi)$ come segue:

$$S(\xi) = \sum_{k=1}^{N} \left(\delta_k^H \log p_k^H + \delta_k^D \log p_k^D + \delta_k^A \log p_k^A\right) \quad , \tag{2.13}$$

in cui le probabilità presenti nella (2.13) sono definite come:

$$\begin{aligned} p_k^H &= \sum_{l>m} Pr(X_k = l; Y_k = m) \\ p_k^D &= \sum_{l=m} Pr(X_k = l; Y_k = m) \\ p_k^A &= \sum_{l<m} Pr(X_k = l; Y_k = m) \quad , \end{aligned} \tag{2.14}$$

mentre δ_k^H, δ_k^D e δ_k^A sono variabili indicatrici sul vero risultato della k-esima partita, ossia:

$$\begin{aligned} \delta_k^H &= \begin{cases} 1 & \text{se ha vinto la squadra di casa} \\ 0 & \text{altrimenti} \end{cases} \\ \delta_k^D &= \begin{cases} 1 & \text{se la partita è terminata in pareggio} \\ 0 & \text{altrimenti} \end{cases} \\ \delta_k^A &= \begin{cases} 1 & \text{se ha vinto la squadra in trasferta} \\ 0 & \text{altrimenti} \end{cases} \end{aligned} \tag{2.15}$$

$S(\xi)$ è una funzione crescente rispetto all'affidabilità delle stime. Essa ha massimo in 0 e ciò si verifica se $p_k^H = \delta_k^H, p_k^D = \delta_k^D$ e $p_k^A = \delta_k^A$, ovvero quando il modello stima perfettamente l'esito di ogni partita di cui viene fatta la previsione.

Come spiegato in precedenza, per implementare la verosimiglianza profilo per la stima di ξ si è seguito un approccio analogo alla cross-validation, in cui una parte dei dati viene utilizzata per stimare i parametri del modello (training-set) e un'altra parte serve per testarne l'accuratezza (test-set). Affinché la stima dei parametri del modello sia abbastanza soddisfacente, è opportuno che il training-set contenga un numero sufficiente di dati. Spesso, nell'analisi statistica di un campionato di calcio, si possono ottenere dei risultati apprezzabili utilizzando i dati relativi all'intero girone di andata per stimare un modello con cui fare previsioni su tutto il girone di ritorno. In questo progetto di tesi abbiamo seguito proprio questa prospettiva, partendo da metà campionato e convalidando i modelli sulle partite della seconda parte di stagione. Inoltre, è chiaro che ξ può assumere infiniti valori in $\mathbb{R}^+$, per cui il processo da implementare potrebbe risultare troppo oneroso in termini computazionali. Al fine di limitare l'intervallo di valori entro cui ricercare il valore ottimale di ξ, si è scelto di basarci sull'ordine di grandezza del risultato già ottenuto in passato da Dixon e Coles (l'unica differenza è stata esprimere le differenze temporali in giorni anziché in mezze settimane). Sulla base di queste valutazioni è stato possibile ottenere per ciascun valore di ξ la rispettiva verosimiglianza profilo $S(\xi)$, ed il risultato ottenuto è riassunto in Figura 2.1.

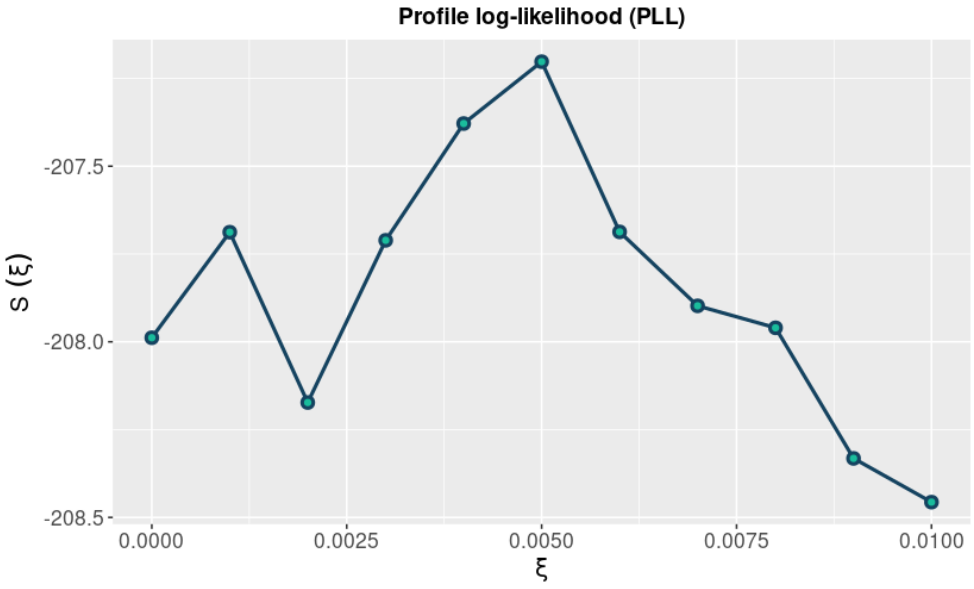

Figura 2.1: Verosimiglianza profilo per la stima di ξ

Come si può osservare dalla Figura 2.1 il valore che massimizza $S(\xi)$ è $\xi = 0.005$, per cui il modello specificato da tale valore sarà quello con la miglior capacità predittiva. Questo risultato consente anche di stabilire che il modello dinamico, sui dati considerati, ha un'accuratezza maggiore rispetto alla sua versione statica ($\xi = 0$). Ciò non era affatto scontato: analizzando singole stagioni sportive è molto frequente ottenere un valore ottimale di ξ pari a 0, cosicché il modello migliore risulterebbe proprio quello statico. In linea generale, infatti, ci si potrebbe aspettare che la ponderazione temporale diventi più significativa man mano che l'arco temporale dei dati si amplia, mentre quando si dispone di dati per una sola stagione non vale la pena abbandonare la versione statica del modello. Nonostante ciò, siccome l'analisi effettuata sui dati ha "premiato" il modello specificato da $\xi = 0.005$, ci baseremo su di esso per costruire alcune applicazioni al campionato di Serie A 2021-2022 (Capitolo 3).

Il modello ottimale, essendo dinamico, associa alle squadre dei coefficienti di attacco e di difesa che evolvono durante il corso della stagione. Grazie al processo svolto per la determinazione di ξ siamo riusciti a ricavare, per ciascuna squadra, le stime dei coefficienti riferite ad ogni giornata del girone di ritorno. Nello specifico, i valori dei coefficienti stimati al termine della stagione (usando come training-set tutte le partite) sono riportati nelle Tabelle 2.6 e 2.7:

Squadra	$\boldsymbol{\alpha}$	**s.e.**	**IC (95%)**
Inter	0.496	0.152	[0.481 ; 0.512]
Lazio	0.469	0.158	[0.453 ; 0.485]
Napoli	0.423	0.158	[0.407 ; 0.439]
Sassuolo	0.283	0.174	[0.266 ; 0.301]
Udinese	0.283	0.167	[0.267 ; 0.300]
Milan	0.259	0.171	[0.241 ; 0.276]
Verona	0.233	0.177	[0.215 ; 0.251]
Atalanta	0.197	/	/
Roma	0.103	0.186	[0.085 ; 0.122]
Fiorentina	0.098	0.184	[0.079 ; 0.116]
Juventus	0.092	0.184	[0.073 ; 0.110]
Empoli	-0.044	0.203	[-0.065 ; -0.024]
Torino	-0.124	0.206	[-0.144 ; -0.103]
Sampdoria	-0.152	0.210	[-0.173 ; -0.131]
Spezia	-0.188	0.216	[-0.210 ; -0.167]
Bologna	-0.192	0.213	[-0.213 ; -0.170]
Salernitana	-0.390	0.232	[-0.413 ; -0.366]
Venezia	-0.422	0.236	[-0.446 ; -0.398]
Cagliari	-0.509	0.251	[-0.534 ; -0.484]
Genoa	-0.915	0.303	[-0.946 ; -0.885]

Tabella 2.6: Stime dei coefficienti d'attacco (ordinati dal migliore al peggiore)

Squadra	α	**s.e.**	**IC (95%)**
Milan	-0.781	0.279	[-0.809 ; -0.753]
Inter	-0.479	0.248	[-0.504 ; -0.455]
Napoli	-0.469	0.244	[-0.494 ; -0.445]
Juventus	-0.300	0.227	[-0.323 ; -0.277]
Roma	-0.245	0.217	[-0.267 ; -0.223]
Torino	-0.223	0.211	[-0.245 ; -0.202]
Fiorentina	-0.008	0.193	[-0.028 ; 0.011]
Atalanta	0.006	/	/
Bologna	0.013	0.191	[-0.006 ; 0.032]
Genoa	0.052	0.187	[0.033 ; 0.070]
Udinese	0.121	0.183	[0.102 ; 0.139]
Verona	0.125	0.183	[0.302 ; 0.143]
Lazio	0.157	0.184	[0.138 ; 0.175]
Sampdoria	0.171	0.178	[0.153 ; 0.189]
Cagliari	0.238	0.172	[0.221 ; 0.256]
Empoli	0.300	0.167	[0.283 ; 0.317]
Spezia	0.317	0.165	[0.301 ; 0.334]
Venezia	0.318	0.165	[0.302 ; 0.335]
Sassuolo	0.326	0.167	[0.309 ; 0.343]
Salernitana	0.364	0.159	[0.348 ; 0.380]

Tabella 2.7: Stime dei coefficienti di difesa (ordinati dal migliore al peggiore)

Le Tabelle 2.6 e 2.7 contengono i coefficienti del modello dinamico stimati sull'intero campionato. Come già discusso nella Sezione 2.2.2 per le Tabelle 2.3 e 2.4, mancano gli standard errors e gli intervalli di confidenza associati ai parametri dell'Atalanta.

Essendoci posti al termine della stagione, in realtà, i coefficienti stimati non risultano così dissimili da quelli già ottenuti per la versione statica del modello di Dixon e Coles. Tuttavia, il fattore temporale (espresso dalla funzione (1.16)) consente di dare maggior peso alle prestazioni più recenti, e ciò può essere apprezzato nei valori assunti da alcuni coefficienti.

Ad esempio, nella parte finale della stagione l'Inter ha avuto una media goal impressionante, superiore ai 3 goal a partita. Basti pensare che in sole cinque partite (tra la 33esima e la 37esima giornata) la squadra ha marcato il tabellino per un totale di ben sedici goal. Il Milan, al contrario, ha vinto di misura gran parte delle ultime partite, per cui il coefficiente di attacco non appare più tra i migliori: se nella Tabella 2.3 risultava il quarto miglior attacco, con un valore pari ad $\alpha = 0.291$, ora nella Tabella 2.6 compare in sesta posizione, con un valore diminuito ad $\alpha = 0.259$. Allo stesso tempo, la squadra non ha concesso praticamente nulla dal punto di vista difensivo: tra la 33esima e la 37esima giornata di campionato i rossoneri

hanno subito solamente due goal. Ciò ha un effetto significativo sul valore del coefficiente difensivo, che nel modello statico risultava $\beta = -0.584$, mentre ora è notevolmente migliorato fino a $\beta = -0.781$. Simmetrico è invece il caso del Sassuolo. La squadra emiliana durante il campionato non ha mai dimostrato una grande solidità difensiva, ma il modello dinamico, rispetto al modello statico, evidenzia un peggioramento: β aumenta da 0.271 a 0.326, e tale effetto potrebbe derivare dalla batosta per 6-1 subita contro il Napoli alla 35esima giornata.

I coefficienti ρ e γ, infine, non hanno fatto emergere particolari differenze rispetto la versione statica del modello. Le stime ottenute per tali parametri (oltre i rispettivi standard errors ed intervalli di confidenza) sono riassunte nella Tabella 2.8 :

Parametro	**Stima**	**s.e.**	**IC(95 %)**
γ	0.274	0.061	[0.268 ; 0.281]
ρ	-0.076	0.103	[-0.087 ; -0.066]

Tabella 2.8: Stime dei coefficienti γ e ρ

Il parametro di dipendenza (ρ) continua ad esprimere un effetto "negativo" (già analizzato nelle sezioni precedenti), anche se leggermente maggiore (in valore assoluto) rispetto a quello del modello statico (in cui si aveva $\rho = -0.071$). Parallelamente, il coefficiente relativo all'*home effect* esprime ancora un effetto "positivo", ed il suo valore non risulta particolarmente lontano da quello stimato per il modello statico (in cui si aveva $\gamma = 0.305$). Riguardo quest'ultimo aspetto, nella Sezione 3.3 avremo modo di osservare come l'*home effect*, oltre ad essere stato considerato uguale per tutte le squadre, abbia assunto valori pressoché costanti durante tutto il corso del campionato considerato.

Capitolo 3

Applicazioni del modello al campionato di Serie A 2021-2022

Nella Sezione 2.2.3.1 è stato proposto un approccio per determinare il modello con la maggior capacità predittiva complessiva. Tale analisi ci aveva portato a "premiare" il modello dinamico di Dixon e Coles specificato dal parametro $\xi = 0.005$. In questo capitolo partiremo proprio da quel modello ed approfondiremo alcune sue applicazioni concrete, le quali ci permetteranno di analizzare e comprendere meglio il campionato di Serie A 2021-2022.

3.1 Previsioni di probabilità

Consideriamo la solita partita di calcio tra le squadre i (squadra di casa) e j (squadra in trasferta), in cui i goal segnati dalle rispettive squadre sono descritti dalle variabili aleatorie $X_{ij} \sim Pois(\lambda_{ij} = \alpha_i \beta_j \gamma)$ e $Y_{ij} \sim Pois(\mu_{ij} = \alpha_j \beta_i)$. Sotto l'ipotesi di indipendenza, la probabilità congiunta può essere calcolata come:

$$Pr(X_{ij} = x; Y_{ij} = y) = Pr(X_{ij} = x) \cdot Pr(Y_{ij} = y) = \frac{(\alpha_i \cdot \beta_j \cdot \gamma)^x}{x!} e^{-\alpha_i \cdot \beta_j \cdot \gamma} \cdot \frac{(\alpha_j \cdot \beta_i)^y}{y!} e^{-\alpha_j \cdot \beta_i} \quad .$$

Corregendola tenendo conto della funzione (1.9), si ottiene:

$$Pr(X_{ij} = x; Y_{ij} = y) = \tau_{\lambda_{ij}, \mu_{ij}}(x, y) \cdot \frac{(\alpha_i \cdot \beta_j \cdot \gamma)^x}{x!} e^{-\alpha_i \cdot \beta_j \cdot \gamma} \cdot \frac{(\alpha_j \cdot \beta_i)^y}{y!} e^{-\alpha_j \cdot \beta_i} \quad . \tag{3.1}$$

Per una data partita, la (3.1) ci permette di stimare la probabilità che si verifichi il risultato $(x; y)$. Facendo variare x ed y, allora, potremo associare una misura di probabilità a ciascun possibile esito. In questo modo, prima che una partita venga disputata, sarà possibile farsi un'idea a priori sui risultati più probabili secondo il modello.

A questo proposito, può risultare molto utile organizzare i possibili esiti in forma matriciale. Secondo questa idea ciascuna partita è rappresentabile come una matrice A_{ij}, in cui la generica entrata a_{ij} contiene la probabilità congiunta che la squadra di casa segni i goal mentre la squadra in trasferta ne segni j. Tuttavia, il calcio è uno sport in cui il risultato finale è fortemente variabile, per cui non esiste una dimensione fissata per queste matrici. Allo stesso tempo, però, molti risultati sono statisticamente improbabili, per cui molte entrate della matrice avrebbero valori poco significativi, prossimi allo 0. Per questo motivo considereremo matrici quadrate di dimensione g, in cui g è assunto (convenzionalmente) come il numero massimo di goal che ragionevolmente può segnare ciascuna squadra. Dal momento in cui è piuttosto raro che una squadra segni un numero di goal maggiore o uguale a cinque, in questo progetto si è posto $g = 4$ per semplicità di trattazione. Inoltre, con un leggero abuso rispetto la classica notazione dell'algebra lineare, è ammesso che i e j possano assumere anche il valore 0 (ciò è essenziale per riuscire a descrivere gran parte dei risultati).

Alla luce delle precedenti valutazioni, diventa dunque possibile rappresentare la partita tra le squadre I (casa) e J (trasferta) con una matrice 5x5 di questo tipo:

$$A_{ij} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \tag{3.2}$$

In particolare, notiamo che la diagonale principale della matrice contiene le probabilità dei risultati di pareggio tra le due squadre, mentre la sottomatrice "triangolare" (di cui però si esclude la diagonale) inferiore/superiore contiene le probabilità associate alle vittorie della squadra di casa/trasferta. La matrice risulta così divisa in 3 zone, per cui sommando tutti i valori appartenenti alla stessa zona possiamo ottenere una stima della probabilità

dell'esito complessivo della partita (senza considerarne il risultato preciso). Dal punto di vista della squadra di casa, si otterrà la probabilità degli eventi vittoria-pareggio-sconfitta che, in ambito di scommesse sportive, vengono comunemente indicati con 1-X-2. Ovviamente vale il discorso simmetrico se ci poniamo dal punto di vista della squadra in trasferta.

Consideriamo ora il modello dinamico di Dixon e Coles con i coefficienti stimati (a massima verosimiglianza) prima che venga disputata l'ultima giornata di campionato. Le previsioni dei risultati per la giornata in questione sono riportati in Figura 3.1:

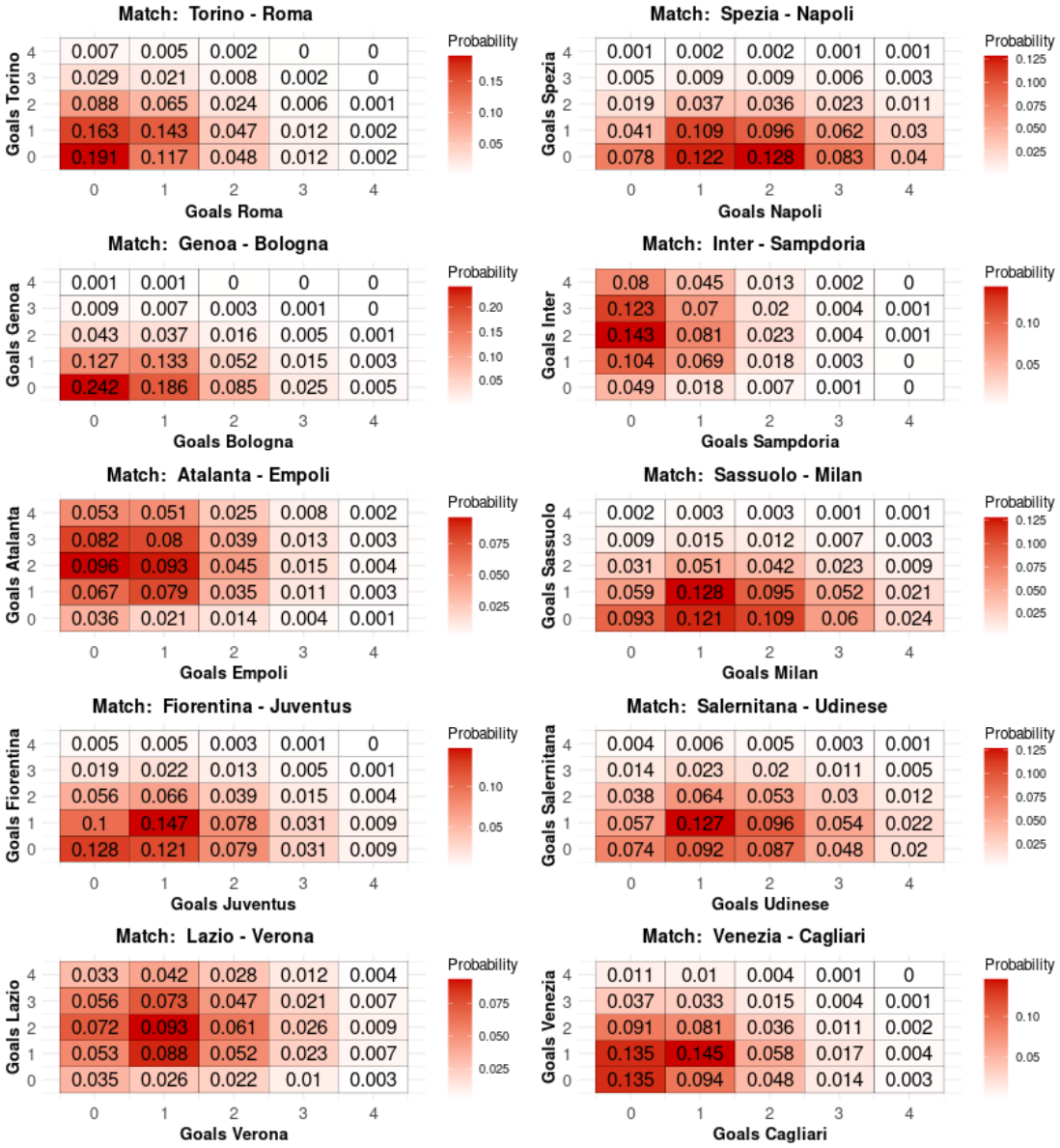

Figura 3.1: Previsioni dei risultati della 38esima giornata (Serie A 2021-2022)

I grafici presenti nella Figura 3.1 seguono la stessa identica logica della matrice (3.2); l'unica differenza sta nell'ordine in cui appaiono gli indici che identificano le righe. Oltre ad esserci i valori numerici delle probabilità congiunte, il grafico consente di visualizzare i risultati più probabili anche in maniera grafica: il colore di ciascuna casella appare con una tonalità più o meno scura a seconda del valore contenuto in essa. È poi possibile aggregare le probabilità dei risultati che determinano lo stesso esito finale, così da ottenere una probabilità di tipo 1-X-2 (Home-Draw-Away). Per la partita descritta dalla matrice A_{ij} :

$$\begin{aligned} Pr(H) &= \sum_{i>j} Pr(X_k = i; Y_k = j) = \sum_{i>j} a_{ij} \\ Pr(D) &= \sum_{i=j} Pr(X_k = i; Y_k = j) = \sum_{i=j} a_{ij} = tr(A_{ij}) \\ Pr(A) &= \sum_{i<j} Pr(X_k = i; Y_k = j) = \sum_{i<j} a_{ij} \quad . \end{aligned} \tag{3.3}$$

In riferimento alla 38esima giornata di campionato, le probabilità aggregate sono risultate:

Home Team	**Away Team**	**Pr(H)**	**Pr(D)**	**Pr(A)**
Torino	Roma	0.389960	0.359143	0.246317
Genoa	Bologna	0.228837	0.392060	0.376741
Atalanta	Empoli	0.592758	0.175297	0.109903
Fiorentina	Juventus	0.290671	0.319969	0.378773
Lazio	Verona	0.509288	0.208244	0.184954
Spezia	Napoli	0.124173	0.228831	0.598496
Inter	Sampdoria	0.680945	0.145528	0.053177
Sassuolo	Milan	0.185364	0.269366	0.517160
Salernitana	Udinese	0.234595	0.266141	0.465997
Venezia	Cagliari	0.417889	0.319849	0.251660

Tabella 3.1: Previsioni degli esiti finali della 38esima giornata (Serie A 2021-2022)

Confrontando i valori contenuti nella Tabella 3.1 con gli esiti reali, il modello avrebbe previsto correttamente solo quattro partite. Teniamo presente però che il calcio è uno sport imprevedibile e che un modello statistico -purtroppo o per fortuna- non potrà mai restituire una verità assoluta. Inoltre, l'ultima giornata di campionato costituisce un caso un po' a sé stante, in cui spesso le partite non sono più determinanti per le squadre. Tuttavia, non dimentichiamo che in ambito di scommesse sportive esistono anche quote di tipo 1X ed X2, per cui applicare il modello in quest'ottica avrebbe permesso (ex-post) di indovinare ben otto risultati su dieci. In ogni caso, trarre conclusioni sulla base di solo dieci partite sarebbe troppo riduttivo e privo di fondamenta statistiche (lo approfondiremo nella Sezione 3.4).

3.2 Evoluzione delle abilità delle squadre durante la stagione

Come spiegato nella Sezione 1.5.3, la versione (1.14) del modello di Dixon e Coles è caratterizzata da parametri dinamici nel tempo. Essi sono necessari per poter tenere conto dello stato di forma delle squadre, un aspetto fondamentale dal momento che ogni squadra, durante il corso della stagione, vive momenti più o meno prolifici in fase offensiva e/o difensiva. Stimando i parametri del modello in riferimento a date diverse, possiamo ottenere una serie storica dei coefficienti di attacco e di difesa di ciascuna squadra. In particolare, il loro andamento dinamico consente di individuare eventuali trend di crescita, decrescita o di stabilità.

Di seguito la serie storica dei coefficienti durante tutto il girone di ritorno (Serie A 2021-2022):

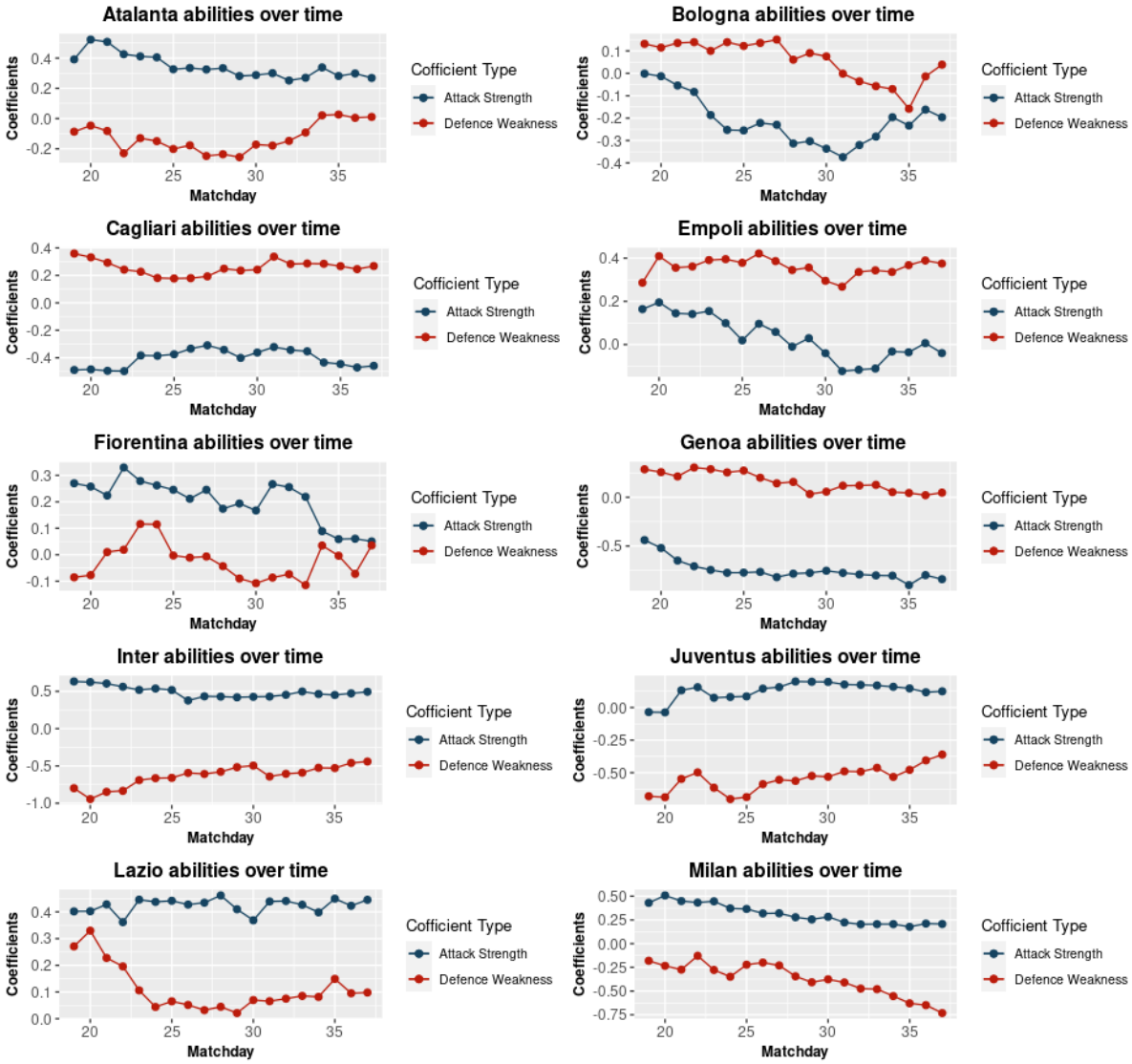

Figura 3.2: Serie storica dei coefficienti (prima parte)

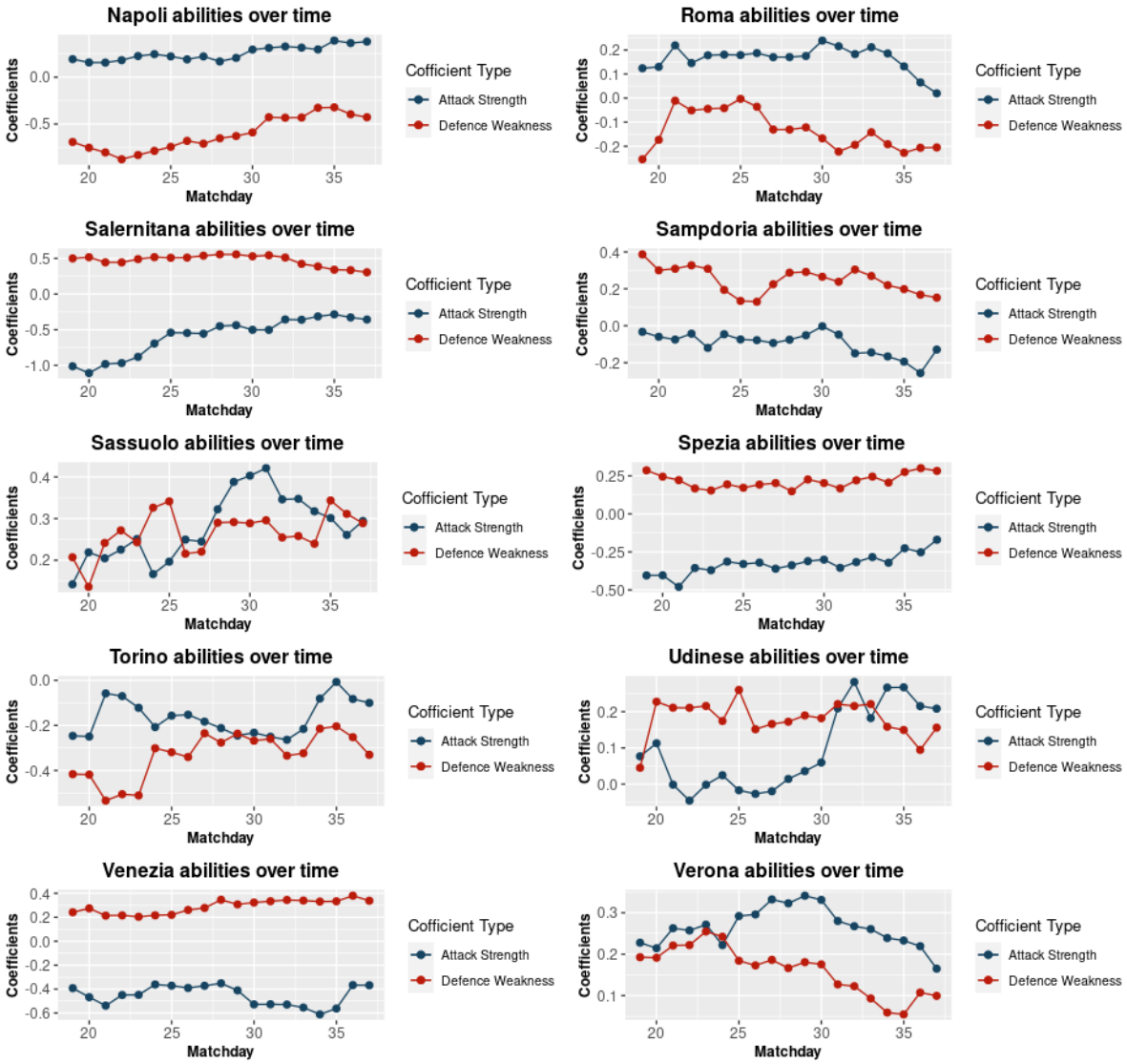

Figura 3.3: Serie storica dei coefficienti (seconda parte)

Queste serie storiche ci permettono di apprezzare graficamente alcuni esempi che avevamo già considerato nel Capitolo 2 per l'interpretazione dei coefficienti delle Tabelle 2.6 e 2.7. Tra questi, ricordiamo come nell'edizione 2021-2022 della Serie A l'Inter abbia dimostrato una schiacciante superiorità dal punto di vista offensivo, totalizzando il numero di goal più alto in assoluto rispetto tutte le altre squadre. I nerazzurri hanno mantenuto una media goal impressionante durante tutto il corso del campionato, e ciò viene messo in evidenza dalla stabilità della serie storica del coefficiente d'attacco. Anche il Milan ha avuto una buona media realizzativa, ma si è distinto dall'Inter per una maggior solidità difensiva: se nel campionato in questione i rossoneri si sono laureati campioni d'Italia, lo devono principalmente ad eccellenti prestazioni difensive, soprattutto nel finale di stagione (in cui è visibile un

chiaro trend negativo della debolezza difensiva). Tra gli altri esempi avevamo menzionato il coefficiente difensivo del Sassuolo, che consentiva di apprezzare concretamente l'impatto della ponderazione temporale verso le stime dei parametri. In quel contesto, avevamo osservato come la batosta subita contro il Napoli alla 35esima giornata (per 6-1) avesse avuto un effetto significativo verso il coefficiente β. Ciò viene suggerito anche dal grafico della serie storica, in cui è lampante il peggioramento improvviso del coefficiente difensivo. Riguardo le altre squadre, citiamo le difficoltà in fase offensiva della Roma nelle ultime partite, il cui parametro d'attacco segue un vistoso trend negativo: tra la 33esima e la 37esima giornata di campionato, i giallorossi hanno messo a segno solamente tre goal. Ricordiamo infine il percorso della Salernitana, che fino alla 32esima giornata si trovava in ultima posizione, destinata alla retrocessione. La serie storica dei coefficienti ci permette di comprendere più a fondo i meccanismi dietro la salvezza della squadra campana sotto la guida di Davide Nicola, che era stato nominato come tecnico prima della 26esima giornata di campionato. Da quel momento in poi il grafico non mostra particolari progressi dal punto difensivo, ma al contrario evidenzia un netto miglioramento in fase offensiva. I più appassionati di calcio ricorderanno che la squadra iniziò a giocare con più coraggio, proponendo un calcio intenso e più propositivo, il quale portò al compimento di un vero e proprio miracolo sportivo.

3.3 Evoluzione dell'*home effect* durante la stagione

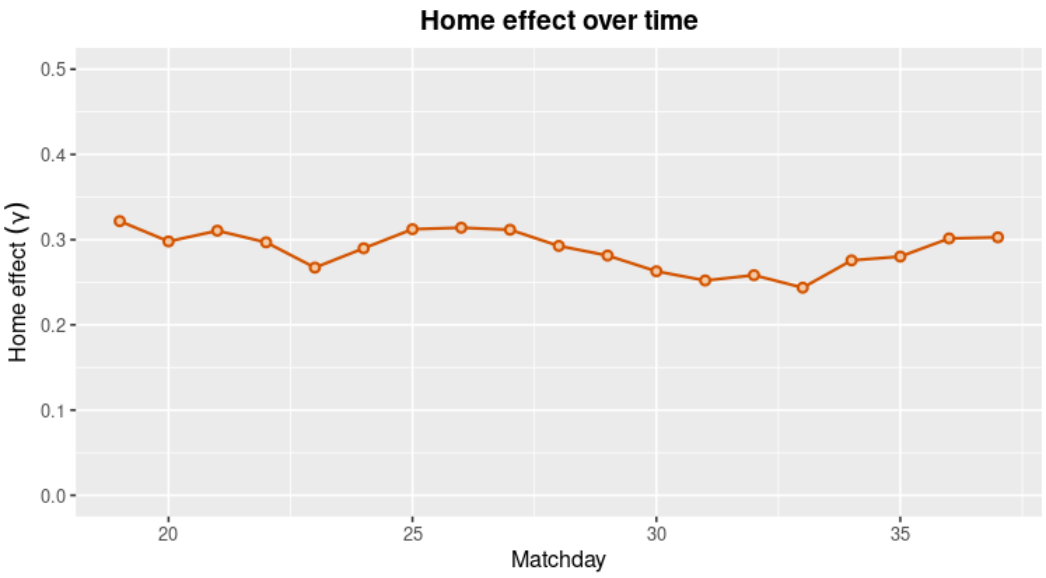

Figura 3.4: Serie storica dell'*home effect*

Da un punto di vista teorico, anche il parametro che descrive l'*home effect* (γ) è dinamico e può cambiare nel corso del tempo. Nonostante ciò, la Figura 3.4 presenta un andamento tendenzialmente costante, il quale fa presupporre una caratteristica di stazionarietà del coefficiente (sia in media che varianza). Il fatto che le stime di γ rimangano sostanzialmente stabili nel tempo è piuttosto ragionevole, ed è assolutamente sensato assumere che il beneficio di giocare in casa non sia in alcun modo correlato ai diversi periodi dell'anno.

In maniera analoga, anche il parametro ρ evolve nel corso del tempo. Tuttavia, la serie storica di questo coefficiente non appare così interessante, oltre ad essere particolarmente complicata dal punto di vista interpretativo. Per completezza, la riportiamo in Figura 3.5:

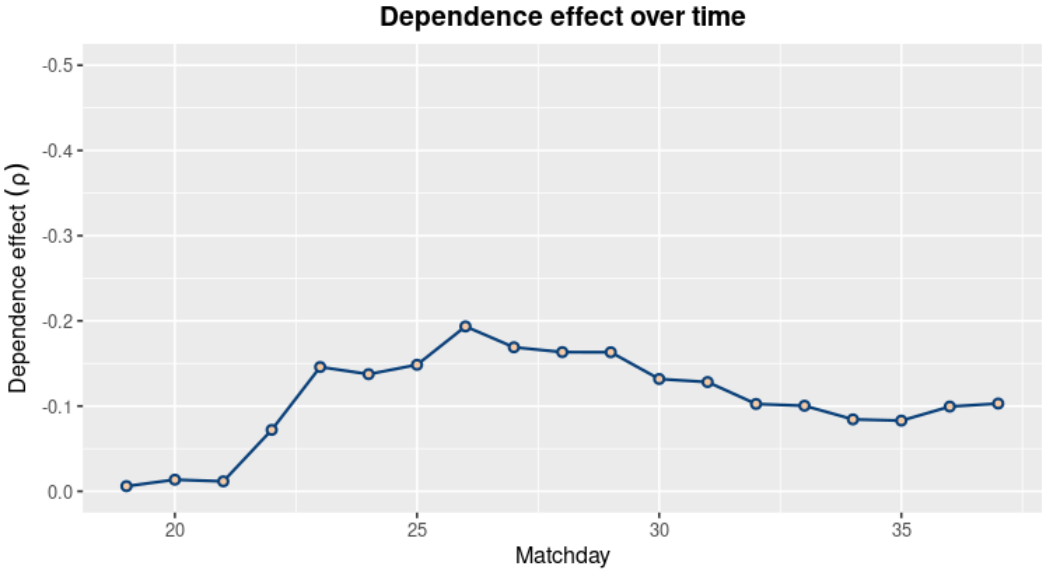

Figura 3.5: Serie storica del parametro di dipendenza ρ

Nota: nelle sezioni precedenti è stato più volte messo in luce l'effetto "negativo" del coefficiente di dipendenza ρ. Con l'intenzione di rappresentarne l'effetto in maniera più chiara, si è scelto di ordinare i valori dell'asse y in termini di valore assoluto (tralasciando il segno -).

3.4 Valutazione e confronto tra modelli

In questa sezione valuteremo le prestazioni dei diversi modelli statistici che sono stati presentati nel Capitolo 1 ed implementati utilizzando il software R (Capitolo 2). Nello specifico, confronteremo il modello di Maher ed il modello di Dixon e Coles, nella sua versione sia statica che dinamica. In realtà, nella Sezione 2.2.3.1 avevamo già effettuato una prima indagine sul modello (dinamico) con la miglior capacità predittiva, grazie a cui avevamo poi stabilito il valore ottimale del parametro ξ. Le valutazioni in termini di verosimiglianza profilo, inoltre, ci avevano permesso di stabilire come il modello dinamico, sulla base dei dati analizzati, avesse un'accuratezza maggiore rispetto alla versione statica (che d'altronde è un caso particolare del modello dinamico, in cui $\xi = 0$). Per fornire una valutazione più completa sulle capacità predittive dei modelli, presenteremo ora due metriche ben consolidate: il Brier Score e lo Pseudo-R^2. Queste metriche sono ampiamente riconosciute ed utilizzate nel campo della statistica sportiva, ma anche nella modellazione predittiva più generale. Sfruttando queste misure di valutazione potremo allora quantificare l'accuratezza e l'affidabilità delle previsioni di ciascun modello, così da trarre solide conclusioni sulle loro rispettive prestazioni.

3.4.1 Brier Score

Il Brier Score [8] è una misura di valutazione della precisione e dell'accuratezza delle previsioni probabilistiche di un modello statistico. Di seguito proponiamo la sua definizione.

Definizione 3 (**Brier Score**)**.** *Consideriamo un insieme di n eventi aleatori, ognuno dei quali può avere un numero r di determinazioni possibili. Per il generico evento i (con $i \in \{1,...,n\}$) indichiamo con p_{ij} la probabilità associata all'esito j (con $j \in \{1,...,r\}$). Inoltre, supponiamo che l'insieme delle r classi sia mutualmente esclusivo ed esaustivo (ossia una partizione dell'evento certo Ω), per cui si richiede che valga: $\sum_{j=1}^{r} p_{ij} = 1 \quad \forall i \in \{1,...,n\}$.*
Sotto queste ipotesi, il Brier Score (BS) viene definito come la seguente quantità:

$$BS = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} (p_{ij} - d_{ij})^2 \quad , \tag{3.4}$$

con d_{ij} variable indicatrice (dummy) sulla j-esima determinazione nell'evento i-esimo

Per l'interpretazione del Brier Score di un modello vale il seguente principio: più basso è il suo valore, migliore è l'accuratezza predittiva del modello. Il caso estremo di accuratezza si ottiene quando il modello, per ciascuno degli n eventi, assegna probabilità 1 alla determinazione che effettivamente si verifica (e, di conseguenza, probabilità nulla a tutte le altre). In questo caso è abbastanza immediato verificare che la (3.4) restituirebbe un Brier Score pari a 0. Per quanto riguarda una limitazione superiore del Brier Score, invece, è possibile ricavarla ragionando sul caso peggiore in termini di accuratezza delle previsioni. In particolare, esso si ottiene quando il modello assegna, per ciascuno degli n eventi, probabilità 0 alla determinazione che si verifica e probabilità 1 ad una qualsiasi delle altre (alle rimanenti verrà assegnata chiaramente probabilità 0). In questo caso la (3.4) restituirebbe un Brier Score pari a 2.

La Definizione 3 può essere adattata facilmente al nostro fenomeno d'interesse, ossia le partite di calcio. Consideriamo quindi un insieme di M partite, in cui indicheremo con m la generica partita. Facendo riferimento alla notazione già introdotta nella (3.3), l'evento "partita" presenta $r = 3$ possibili determinazioni (mutualmente esclusive ed esaustive): Home, Draw e Away (o, equivalentemente 1,X,2). La probabilità che si realizzi l' esito i-esimo ($i \in \{H, D, A\}$) per la partita m-esima la indicheremo come p_{mi}, anche se si tratta semplicemente delle probabilità che avevamo già incontrato nella (3.3). Infine, il ruolo delle variabili indicatrici (d_{mi}) lo faranno le variabili δ_m^H, δ_m^D e δ_m^A, che avevamo precedentemente definito nella (2.15). Il Brier Score per i nostri modelli calcistici avrà dunque la seguente formulazione:

$$BS = \frac{1}{M} \sum_{m=1}^{M} \sum_{i} (p_{mi} - d_{mi})^2 \qquad \text{con } i \in \{H, D, A\} \quad . \tag{3.5}$$

Il calcolo del Brier Score è stato implementato in R ed i risultati sono riportati in Tabella 3.2:

Model	Maher model	Static Dixon-Coles	Dinamic Dixon-Coles ($\xi = 0.005$)
Brier Score	0.67686	0.61288	0.61234

Tabella 3.2: Brier Score dei modelli

Dalla Tabella 3.2 emerge come il peggior modello dal punto di vista del Brier Score sia quello di Maher. Considerata la semplicità di tale modello ed i forti limiti nella sua formulazione generale, questo risultato era piuttosto prevedibile. Un confronto più interessante emerge invece tra il modello statico di Dixon e Coles e quello dinamico. L'analisi sulla verosimiglianza

profilo aveva già attributo un'accuratezza maggiore alla versione dinamica del modello, ed in effetti il Brier Score sembrerebbe confermarlo. Infatti, anche se in leggera misura, il Brier Score associato al modello dinamico risulta minore rispetto a quello del modello statico.

Il Brier Score, per com'è stato definito, riassume con un unico valore la precisione di un modello predittivo in riferimento ad un insieme di dati (nel nostro caso, le $M = 190$ partite), e ci permette di confrontarlo con altri modelli (applicati agli stessi dati). Oltre a fare questo, ci è sembrato interessante calcolare un Brier Score a livello di giornata di campionato ($M = 10$) ed ottenere una serie storica dei valori. Così facendo è stato possibile confrontare la capacità predittiva dei modelli nel corso del tempo. Il risultato è presentato in Figura 3.6 :

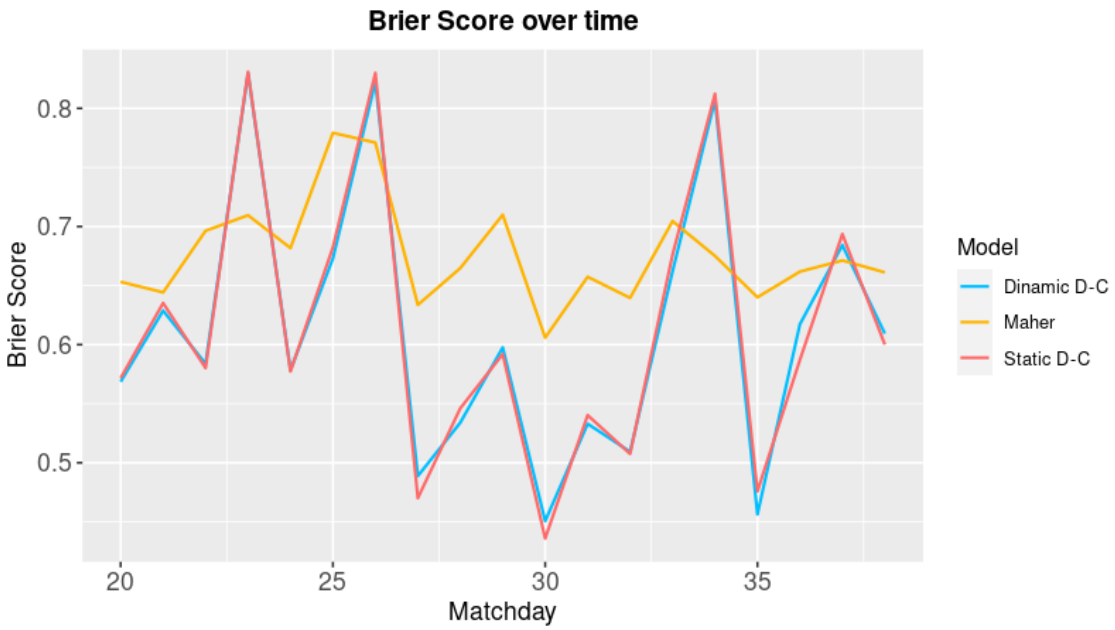

Figura 3.6: Brier Score: confronto temporale tra modelli

La Figura 3.6 rivela che il modello di Maher, oltre ad essere il peggiore a livello complessivo, mediamente lo è anche a livello di singola giornata. Eccetto quattro punti corrispondenti alle giornate 23,26,34 e 37, il suo Brier Score è sempre risultato peggiore (ossia più alto) rispetto agli altri due modelli. Riguardo invece le due versioni (statica e dinamica) del modello di Dixon e Coles, le serie storiche dei rispettivi Brier Scores seguono un'andamento pressoché identico, che rende difficile stabilire il più preciso. In effetti, anche i Brier Scores complessivi dei due modelli non presentavano una differenza così accentuata (si veda la Tabella 3.2).

3.4.2 Pseudo-R^2

Un altro strumento per valutare le prestazioni predittive su un numero qualsiasi di partite è rappresentato dallo pseudo-R^2. Esso può essere definito come la media geometrica delle probabilità assegnate al risultato effettivo (sempre inteso come H-D-A) di ciascuna partita. Definiremo ora la sua formulazione, adattandola già al contesto delle partite di calcio.

Definizione 4 (**Pseudo-R^2**)**.** *Consideriamo un modello statistico per la previsione di risultati calcistici ed un insieme di M partite, in cui indicheremo con m la generica. Indicata con p_m la probabilità associata al risultato effettivo della partita m, lo pseudo-R^2 è dato dalla quantità:*

$$pseudo\text{-}R^2 = \left(\prod_{m=1}^{M} p_m \right)^{\frac{1}{M}} \quad . \tag{3.6}$$

A partire dalla (3.6) si deduce banalmente che lo pseudo-R^2 assume valori nell'intervallo reale $[0,1]$. Altrettanto intuitiva risulta la sua interpretazione, per la quale vale il seguente principio: più alto è il valore dello pseudo-R^2, migliore è l'accuratezza predittiva del modello. Il calcolo dello pseudo-R^2 è stato implementato in R ed è riassunto in Tabella 3.3 :

Model	Maher model	Static Dixon-Coles	Dinamic Dixon-Coles ($\xi = 0.005$)
Pseudo-R^2	0.24385	0.33465	0.33604

Tabella 3.3: Pseudo-R^2 dei modelli

Dalla Tabella 3.3 emerge come il peggior modello dal punto di vista dello pseudo-R^2 sia quello di Maher, che era risultato il meno accurato anche dal punto di vista del Brier Score (Tabella 3.2). D'altronde, come ribadito nella Sezione 3.4.1, la semplicità di tale modello ed i suoi forti limiti rendevano questo risultato piuttosto prevedibile. Per quanto riguarda il modello di Dixon e Coles, invece, l'analisi sulla verosimiglianza profilo ed il Brier Score suggerivano di preferirne la versione dinamica. Questo risultato trova ora una conferma anche dal punto di vista dello pseudo-R^2, il cui il miglior valore (ossia il più alto) è associato proprio al modello dinamico di Dixon e Coles. Inoltre, come valeva per il Brier Score, la versione statica e dinamica del modello non evidenziano una differenza netta di pseudo-R^2.

Anche lo pseudo-R^2, per com'è stato definito, riassume con un unico valore la precisione di un modello predittivo (in riferimento ad un insieme di dati) e ci permette di confrontarla con quella di altri modelli. In analogia a quanto fatto per il Brier Score, ci è sembrato interessante calcolare uno pseudo-R^2 a livello di singola giornata di campionato (M = 10), così da ottenere una serie storica dei suoi valori. In questo modo stato possibile confrontare la capacità predittiva dei 3 modelli nel corso del tempo, come riportato in Figura 3.7:

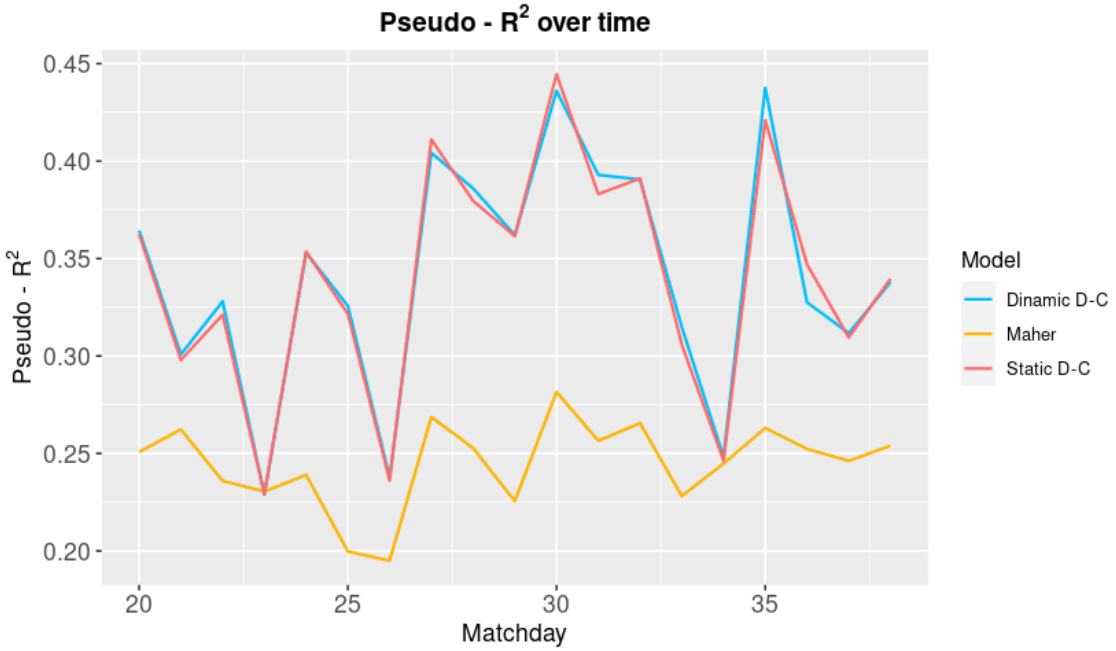

Figura 3.7: Pseudo-R^2: confronto temporale tra modelli

La Figura 3.7 permette di fare alcune osservazioni del tutto analoghe a quelle suggerite dalla Figura 3.6 in riferimento al Brier Score. In particolare, il grafico rivela che il modello di Maher, oltre ad essere il peggiore (pseudo-R^2 più basso) a livello complessivo, lo è anche a livello di singola giornata (tranne nella 23esima). Riguardo invece le versioni statica e dinamica del modello di Dixon e Coles, le serie storiche dei rispettivi pseudo-R^2 seguono un'andamento pressoché identico, che anche in questo caso rende difficile stabilire il più preciso tra i due.

3.4.3 Matrici di confusione

La valutazione quantitativa dei modelli tramite il Brier Score (Sezione 3.4.1) e lo Pseudo-R^2 (Sezione 3.4.2) ha confermato le lacune (fin da subito evidenti) del modello di Maher. Tuttavia, tra la versione statica e quella dinamica del modello di Dixon e Coles non è ancora emersa una sostanziale differenza in termini di prestazioni. Considereremo allora un altro strumento statistico per tentare di fare più chiarezza sulla questione: le matrici di confusione.

In generale, le matrici di confusione sono molto utili per valutare le prestazioni dei modelli di classificazione, ossia quei modelli statistici che vengono utilizzati per assegnare etichette o classi a diverse istanze di dati. In una matrice di confusione gli elementi sulla diagonale (principale o secondaria a seconda dei casi) rappresentano le previsioni corrette, mentre gli elementi fuori dalla diagonale rappresentano gli errori di classificazione. A partire da una matrice di confusione binaria si possono costruire alcune metriche significative, poi generalizzabili al caso multi-classe. Consideriamo allora una matrice di confusione binaria:

		Actual	
		P	N
Predicted	P	TP	FP
	N	FN	TN

La Tabella 3.4 riporta alcune tra le metriche più utilizzate [9]:

Metrica	**Formula**
Precision (Positive-Predicted-Value)	$\frac{TP}{TP+FP}$
Negative-Predicted-Value (NPV)	$\frac{TN}{TN+FN}$
Recall (sensitivity)	$\frac{TP}{TP+FN}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1-score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Tabella 3.4: Principali metriche di valutazione per un modello di classificazione

Come anticipato, tali metriche possono essere poi generalizzate al caso multi-classe. L'idea di fondo è quella di ragionare per singole classi e calcolare la metrica a livello di classe, per poi combinare tutto insieme ed ottenere un'unica misura rappresentativa. In particolare, per il calcolo della metrica relativa alla i-esima classe sarà prima necessario convertire la matrice di partenza in una matrice "one-vs-all" per la classe i, in modo da ricondursi alla struttura di una matrice di confusione binaria e poter applicare le formule della Tabella 3.4.

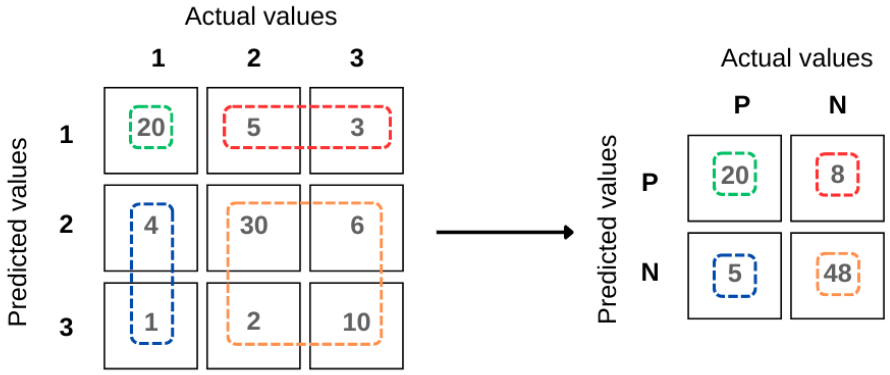

Figura 3.8: Esempio di matrice *one-vs-all* per la classe $i = 1$

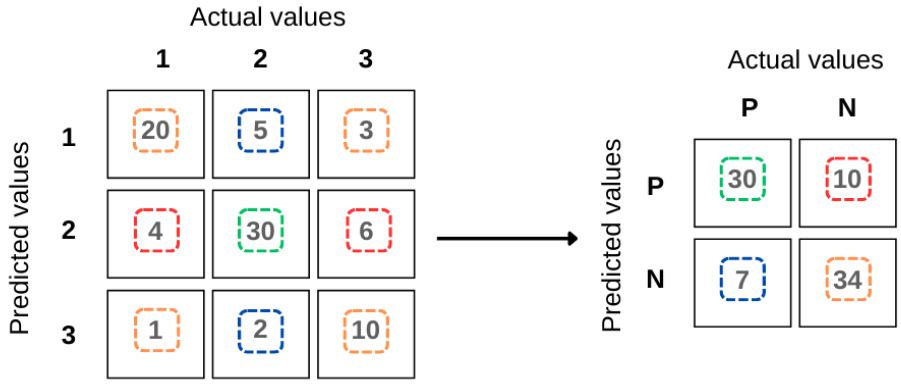

Figura 3.9: Esempio di matrice *one-vs-all* per la classe $i = 2$

Dopo aver calcolato le metriche in ciascuna classe, si possono adottare due diversi approcci:

1. Approccio *Macro-Average*, in cui la metrica "globale" viene calcolata come media non ponderata delle misure calcolate a livello di singole classi;

2. Approccio *Micro-Average*, in cui la metrica "globale" viene calcolata come media ponderata delle misure calcolate a livello di singole classi. In particolare, i pesi di ciascuna misura sono dati dalla frequenza della rispettiva classe sull'intero dataset.

La costruzione delle matrici di confusione ed il calcolo di tutte le metriche possono essere facilmente automatizzate in R grazie alla funzione *confusion_matrix()*. Essa fa parte del pacchetto *cvms*, una libreria per la cross-validation di modelli di regressione e classificazione.

Nel caso delle partite di calcio, coerentemente alla notazione già adottata nelle sezioni precedenti, è possibile costruire delle matrici di confusione multi-classe di dimensione 3x3, in cui le classi a cui faremo riferimento sono: Home-Draw-Away. In Figura 3.10 sono proposte le matrici di confusione relative alle versioni statica e dinamica del modello di Dixon e Coles.

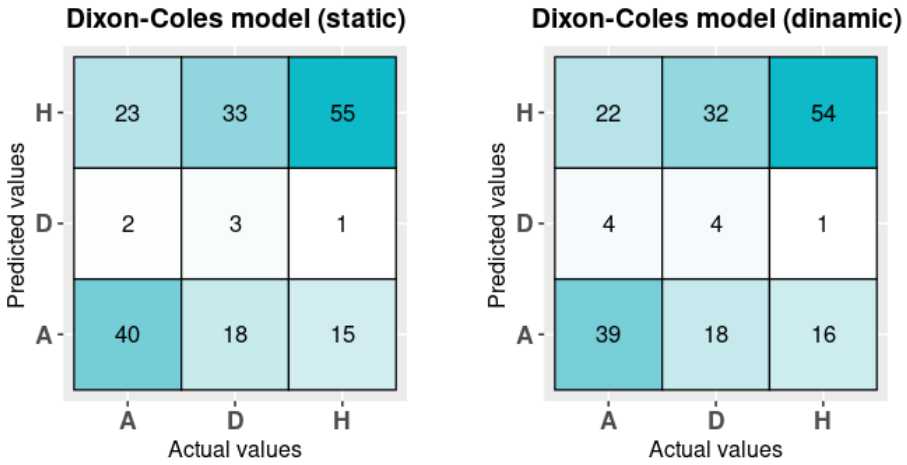

Figura 3.10: Matrici di confusione H-D-A

Con la funzione *confusion_matrix()* siamo riusciti a ricavare facilmente le metriche in ciascuna classe. Esse sono riportate nelle Tabelle 3.5 e 3.6:

Class	Precision (PPV)	NPV	Sensitivity	Accuracy	F1-score
A	0.548	0.786	0.615	0.676	0.580
D	0.500	0.723	0.056	0.517	0.100
H	0.495	0.797	0.775	0.652	0.604

Tabella 3.5: Modello Dixon-Coles statico: metriche a livello di classe

Class	Precision (PPV)	NPV	Sensitivity	Accuracy	F1-score
A	0.534	0.778	0.600	0.664	0.565
D	0.444	0.724	0.074	0.519	0.127
H	0.500	0.793	0.761	0.653	0.603

Tabella 3.6: Modello Dixon-Coles dinamico: metriche a livello di classe

La Tabella 3.7, invece, contiene le metriche globali per ciascuno dei due modelli:

Model metrics	**Static Dixon-Coles**	**Dinamic Dixon-Coles**
Precision(PPV)	0.514	0.493
NPV	0.769	0.765
Sensitivity	0.482	0.478
Overall accuracy	0.516	0.511
Balanced accuracy	0.615	0.612
F1-score	0.428	0.432

Tabella 3.7: Confronto tra modelli: metriche globali

Ricordiamo inoltre che nell'ambito delle scommesse sportive, oltre alle classiche quote 1-X-2, sono molto frequenti le quote miste 1X ed X2. Potrebbe quindi essere interessante valutare come si comportano i nostri modelli in riferimento a previsioni di tipo binario. Chiaramente non è possibile costruire una matrice di confusione utilizzando come classi 1X ed X2, in quanto non risulterebbe valida la mutua esclusività (la loro intersezione fornisce i pareggi).

Tuttavia, ponendoci dal punto di vista della squadra di casa e considerando l'evento "la squadra di casa non perde", ecco che si potranno applicare i modelli per una classificazione binaria delle partite, in cui le classi saranno 1X (la squadra di casa non perde) e 2 (perde).

Le matrici di confusione relative a questo problema di classificazione binaria sono presentate di seguito, nella Figura 3.11:

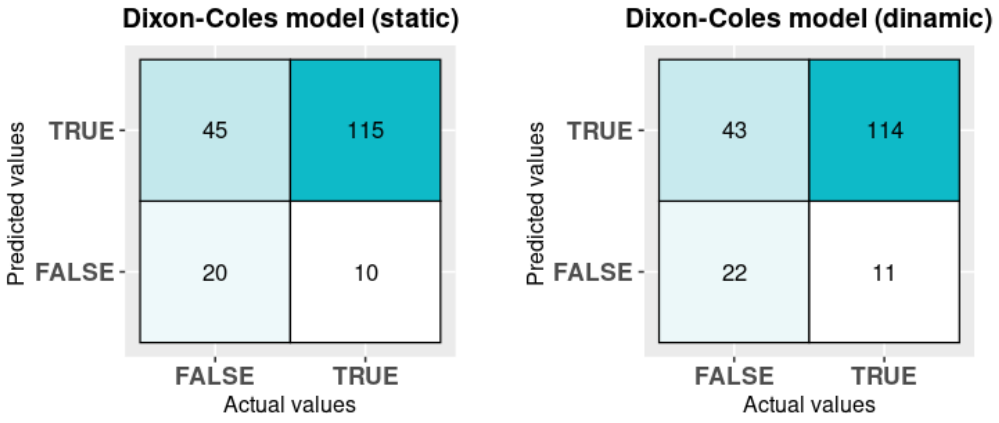

Figura 3.11: Matrici di confusione 1X - 2

Inoltre, dato che le matrici sono binarie, è stato possibile calcolare direttamente le metriche della Tabella 3.4 senza alcuna conversione preliminare (come nel caso multi-classe):

Model metrics	**Static Dixon-Coles**	**Dinamic Dixon-Coles**
Precision(PPV)	0.719	0.726
NPV	0.667	0.667
Sensitivity	0.920	0.912
Accuracy	0.711	0.716
F1-score	0.807	0.809

Tabella 3.8: Confronto metriche tra modelli

Un risultato analogo si poteva ottenere anche partendo dall'evento "la squadra in trasferta non perde" ed effettuando una classificazione binaria con le classi 1 ed X2. In particolare, tale risultato è riportato in Figura 3.12 ed in Tabella 3.9:

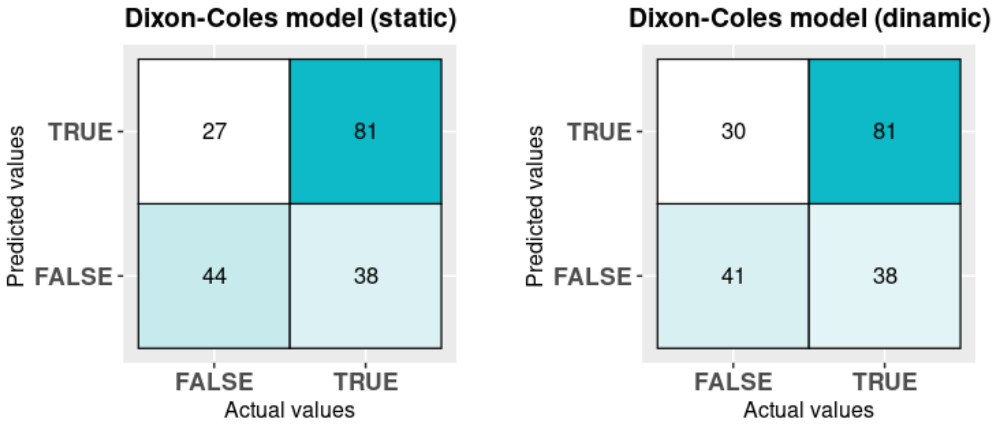

Figura 3.12: Matrici di confusione 1 - X2

Model metrics	**Static Dixon-Coles**	**Dinamic Dixon-Coles**
Precision(PPV)	0.750	0.730
NPV	0.537	0.519
Sensitivity	0.681	0.681
Accuracy	0.658	0.642
F1-score	0.714	0.704

Tabella 3.9: Confronto metriche tra modelli

Conclusioni

Il modello di Dixon e Coles presenta svariati ambiti di applicazione ed ha il vantaggio di essere piuttosto semplice, sia dal punto di vista teorico che implementativo. Esso risulta un prezioso punto di riferimento nel panorama dei modelli calcistici, in cui vengono spesso proposte leggere modifiche o generalizzazioni del modello "base" presentato in questo progetto.

Durante l'implementazione in R, aver adottato un approccio *from scratch* ha consentito una conoscenza più profonda dei concetti e delle dinamiche sottostanti i vari modelli, oltre che ad aver messo in luce alcune loro limitazioni. Tra queste, il processo di ottimizzazione per la verosimiglianza profilo è stato molto oneroso in termini di tempo, nonostante la dimensionalità del problema non fosse particolarmente esagerata. Un possibile sviluppo futuro per il progetto potrebbe dunque consistere nella parallelizzazione del codice, così da distribuire il carico di lavoro su più core (o macchine) ed accelerare il processo di ottimizzazione. Altrettanto interessante potrebbe essere indagare se i tempi di stima dei parametri si riducano utilizzando altri linguaggi di programmazione (o altre strutture dati). In *Python*, ad esempio, librerie come *NumPy* sono ottimizzate per svolgere efficientemente operazioni su arrays.

Infine, riguardo il confronto tra la versione statica e dinamica del modello di Dixon e Coles, le analisi statistiche tramite Brier-Score, Pseudo-R^2 e matrici di confusione non hanno fatto emergere particolari differenze tra i due modelli. Al contrario, esse sembrerebbero convergere verso la seguente conclusione: la ponderazione temporale del modello dinamico diventa apprezzabile per dati che coinvolgono più stagioni di campionato (come nell'articolo originale di Dixon e Coles), mentre quando si analizza una singola stagione sportiva non vale la pena abbandonare la versione statica del modello.

Bibliografia

[1] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[2] Mark J. Dixon and Stuart G. Coles. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 46(2):265–280, 1997.

[3] Leonardo Egidi and Nicola Torelli. Comparing goal-based and result-based approaches in modelling football outcomes. *Social Indicators Research*, 156, 08 2021.

[4] M.J. Moroney. *"Facts from figures"*. Penguin, 3rd edition, 1956.

[5] Dane McCarrick, Merim Bilalic, Nick Neave, and Sandy Wolfson. Home advantage during the covid-19 pandemic: Analyses of european football leagues. *Psychology of Sport and Exercise*, 56:102013, 2021.

[6] Richard Pollard. Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, 1, 06 2008.

[7] David Dandolo. Modellazione statistica di risultati calcistici. 2017.

[8] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950.

[9] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *ArXiv*, abs/2008.05756, 2020.