Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

Relazione finale

La previsione dei risultati delle partite di calcio: analisi dei dati e stima di modelli

Relatore Prof. Luigi Grossi

Dipartimento di Scienze Statistiche

Laureando: Alberto Franzò

Matricola n. 1138117

Anno Accademico 2021/2022

Indice

1. LA STATISTICA NELLO SPORT

1.1 Il trattamento dei dati nello sport

1.1.1 Cenni storici

I numeri sono sempre più citati ed analizzati nell'ambito degli sport di squadra: gol segnati, tiri da tre punti andati a canestro o meno, basi conquistate, muri riusciti sono sulla bocca di tutti gli appassionati. Per molto tempo però le analisi si sono limitate a poche variabili, le quali erano ritenute (erroneamente) riassuntive delle prestazioni di giocatori e squadre e dell'operato degli allenatori.
Una prima svolta è avvenuta negli Stati Uniti, in relazione al baseball, sport che si presta particolarmente all'analisi statistica grazie al ristretto numero di eventi che avviene in campo e alle competizioni "uno contro uno" facilmente isolabili. La cosiddetta *sabermetrica* (Nome derivante dall'acronimo SABR, "Society for American Baseball Research") è nata fin da quando il baseball è diventato sport professionistico e si è evoluta definitivamente negli anni '90 grazie a Billy Beane, general manager degli Oakland Athletics. Non avendo un budget paragonabile a quello di altre squadre, Beane assunse degli statistici nel suo team di gestione, arrivando alla conclusione che dati considerati fondamentali fino ad allora non erano così importanti nell'analisi delle prestazioni complessive. Vennero presi in considerazione nuovi indicatori con cui individuare i giocatori per assemblare la squadra e, alla fine, i risultati diedero ragione a questo approccio innovativo. Oggi questa rivoluzione viene detta "effetto Moneyball", dal titolo dell'omonimo libro poi divenuto anche un film che racconta la storia di Beane e della sua squadra, e che ha portato l'analisi dei dati a un livello superiore in tutti gli sport.

1.1.2 I dati nello sport oggi

Negli ultimi decenni, grazie allo sviluppo di sofisticate tecnologie, è stato possibile raccogliere sempre più dati su molteplici aspetti del mondo che ci circonda, inclusi

gli sport di squadra ed individuali. Sempre più società sportive, specie nel mondo anglosassone, hanno cominciato ad affidarsi a professionisti dei *big data* e a piattaforme di società specializzate nella raccolta dati.
Sensori, telecamere e computer ormai pervadono i campi da gioco non solo durante le partite, ma anche negli allenamenti, e non mancano nemmeno durante le sedute di fisioterapia o di recupero dagli infortuni.

1.2 I dati nel calcio

1.2.1 Breve storia

Il calcio è stato fin da subito tra gli sport più diffusi e popolari, ma forse anche per questo più tradizionalisti e meno inclini all'innovazione. È dovuto infatti passare un secolo dalla sua nascita prima che qualcuno cominciasse a registrare e contare gli eventi che avvenivano su un campo di gioco. È ciò che fece Charles Reep, che dal secondo dopoguerra raccolse dati su centinaia di partite con il solo ausilio di carta, penna e i suoi occhi. Cercando una "correlazione" tra gol segnati e numero di passaggi, Reep giunse ad affermare (erroneamente) che le reti arrivavano a seguito di azioni con pochi passaggi, e il suo pensiero influenzò lo stile di gioco inglese di quei tempi.

Il sopra citato "Effetto Moneyball" non fu però osservato immediatamente nel calcio come in altri sport: molti allenatori e manager sono stati inizialmente scettici, lo stesso Reep affermava che il gioco del calcio fosse troppo dominato dal caso per essere inquadrato da freddi numeri.
Alcune società "illuminate", probabilmente spinte dal budget minore rispetto alle concorrenti, iniziarono ad affidarsi ai numeri per acquistare giocatori a prezzo basso. Il primo esempio è stato il Brentford, gestito a partire dal 2012 da Matthew Benham, proprietario di una società di consulenza per scommettitori, la quale raccoglieva dati per fini di *betting*. Lo seguirono squadre più blasonate come il Liverpool e gli olandesi dell'AZ Alkmaar. Negli ultimi anni tutte le squadre di massima caratura, anche in Italia, hanno alle loro dipendenze almeno un *data*

analyst che partecipa a molte scelte nella gestione tecnica e sportiva, e si appoggiano a società esterne di raccolta dati.

1.2.2 Come vengono utilizzati i dati

I numerosi dati raccolti durante le partite (in un match di calcio si registrano alcune migliaia di eventi) vengono utilizzati per reclutare giocatori sconosciuti ai più, i quali eccellono in particolari caratteristiche che spesso sfuggono all'occhio di osservatori anche esperti. L'utilità può essere rilevante anche per i calciatori già presenti nella rosa, in particolare per evidenziare dei cali di prestazione o per cambiare tipo di preparazione nelle settimane seguenti.
I dati sono diventati fondamentali anche per la prevenzione degli infortuni, o per la gestione degli stessi. Una struttura all'avanguardia è stata per molti anni *Milan Lab*, un vero e proprio laboratorio fondato dalla società Milan che, raccogliendo milioni di dati, personalizza la preparazione fisico-atletica e pianifica efficientemente il recupero dagli infortuni.
Un terzo utilizzo, che verrà trattato in questa tesi, è improntato alla previsione di risultati ed eventi all'interno delle partite. Intere agenzie, che si occupano principalmente di scommesse sportive e sono quindi esterne al sistema delle società calcistiche, raccolgono i dati (o li comprano da altre aziende, anche dalle stesse società che li forniscono alle squadre), li elaborano e li rivendono a scommettitori professionisti.

1.2.3 La rilevazione dei dati

Dai tempi di Charles Reep molto è cambiato: si è passati da persone che raccoglievano i dati con carta e penna, a persone che "in differita" visionano istante per istante i video delle partite e registrano ogni evento in un computer. L'ultima frontiera, quella dell'intelligenza artificiale, utilizza sensori e telecamere che rilevano i dati automaticamente e li forniscono in modo istantaneo agli analisti.
Esistono anche delle piattaforme con questi dati accessibili al pubblico, o con un abbonamento o completamente *open source*. Tra le seconde, figura *fbref.com*, sito

che riporta i parametri raccolti dalla piattaforma *Statsbomb* e che sarà la fonte da cui verranno attinti i dati analizzati nel capitolo 3.

1.2.4 Metriche probabilistiche: gli *Expected Goals*

L'analisi statistica del calcio si è ulteriormente evoluta negli ultimi anni, non limitandosi a ciò che è visibile come tiri, passaggi o km percorsi. Sono nate infatti nuovi tipi di metriche, tra cui la più nota è detta *Expected Goals o xG* (Brechot e Flepp, 2020). Ad ogni tiro effettuato in una partita viene assegnata una probabilità di finire in rete, che corrisponde al suo valore xG. Questo valore viene definito analizzando gli esiti di centinaia di migliaia di tiri avvenuti in passato con diverse caratteristiche (distanza dalla porta, posizione del corpo, posizione del portiere e dei difensori e molte altre). Sommando gli *Expected Goals* dei tiri divisi per squadra si ottiene un "risultato oggettivo", cioè l'esito atteso di una partita su base probabilistica. Essendo il calcio però uno sport a punteggi bassi e deciso da pochi eventi, spesso il risultato reale non coincide con quello stabilito da questa metrica, la quale però dà un'idea di chi ha creato più azioni offensive.
Oltre agli xG diverse nuove metriche si stanno sviluppando, con aziende che si occupano interamente di comporre algoritmi adibiti a combinare ed elaborare le variabili raccolte in campo per creare dei veri e propri indici di prestazione.

1.3 Obiettivo della tesi

Questa tesi si propone di passare in rassegna i modelli per la previsione dei risultati delle partite di calcio già presenti in letteratura (capitolo 2), per poi applicarne e valutarne uno sui dati più recenti del campionato italiano (capitolo 3).

2. MODELLI PRESENTI IN LETTERATURA PER LA PREVISIONE DEI RISULTATI SPORTIVI

In questo capitolo verrà presentata una rassegna dei principali modelli presenti in letteratura per la previsione dei risultati di incontri di calcio.

2.1 La distribuzione di Poisson

2.1.1 Gol segnati da una squadra e distribuzione di Poisson

Il calcio è uno sport "a punteggio basso", in quanto il numero di gol segnati da una squadra in una partita è, nella stragrande maggioranza dei casi, compreso tra 1 e 2, come si può vedere dalla Tabella 1.

Gol segnati	0	1	2	3	4	5	6 o più
Frequenza assoluta	732	1023	727	359	145	37	17

Tabella 1: *Gol segnati da una squadra e relative frequenze assolute nelle ultime 4 stagioni della Serie A italiana.*

Risulta ragionevole quindi pensare che tale processo possa essere approssimato con un processo di Poisson di parametro λ. La funzione di probabilità di una variabile casuale X di Poisson è la seguente:

$$p_x(X) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0,1,2,\ldots \quad \text{e} \quad \lambda > 0 \qquad (1)$$

Dove X rappresenta il numero di eventi accaduti in un certo periodo di tempo (in questo caso, reti segnate da una squadra in una partita) e il parametro λ riassume valore atteso e varianza della distribuzione.

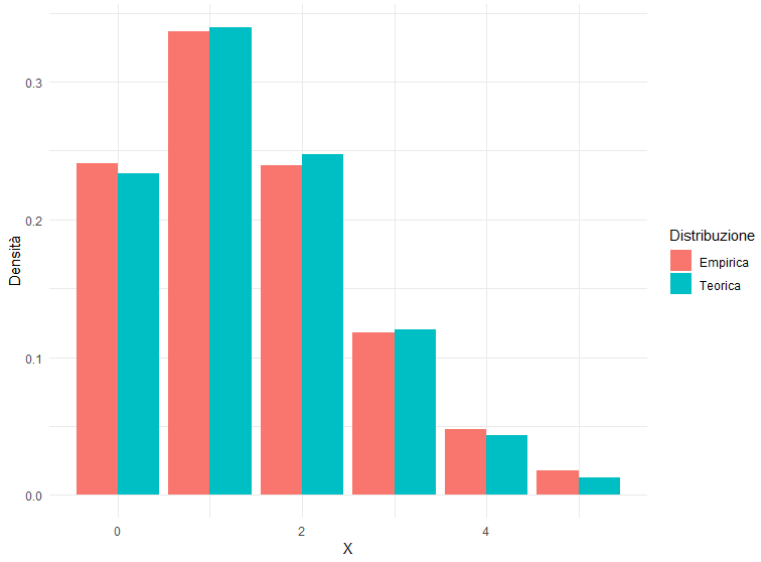


Figura 1: *Distribuzione empirica (dati Serie A, ultime 4 stagioni) e distribuzione teorica di Poisson (con lambda=1.456) dei gol segnati, colonne affiancate.*

La Figura 1 mostra che l'adattamento dei dati sulle reti segnate alla distribuzione ipotizzata è ottimo: nel capitolo 3, riguardante l'applicazione, si scenderà più nel dettaglio andando a vedere se il modello scelto approssimi bene anche dati di una singola squadra.
L'assunzione distributiva di Poisson per questo argomento è comunque universalmente riconosciuta, fin da quando questi dati sono stati analizzati (Maher, 1982), e quindi costituirà un punto di partenza per l'intera tesi.

2.1.2 La distribuzione Double Poisson: modello di Maher

Com'è noto, in una partita di calcio si affrontano due squadre; quindi, sono due le distribuzioni di Poisson da utilizzare per modellare il risultato previsto. A tal proposito un primo modello è stato proposto da Maher (1982) e prevede l'utilizzo di due Poisson indipendenti. Il numero di reti della i-esima squadra di casa, X_{ij}, e I gol segnati dalla j-esima squadra in trasferta, Y_{ij}, vengono modellati in questo modo:

$$X_{ij} \sim Poisson(\alpha_i \beta_j) \qquad Y_{ij} \sim Poisson(\alpha_j \beta_i) \qquad con\ X_{ij} \perp\!\!\!\perp Y_{ij} \tag{2}$$

In cui il parametro α_i indica la forza offensiva in casa della i-esima squadra, β_i la forza difensiva in casa della i-esima compagine e analogamente α_j e β_j per la j-esima squadra in trasferta. Tali parametri sono stimati da un modello di regressione (argomento del paragrafo 2.1.4), e la loro unicità è garantita dalla condizione indicata nella (3).

$$\sum_i \alpha_i = \sum_j \beta_j \,; \qquad \sum_i \beta_i = \sum_j \alpha_j \tag{3}$$

Il modello teorico diventa, per esteso:

$$p_x(X_{ij}) = \frac{e^{-\alpha_i\beta_j}(\alpha_i\beta_j)^x}{x!}; \quad p_y(Y_{ij}) = \frac{e^{-\alpha_j\beta_i}(\alpha_j\beta_i)^y}{y!};$$
$$X_{ij} \perp\!\!\!\perp Y_{ij};\ x = 0,1,2,\ldots \quad \text{e} \quad \alpha_k\beta_k > 0 \tag{4}$$

2.1.3 La distribuzione Bivariate Poisson

Il modello più plausibile e, come verrà confermato in seguito, più realistico, prevede di considerare congiuntamente le due Poisson in una distribuzione bivariata contenente un parametro di dipendenza. Il modello teorico di base è stato introdotto da Kocherlakota e Kocherlakota (1992), per poi essere applicato al contesto calcistico da Dixon e Coles prima (par. 2.2) e da Karlis e Ntzoufras poi (par. 2.3) con modifiche diverse.

Il modello di Kocherlakota e Kocherlakota per la Bivariate Poisson (BP in breve) prevede l'utilizzo di tre parametri $\lambda_1, \lambda_2, \lambda_3$, in cui λ_3 rappresenta la covarianza tra le due marginali. La funzione di probabilità congiunta è

$$P(X = x, Y = y) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!} \sum_{k=0}^{\min(x,y)} \binom{x}{k}\binom{y}{k} k! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k \tag{5}$$

Le due distribuzioni marginali sono Poisson, $X \sim Poisson(\lambda_1 + \lambda_2)$ e $Y \sim Poisson(\lambda_2 + \lambda_3)$, mentre se $\lambda_3 = 0$ si torna alla Double Poisson.

2.1.4 Regressione di Poisson

Come visto nei paragrafi precedenti, qualsiasi sia il modello di Poisson utilizzato il parametro della distribuzione (λ_i, gol segnati in media) è dipendente da altri parametri, variabili in base alla squadra e/o alla partita affrontata. λ diventa quindi una variabile risposta di un modello di regressione, con i covariate β_i:

$$\lambda_i = \beta_0 + \beta_1 x_i + \cdots \qquad (6)$$

Un semplice modello lineare come il (6) però non soddisfa le condizioni della distribuzione ipotizzata, in quanto λ_i dovrebbe essere sempre maggiore di zero. Per ovviare a questo problema si usa la trasformata logaritmica

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i + \cdots \qquad (7)$$

Per ritrovare λ_i basta applicare la trasformazione inversa al logaritmo, l'esponenziale:

$$\lambda_i = e^{\beta_0 + \beta_1 x_i + \cdots} \qquad (8)$$

Si parla di regressione di Poisson, un modello che deve soddisfare quattro principali assunzioni:

- Variabile risposta di Poisson: deve essere un conteggio e quindi un numero intero, descritto da una distribuzione di Poisson;
- Indipendenza tra le osservazioni;
- Media = varianza, come da ipotesi della distribuzione di Poisson;
- Linearità: $\log(\lambda_i)$ dev'essere una funzione lineare delle x_i.

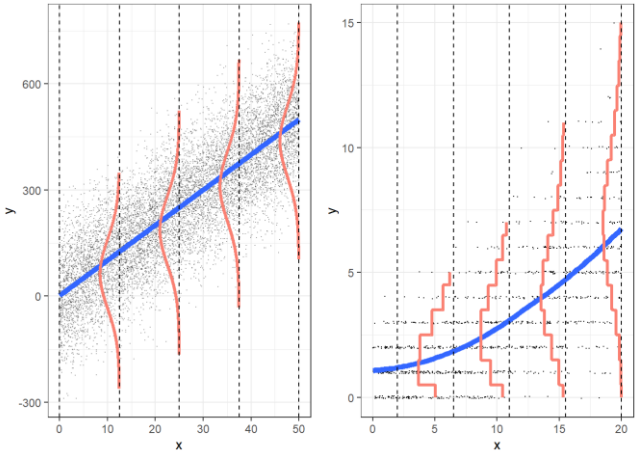


Figura 2: *differenza tra un modello di regressione lineare (a sinistra) e un modello di regressione di Poisson (a destra).*

La stima dei parametri non può essere fatta con il metodo dei minimi quadrati, ma con i metodi di massima verosimiglianza.

2.2 Modello di Dixon-Coles

2.2.1 Oltre la Double Poisson

Dixon e Coles (1996) affermano che il modello Double Poisson utilizzato da Maher (par. 2.1.2), che assume l'indipendenza tra le distribuzioni delle reti segnate dalle due squadre, è inesatto: provando infatti ad utilizzare il modello con i dati dal 1992 al 1995 per calcolare le probabilità di ogni risultato dallo 0-0 al 4-4, si nota che l'ipotesi di indipendenza non può essere accettata per risultati "bassi" (0-0, 1-0, 0-1 e 1-1). Va utilizzato un modello bivariato, con l'inserimento di un parametro di dipendenza.

2.2.2 Bivariate Poisson con parametro di dipendenza per risultati bassi

Secondo Dixon e Coles, le caratteristiche del modello devono essere le seguenti:

- Il modello deve tenere conto delle diverse abilità delle due squadre partecipanti alla partita;

- Deve essere inserito un parametro correttivo per l'*home effect*, in quanto le squadre che giocano in casa generalmente hanno un vantaggio;
- La misura più ragionevole dell'abilità di una squadra è la valutazione delle sue performance recenti;
- Le performance vanno suddivise in offensive (capacità di segnare gol) e difensive (capacità di non subire gol);
- Nel valutare tali performance bisogna tenere conto anche del valore degli avversari affrontati.

Per soddisfare queste caratteristiche, e in particolare l'ultimo punto, il modello non può prevedere l'indipendenza tra le distribuzioni delle due sfidanti in caso di risultati bassi (par. 2.2.1), per questo viene proposta un'alternativa bivariata al modello di Maher:

$$\Pr(X_{ij} = x, Y_{ij} = y) = \pi_{\lambda_1,\lambda_2}(x,y)\frac{e^{-\lambda_1}\lambda_1^x}{x!}\frac{e^{-\lambda_2}\lambda_2^y}{y!} \qquad (8)$$

Con $\lambda_1 = \alpha_i\beta_j\gamma$ parametro per la squadra che gioca in casa e $\lambda_2 = \alpha_j\beta_i$ per la squadra in trasferta, in cui α indica la performance offensiva, β la prestazione difensiva e γ il vantaggio dato dall'*home effect*.

$\tau_{\lambda_1,\lambda_2}(x,y)$ è il parametro che definisce la dipendenza tra le due squadre, che assume valori diversi da 1 in caso di risultati bassi.

$$\pi_{\lambda_1,\lambda_2}(x,y) = \begin{cases} 1-\lambda_1\lambda_2\rho & se\ x = y = 0 \\ 1+\lambda_1\rho & se\ x = 0, y = 1 \\ 1+\lambda_2\rho & se\ x = 1, y = 0 \\ 1-\rho & se\ x = y = 1 \\ 1 & altrimenti \end{cases} \qquad (9)$$

Con $\max\left(-\frac{1}{\lambda_1}, -\frac{1}{\lambda_2}\right) \leq \rho \leq \min\left(\frac{1}{\lambda_1\lambda_2}, 1\right)$ che è il vero e proprio parametro di dipendenza. In caso di $\rho = 0$ ci sarebbe indipendenza per tutti i risultati.

Da notare che le due distribuzioni marginali rimangono comunque Poisson.

2.3 Modello di Karlis-Ntzoufras

2.3.1 Analisi dei pareggi

Così come Dixon e Coles, anche Karlis e Ntzoufras (2003) appoggiano la tesi di utilizzare una distribuzione Bivariate Poisson ed evidenziano la forte differenza di probabilità di pareggio tra BP e Double Poisson. La figura 3 mostra proprio la differenza relativa di probabilità di pareggio tra i due modelli (sull'asse delle ordinate) al variare di λ_2, con λ_1 fissato e con diversi valori di covarianza λ_3.

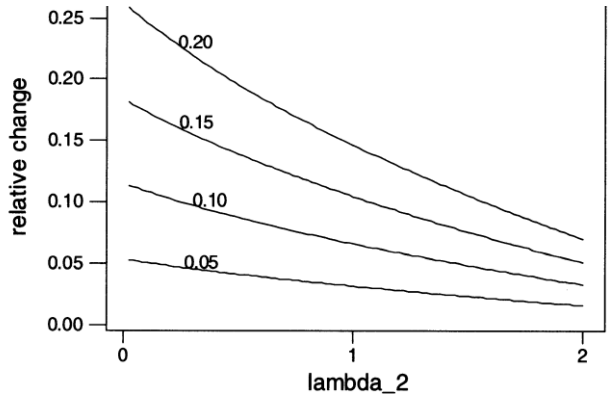


Figura 3: *Differenza relativa di probabilità di pareggio tra I modelli Bivariate Poisson e Double Poisson, al variare di λ_2 e λ_3 e con $\lambda_1 = 1$. Le diverse linee rappresentano differenti valori di λ_3.*

Ad esempio, con $\lambda_1 = \lambda_2 = 1$ e $\lambda_3 = 0.05$ il modello BP prevede il 3,3% in più di pareggi, valore che sale al 14% se la covarianza sale e $\lambda_3 = 0.20$. Anche le evidenze empiriche, come già notato da Dixon e Coles, mostravano più pareggi di quanti previsti con la Double Poisson: la strada della Bivariate Poisson si conferma quella corretta.

2.3.2 Ampliamento del modello Dixon-Coles a tutti i pareggi

Quanto visto al paragrafo precedente ha portato a modificare il modello Dixon-Coles, che prevedeva una correzione per I risultati 0-0, 1-0, 0-1 e 1-1 (par. 2.2), e di virare verso un modello *diagonal inflated*, che rivede i termini "diagonali" riferiti ai

pareggi nella tabella di probabilità aggiungendo un contributo $D(x, \theta)$ che ha a sua volta una distribuzione discreta (geometrica, Poisson o altro):

$$P_D(X = x, Y = y) = \begin{cases} (1-p)BP(x, y|\lambda_1, \lambda_2, \lambda_3) & se\ x \neq y \\ (1-p)BP(x, y|\lambda_1, \lambda_2, \lambda_3) + pD(x, \theta) & se\ x = y \end{cases} \quad (10)$$

Con p pari al valore di *inflation*, ovvero a quanto si vuole che pesi la differenza di stima sulla diagonale.
Modelli come questo sono stimabili con algoritmi EM (Expectation-Maximization).

2.4 Modello di Koopman-Lit

2.4.1 Specificazione con parametri dinamici

Koopman e Lit (2013) concordano sul fatto che il miglior modello per descrivere il risultato di un match calcistico sia la Bivariate Poisson (5), ma fanno un passo oltre a Dixon-Coles e Karlis-Ntzoufras nella definizione dei parametri di regressione.
Nelle precedenti trattazioni infatti i parametri offensivi e difensivi coefficienti $\lambda_1\ e\ \lambda_2$ erano fissati all'inizio della stagione e poi considerati costanti, ma ciò non è realistico: le capacità di una squadra possono migliorare o peggiorare (anche se non di molto) all'interno della stagione, per questo c'è bisogno di un modello di regressione dinamico. Definendo $\lambda_1\ e\ \lambda_2$ come nella (11), considerando una generica partita tra l'i-esima squadra in casa e la j-esima squadra in trasferta avvenuta al tempo t

$$\lambda_{1ijt} = e^{\delta + \alpha_{it} + \beta_{jt}};\quad \lambda_{2ijt} = e^{\alpha_{jt} + \beta_{it}} \quad (11)$$

In cui δ rappresenta l'*home effect*, α_{it} la capacità offensiva e β_t la capacità difensiva, con gli ultimi due dipendenti dal tempo.
α_{kt} e β_{kt} sono quindi delle serie storiche, che vengono modellate come dei processi ARIMA, in questo caso processi autoregressivi AR(1), così definiti per la i-esima squadra:

$$\alpha_{i,t} = \mu_{\alpha i} + \varphi_{\alpha i}\alpha_{i,t-1} + \eta_{\alpha i,t}$$
$$\beta_{it} = \mu_{\beta i} + \varphi_{\beta i}\beta_{i,t-1} + \eta_{\beta i,t} \qquad (12)$$

Dove $\mu_{\alpha i}$ e $\mu_{\beta i}$ sono costanti ignote, $\varphi_{\alpha i}$ e $\varphi_{\beta i}$ sono i coefficienti autoregressivi e $\eta_{\alpha i,t}$, $\eta_{\beta i,t}$ rappresentano dei termini di errore normalmente distribuiti e indipendenti, in breve

$$\eta_{ki,t} \sim NID(0, \sigma_{ki}^2) \quad con\ k = \alpha, \beta \quad e \quad \sigma_{ki}^2 > 0 \qquad (13)$$

Assumendo poi che tutti i processi AR siano tra loro indipendenti e stazionari deve valere che $|\varphi_{ki}| < 1$ con $k = \alpha, \beta$.
Le condizioni iniziali dei processi sono basate sulla media e la varianza delle distribuzioni

$$E(k_{it}) = \frac{\mu_{ki}}{1 - \varphi_{ki}}; \quad V(k_{it}) = \frac{\sigma_{ki}^2}{1 - \varphi_{ki}^2} \qquad con\ k = \alpha, \beta \qquad (14)$$

Tutti i parametri vengono stimati tramite un algoritmo Monte Carlo di massimizzazione della verosimiglianza.

2.4.2 Aggiustamento per risultati bassi

Il modello base presentato al paragrafo 2.4.1 si presta a vari aggiustamenti. Il più importante riguarda la già trattata (par. 2.2.1) sottostima della probabilità di avere dei "risultati bassi", che Koopman e Lit risolvono, analogamente a Dixon e Coles, utilizzando un parametro moltiplicativo sulla funzione della probabilità della Poisson, $\pi_{\lambda_1,\lambda_2}(x, y)$, che però definiscono in modo leggermente diverso rispetto alla (9)

$$\pi_{\lambda_1,\lambda_2}(x, y) = \begin{cases} 1 - \lambda_1\lambda_2\rho & se\ x = y = 0 \\ 1 + \lambda_1\rho & se\ x = 0, y = 1 \\ 1 + \lambda_2\rho & se\ x = 1, y = 0 \\ 1 - \dfrac{\rho}{1 + \dfrac{\lambda_3}{\lambda_1\lambda_2}} & se\ x = y = 1 \\ 1 & altrimenti \end{cases} \qquad (15)$$

2.4.3 Altri aggiustamenti

Koopman e Lit propongono anche altri possibili accorgimenti per rendere il modello più realistico. Il primo riguarda la modifica dei termini d'errore $\eta_{\alpha i,t}$ e $\eta_{\beta i,t}$ in occasione delle finestre di calciomercato estivo e invernale (*summer break* e *winter break*), in cui le squadre acquistano e cedono giocatori potendo potenzialmente cambiare in modo sostanziale le loro capacità offensive e difensive. Si ha quindi che

$$\eta_{ki,t} \sim NID\left(0, \sigma_{ki}^2 + \sigma_{kS}^2 \tau_S(t) + \sigma_{kW}^2 \tau_W(t)\right) \quad con\ k = \alpha, \beta \quad e \quad \sigma_{ki}^2, \sigma_{kS}^2, \sigma_{kW}^2 > 0 \qquad (16)$$

Con $\tau_S(t)$ e $\tau_W(t)$ variabili indicatrici che vengono poste uguali a 1 rispettivamente al termine del *summer break* (fine agosto) e del *winter break* (fine gennaio).
Un ulteriore accorgimento può riguardare il parametro δ che riguarda l'*home effect*, che nel modello base viene considerato uguale per tutte le squadre. È plausibile però pensare che alcune compagini abbiano un maggiore vantaggio a giocare nel proprio campo: si possono quindi suddividere le partecipanti in due o più gruppi, con coefficienti δ diversi per ogni gruppo.

3. APPLICAZIONE CON DATI DELLA SERIE A

3.1 Provenienza dei dati e rilevazione

I dati utilizzati per l'applicazione riguardano le ultime 4 stagioni (da agosto 2018 a maggio 2022) del campionato italiano di Serie A, e sono stati scaricati da *fbref.com*, un sito che a sua volta raccoglie parte dei dati forniti da *Statsbomb*, azienda tra le migliori al mondo nella raccolta e distribuzione di dati calcistici.
Le tabelle contenenti tutti i risultati delle partite del periodo considerato possono essere scaricate in formato .csv e poi importate su R, oppure direttamente caricate su R tramite il comando *read_html()* del pacchetto *rvest*. La schermata viene riportata nella figura 4.

X.U.FEFF.Wk	Day	Date	Time	Home	xG1	Score1	Score2	xG2	Away
1	Sat	2018-08-18	18:00	Chievo	1.0	2	3	2.3	Juventus
1	Sat	2018-08-18	20:30	Lazio	1.1	1	2	1.5	Napoli
1	Sun	2018-08-19	18:00	Torino	0.7	0	1	1.5	Roma
1	Sun	2018-08-19	20:30	Sassuolo	1.6	1	0	1.4	Inter
1	Sun	2018-08-19	20:30	Empoli	0.9	2	0	0.9	Cagliari
1	Sun	2018-08-19	20:30	Parma	1.4	2	2	2.1	Udinese
1	Sun	2018-08-19	20:30	Bologna	0.9	0	1	0.7	SPAL
1	Mon	2018-08-20	20:30	Atalanta	2.3	4	0	0.5	Frosinone
2	Sat	2018-08-25	18:00	Juventus	2.2	2	0	0.2	Lazio
2	Sat	2018-08-25	20:30	Napoli	2.1	3	2	0.7	Milan
2	Sun	2018-08-26	18:00	SPAL	1.2	1	0	0.5	Parma
2	Sun	2018-08-26	20:30	Cagliari	1.6	2	2	1.9	Sassuolo
2	Sun	2018-08-26	20:30	Genoa	1.1	2	1	2.3	Empoli

Figura 4: *Prime righe della tabella estratta da fbref.com. Contiene nell'ordine: numero della giornata, giorno della settimana, data, ora, nome della squadra di casa, xG della squadra di casa, gol della squadra di casa, gol della squadra in trasferta, xG della squadra in trasferta, nome della squadra in trasferta.*

3.2 Adattamento della distribuzione di Poisson ai dati sui gol

3.2.1 Adattamento con dati di tutte le squadre

La prima verifica da eseguire prima di applicare un modello è quella sull'assunzione distributiva: in questo caso i dati sui gol (*Score1* e *Score2* nella Figura 4) devono seguire, almeno approssimativamente, una distribuzione di Poisson.
Prendendo i dati aggregati di tutte le squadre e per le 4 stagioni considerate (n = 3040) il test chi quadro di Pearson per valutare la bontà di adattamento fornisce un ottimo risultato (Figura 5). A conferma di ciò, media campionaria e varianza campionaria sono molto simili, rispettivamente 1.456 e 1.523.

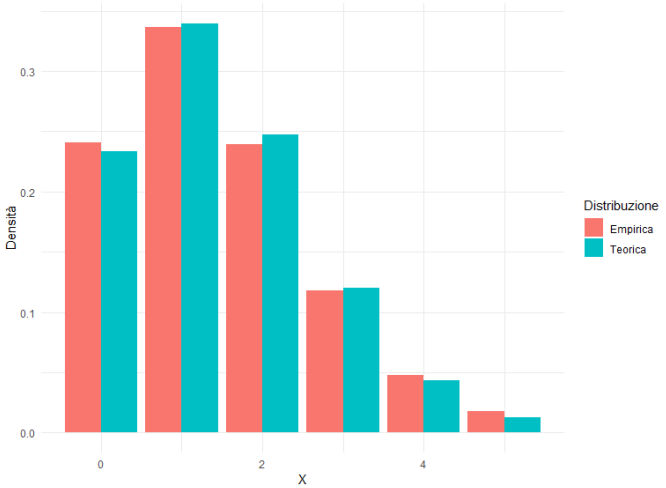


Figura 5: *Distribuzione empirica e distribuzione teorica di Poisson (con lambda=1.456) dei gol segnati da tutte le squadre nelle ultime 4 stagioni della Serie A italiana, e output di R della funzione chisq.test().*

3.2.2 Adattamento con dati di una sola squadra in una stagione

Nel presente paragrafo, si restringe il campo considerando la distribuzione delle reti segnate da una singola squadra in un'unica stagione (n = 38).
Prendendo come esempio due squadre (Milan e Juventus), ognuna con la propria distribuzione, l'adattamento risulta sempre essere buono, anche se con delle ovvie

differenze tra squadra e squadra e con divergenze maggiori tra media e varianza campionarie, rispetto al caso generale, date dalla minore numerosità campionaria.

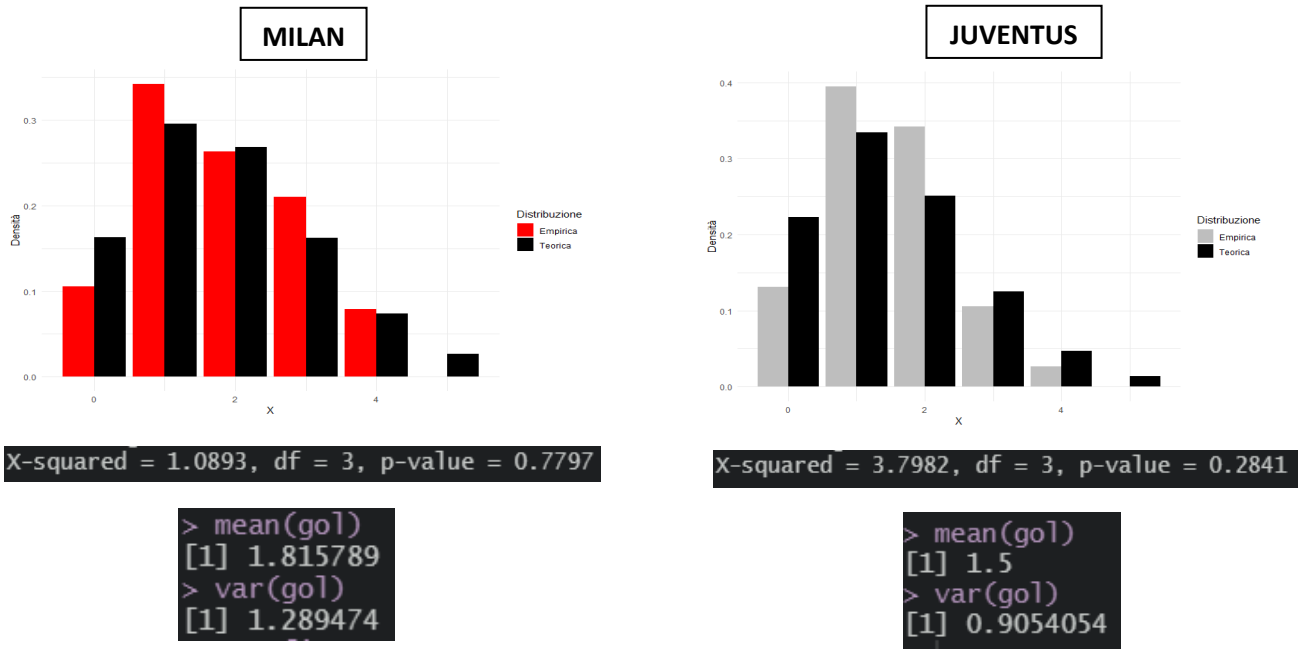


Figura 6: *Distribuzione empirica e distribuzione teorica di Poisson dei gol segnati da Milan e Juventus nella stagione 2021-2022, output di R della funzione chisq.test() e del calcolo di medie e varianze campionarie.*

Si noti che, per l'applicazione del test chi quadro, il numero di classi è stato ridotto, accorpando le classi con $X \geq 3$, per avere una numerosità almeno pari a 5 in ogni classe. Nei grafici invece sono state inserite tutte le classi separate fino a X = 5 per renderli più esplicativi.

3.3 Applicazione del modello previsivo

3.3.1 Modello utilizzato

Il modello applicato è quello di Dixon e Coles introdotto nel par. 2.2, con una variante relativa alla definizione dei parametri λ_i e λ_j che non verranno determinati per moltiplicazione, ma attraverso due modelli di regressione di Poisson che, per la i-esima squadra in casa e la j-esima squadra in trasferta, sono così definiti:

$$\log(\lambda_i) = \varphi_{0ij} + \varphi_{1ij}\alpha_i + \varphi_{2ij}\beta_j \qquad (17)$$

$$\log(\lambda_j) = \varphi_{0ji} + \varphi_{1ji}\alpha_j + \varphi_{2ji}\beta_i \qquad (18)$$

dove α e β sono parametri dipendenti dalla forza offensiva e difensiva delle squadre e φ parametri di regressione (non è stata usata la consueta β per chiarezza di notazione).

Per ottenere le stime dei parametri λ_i e λ_j è sufficiente applicare la trasformazione esponenziale. Si ricordi che λ_i deve essere corretto per tener conto dell'*home effect* γ:

$$\lambda_{ih} = \gamma e^{\log(\lambda_i)}; \quad \lambda_j = e^{\log(\lambda_j)} \qquad (19)$$

Successivamente, si procede ad applicare la distribuzione Poisson bivariata, specificata nella (8) e nella (9), per calcolare le probabilità di 25 possibili risultati compresi tra 0-0 e 4-4.

Nei paragrafi seguenti vengono motivate le scelte e le specificazioni di ogni singolo parametro utilizzato nel modello sopra specificato.

3.3.2 Parametri α e β

I parametri di capacità offensiva (α) e difensiva (β) sono definiti in funzione del tempo: essi sono infatti la previsione a un passo delle serie storiche di gol fatti e subiti utilizzando la tecnica di lisciamento esponenziale di Holt additivo. Le serie sono state preventivamente standardizzate rispetto a media e deviazione standard generali di tutto il dataset, per poter operare con valori meno distanti tra loro tipici di una serie discreta come quella delle reti segnate/subite.

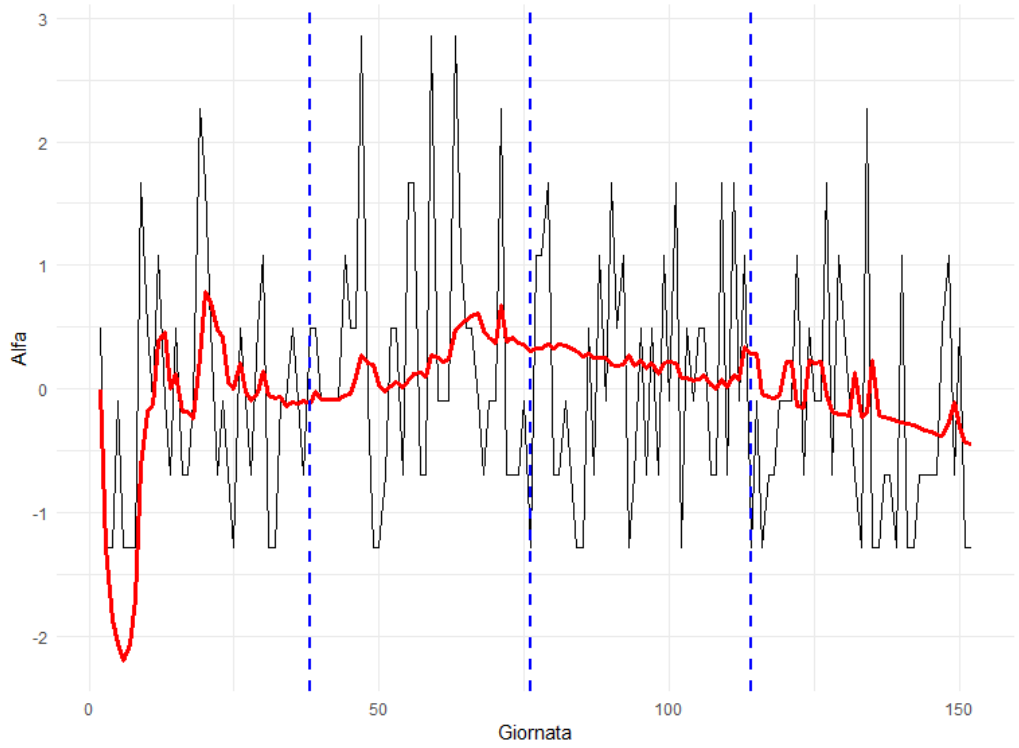
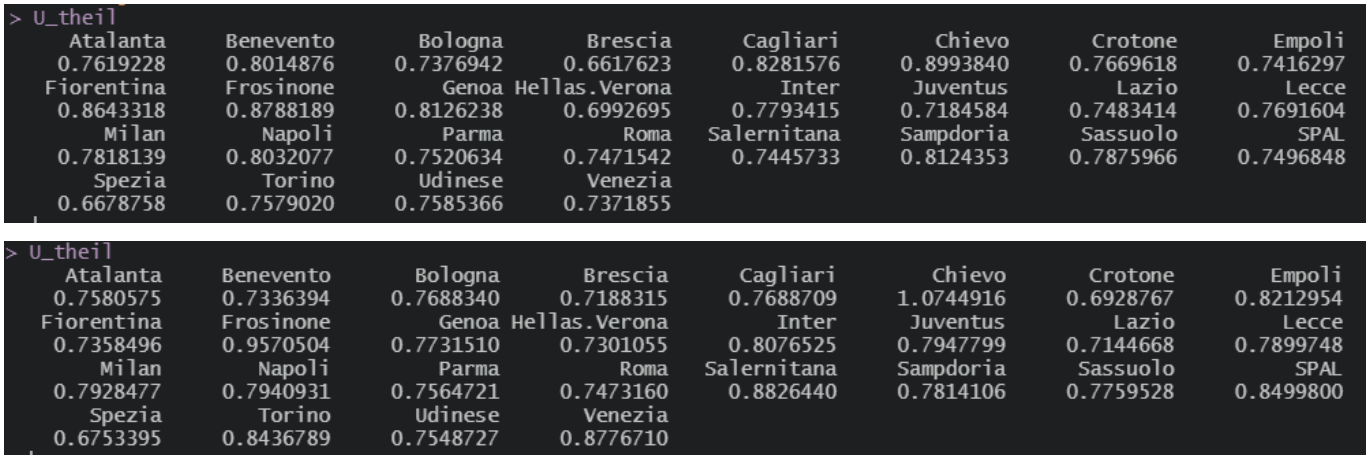


Figura 6: *Andamento del parametro α ottenuto con previsioni ad un passo tramite lisciamento esponenziale di Holt per la squadra Atalanta tra settembre 2018 e maggio 2022 (in rosso) rispetto all'andamento reale della serie standardizzata delle reti segnate (in nero). Le linee blu delimitano le stagioni.*

La bontà della tecnica previsiva viene confermata dall'indice U di Theil, che confronta le previsioni fatte con delle previsioni *naive* ottenute prendendo semplicemente il dato precedente nella serie. Se tale indice assume valori minori di 1 la tecnica scelta è più efficace di quella del cosiddetto "*modello naive*".



```
> U_theil
    Atalanta     Benevento      Bologna      Brescia     Cagliari      Chievo      Crotone       Empoli
   0.7619228     0.8014876    0.7376942    0.6617623    0.8281576   0.8993840   0.7669618    0.7416297
  Fiorentina     Frosinone        Genoa Hellas.Verona        Inter    Juventus       Lazio        Lecce
   0.8643318     0.8788189    0.8126238    0.6992695    0.7793415   0.7184584   0.7483414    0.7691604
       Milan        Napoli        Parma         Roma   Salernitana   Sampdoria   Sassuolo         SPAL
   0.7818139     0.8032077    0.7520634    0.7471542    0.7445733   0.8124353   0.7875966    0.7496848
      Spezia        Torino      Udinese      Venezia
   0.6678758     0.7579020    0.7585366    0.7371855

> U_theil
    Atalanta     Benevento      Bologna      Brescia     Cagliari      Chievo      Crotone       Empoli
   0.7580575     0.7336394    0.7688340    0.7188315    0.7688709   1.0744916   0.6928767    0.8212954
  Fiorentina     Frosinone        Genoa Hellas.Verona        Inter    Juventus       Lazio        Lecce
   0.7358496     0.9570504    0.7731510    0.7301055    0.8076525   0.7947799   0.7144668    0.7899748
       Milan        Napoli        Parma         Roma   Salernitana   Sampdoria   Sassuolo         SPAL
   0.7928477     0.7940931    0.7564721    0.7473160    0.8826440   0.7814106   0.7759528    0.8499800
      Spezia        Torino      Udinese      Venezia
   0.6753395     0.8436789    0.7548727    0.8776710
```

Figura 7: *Indici U di Theil di ogni singola squadra per le previsioni tra settembre 2018 e maggio 2022, utilizzando il lisciamento esponenziale di Holt. Output di R.*

Un solo indice supera la soglia di 1: si tratta del parametro difensivo del Chievo, squadra che ha partecipato al campionato solo nella prima stagione considerata (2018-2019), stagione che non verrà utilizzata in questa applicazione per fare previsioni.

3.3.3 Parametro γ di home effect

Si nota che il numero di goal segnati dalle squadre che giocano in casa è superiore rispetto al numero di reti segnate dalle squadre in trasferta: nel periodo considerato nel dataset sono rispettivamente 2372 e 2053, per cui i goal segnati in casa sono superiori del 15.54% rispetto ai goal segnati in trasferta.
Per questo motivo il parametro γ di home effect, che moltiplica λ_i, viene posto pari a 1.1554.

3.3.4 Parametro ρ di dipendenza

Come spiegato nel par. 2.3.1 e mostrato in Figura 3, il modello necessita di un parametro di dipendenza ρ per aumentare la probabilità che si verifichino risultati con un numero contenuto di goal. Verranno analizzati i risultati ottenuti con i quattro valori proposti in Figura 3 (0.05, 0.10, 0.15 e 0.20) per valutare quale sia il migliore da utilizzare.

3.3.5 Risultati

Il modello definito in base alle scelte illustrate nei paragrafi precedenti è stato applicato per previsioni in-sample alle partite della 8ª, 15ª e 29ª giornata della Serie A 2021-2022. La scelta di giornate relativamente distanti nel tempo è stata fatta per apprezzare dei cambiamenti significativi nei parametri α e β; inoltre, non sono state prese le giornate iniziali per la scarsità di dati disponibili riguardo a squadre esordienti nel campionato (Salernitana e Venezia), e nemmeno giornate troppo a ridosso della fine del torneo (38ª giornata), in cui alcune squadre senza più obiettivi di classifica non partecipano al pieno delle proprie capacità. In ultima instanza, non si è preso in considerazione neanche il periodo tra la 19ª e la 23ª in cui la pandemia di COVID-19 ha portato al rinvio di molte partite, che quindi sono state disputate in momenti diversi rispetto alle altre.
Per ogni partita vengono calcolate le probabilità che si verifichino i risultati tra 0-0 e 4-4, costruendo una matrice come quella mostrata in Figura 8.

	0	1	2	3	4
0	0.0263	0.0327	0.0101	0.0026	0.0005
1	0.0936	0.0612	0.0266	0.0069	0.0014
2	0.1140	0.0893	0.0349	0.0091	0.0018
3	0.0998	0.0782	0.0306	0.0080	0.0016
4	0.0656	0.0513	0.0201	0.0052	0.0010

Figura 8: *Matrice di probabilità di 25 risultati per la partita Atalanta-Venezia, disputata nella 15ª giornata di Serie A 2021-2022, calcolata con ρ = 0.10. Sulle righe i goal della squadra di casa (Atalanta), sulle colonne i goal della squadra in trasferta (Venezia).*

Da tale matrice viene poi estratto il risultato più probabile (nel caso di Figura 8, 2-0) e confrontato con ciò che si è realmente verificato, sia in termini di reti, che per quanto riguarda l'esito della partita (vittoria casa, pareggio o vittoria trasferta, di seguito denominate rispettivamente 1-X-2 come nel gergo delle scommesse).

Partita	ρ = 0.05		ρ = 0.10		ρ = 0.15		ρ = 0.20		Osservato	
	Risultato (prob)	Esito (prob)	Risultato (prob)	Esito (prob)	Risultato (prob)	Esito (prob)	Risultato (prob)	Esito (prob)	Risultato	Esito
Spezia-Salernitana	2-0 (0.1736)	1 (0.7253)	2-0 (0.1736)	1 (0.7280)	2-0 (0.1736)	1 (0.7311)	2-0 (0.1736)	1 (0.7340)	2-1	1
Lazio - Inter	2-2 (0.0673)	2 (0.5630)	2-2 (0.0673)	2 (0.5650)	2-2 (0.0673)	2 (0.5680)	2-2 (0.0673)	2 (0.5700)	3-1	1
Milan - Verona	2-1 (0.0816)	1 (0.4480)	2-1 (0.0816)	1 (0.4520)	2-1 (0.0816)	1 (0.4550)	2-1 (0.0816)	1 (0.4580)	3-2	1
Cagliari - Sampdoria	0-1 (0.1235)	2 (0.4560)	0-1 (0.1298)	2 (0.4620)	0-1 (0.1362)	2 (0.4690)	0-1 (0.1425)	2 (0.4750)	3-1	1
Empoli - Atalanta	1-2 (0.0905)	2 (0.5680)	1-2 (0.0905)	2 (0.5720)	1-2 (0.0905)	2 (0.5770)	1-2 (0.0905)	2 (0.5810)	1-4	2
Genoa - Sassuolo	1-1 (0.1051)	2 (0.4560)	1-1 (0.0996)	2 (0.4610)	0-1 (0.0969)	2 (0.4670)	0-1 (0.1025)	2 (0.4720)	2-2	X
Udinese - Bologna	1-1 (0.1109)	1 (0.4020)	1-1 (0.1051)	1 (0.4080)	1-0 (0.1047)	1 (0.4140)	1-0 (0.1105)	1 (0.4200)	1-1	X
Napoli - Torino	2-1 (0.0864)	1 (0.5540)	2-1 (0.0864)	1 (0.5570)	2-1 (0.0864)	1 (0.5600)	2-1 (0.0864)	1 (0.5630)	1-0	1
Juventus - Roma	2-1 (0.0782)	1 (0.4110)	2-1 (0.0782)	1 (0.4150)	2-1 (0.0782)	1 (0.4180)	2-1 (0.0782)	1 (0.4220)	1-0	1
Venezia - Fiorentina	0-1 (0.1615)	2 (0.4960)	0-1 (0.1678)	2 (0.5020)	0-1 (0.1742)	2 (0.5080)	0-1 (0.1805)	2 (0.5150)	1-0	1

Tabella 2: *Risultati previsti dal modello ed effettivi per l'ottava giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette.*

Analizzando la Tabella 2, le previsioni non sembrano molto diverse al variare dei valori del parametro di dipendenza, anche se valori alti sembrano "polarizzare" le probabilità sulla vittoria di una delle due squadre. Il giusto compromesso sembra essere $\rho = 0.10$, dato che verrà confermato anche nel paragrafo 3.3.6.
Per le successive analisi si utilizzerà quindi solo $\rho = 0.10$.

Partita	ρ = 0.10		Osservato	
	Risultato (prob)	Esito (prob)	Risultato	Esito
Atalanta - Venezia	2-0 (0.1140)	1 (0.6480)	4-0	1
Fiorentina - Sampdoria	1-1 (0.1038)	1 (0.4110)	3-1	1
Verona - Cagliari	1-0 (0.1353)	1 (0.5360)	0-0	X
Salernitana - Juventus	0-1 (0.1487)	2 (0.5580)	0-2	2
Bologna - Roma	1-1 (0.0974)	2 (0.4620)	1-0	1
Inter - Spezia	2-1 (0.0822)	1 (0.5500)	2-0	1
Genoa - Milan	1-2 (0.0970)	2 (0.6000)	0-3	2
Sassuolo - Napoli	1-2 (0.0832)	2 (0.5070)	2-2	X
Torino - Empoli	1-1 (0.1104)	2 (0.4100)	2-2	X
Lazio - Udinese	1-2 (0.0967)	1 (0.4960)	4-4	X

Tabella 3: *Risultati previsti dal modello ed effettivi per la quindicesima giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette*

Partita	ρ = 0.10		Osservato	
	Risultato (prob)	Esito (prob)	Risultato	Esito
Salernitana - Sassuolo	0-1 (0.1292)	2 (0.5310)	2-2	X
Spezia - Cagliari	0-1 (0.1778)	2 (0.3720)	2-0	1
Sampdoria - Juventus	1-1 (0.0927)	2 (0.4670)	1-3	2
Milan - Empoli	2-1 (0.0929)	1 (0.4550)	1-0	1
Fiorentina - Bologna	1-0 (0.1459)	1 (0.5610)	1-0	1
Verona - Napoli	1-2 (0.0781)	2 (0.4550)	1-2	2
Atalanta - Genoa	1-0 (0.1616)	1 (0.6460)	0-0	X
Udinese - Roma	1-1 (0.0967)	2 (0.4270)	1-1	X
Torino - Inter	1-1 (0.0964)	2 (0.4960)	1-1	X
Lazio - Venezia	2-1 (0.0903)	1 (0.4630)	1-0	1

Tabella 4: *Risultati previsti dal modello ed effettivi per la ventinovesima giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette*

3.3.6 Indici di precisione

Per valutare la bontà del modello utilizzato si costruiscono degli "indici di precisione" della previsione, valutando lo scarto da 1 della probabilità, data dal modello, del risultato/esito effettivamente osservato (in seguito p_{oss}), inserite nella terza e quarta colonna delle tabelle 2, 3 e 4. Tale valore viene poi diviso per lo scarto da 1 della probabilità del risultato/esito indicata come più alta, e quindi quella che

rappresenta le previsioni inserite nelle prime due colonne delle medesime tabelle (in seguito p_{prev}):

$$I = \frac{1 - p_{oss}}{1 - p_{prev}} \qquad (20)$$

Il denominatore viene inserito per tenere conto di distribuzioni molto disperse con molti risultati con probabilità simili, e quindi in questo caso di partite particolarmente imprevedibili. Gli indici vengono denominati I_{ris} per il risultato esatto e I_e per l'esito, ma hanno la medesima formulazione. Entrambi hanno come valore minimo 1, che indica che il risultato/esito previsto si è effettivamente verificato. In linea teorica, ambedue gli indici potrebbero assumere valore massimo infinito, ma di fatto I_{ris} non può superare 1.25 (p_{prev} in una Poisson bivariata è quasi sempre < 0.20) mentre I_e può assumere anche valori più alti, poichè è relativo ad una distribuzione con soli tre esiti.

Si fissano delle soglie per ritenere buone le previsioni:

- I_{ris} deve essere minore di 1.10;
- I_e deve essere minore di 1.50.

Nelle tabelle successive vengono mostrati i risultati dei calcoli per gli indici delle giornate già analizzate al par. 3.3.5.

Partita	Risultato osservato	ρ = 0.05		ρ = 0.10		ρ = 0.15		ρ = 0.20	
		I_{ris}	I_e	I_{ris}	I_e	I_{ris}	I_e	I_{ris}	I_e
Spezia-Salernitana	2-1	1.1245	1.0000	1.1245	1.0000	1.1245	1.0000	1.1245	1.0000
Lazio - Inter	3-1	1.0335	1.6895	1.0335	1.6972	1.0335	1.7090	1.0335	1.7170
Milan - Verona	3-2	1.0288	1.0000	1.0288	1.0000	1.0288	1.0000	1.0288	1.0000
Cagliari - Sampdoria	3-1	1.1129	1.3051	1.1210	1.3085	1.1293	1.3126	1.1376	1.3162
Empoli - Atalanta	1-4	1.0585	1.0000	1.0585	1.0000	1.0585	1.0000	1.0585	1.0000
Genoa - Sassuolo	2-2	1.0477	1.4173	1.0413	1.4508	1.0382	1.4878	1.0447	1.5227
Udinese - Bologna	1-1	1.0000	1.2759	1.0000	1.3091	1.0060	1.3413	1.0192	1.3759
Napoli - Torino	1-0	1.0355	1.0000	1.0321	1.0000	1.0287	1.0000	1.0173	1.0000
Juventus - Roma	1-0	1.0406	1.0000	1.0368	1.0000	1.0331	1.0000	1.0384	1.0000
Venezia - Fiorentina	1-0	1.0714	1.5179	1.0719	1.5241	1.0725	1.5305	1.0730	1.5381
Media di giornata		**1.0553**	**1.2206**	**1.0548**	**1.2290**	**1.0553**	**1.2381**	**1.0576**	**1.2470**

Tabella 5: *Indici di precisione della previsione per l'ottava giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette e che hanno quindi indice pari a 1.*

Osservando la Tabella 5, viene confermato che per $\rho = 0.10$ si hanno le previsioni migliori, ovvero quelle che minimizzano entrambi gli indici. Le successive analisi si svolgeranno quindi solo per tale valore del parametro di dipendenza.
I risultati in generale (Tabelle 6-7) sono soddisfacenti, alcune partite sforano le soglie fissate, ma la media di giornata è molto soddisfacente.

Partita	Risultato osservato	ρ = 0.10	
		I_{ris}	I_e
Atalanta - Venezia	4-0	1.0546	1.0000
Fiorentina - Sampdoria	3-1	1.0646	1.0000
Verona - Cagliari	0-0	1.0848	1.6853
Salernitana - Juventus	0-2	1.0469	1.0000
Bologna - Roma	1-0	1.0240	1.2602
Inter - Spezia	2-0	1.0181	1.0000
Genoa - Milan	0-3	1.0450	1.0000
Sassuolo - Napoli	2-2	1.0128	1.6389
Torino - Empoli	2-2	1.0617	1.2983
Lazio - Udinese	4-4	1.1043	1.5873
Media di giornata		**1.0517**	**1.2470**

Tabella 6: *Indici di precisione della previsione per la quindicesima giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette e che hanno quindi indice pari a 1.*

Partita	Risultato osservato	ρ = 0.10	
		I_{ris}	I_e
Salernitana - Sassuolo	2-2	1.0932	1.6524
Spezia - Cagliari	2-0	1.1451	1.0923
Sampdoria - Juventus	1-3	1.0451	1.0000
Milan - Empoli	1-0	1.0147	1.0000
Fiorentina - Bologna	1-0	1.0000	1.0000
Verona - Napoli	1-2	1.0000	1.0000
Atalanta - Genoa	0-0	1.1089	2.2966
Udinese - Roma	1-1	1.0000	1.3682
Torino - Inter	1-1	1.0000	1.5635
Lazio - Venezia	1-0	1.0270	1.0000
Media di giornata		**1.0434**	**1.3973**

Tabella 7: *Indici di precisione della previsione per la ventinovesima giornata della Serie A 2021-2022. In verde le previsioni che si sono poi rivelate corrette e che hanno quindi indice pari a 1.*

Anche le altre due giornate scelte come campione hanno indici inferiori alle soglie fissate.

Conclusioni

Limiti del modello utilizzato

Il limite principale del modello di Dixon-Coles utilizzato nel capitolo 3 è già stato trattato da Karlis e Ntzoufras (par. 2.3) e riguarda la sottostima dei pareggi con ogni numero di gol subiti, quindi anche 2-2, 3-3, 4-4, …, risolta con il modello *diagonal inflated*. Come si può notare dalle tabelle 2, 3 e 4 nel paragrafo 3.3.5 l'esito "X" non è mai quello con maggiore probabilità. Si può però affermare che il pareggio sia effettivamente l'esito di gran lunga meno probabile: nel periodo considerato nel dataset i pareggi sono stati 398 a fronte delle 1520 partite disputate, circa il 26%.
Altro limite del modello risiede nella stima dei parametri α e β, molto precisa con il passare delle giornate o per squadre che partecipano al campionato da molto tempo, ma estremamente fuorviante per compagini esordienti su cui non si hanno dati a disposizione. Per questo motivo conviene utilizzare il modello quando si hanno sufficienti dati a disposizione (a partire dalla quinta/sesta giornata di ogni campionato).

Conclusioni generali

Con questa tesi si è voluto dimostrare che nel calcio, per quanto domini l'imprevedibilità, fare buone previsioni è possibile. Anche in questo ambito si deve seguire con maggiore frequenza l'approccio *data driven* per evolversi e sganciarsi dall'alone di soggettività che da sempre circonda le valutazioni e le previsioni sportive. Tale considerazione è ancora più urgente nell'ambito del calcio moderno in cui tra una vittoria e una sconfitta o tra un buon giocatore e un giocatore non funzionale possono decidersi le sorti di vere e proprie aziende con milioni di euro di fatturato.

Bibliografia

Brechot, M. e Flepp, R. (2020), "Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals", *Journal of Sports Economics*
2020, Vol. 21(4) 335-362

Dixon, M. e Coles, S. (1996), "Modelling Association Football Scores and Inefficiencies in the Football Betting Market", *Appl. Statist.* (1997) 46, No. 2, pp. 265^280

Karlis, D. e Ntzoufras, I. (2003), "Analysis of sports data by using bivariate Poisson models", *The Statistician* (2003) 52, Part 3, pp. 381-393

Kocherlakota, S. e Kocherlakota, K. (1992) Bivariate Discrete Distributions, *New York: Dekker*

Koopman, S. e Lit, R. (2013), "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League", *J. R. Statist. Soc. A* (2015) 178, Parti, pp. 167-186

Maher, M. J. (1982) Modelling association football scores. *Statist. Neerland.*, 36, 109-11

Nguyen, Q. (2020), "Predictive Models of English Premier League Goal Scoring", Wittenberg University, cap. 4

Roback, P. e Legler, J. (2021), "Beyond Multiple Linear Regression. Applied Generalized Linear Models and Multilevel Models in R", *Chapman & Hall*

Sitografia

Manisera, M. e Zuccolotto, P. (2018), "Gli statistici hanno i numeri… anche nello sport", *Statistica & Società.* http://www.rivista.sis-statistica.org/cms/?p=334

Pearse, W. (2020), "La statistica nello sport", *Inomics*. https://inomics.com/it/blog/la-statistica-nello-sport-1291601

Documentazione sul test chi quadro, https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test

Per i dati utilizzati nel capitolo 3, https://fbref.com/en/comp/11/Statistiche-di-Serie-A